Next Generation Sequencing Technologies: Data Analysis and Applications

Data Quality

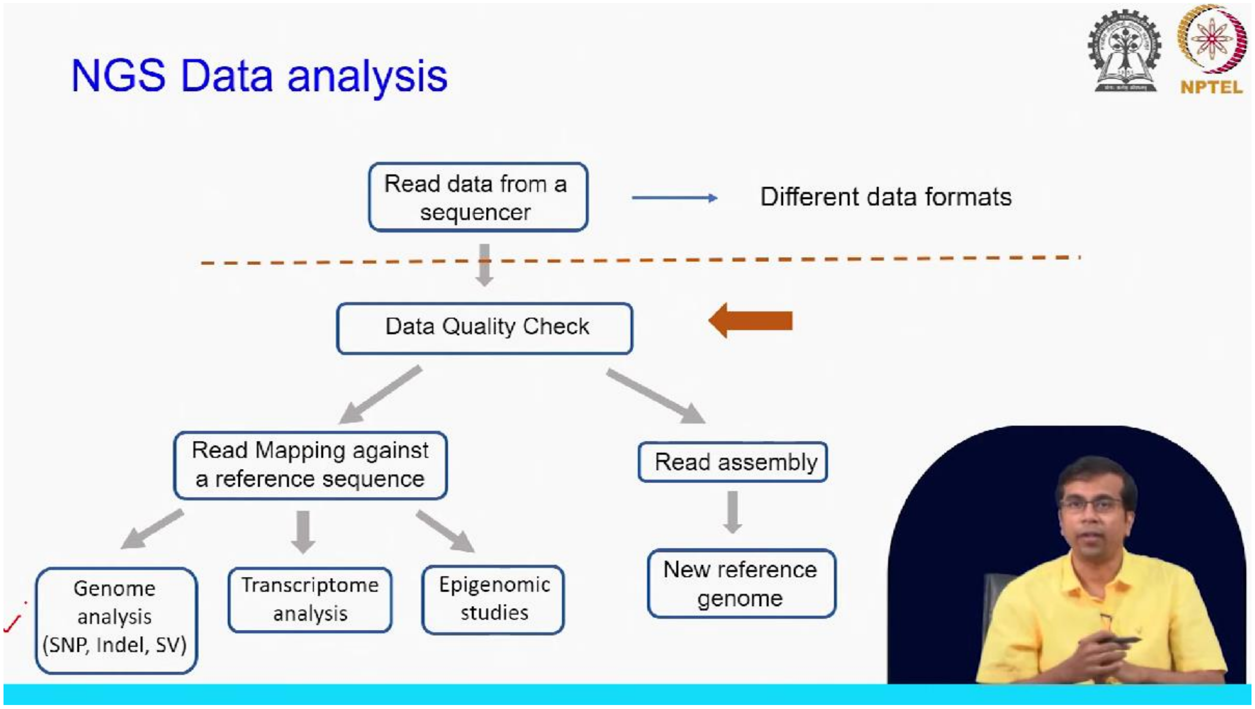
Dr. Riddhiman Dhar, Department of Biotechnology

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 data formats from different sequencing platforms. Especially, we will talk about the FASTQ format, which is used by the Illumina platform. We talked about the FAST5 format used by Nanopore. We talked about other formats that can be converted to FASTQ, for example, SMRT as well as Ion Torrent.

In today's class, we will talk about the quality of the data. Once you have the data, we have understood the data, how it looks, etc., and we want to check how good the data actually is. This is a very important step for subsequent analysis.

So, if the data is not of good quality, you will have problems with all the inferences that you make at the end. So, you might make erroneous inferences, you might make some mistakes, etc. So, these are the concepts that we will be covering today. So, what do you mean by data quality, and how do you check the data quality? So, sometimes we call it quality control, quality check, or, in short, QC. So, these are the keywords: quality control and FASTQC.



NGS Data analysis

Read data from a sequencer → Different data formats

Data Quality Check

Read Mapping against a reference sequence

Read assembly

Genome analysis (SNP, Indel, SV)

Transcriptome analysis

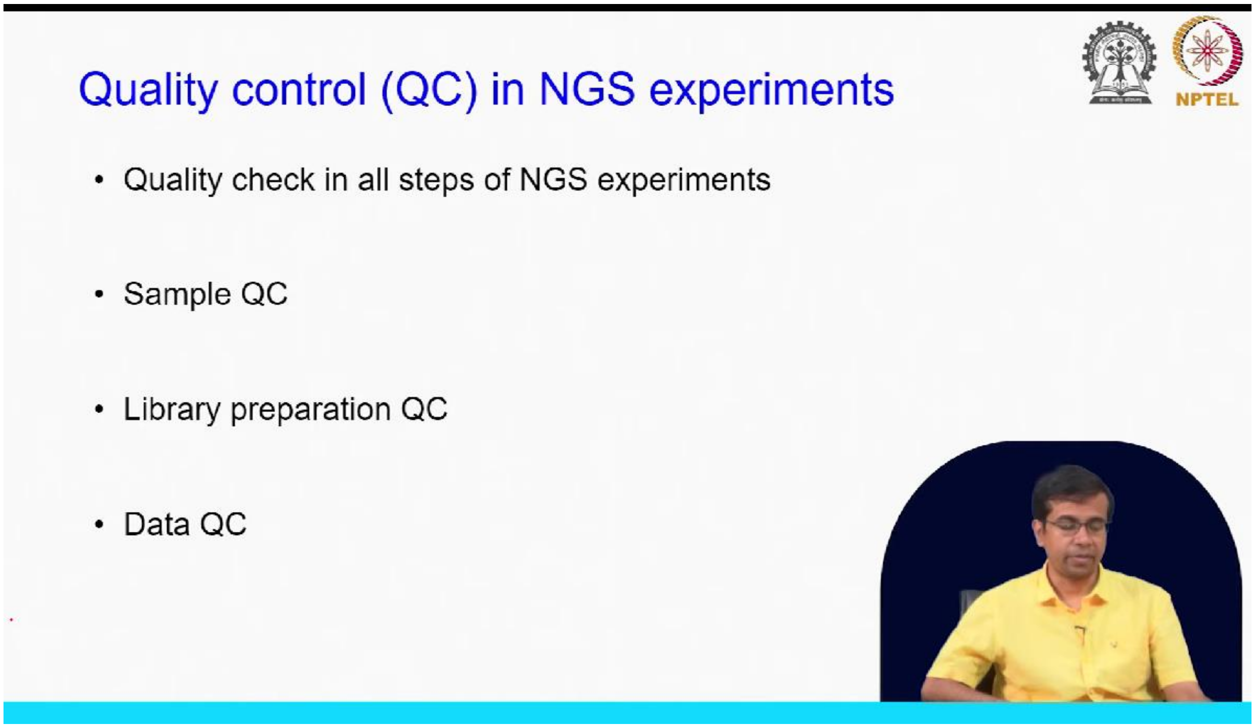
Epigenomic studies

New reference genome

So, going back to the flowchart that I have already shown you, we have the read data from a sequencer, and it comes in different data formats. The next step is the data quality check. Here we are now, and you will see that as we progress, we will go down this path, and we will probably add a few more when we come to these stages here because these are much more complex steps. So, in today's class, we will be talking about this data quality check. How do you check the data quality? So, quality control in NGS experiments is a very important point, and this check is done in all steps of NGS experiments, right?

So, not just a data quality check, but it is also done at the experimental steps, which we have I will briefly mention those that we have not discussed in detail. So, there is something called sample quality control: the input material that is coming in, the DNA, genomic DNA that you are working with, or RNA molecules that you are sequencing, right? You need to do quality control for those samples, right? So, if you are giving degraded material or some bad quality material, of course, the data will not be good, right? So, these are experimental processes that are used to determine the sample quality.

You also have a library preparation quality check, right? So, this is something we talked about: the library preparation steps we have talked about. After the library preparation, you check through experimental methods how good the library is and whether there are any issues with it. And of course, the final check is the data quality check, right? So, at each step, you see this quality check, and the data quality check looks at the final data that has come out of the sequencer.



Quality control (QC) in NGS experiments

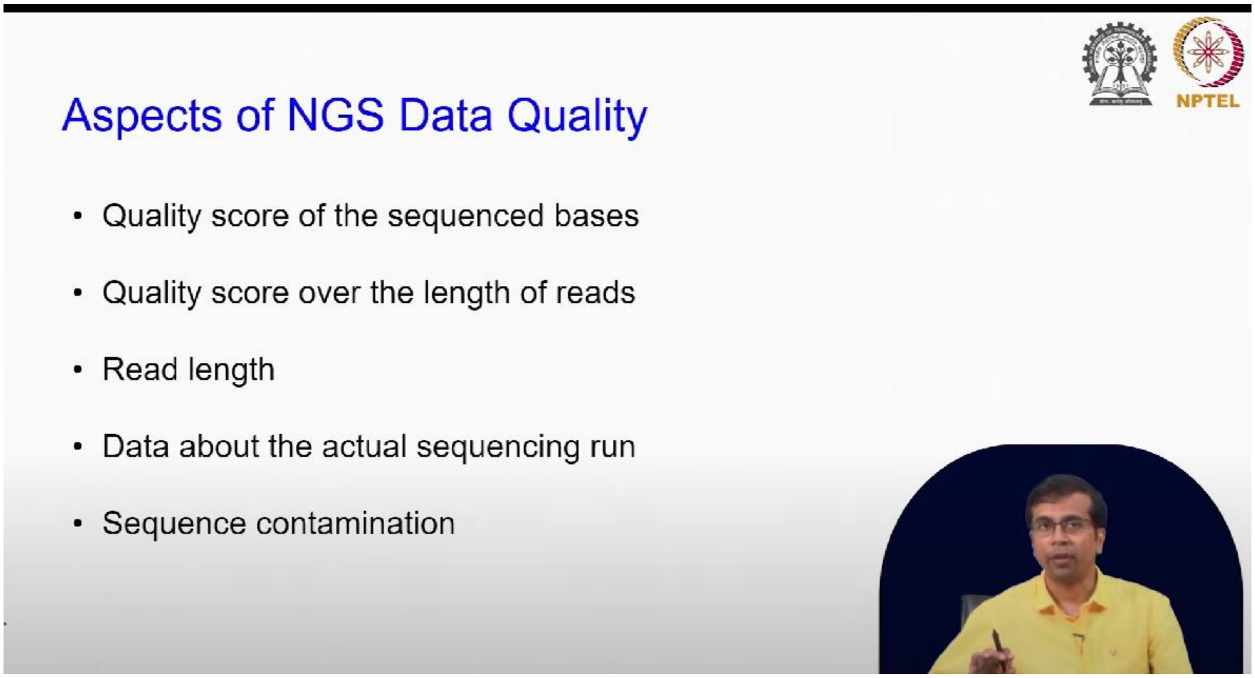
- Quality check in all steps of NGS experiments
- Sample QC
- Library preparation QC
- Data QC

So, we will talk about this data quality control or quality check in much more detail, ok? So, this is a critical step; we do it for the raw data that comes out of the sequencer, and this can help us identify some potential issues in library preparation or in sequencing, right? So, as much as you try to do your library preparation well, there could be some factors that could affect library preparation, and those might not be visible when you are doing this library QC or sample QC, right? When you do the sequencing and the data QC, they turn off, right? So, this is a key issue that we have to address, right?

In some cases, you might have some issues with the sequencing process itself, right? There might be some sort of issue with the adapter ligation etcetera, all those things, and the sequencing process did not work for whatever reason and those will be identified here at this step as well, right? You can also identify whether there is some sort of potential contamination in your sample, right? And this is a problem that we should be aware of, right? When you are processing, for example, some genomic DNA from some organism, you might also get some human DNA contamination, right?

So, we are working with these DNA samples ourselves, and maybe some DNA can get into that sample, ok? So, from this data quality check, we can also identify these potential contaminations. And this is a critical step because you have a huge impact on the downstream analysis and inferences that you make, right? So, if you can identify that maybe there are some errors or issues in library preparation, we need to design our downstream analysis pipeline in an appropriate manner. So, as to taking care of the issue if we can, ok.

So, this is something that this step helps in making those decisions later on, ok? So, what are the aspects of NGS data quality? So, we say, "Okay, we do quality checks on what we are doing and what we are looking at."So, is it just the quality score? So, of course, it is the quality score of the sequence basis, right? How so? This is measuring how well the sequencing library and the sequencing process work, right?

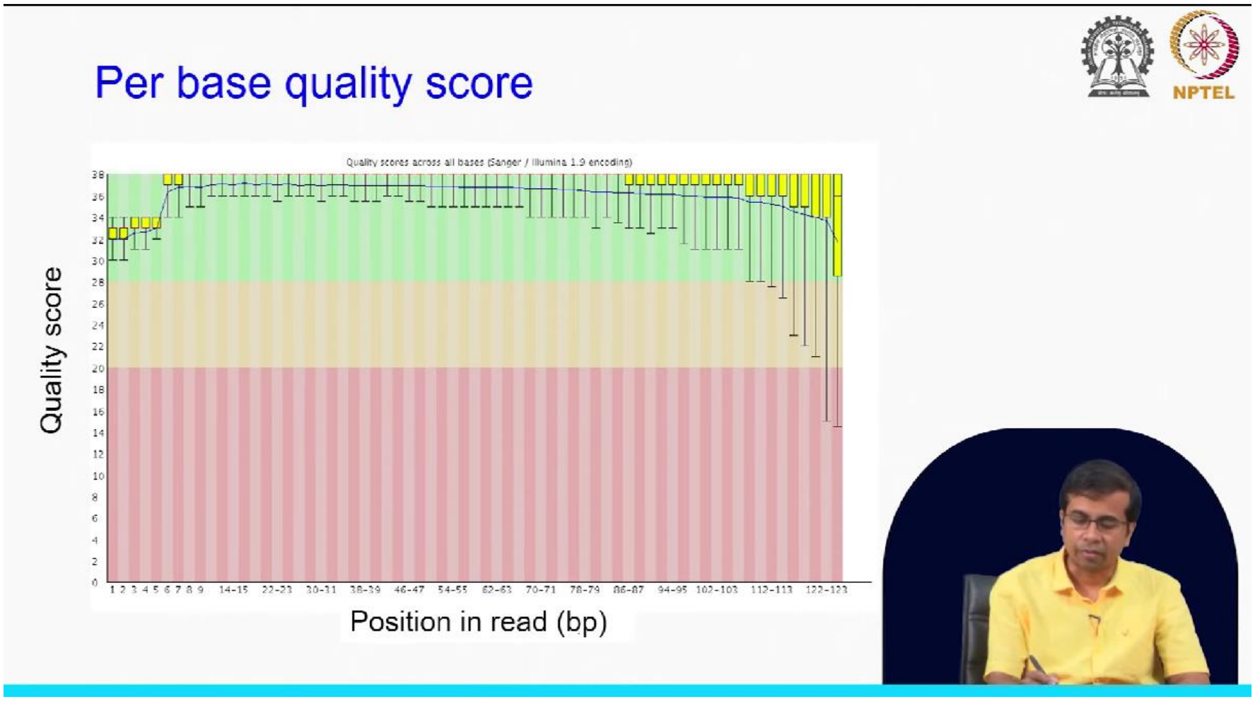


Aspects of NGS Data Quality

- Quality score of the sequenced bases
- Quality score over the length of reads
- Read length
- Data about the actual sequencing run
- Sequence contamination

So, this is the quality score of the sequence basis, the quality score of the length, the length of the reads, and whether the sequencing works fine. And then there are also other aspects, right? For example, read length, you have the data about the actual sequencing run, how the sequencing went, some statistics that will be gathered, whether there is some sort of contamination, that kind of aspect will come there, ok? And there is a tool that we can use for this quality control called FastQC, ok? This is a very popular, widely used, and freely available tool that can take FASTQ data and generate a lot of statistics that can tell us about different aspects of quality control.

So, we will give you some examples from this tool using the Illumina dataset, ok? So, this is something that can be applied to other datasets in FASTQ format, but the example that we will discuss today here will be mostly from the Illumina dataset, ok? We will do this in hands-on classes in the hands-on tutorials where we will be actually analyzing datasets using these FASTQ tools, ok? So, the inputs are FASTQ files, right, and here in this example, this would be mostly Illumina data. So, the first statistic that you get is something called a par-basis quality score, ok?



Per base quality score

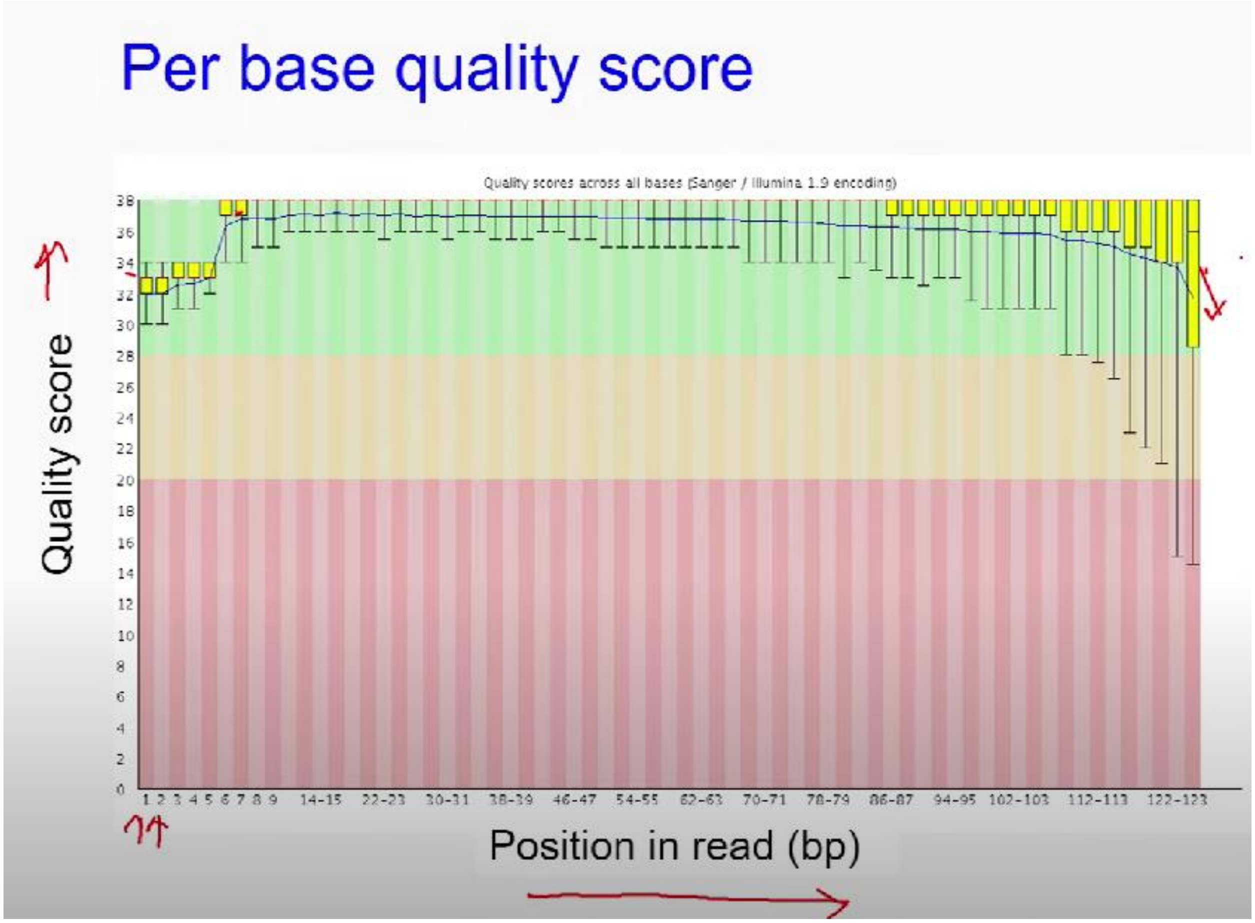
So, what is this par-based quality score? So, here you have along x axis position in read and along y axis you have the quality score, ok. So, this quality score comes from the encoding, right? We talked about SK encoding. This encoding represents the quality score. So, what we do here is, for each base at this position (position 1, position 2, etcetera), we calculate the distribution of the quality score.

So, this tool calculates this distribution of quality scores for reads about 10000 reads or 100000 reads for this first base. What is the distribution of quality score? So, this is something that is represented by this box plot, right? So, we have this here; you have this box plotted in yellow, right; it is showing the distribution, right. So, here the distribution is around, let us say, 32 around 32, right? So, which is probably quite decent if you remember the probability of error and how it is related to the quality score.

And as we move along the length of the read, the quality score changes, right? So, here you see the quality score for positions 2, 3, etcetera, and these reads were of 125 bases per length. So, you see the quality score up to 125, ok? So, we move along, right, and we see the quality score changing, and in the middle the quality score is quite high, right; it is about 38 here, which is the limit here, and then slowly goes down towards the end here, right. You can see this quality score going down a little bit.

So, this is the statistic that is generated by this program, ok? In addition, you see some color coding, right? So, you have these green zones, you have this yellow zone, and you have the red zone, right? It is kind of like a decision-making thing, right? So, if you are in the green zone, you are kind of doing all right.

So, your data is of good quality. In the yellow zone, perhaps you need to decide whether you want to use that kind of data or not. Red Zone: maybe it is really bad quality and maybe you do not want to use this, ok? So, that kind of helps in kind of visually making some decisions here.

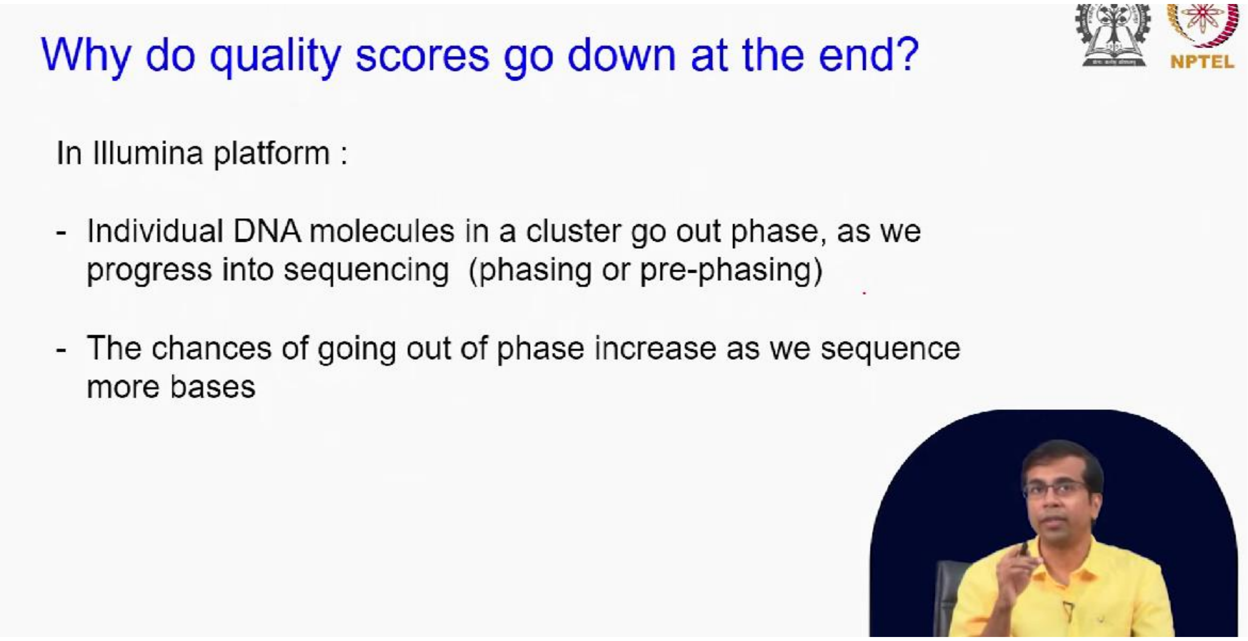


So, this color code. Now, the question again, right, like Sanger sequencing, where the quality scores low in the beginning. This is something that we have seen in Sanger sequencing, right? So, we will come to that in a moment. And it turns out that, on in Illumina platform, there is no electrophoresis, right? It is not like Sanger sequencing, where you have this capillary electrophoresis.

So, on the Illumina platform, what happens is that the first few cycles are used for detection on these clusters, right, and for setting up the parameters for base calling, right. So, what would be the signal thresholds where the system will call these bases, etcetera? So, the first few cycles are utilized for that, and in those cycles, the system runs with some preset values, and it usually reports conservative quality scores. So, that is why you see this lower quality score in the beginning, ok? And then it slowly goes up, right, and again the quality scores go down at the end, ok.

So, here you see that the quality score trend is going down. Again, the question is why that is the case. Again, here there is no electrophoresis, etcetera. So, it is actually interesting why this goes down. So, this actually goes on because of a problem called phasing or prefacing, ok?

So, in the Illumina platform, if you remember the principle, right, so the sequencing is done, right, in clusters, right. You have these clusters of molecules; they are of the same clone, denoted by the same clone, right? They are sequenced at the same time, and the signal that we detect is from the cluster, right? So, it is the sum total signal, right, that comes from the cluster. So, what happens as we proceed along the sequencing? These DNA molecules go out of phase in the synthesis process, ok?



Why do quality scores go down at the end?

In Illumina platform :

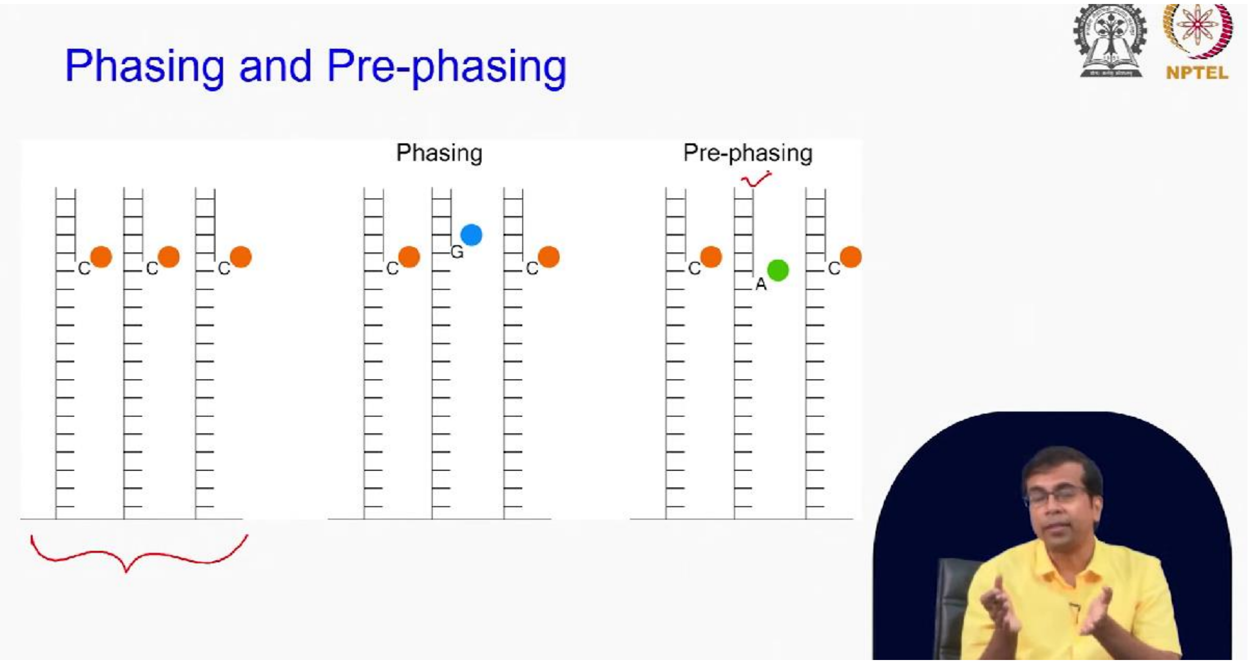
- Individual DNA molecules in a cluster go out phase, as we progress into sequencing (phasing or pre-phasing)
- The chances of going out of phase increase as we sequence more bases

So, once they go out of this phase, right, this chance of going out of phase actually increases, right, as we sequence more and more bases. So, we start with the same base. So, initially, when they are in phase, they are kind of going in sync, but then later on, when they go in, they have this problem. They go out of phase and this asynchronous signal actually increases noise and reduces the quality score. So, what is this phasing, prefacing, right? So, let us look at this schematic, right?

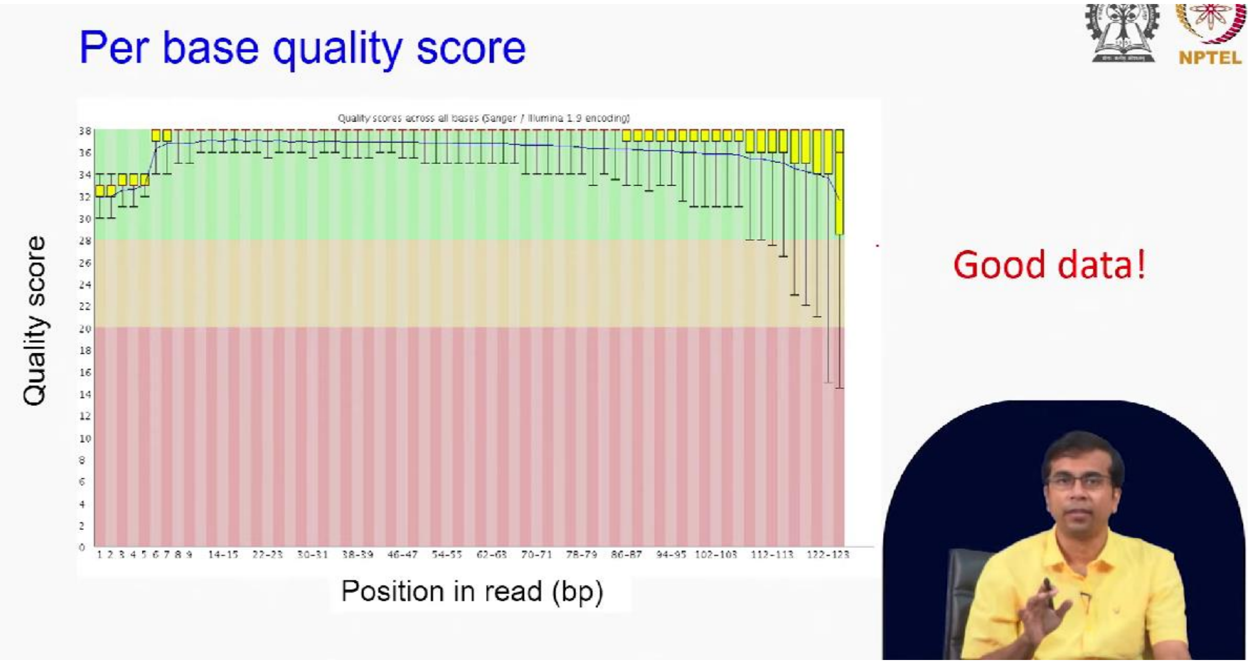
So, as I mentioned, right in one cluster you have these DNA molecules; they are from the same clone, the same molecule; they are being sequenced, and the signal should be identical, right? They will have the same pattern of nucleotides. Now, what happens sometimes? Maybe some of these nucleotides are not cleaved properly. So, they kind of stay back. So, the fluorescence signal does not this fluorescence tag does not wash away, right, in one strand perhaps and it stays back, right.

And what it means is that it is read twice, right? It will read twice the G, and then it will kind of go out of phase, right? So, in the next cycle, it will then do the C when the C bases have moved on to the next one, right? So, it will kind of lag behind the other strands. The moment this happens, you can see there are confusing signals: two strands giving an orange signal and one strand giving a blue signal, right? So, if you have this confusing signal then your signal-to-noise ratio will fall, which means the quality score will fall.

So, this problem is called phasing. The other problem is prefacing. So, maybe there is some sort of defect in the level of nucleotides, and at some point there might be the addition of two bases at once, ok? So, in this molecule here in the middle, there are two additions, right, two bases added at once, and this one will be ahead, right, of the rest of the molecules in the cluster, the rest of the DNA molecules. So, this problem is called preplanning. Again, the effect is the same, right? You have this mixed signal, right?


Phasing and Pre-phasing

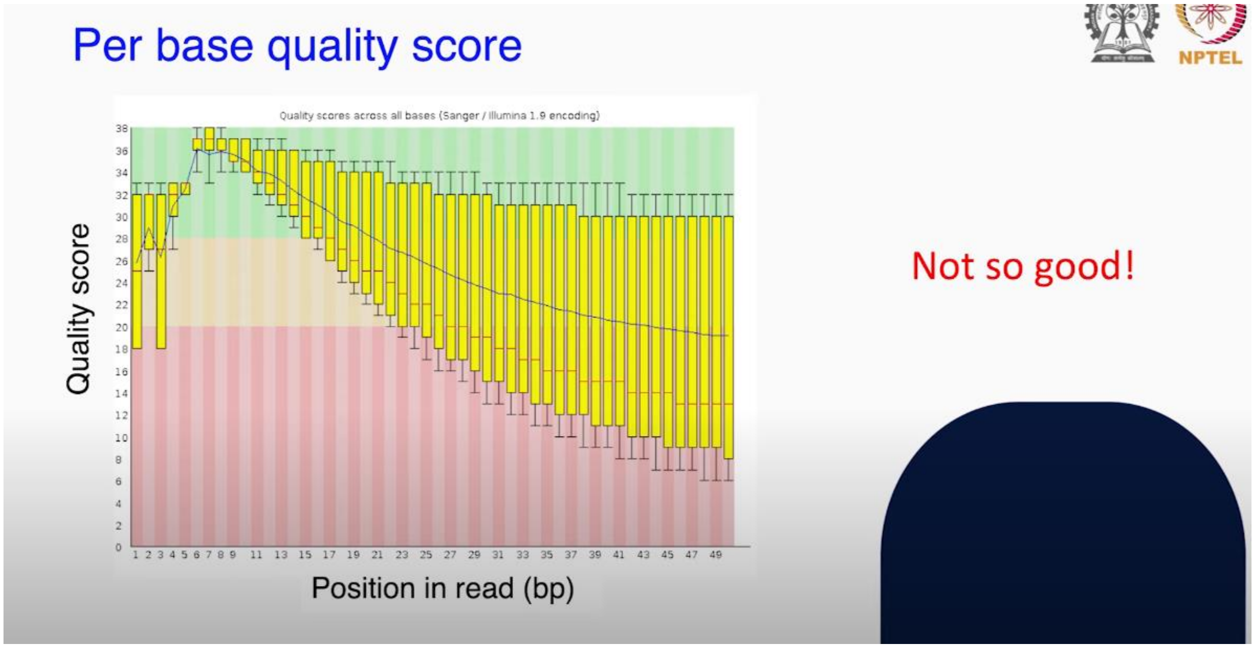
So, you have here orange and green, which will mean there will be a lot more noise and the signal-to-noise ratio will fall, right? And this will increase as we sequence more and more, right? So, as we go on, this out-of-sync phenomenon will happen and you will see this drop in quality score. However, in this case, the quality score drop is still well controlled, right? You see, most of this quality score is within the green region.



Per base quality score

Good data!

So, this is quite good data, right? So, we do not have to worry about the quality score aspect here, at least, ok? So, this is something that is good, but in some cases, right, so just to give you an example in some cases you might have this kind of distribution of quality scores, right? And we are looking at the same thing you have the position to read, right? We are starting from position 1, and here is the T basis only, and the y axis is the quality score, right? And here you see, we start in the green region, and then the quality score drops to the orange region and the red region, ok?

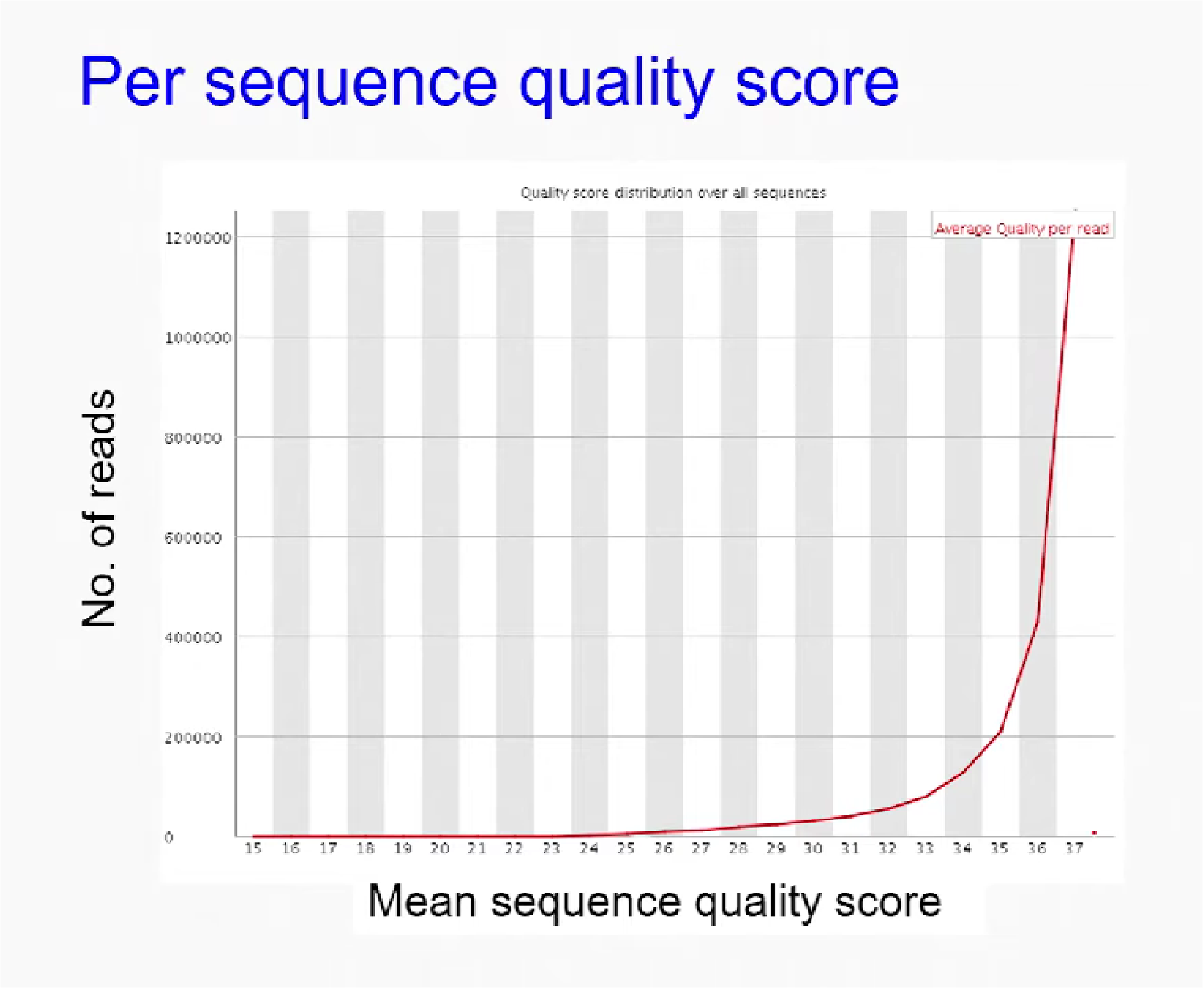


Per base quality score

Not so good!

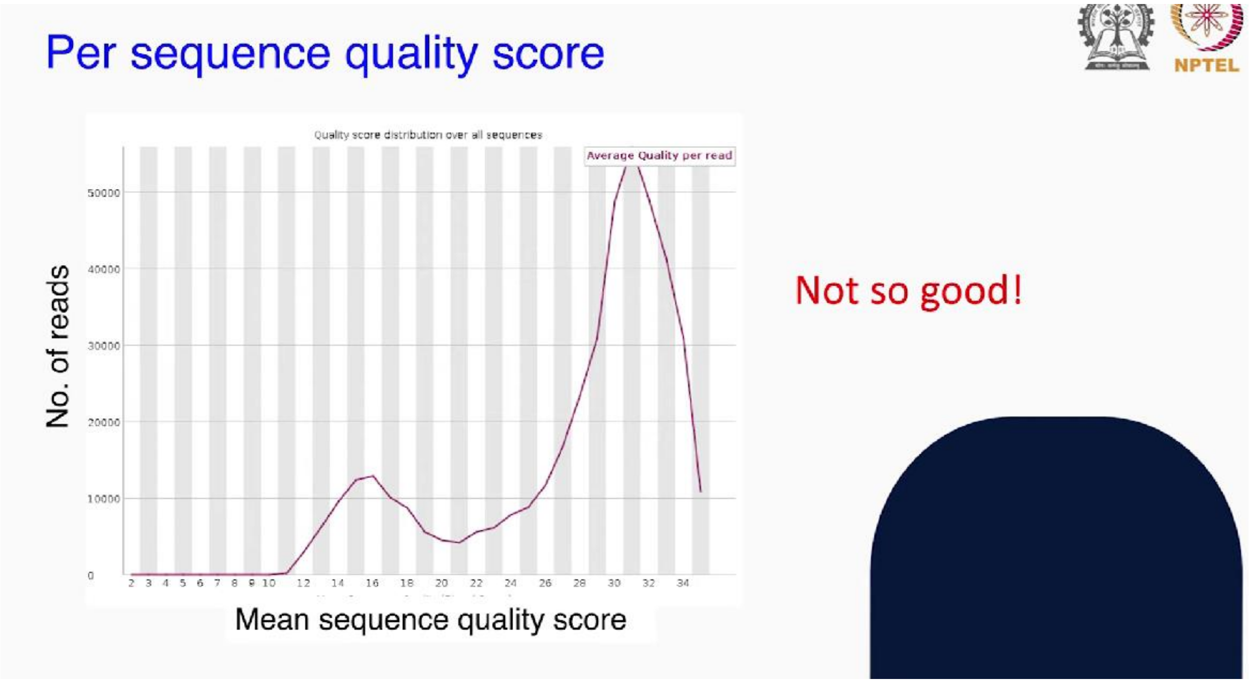
And this is not so good, right? So, this is where we have to make some decisions, right, whether we should use this read or whether there are other processes that we can utilize. Clearly, in this region here, right, towards the last part, right, this region, the quality score is quite low, and maybe whatever the base calls are, they are not probably reliable, ok? So, this is something that will come when you do the data analysis ourselves, right? This decision-making we would have to do, ok? The next parameter that we look at is something called the par sequence quality score, ok?

So, the earlier one was part of the basic quality score; we are looking at the quality score distribution on an individual basis. Here we are looking at the average quality score of the read, ok? So, this is something again shown in the x-axis; this is the mean sequence quality score. So, if you take the quality score of all basis in a read and just average, you will get the mean sequence quality score, ok?



Per sequence quality score

And along the y-axis, you have a number of reads, ok? So, in this data, most of these reads, so the numbers on the y-axis you see are really large numbers, 1.2 million, etcetera. You see most of the reads have very high average quality, right, 36, 37, right. And there are few reads that have quality score average quality score less than 30, right, very few.



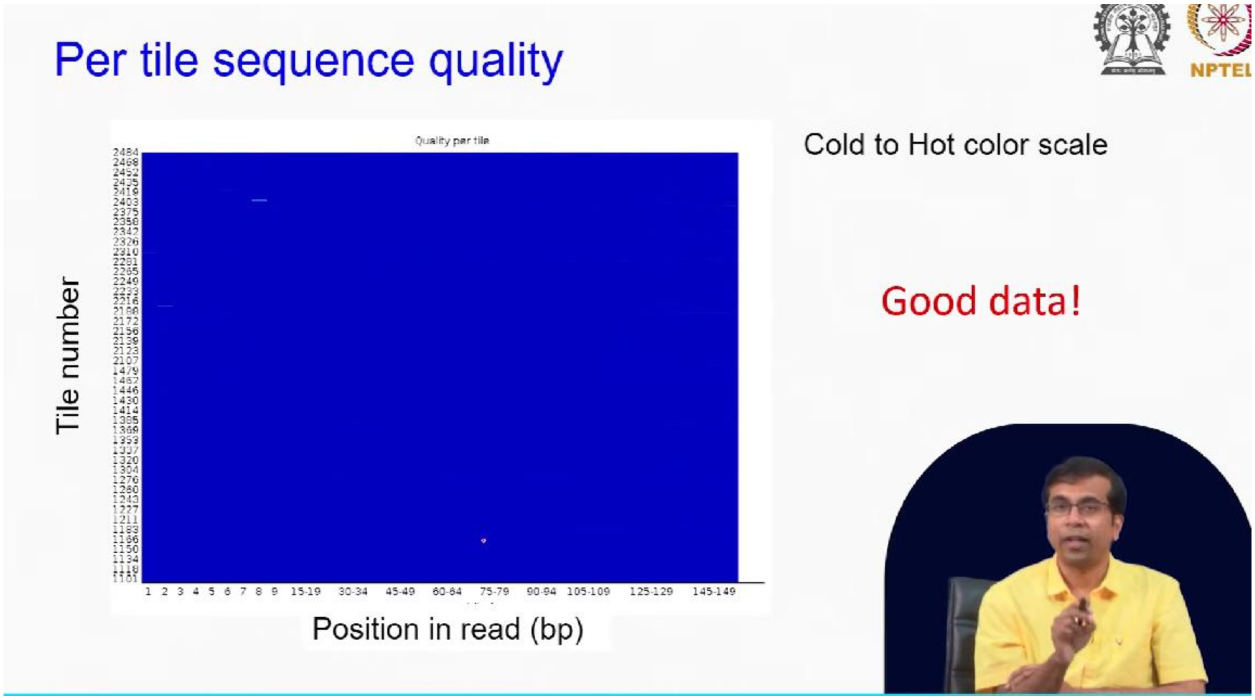
Per sequence quality score

Not so good!

So, which means this is really good data, ok? Now, why do you want to analyze this part-sequence quality score? Because it may turn out right, there are some reads that are of poor quality, right? And this, so this is good data, right, as I mentioned. But in some cases, you might have this kind of situation, where most of the reads are showing a good average quality score, so around 30 or so, which, again, if you look at the probability of error and accuracy, is quite acceptable. But then you have a bunch of reads that have a quality score below 20, ok? And this is where you might have more errors in the reads, and this is something that you probably want to avoid in your downstream analysis, right?

So, when you look at this means par sequence quality score, this is what will be beneficial, right? You want to know whether there are certain reads that are of low quality, right? And this actually helps us with downstream analysis, right? So, if you know there are some of these reads that are of low quality, you can simply discard these reads and go ahead with the good-quality reads, ok? So, this decision is very simple: you simply discard reads that are of low quality.

The next parameter that we look at is something called par tile sequence quality, ok? So, this is something that you have to go back to the flow cell and the tiles, right, that we also talked about in the data format, right. In the header, you have this information. So, FASTQC can utilize that information and generate these statistics, right? So, here it seems all blue, right? It does not make any sense.



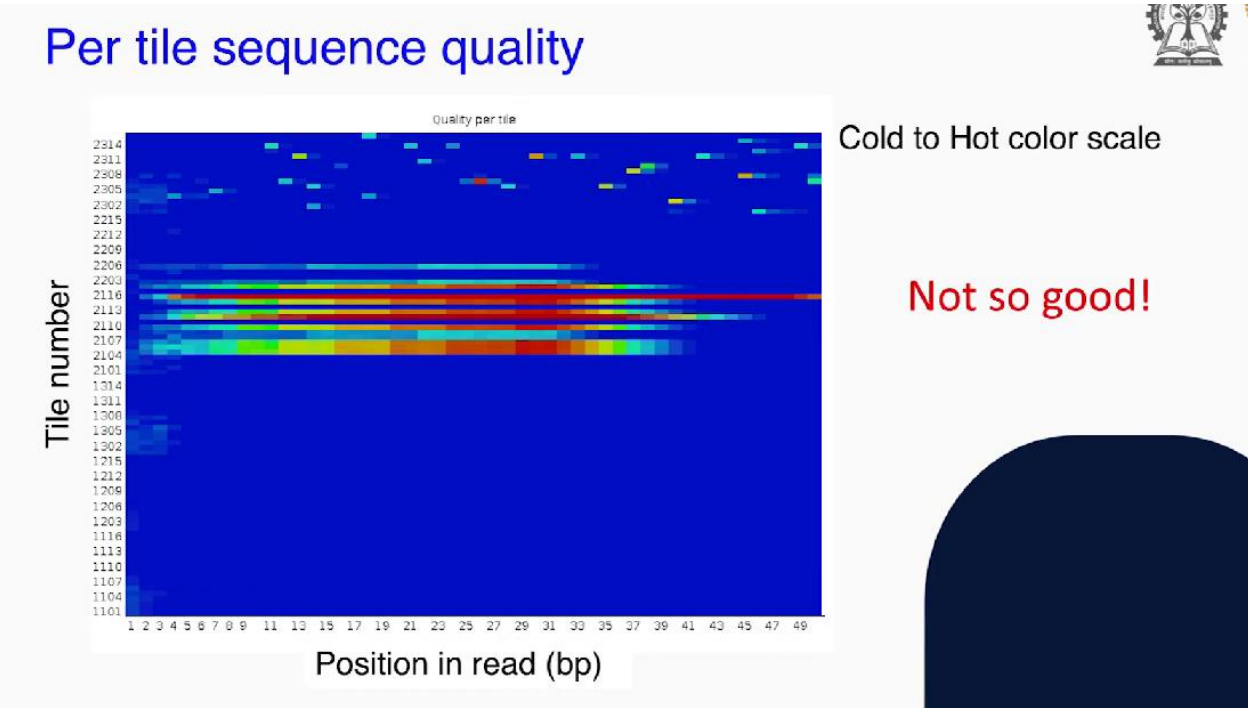
Per tile sequence quality

Cold to Hot color scale

Good data!

So, I will explain what it actually is. So, in the x axis, you have position in terms of base pairs, and in the y axis, you have the tile number, ok? You remember the tile number that was present in the header, ok? And what you see here is the quality of the tile, ok? So, the blue actually gives you quality par tile, and this is in a color scale, ok? So, this color scale goes from cold to hot, which means cold color scales will be in blue and hot will be in red, ok?

So, cold to hot color scale. So, if you see some red, that means you have this hot color, right? So, usually, we expect that this tile sequence quality is blue, which means we have very good quality; we do not have to worry about the data. So, here it is mostly blue, right, except I think you can see maybe one kind of slightly lighter color, but that is ok and most of the things are ok, right. So, this is good data. But in some cases this might happen, right, where you have some of these regions that are in red, ok, again following the cold to hot color scale, ok.

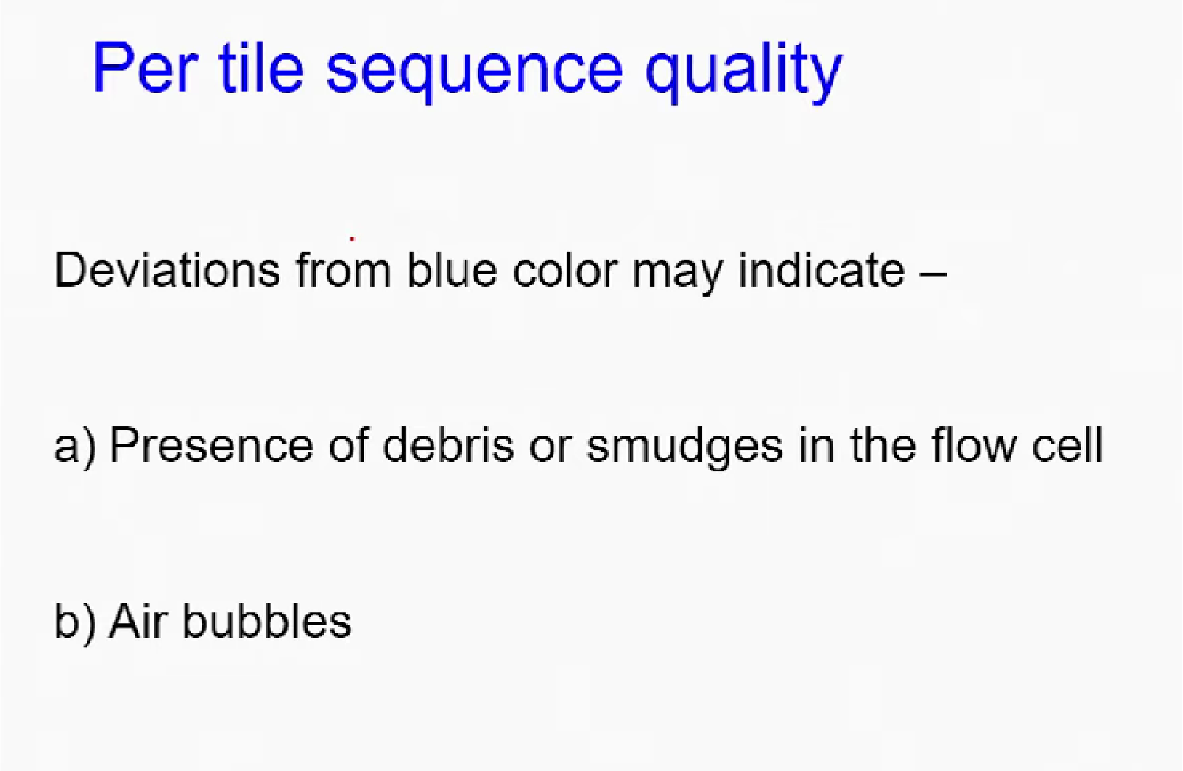


Per tile sequence quality

Cold to Hot color scale

Not so good!

So, this region that is in red means these tiles have some problems, right? So, their quality score is low, ok? So, now going back to the data format, when talking about the header why do you have these coordinates, this tile number, etcetera? This is where this information becomes useful, right? So, one of the things you probably want to check, right, is whether you see a bunch of low-quality reads, ok, in your data, and you want to check whether these low-quality reads come from the same place in the flow cell, right, whether they cluster together in the flow cell from the same tile, right, or in neighboring regions, ok.



Per tile sequence quality

Deviations from blue color may indicate –

a) Presence of debris or smudges in the flow cell

b) Air bubbles

This will tell us two things, right? Maybe there is some sort of issue with the flow cell, right? In those regions, there might be some sort of, let us say, dirt or dust, which probably inhibited or hindered the signal detection process. So, this is something that is quite useful, ok? So, as I said, a deviation from the blue color may indicate, right, that you have some sort of debris in there or smudges in the flow cell, or maybe you can have air bubbles. So, during the experiment, right, when you are adding those dNTPs, etcetera, you might have some air bubbles. So, the signal detection is not as good, okay, and the quality will drop because of that.

Again, it depends on the signal-to-noise ratio, right? So, these are the references that we have followed in this part. So, just to conclude, we have mentioned that quality control (QC of NGS raw data is a critical first step in the NGS data analysis pipeline. And this, as we have seen, right, we have this quality score-based quality control, right? So, we have talked about a few measures. So, the first one is the par-basis quality score, which is looking at the distribution of the quality score on an individual basis, and we can figure out whether there are certain bases that show low quality, right?

So, just going through the example data, we have seen in some cases, right, that the initial part has low quality and also the last part has low quality. So, these two regions have low quality. We have discussed the reason why we see low quality in the beginning, ok, and this is because the signal detection parameters are being set up there, ok. And towards the end, we have talked about problems called phasing or prephasing, right? So, towards the end of the read, when you are detecting these clusters of sequences, right, so you are sequencing in clusters, you are getting the average signal, some of the molecules might go out of sync, ok, and this is a problem that we call phasing or prefacing and this gives you this low signal-to-noise ratio that affects the quality score, right?

So, we have talked about this quality score, the par-based quality score. The second measure that we have used is something called a par sequence quality score, right? So, this is an average quality score over the read sequence, ok, and this helps us identify the reads that are of low quality, ok. This is something very useful because if we have a bunch of reads that are of low quality, we can simply take them out or discard them out before we proceed with the analysis. So, as we have seen now, right, we are kind of getting into that, right, this data analysis pipeline, right, we are preprocessing the data, right. We are seeing whether the data that we are going to use is of good quality, and if not, maybe if we can discard some of it, the rest will be of very good quality, right?

We also talked about the tiles per tile read quality, right? So, as you have seen, if there are any specific problems with the data in certain tiles, this might tell us something about the flow cell itself; there might be some sort of issue, some sort of air bubble, or maybe there are some sort of dirt particles that are interfering with the signal detection. So, why is it important, right? Why are these steps important, right? So, as we have mentioned, right, this data analysis pipeline, right, we are going to do much more complex analysis after this, ok? As I have shown in the beginning, we have this flow chart, right? We have so many complex analyses downstream that if we start with bad quality data or if there are certain issues in the data, the output that we get and the conclusions that we will draw will be flawed, ok?

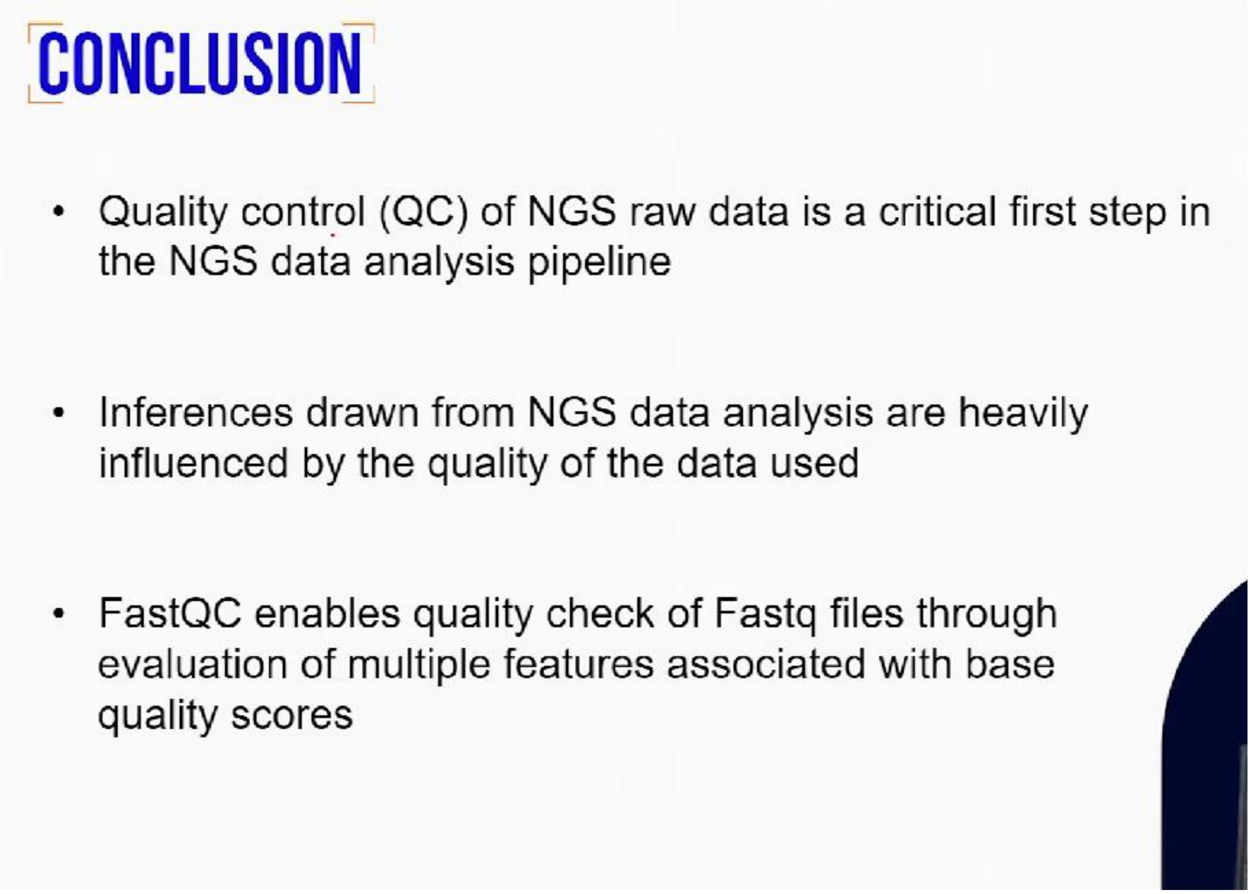
So, for example, I can give you an example. For example, we can take single nuclear polymorphism detection, ok? So, this is something that is very important, right? For example, we are screening patient samples, right? We are looking at mutations. We are looking for mutations in a gene or maybe a few genes that might tell us, right, whether that patient will be responding to certain treatments or not.

Now, if you are dealing with low-quality data, by mistake, you might end up detecting certain mutations there, ok, because the quality score is low, which means the signal-to-noise ratio is low. So, by mistake, you might end up detecting certain mutations there, right? So, this will actually impact patients lives, right? So, we are making decisions on therapy, etcetera, and if you are making wrong decisions that will impact patients lives, ok.

So, this is something that is very important, right? So, as I said, the inferences drawn are heavily influenced by the quality of the data used, especially in cases of mutation analysis, indel analysis, etcetera, and in many cases, these would have an impact; this could have a clinical impact, right? So, we want to be very careful in that case, right? We would rather throw away data than work with bad-quality data, ok?

So, this is one thing. So, all sorts of inferences will depend on the data used. Now, for some of the algorithms, we also have some applications for example, that read assembly. So, when we are building a reference genome, we kind of apply these assembly algorithms. There are different classes of assembly algorithms, and if we have low-quality or bad-quality data, the assembly algorithms might give us the wrong genome sequences. They might not be able to assemble sequences well, ok? So, this might actually give us the wrong reference genome assembly, or it might not be able to assemble the data properly, right?

So, this is a very important step. Now, we have also seen that FASTQC enables quality checks of these FASTQ files through the evolution of multiple features. So, here in this class, we talked about these features associated with only base quality scores, ok? So, we have talked about the quality check only based on base quality scores. There are other types of quality checks that we will talk about, and FASTQC reports those quality checks as well. Now, one of the questions that we might have is: What do we do once we have detected this bad quality?



CONCLUSION

- Quality control (QC) of NGS raw data is a critical first step in the NGS data analysis pipeline
- Inferences drawn from NGS data analysis are heavily influenced by the quality of the data used
- FastQC enables quality check of Fastq files through evaluation of multiple features associated with base quality scores

So, for per-sequence quality, right? In that case, we can leave out those reads, but what about the reads where the initial part is of good quality but the rest of it is not? So, the quality drops as you go along, and maybe the quality drops so much that it goes into that red zone, and we probably do not want to use that region, ok? So, this region may contain certain mutations that will give us the wrong results, right? So, we want to discard those regions. So, it turns out that there is a way, and we will discuss this in the next few classes, and we also do that, and so on, with something called read trimming, ok?

So, in this part, we can actually trim the reads of the low-quality regions and just work with the good-quality regions, right? So, this is something that is possible, ok? So, in the next class, we will be talking about the sequence of best quality evaluation features, and that, combined with this quality score-based evaluation, will form the full quality control, ok? So, once we have done this quality control, we can then decide on the next course of action, right?

So, FASTQC simply gives us this statistic. It does not do anything; it just generates all the statistics. The next steps, whatever we want to do, will have to be decided by us, right? So, we will make that decision. Thank you.