**Next Generation Sequencing Technologies: Data Analysis and Applications**

**Variant Calling**

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

**Indian Institute of Technology, Kharagpur**

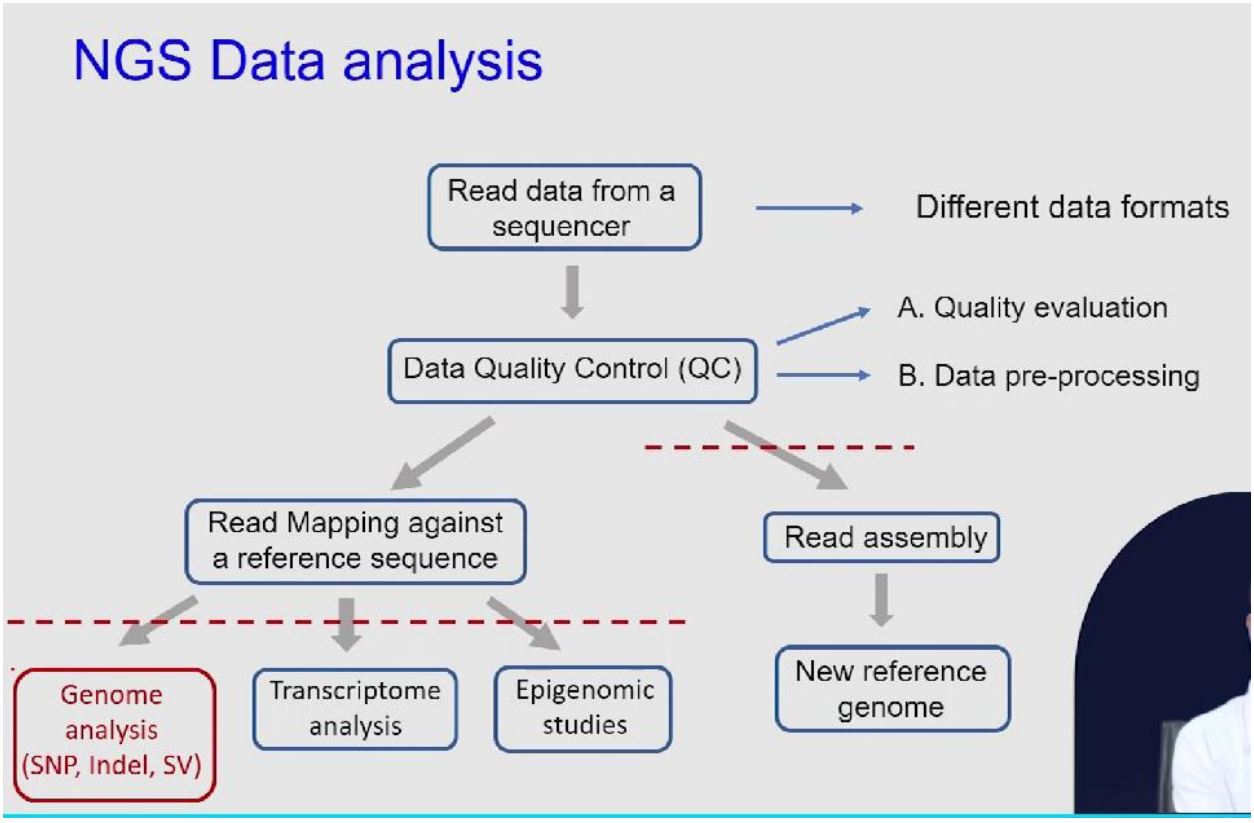
Good day, everyone. Welcome to the course on Next Generation Sequencing Technologies, Data Analysis, and Applications. In the last few classes, we have talked about read mapping, we have talked about the algorithms, and we have seen how we do the read mapping hands-on. So, once we have done the read mapping, the next step is to do variant calling. So, this is what we will be focusing on today. So, the agenda for today's class is that we will discuss the different variant types that are possible, and we will actually pick only certain types for this class, and then in the subsequent classes, we will look into other variant types and how we can call these variant types from the map data.

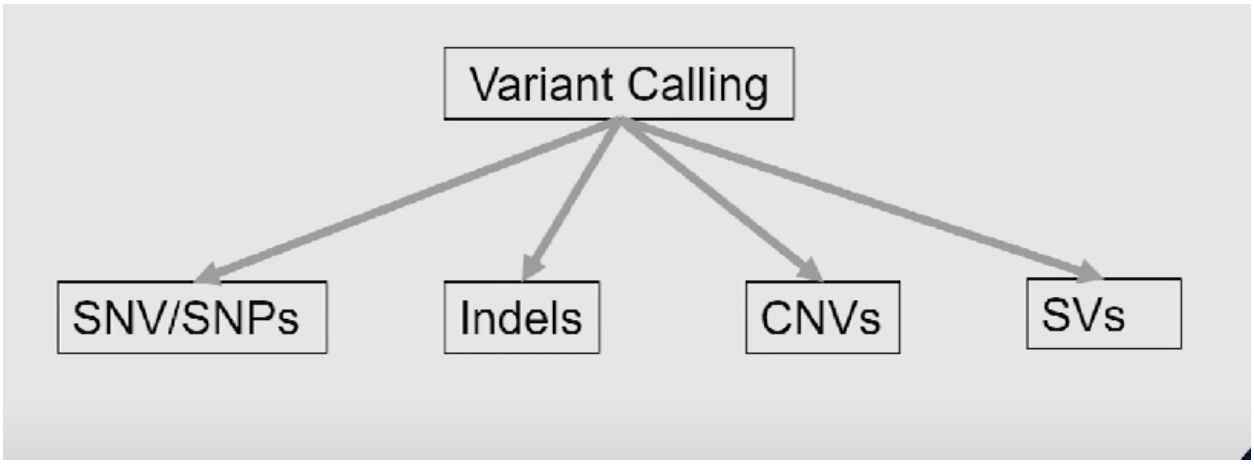
And then we will also talk about some of the applications of variant calling. So, this will help you understand why this is a very important problem, ok? So, let us begin and start with the different variant types. So, these are the keywords that we will come across today.

So, variants, SNP, and Intel. Going back to the flow chart for NGS data analysis, we have looked into different types of data in different data formats from different sequences. Then we have talked about data quality control, and we have done that hands-on. In the last few classes, we have talked about read mapping against a reference sequence. So, once you have done this read mapping, there are three applications that you can see and three types of analysis that you can do after read mapping. One is genome analysis.

So, this involves the identification of single nucleotide polymorphism, or SNP, indels. So, we have insertions and deletions, especially small ones, and then we also have structural variants. So, we will talk about each of them one after another, starting today. So, these lines actually mark where we are now. So, we have learned about this mapping, and also, on the other side, we have learned about the quality control up to that point we have reached, but we have not talked about

the assembly.

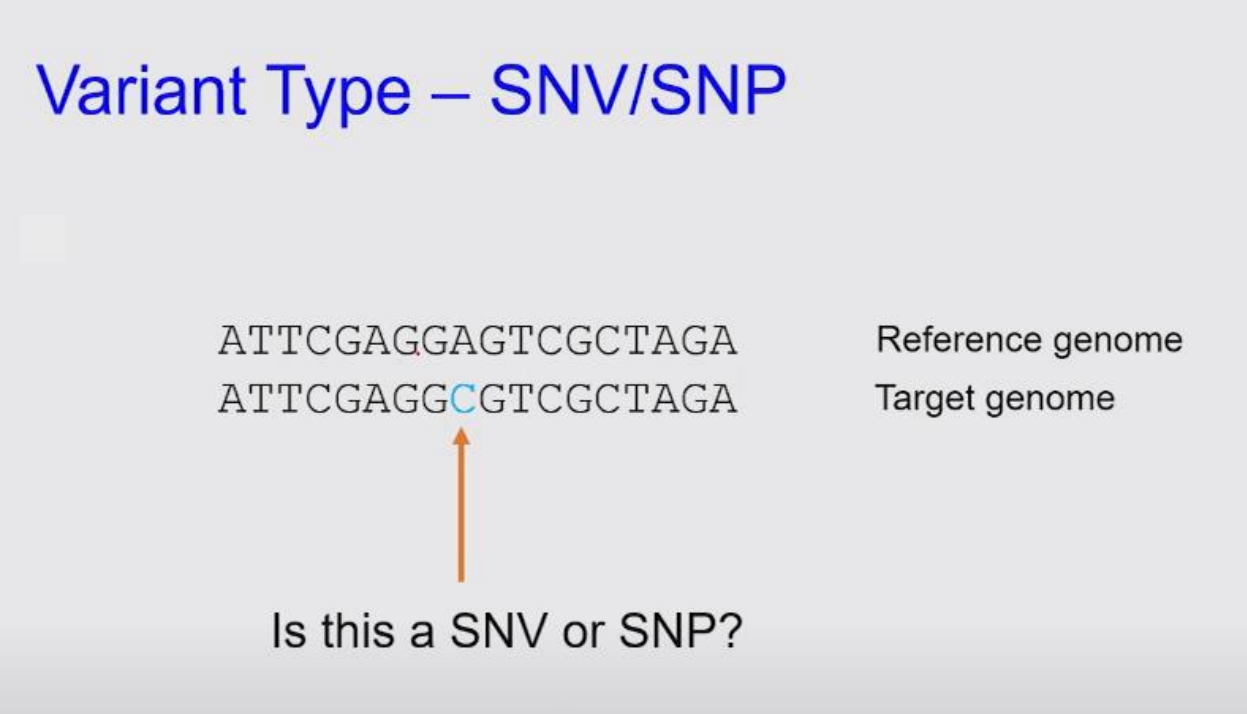



So, we will now focus on this genome analysis, which is highlighted here. So, the problem of identification of these variants is called the variant calling OK, and there are different types of variants that are possible. You have SNV or SNP; we have indels; we have CNVs; and we have SVs. So, what are these SNVs? So, these are the full forms of expression. So, single nucleotide variation, or SNV, then you have single nucleotide polymorphisms, or SNPs, then you have insertions and deletions, which are called indels in short, and then we have copy number variations,

or CNVs, and we have structural variations, or SVs. So, we will come across all of these one after another, but today's discussion will be mostly focusing on these SNVs and SNPs, and I will also introduce these SVs and CNVs in very brief terms. So, what is variant-calling? So, variant calling is a process by which we identify these variants from the sequence data that we have against a reference genome sequence. So, we need the reference data, we do the mapping, and then we do the variant calling OK, and as you can probably understand from this, this is a comparative analysis, which means we do a comparison of a target genome against a reference genome. So, let us now discuss these different variant types. What are the major variant types that we can see when we do this variant analysis for different types of genome? So, researchers have done this kind of analysis for a long time now, and they have seen different types of variants, which are actually quite common when you do this kind of analysis.

So, we also call them mutations, as you are familiar with this term. So, the first variant type is the SNV, or SNV OK. So, here is a very small example of the reference genome, as you can see on top. So, we have the reference genome here, and then we have the target genome, or the genome that we are working with, which we have mapped against the reference genome. So, here you can see this space in blue that is actually different from the reference genome, right? So, this is what we call SNV or SNP, right? So, single nucleotide variation or single nucleotide polymorphism.




Now, you can have more than just one such change; you can have multiple changes in nearby regions, and we call them MNVs, or multiple nucleotide variations. These are a very small number of changes that may be 3–4 at most and not more than that. So, we have used these two terms, SNV and SNP, and you might be wondering why these two terms mean the same.

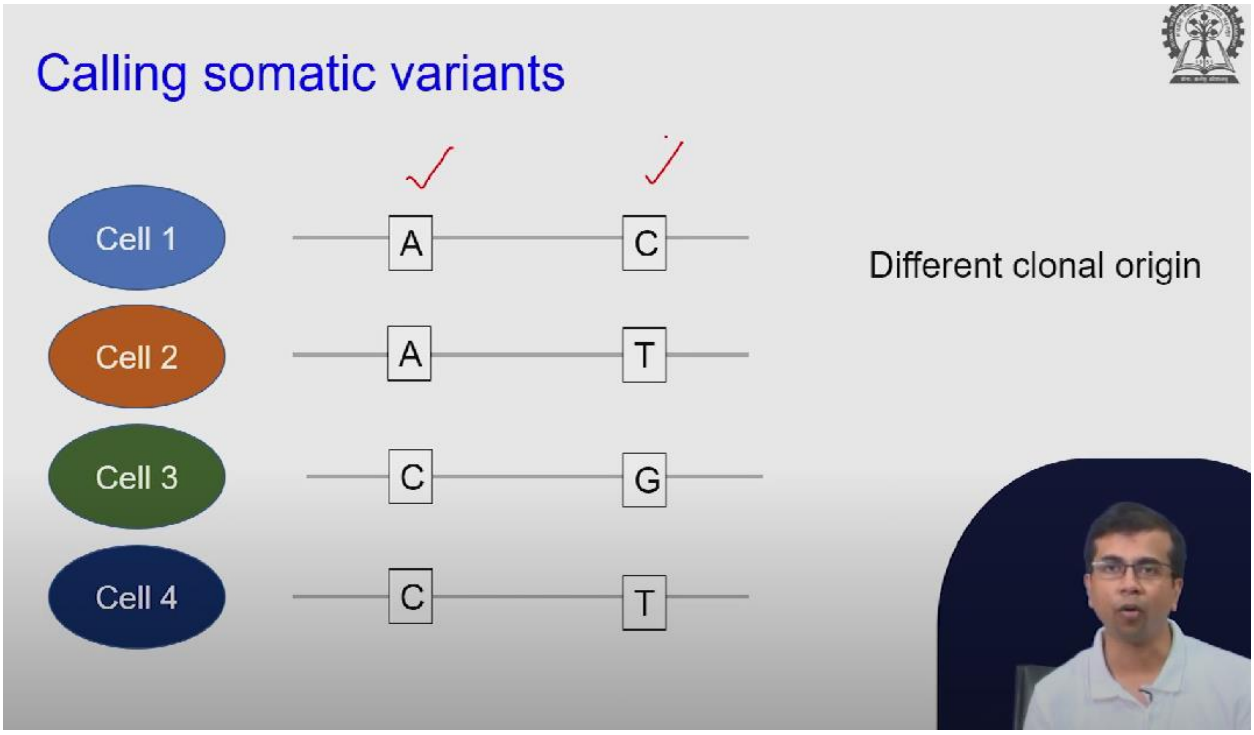
So, we use these terms actually often interchangeably, and people think they are the same, but there is a slight difference. Again, it depends on the person you are asking. So, again, there are different conventions depending on the researcher. So, some will call SNP a variant that is present in at least 1 percent of the population. So, let us say you are studying a variant like this and you observe this variant in some individuals in the population, whether this variant is present or not, and then you calculate the frequency. And if that frequency is greater than 1 percent, then you call this a single nucleotide polymer.

So, right, this comes from a population genetic perspective, but on the other hand, there is another convention. So, SNP is used for germline variants, and SNV is used for somatic variants. So, what are these germline variants and somatic variants? We will discuss them in a moment in the next slide, but again, how do you know which meaning is being used? So, this is something that, again,

you would have to ask the person who is using these terms. So, again, when you are using these terms, it is actually better to define them yourselves. What do you mean by SNV, or what do you mean by SNP? So, because of this confusion that you have out there, it is better to define that before you start writing or analyzing anything with SNP or SNV. So, now, the question is: is this SNV or SNP again? As I said, this depends on the convention; it depends on the researcher that you asked about. So, we have introduced these two terms: germline and somatic variants. So, what are these germline and somatic variants? So, let us define those. So, germline variants are variants that are inherited from parents.

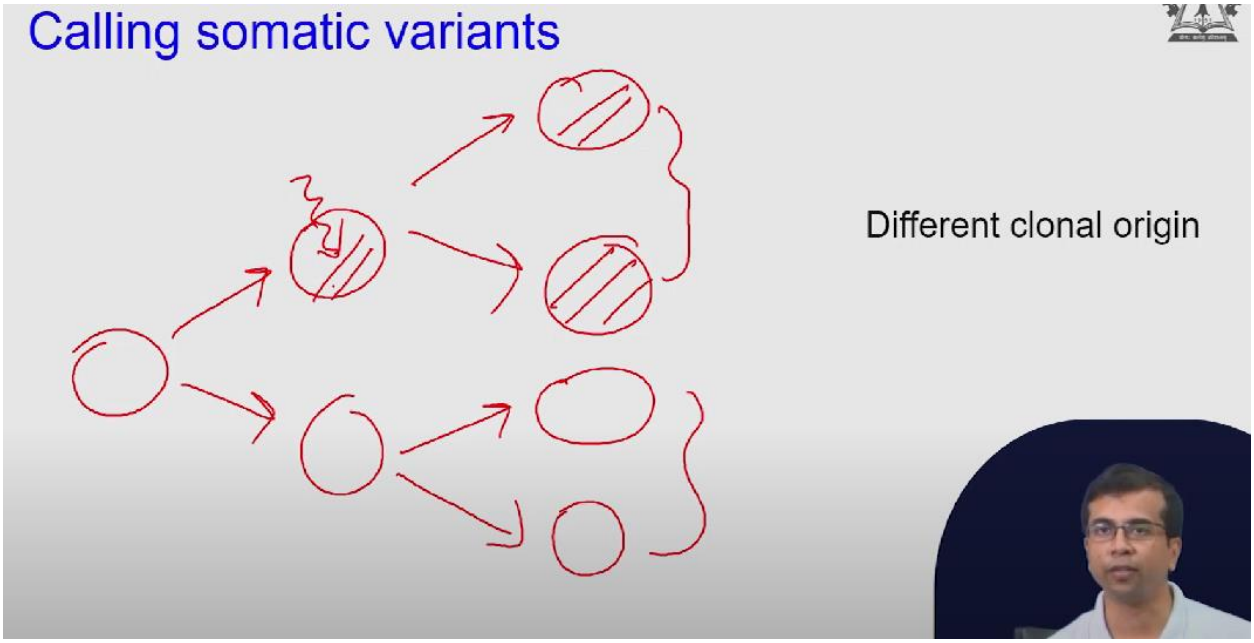
So, as you know, right in the zygote we have one copy of the chromosome from the father and one copy of the chromosome from the mother, and these variants are actually inherited from the parents. So, these will be present in the zygote, and they will be subsequently present in all cells that come out that are arising from the zygote. Somatic variants, on the other hand, are variants that are generated during the development process during divisions of the cells from the zygote. So, it is not that these variants are not present in the zygote; they are originating during the cell division process, right? Because these changes are happening, these mutations are happening. So, these variants do not affect the germline, and that is why they are not inherited. So, if you see, let us say you sequence the genome from tissue samples. Some tissues might have some variants, whereas other tissues might not have those variants. So, it is because, again, these tissues originated from different cells, and that is why some cells have these variants and some cells do not have those variants. So, I hope this is clear: what are germline variants and what are somatic variants? So, in our body, germline variants will be present in all cells, whereas somatic variants will be present in only certain cells or certain tissues. So, and again, these variations can be different from one tissue to another, ok? So, how do you call these somatic and germline variants? So, this is important when you are doing the variant calling: how do you differentiate between these two? So, we can actually remember the definition that we just used. So, let us look at this, like, how do you actually call somatic variants? So, in somatic variants, what you will see is that, if you consider these 4 cells, if these 4 cells have different variants in these certain positions, these variants will be called somatic variants. So, you can see this, we have a kind of giving examples of two positions here, and in one position, cell 1 and cell 2 have this A right compared to the reference sequence; this is a change right, that is what we are assuming, and cell 3 and cell 4 show C base ok. So, again,

this is different between this cell 1 cell 2 and cell 3 cell 4, and similarly, if you look at the other position again, you have a different right. So, you have C, T, or G, and this arises because these cells perhaps come from different clones, right? So, what do I mean by this different clone origin? So, during the cell division process, you can imagine this.




So, you have this cell 1 dividing into 2 right, and this will again go on right. Now, imagine this mutation process or this variant arises here in this cell. Okay, this cell gets this mutation. So, what we will see is that this variant will be present in this one as well as this one. So, they are originating from this cell. So, these two cells here will also carry the same mutation, whereas the other two here right at the end here will not carry this mutation. So, they are arising from two different clones. So, that is why you see this difference, ok, and this is happening inside a body right where the cells are originating. There is this cell division happening right. So, some cells will have certain mutations, whereas other cells will not have this mutation. So, that is why you will see these different variants for the same position within an individual, and this is called the somatic variant. So, somatic variants are very often seen in cancers and tumors because there is a lot of variant mutation going on.

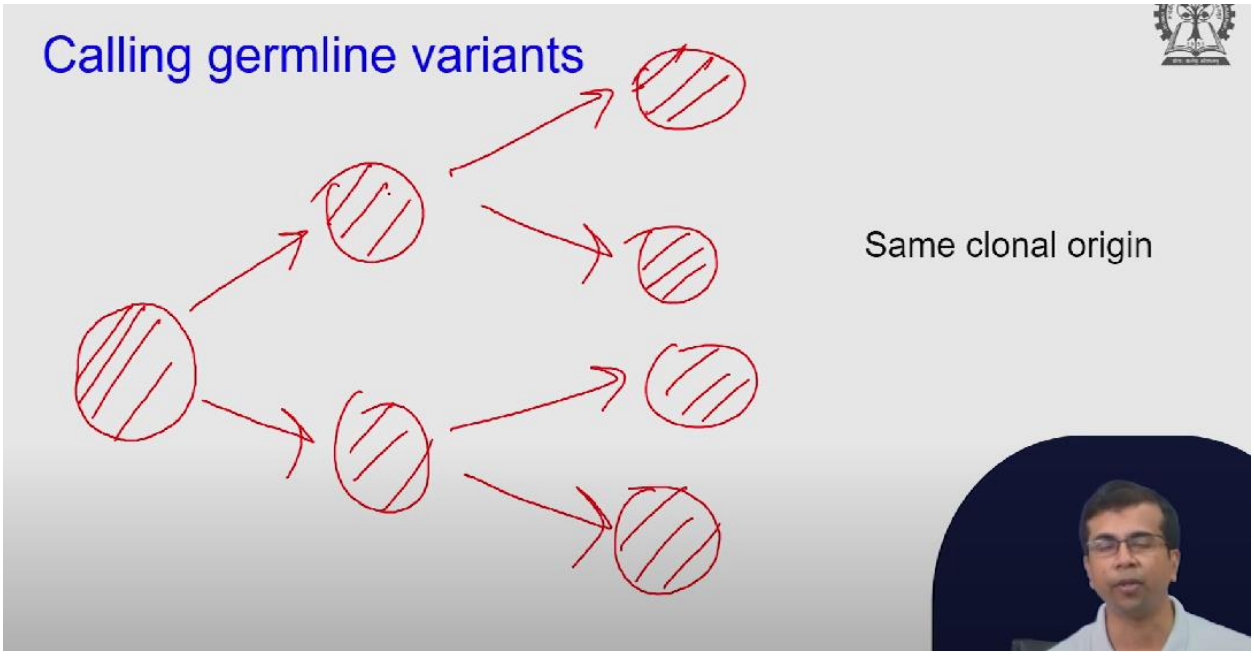
So, they generate a lot of variants, and again, these variants originate from different clones inside the tumors. So, you will see a lot of these somatic variations in tumor cells if you sequence them.



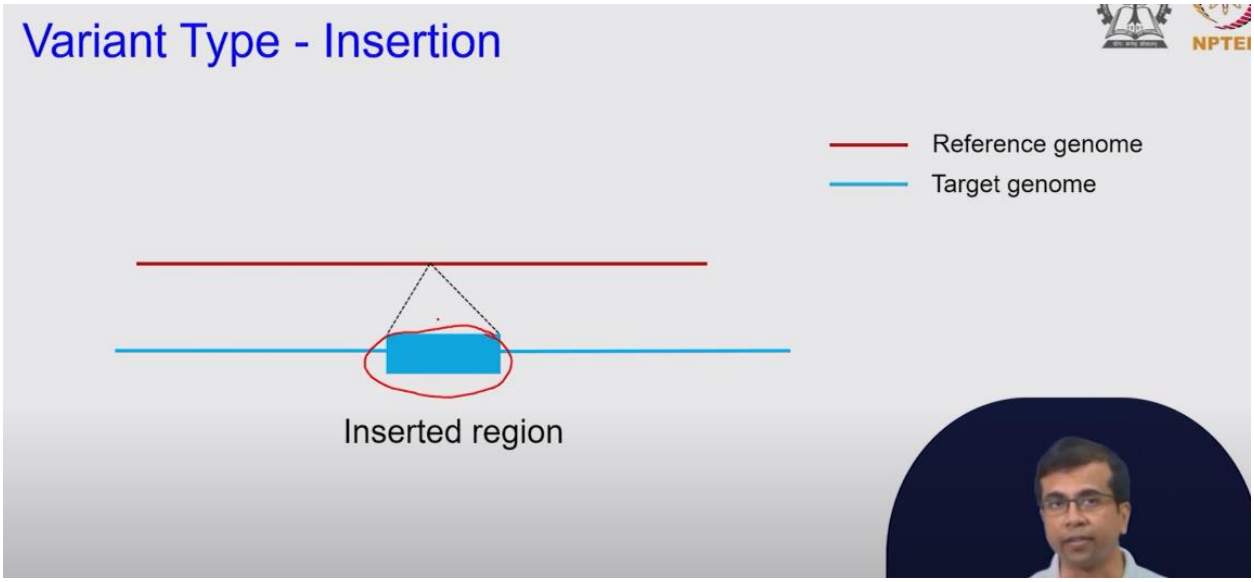

So, now, we are moving to germline variants. So, you can probably now understand how to identify these germline variants. So, in germline variants, they will show the same mutation compared to the reference sequence.

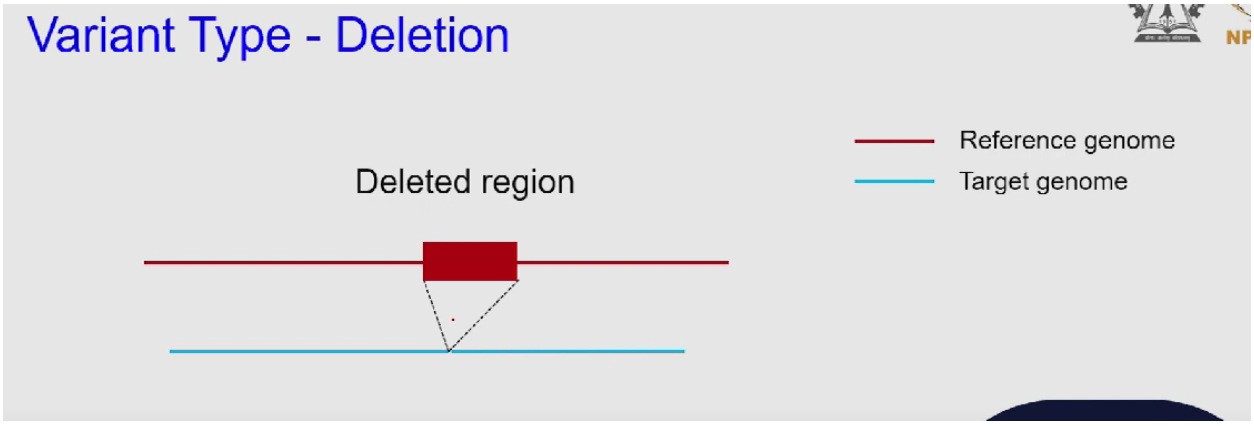
So, if you take all cells, they will show the same mutation. So, again, we are taking these examples of two positions, and we are assuming these are different from the reference sequence. As you can see all cells show A in this position and the G variant in the other position. So, this is because they are originating from the same clone right? So, as I mentioned, these cells originate from these variants that are present in the zygote. So, they are originating from the parents.

So, this cell will have this variant presence right, and it is dividing them ok, and what this means is that all cells that are coming from this one here right will be carrying this variant or this mutation ok. So, you can now imagine the Zygote. So, thus, at the single cell set, this mutation is present, and as the development process is going on, the cells are generated right. So, you are getting the cell division. So, all the cells originating from this one will have these mutations, okay? So, that is why these are called germline variants. So, these will again be passed on to the next generation, ok? So, now we talk about the next variant type, which is small indels. So, we can see insertions and deletions.

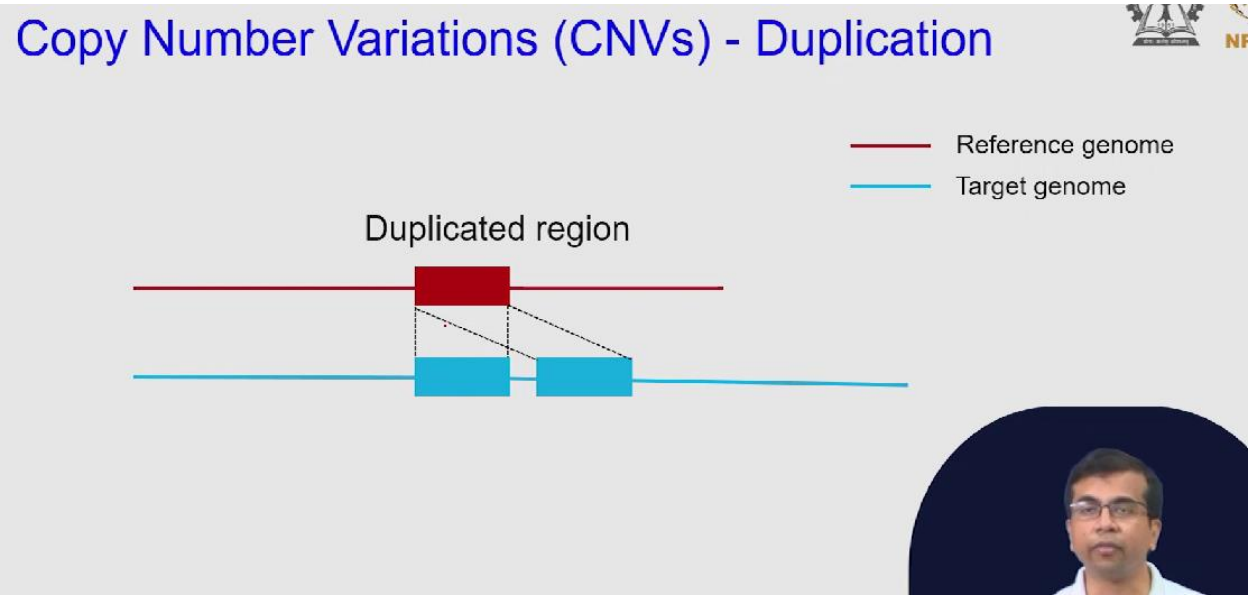



So, let us talk about insertions. So, we are talking about mostly small insertions and deletions. So, you can have very large-scale ones that will actually fall under structural variations. So, this small means 8-10, maybe 8-10 base pairs, or at most not more than that. So, what is insertion? So, you can take this example of a reference genome and a target genome. So, target genome means the genome that we are sequencing or analyzing. So, in red, we have the reference genome, and in blue, we have the target genome. So, insertion means there is a sequence of DNA that is present in the target genome and not present in the reference genome. So, you can see this blue region here; this is extra in this target genome, and it is not present in the reference genome.



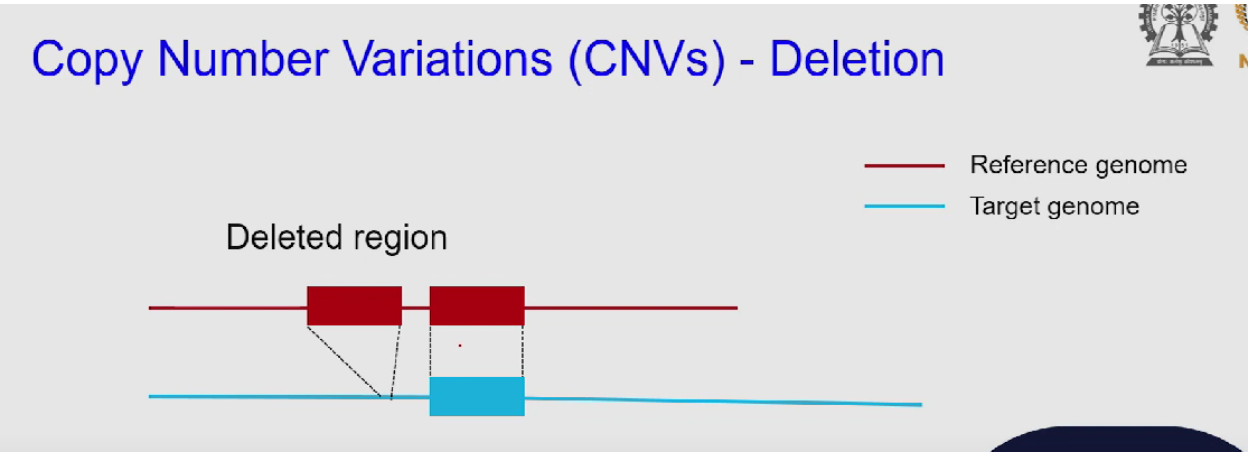

So, this extra sequence has been inserted into the reference genome, okay? So, that is why it is called insertion. Then you also have something called deletion, and this looks like something like this. So, we have the reference genome in red and the target genome, or sample genome, in blue. What happens in deletion is that this region from the reference genome is not present in the target genome. So, this region is deleted from the target genome, okay? So, this is what we call a deletion. Then you also have the structural variance, which we will actually talk about later, and under this structural variance, we have copy number variations, non-CNV structural variance and we also have complex SVs. So, we will talk about these different types of structural variance later on when you go to the next classes. So, just very briefly, what are copy number variations? So, we have something called duplication. Now, this duplication can be gene duplication. So, where a gene is duplicated, it is supposed to be present in one copy, but instead of that, it is present in two copies.




So, this is what we call gene duplication, or you can have segmental duplication. So, what segmental duplication means is that it involves multiple genes or a larger region of the genome that is duplicated instead of a single gene. And similarly, you can have deletions. So, again, do not confuse this with the indel ones it is a large-scale deletion. So, which means you have a gene or segment that can be deleted? So, just to illustrate this copy number variation, So, you can imagine this is a region, a gene, or a segment of the genome. Right in the red, this is the reference genome, and in the blue target genome you have this region present in two copies. So, this is what we will call duplication, okay? So, the general term for duplication is amplification.

So, where you can see even more than two copies of the same region, So, this can happen if we have three to four copies of the same gene in the target gene. So, that will be what we will call amplification. Whatever deletion you make, you can probably guess right.

So, let us imagine you had these two regions, right? So, two copies of the same gene were present in the reference genome, but one copy is missing in the target genome. So, this actually deleted one copy. So, you are left with just one copy in the target genome, okay? So, this is what you call deletion again. This can happen for a single gene, or it could be multiple segments that are deleted.




Then you have different SVs, right? So, we will talk about them later on; we will not be discussing in this class inversion, translocation, and complex SVs. So, complex SVs are actually combinations of these copy number variations, and the other types of SVs are okay. So, this will be discussed in the next classes. Now, we will introduce a couple of terms that we often use when you are analyzing this variance.

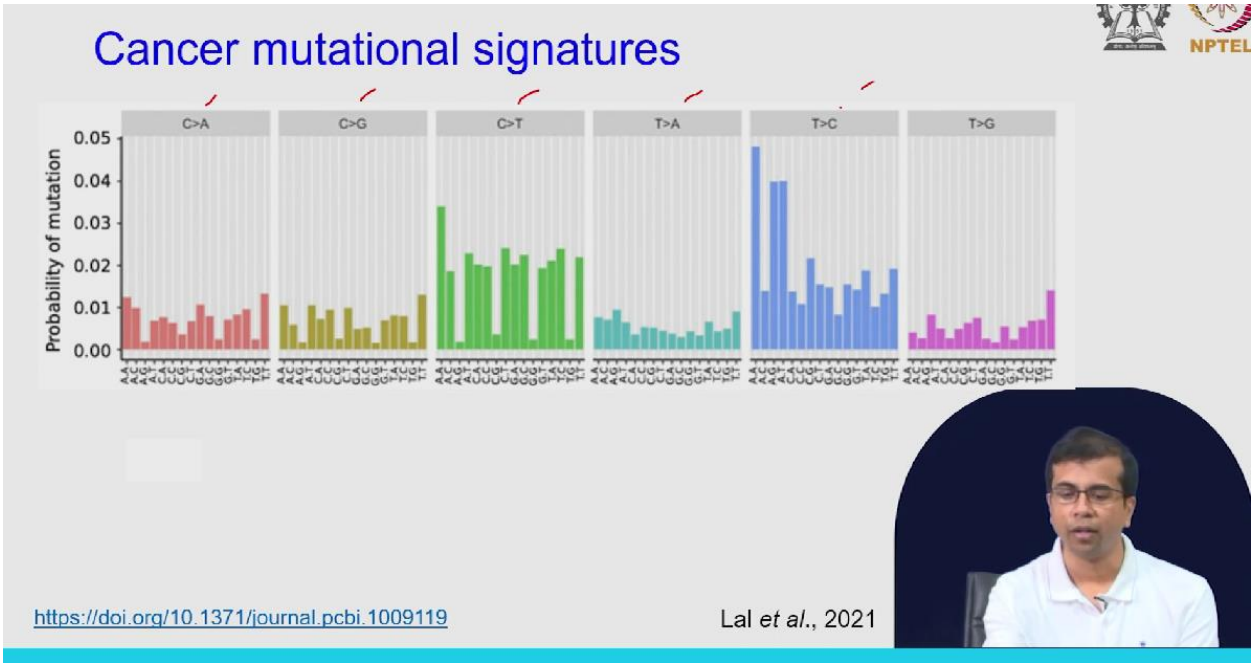
So, one is called targeted resequencing. So, what we do in targeted resequencing is that instead of sequencing the full genome we only target a subset of genes or the regions of the genome that we sequence. So, why would you do that? So, this is important. Sometimes what you will see is that these regions are actually of interest to us. Now, these targeted regions could be exomes, mostly the protein-coding portion of the genome, or they could be some specific gene of interest. So, let us say you are a typing patient and you know that a couple of genes are involved in certain diseases. So, they can give rise to certain disorders. So, in that case, if you are screening patients for that disease or disorder, you will sequence only those couple of genes, right? And this is actually very effective because, instead of doing the full genome, you are just sequencing only a small part of the genome, which is also cost-effective. If you need only the sequence of a part of the genome, you just sequence that region. And also, one advantage of targeted resequencing is that you can actually do deep sequencing of that region to identify the rare variants. So, what are these rare variants? So, imagine you have a population of, let us say, 100 individuals. Let us say a couple of individuals or two individuals carry this variant. So, this will be a rare variant, and only a very small number of the population is carrying this variant. So, once you do deep sequencing, you can identify this rare variant in the populations. Similarly, there is a related term, which is called

amplicon sequencing. So, this is again related to targeted resequencing, where we actually amplify only a specific region of interest and then sequence that region. And again, this is a focused analysis; we are looking at only one or two genes, maybe a small segment of the genome, and we do deep sequencing.

So, this again helps in getting to the rare variants. So, we will now talk about some of the applications of variant calling before we actually jump into the variant calling pipeline in the subsequent classes. This will give you a better understanding of why variant calling is important. So, one of the first applications is clinical cancer genome profiling. So, what do you mean by this cancer genome profiling? So, as we have seen and as I have discussed, in tumors, a lot of these mutations have accumulated. So, this is something that we see: there are a lot of genetic variants present. So, once we want to do what we want to do in cancer genome profiling, we want to study these variants. What are these variants that are accumulated inside tumor cells? Because some of these variants might be important for some of the phenotypes of cancer, maybe they might help in immunization, or they might help in drug resistance or therapy resistance. So, you know, many times the anti-cancer therapy does not work on a tumor, and we now know that there are some variants that kind of are responsible for this resistance. So, we can study this germline and somatic mutations from tumors. So, as I have told you, tumors actually accumulate a lot of somatic mutations during the cell division process, and we can study these somatic mutations, and by studying them, we can actually identify the driver mutations. So, what are the driver mutations that are actually driving the progression of tumors through different stages and driving those phenotypes right—the drug resistance or immunization phenotypes, etcetera? We can identify different subtypes of tumors. So, different subtypes will show certain characteristics, and they will also respond to different therapies. So, by studying these variants, we can actually do something called subtype classification, and that can help in selecting the appropriate therapy. So, there is this term we introduced earlier, which is called precision medicine. So, in the first few classes, we talked about applications of NGS and precision medicine. So, precision medicine means we had to develop therapy depending on the genotype. So, if certain mutations are present in a tumor, we give certain therapies; if other types of mutations are present, we give other types of therapies. So, this kind of variant-calling process is very important if we are doing this type of identification of the genetic basis of these diseases and also identifying the therapy. So, here is a paper that you can

read that discusses the validation of this clinical cancer genome profiling based on DNA sequencing methods, and this would give you a good overview of the whole process. We can also study cancer mutational signatures. So, what do you mean by this? So, here is an example of mutational signatures.

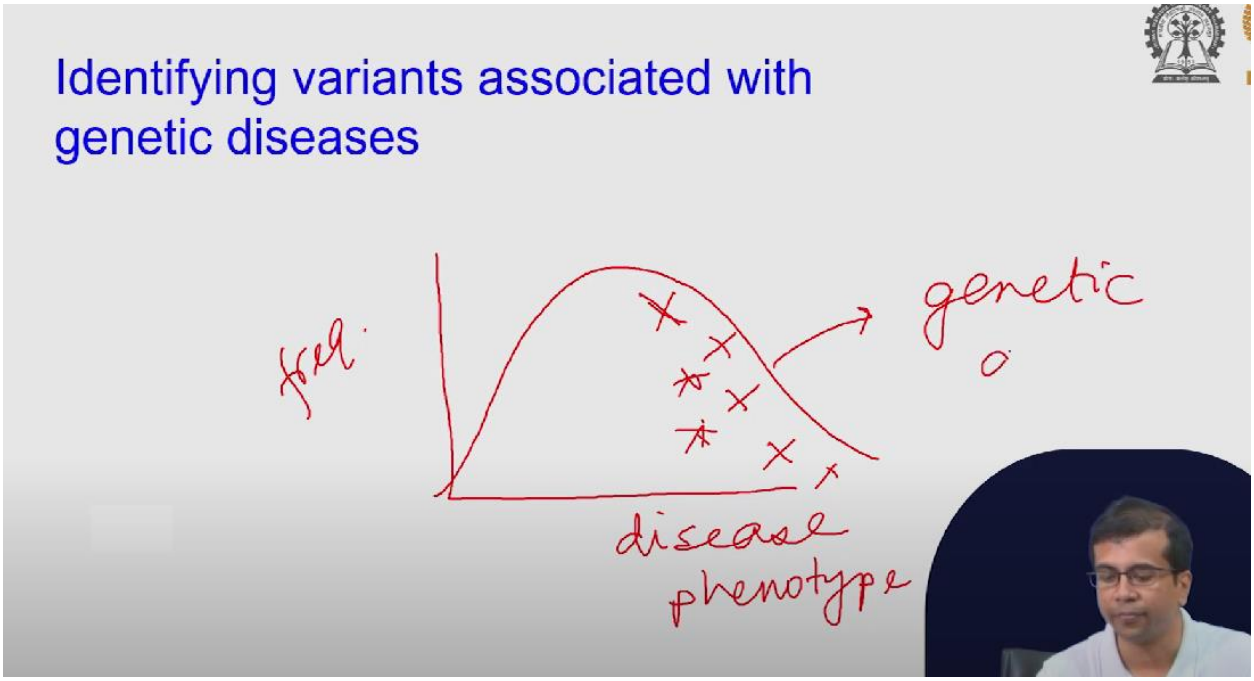
So, we see certain different types of mutations in cancer, and we study them in the context of the genomic sequence. So, what you see here is that we have these different types of mutations, C to A, C to G, etcetera, and what we are looking at is the frequency or number of occurrences of these mutations in certain types of cancer. So, this is for one type of cancer, and as you can see, we see a lot of this C to T mutation, and we also have a lot of this T2C mutation. What else you can observe here is that this C2T mutation, or T2C mutation, occurs in certain contexts much more often than others. So, for example, you see here that this T to C mutation is much more common if you are looking at this AA context, right? So, you have this ATA sequence, and this T is often mutated to C, ok?




So, again, we can study these context-specific occurrences and identify the different mutational signatures that are present in different cancers, and again, we have a lot of data. A lot of analysis has already been done, and we know these mutational signatures for different types of cancers across different patients. We can also identify germline variants that are present in individuals by

just comparing the parent's genome to the child's genome. So, if you are studying one individual and you compare the variants that are present in the parents, you can identify the germline variants, and of course, you can also identify the somatic variants. Now, one important application of identifying germline variants is actually to identify variants that are associated with genetic diseases. So, there are several genetic diseases for which we know that different types of mutations in certain genes are responsible.

So, by analyzing these variants, we can actually associate some of them with genetic diseases. So, we can imagine that you have a situation where you have specific variants we see associated with a disease. So, I can do what I can do; I can draw something like this, ok? So, let us say you have this is the frequency right and this is the disease phenotype right, and maybe this is something that the distribution looks like this ok and some of these mutations right. So, so frequency of individuals right, they are showing this different, this is phenotypes, and we see certain mutations are associated with very severe disease phenotypes, ok, and that might mean ok, these mutations, these variants are perhaps associated with this genetic disease, ok.




So, studying these germline variants can also help us study this genetic disease. So, here is again some reference for your better understanding. You can go and read looking at the disease gene

discovery right association of certain gene mutations in certain genes with certain genetic diseases. So, another application of this variant calling is that we can study genetic diversity in populations, infer population history, how the population has evolved, what kinds of changes have happened, and also associate them with certain events in the history. So, for example, you can study population migration through different continents, etcetera, when you study this genetic diversity.

We will not go into it, but you can, of course, explore more about it. We can also study different cell type diversity using the next generation sequencing variant application a variant calling application. So, for example, we can study T cell receptor diversity again; this is very important for the identification of different pathogens. So, T cells, as you probably know from immunology, have a lot of receptors, and this helps in identification of certain pathogens. So, with this next-generation sequencing and combining it with variant calling, you can study this T cell receptor diversity in great detail. So, here is again a reference for you to read that actually talks about these T cell receptor sequencing technologies and tries to quantify this diversity.

So, here are the references that we have used for this class and to summarize, we have talked about different types of variants, but we have focused mostly on SNV, SNP, and small indels in this class, and there are many more structural variants out there, especially those that we will talk about in the next few classes. So, we will come to the structural variant analysis, and then we will discuss that in much more detail. So, very briefly, we talked about copy number variations, or CNVs, but apart from that, we also have non-CNVs SVs structural variants, and we also have complex structural variants, which are combinations of CNVs as well as non-CNVs SVs. So, we will talk about that in much more detail in the next classes when we talk about SVs, but this is probably something that is clear for you right: there are a huge number of variants that we need to identify when we are analyzing genome sequencing data. Then we talked about these variants and the applications of these variants across different fields. So, we have shown you that there are different applications, and we have very briefly discussed some of these applications, but I encourage you to go into more detail. Read those references, and you will get a better idea of these applications. And we also talked about the clinical implications of this right. So, we can apply variant calling for clinical applications such as mutational signatures in cancer or disease gene discovery for genetic diseases. Thank you.