Next Generation Sequencing Technologies: Data Analysis and Applications

Analysis of CNVs and SVs

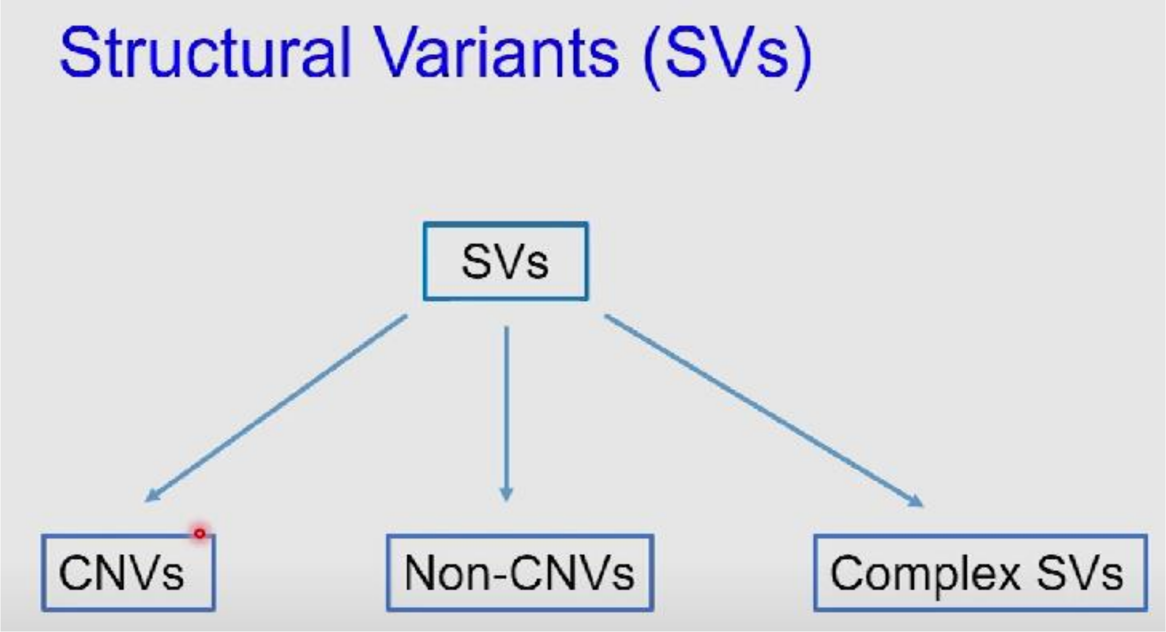
Dr. Riddhiman Dhar, **Department of Biotechnology**

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 looked into the analysis of variance and the different types of variance. So, in the last few classes, we have focused mostly on single nucleotide polymorphisms or SNPs, and small indels. So, in this class, we will look into the analysis of CNVs, or copy number variations, and structural variance. So, let us get started.

So, this is the topic that we will be discussing today: methods for analyzing CNVs and SVs, and we will talk about this in very brief terms, including some of the methodologies and the idea behind these detections. So, these are the keywords that we will come across: CNVs, CNAs, and SVs. So, because of the structural variance, they can be classified into different subclasses. So, these SVs differ from SNPs and small indels. So, in these SVs, they have a part called CNV, or copy number variant. So, here we see only a change in copy number, ok? So, we will talk about this in a bit more detail and what we mean by these CNVs. Then you have the other class, which is non-CNVs. So, again, you have some of these different non-CNVs; we will discuss them in the next few slides and then you have some complex SVs, or complex structural variance.

They can be a combination of these non-CNV SVs or CNVs as well, ok? So, they are combinations of the CNVs, non-CNVs, or different combinations of non-CNV SVs as well, ok?



Structural Variants (SVs)

SVs

CNVs

Non-CNVs

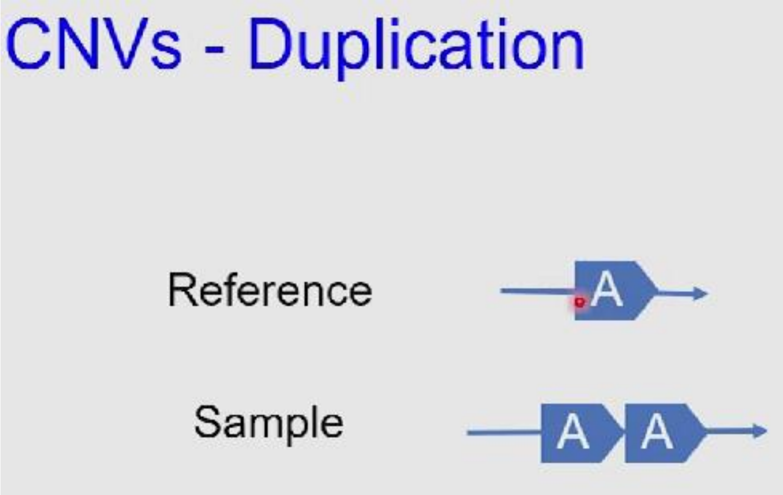
Complex SVs

So, we call them complex SVs, okay? So, what are these copy number variations, or sometimes we refer to them as also copy number alterations or CNAs? So, they are actually changes in the copy number of a segment of 0. So, this segment could be a gene or could be multiple genes in a small region of a genome. So, there could be gain or amplification. So, this is the general term. So, we can use the term duplication. So, where we have increased by one more copy, instead of having one copy, you have two copies now.

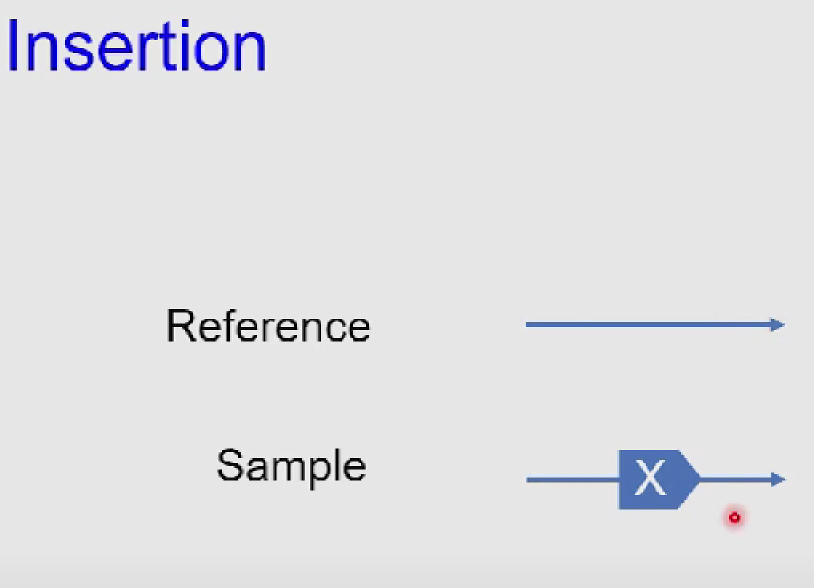
So, that will be duplication, but the general term is gain or amplification. So, you can increase by more, right? So, it is not just 1 to 2; it can be 1 to 3 or 1 to 4, or you can have loss or deletion. So, maybe the gene copy or a region of a genome can be lost, ok? So, here we see a decrease in gene copy number.

So, how do you identify this, and what do they look like? So, for CNV deletion, So, imagine this scenario. So, here you have the reference genome sequence. Here you have this gene fragment A, or gene A, and then in the sample that you are working with, you compare that with the reference, and you see that this gene is not present or that this segment A is not present in your sample. Okay, this is deleted. So, this would be a deletion event ok? So, we will talk about how we identify these different SVs in a moment and what the methods are that we use.

Then you have duplication. So, here, as I have mentioned duplication means you have from 1 to 2. So, here, in the reference, you have a single copy, which is A, that is shown, and in the sample, you have a double copy, right? So, you have two copies.

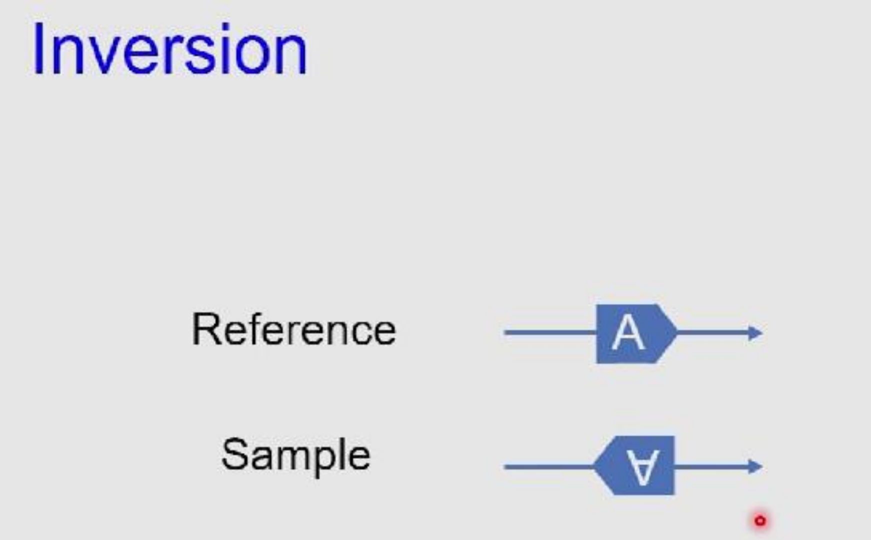

CNVs - Duplication

So, region A has duplicated, okay? So, this could be generalized. So, instead of 1 to 2, you can have 1 to 3 or 1 to 4, ok? So, that will be general. You generally call them amplifications. So, why are we looking at these CNV CNAs? Because these are actually very important events in the genome. So, there are many examples of these in disease samples that also have non-CNB SVs. These are not copy number variants these are other types of structural variants. So, let us talk about those briefly. So, the first one is insertion, okay? So, what happens in insertion? So, you have the reference sequence you can imagine, and in the sample, you have another fragment of the genome inserted. So, that was not present in the original genome; it has been inserted in x region or x gene has been inserted from somewhere else, ok? So, this will be an insertion event.



Insertion

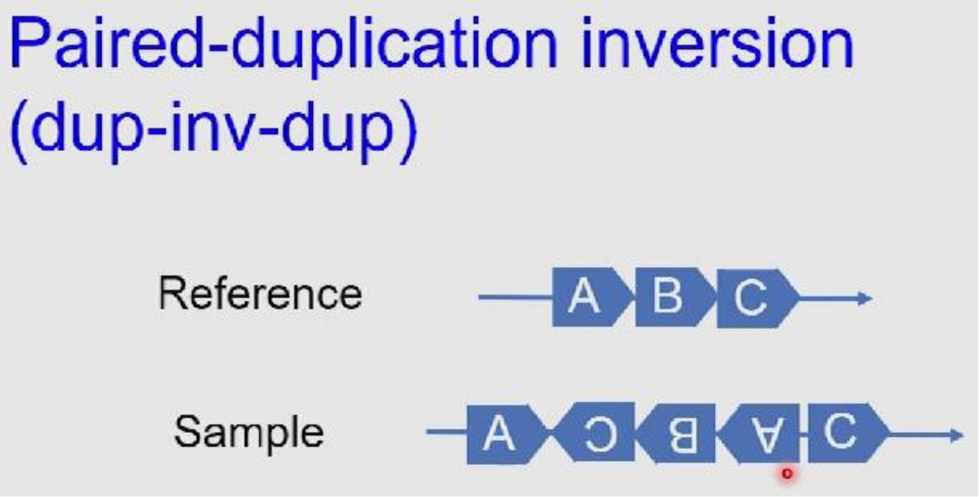
You can have something called an inversion. So, here you can probably see the images, and you will probably understand this clearly. So, you have this reference with this gene A in this direction, and in your sample, you see the gene A direction has reversed right. So, you have changed the orientation, right? So, the nucleotide sequence that has changed is okay. So, the direction has been reversed. So, that is why it is called inversion.



Inversion

The next one is the translocation, okay? So, here is the reference. Now, translocation is a bit more complex.

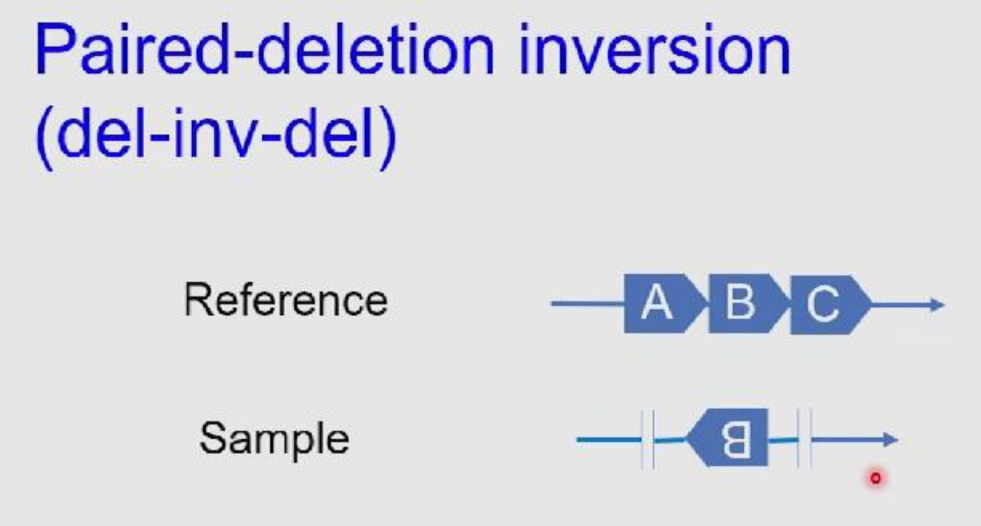
So, imagine you have these two genomic regions, one in blue and one in red. Translocation means one part of the genome has been transferred from this blue region to the red region, and one part from the red region to the blue region. So, that could be the translocation event, and you can see this here in your sample. You can see these blue regions right they are kind of separated out now, and you have the red regions, ok. These are simple S-FIS; now there could be many more complex S-FIS, and these are also found in the human genome. So, the first one is called paired duplication inversion. So, in short, we call it a dup-inv-dup. So, you can kind of remember what is actually happening, right? So, you will have duplication towards the end right, which means region A will be duplicated, and region C will also be duplicated. So, you have two duplication events per pair. So, that is why you have paired duplication, and then you have an inversion event, okay? So, this you have, you have created a copy of A here, you have an extra copy of C here, and this region in between this A B C region is inverted, ok? So, this direction changes. So, that is what we see here, right? The direction has changed. So, if you read the nucleotide sequence, you will see the direction has changed. So, this will be your paired duplication inversion. You have duplication events as well as inversions, ok? So, the duplication is happening at the ends, and then the whole fragment is inverted in between.



Paired-duplication inversion (dup-inv-dup)

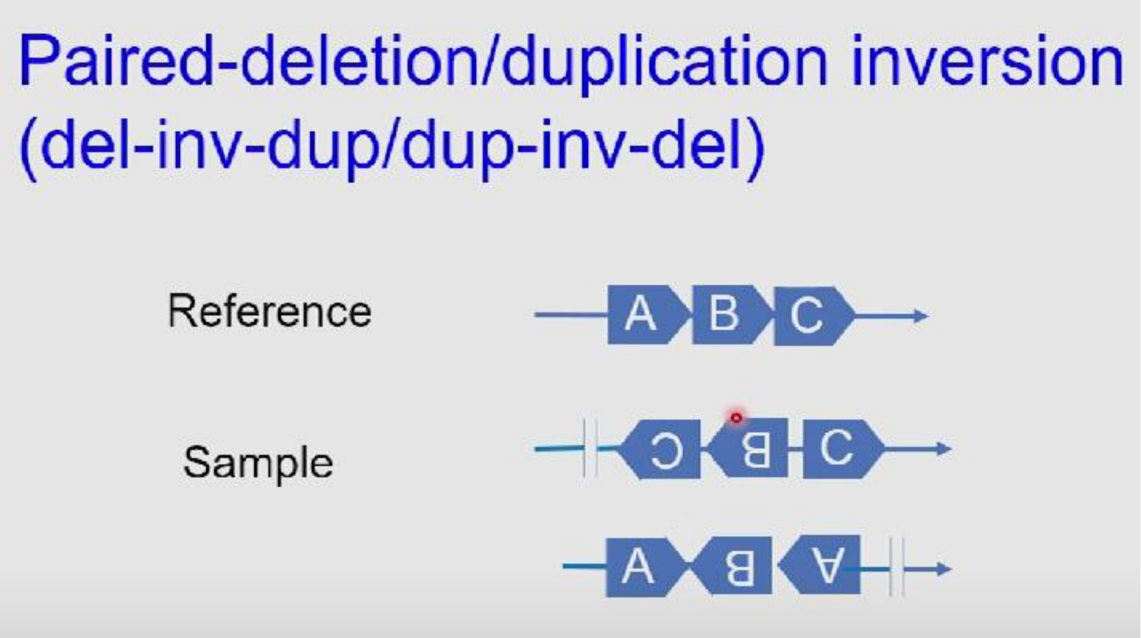
So, similarly, you can have something called paired deletion inversion, ok? So, again, we have a short form. So that you can kind of remember what is actually happening or connect this. So, it is the opposite of what we discussed now right? So, paired duplication inversion.

So, here you have deletion events, right? So, at the ends right, you see that region A is deleted and region C is deleted, whereas the intermediate region P is inverted. So, there are two deletion events and one inversion event right?



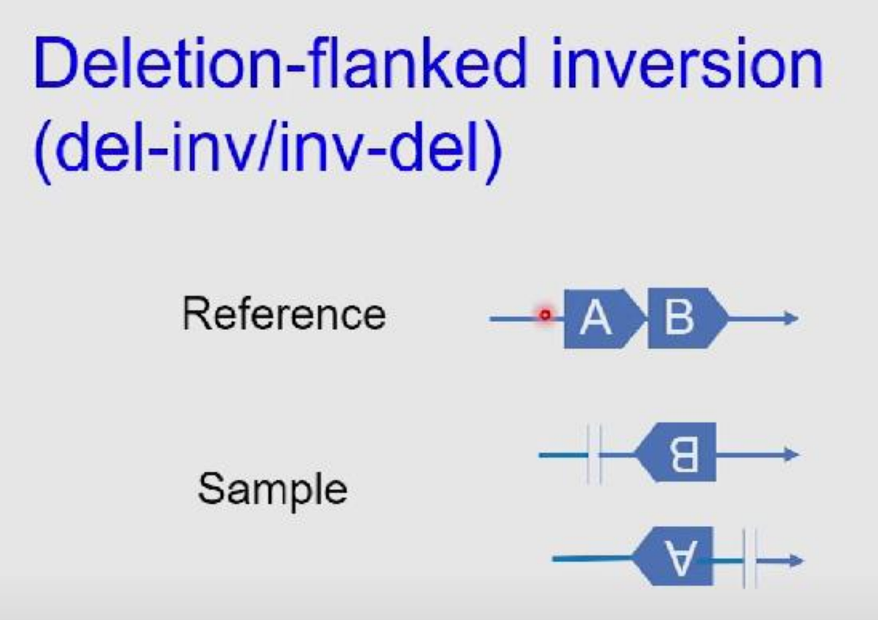
Paired-deletion inversion (del-inv-del)

So, what you see is that this is a combination of this copy number variation, and the earlier one as well is a combination of this copy number variation with a non-CNV SV. So, it is a combination of both. Now, you can also have something called paired deletion duplication inversion, right? So, here what you have is the reference: is this A B C gene okay? Now, compared to the earlier cases where you had either duplication at both ends or deletion at both ends, here at one end you had deletion and at the other end you had duplication. So, if you see this in the reference, you have A, B, C C. The A region is deleted, right, but the C is duplicated. So, you have created two copies of C by this event, but A region is deleted, so you cannot find any A here, ok? So, this is what is shown here, or you have another possibility, ok? So, A is duplicated and C is deleted, ok? So, here in the lower part, you see that A is duplicated and C is deleted, and then, of course, in between, you have the inversion, ok?



Paired-deletion/duplication inversion (del-inv-dup/dup-inv-del)

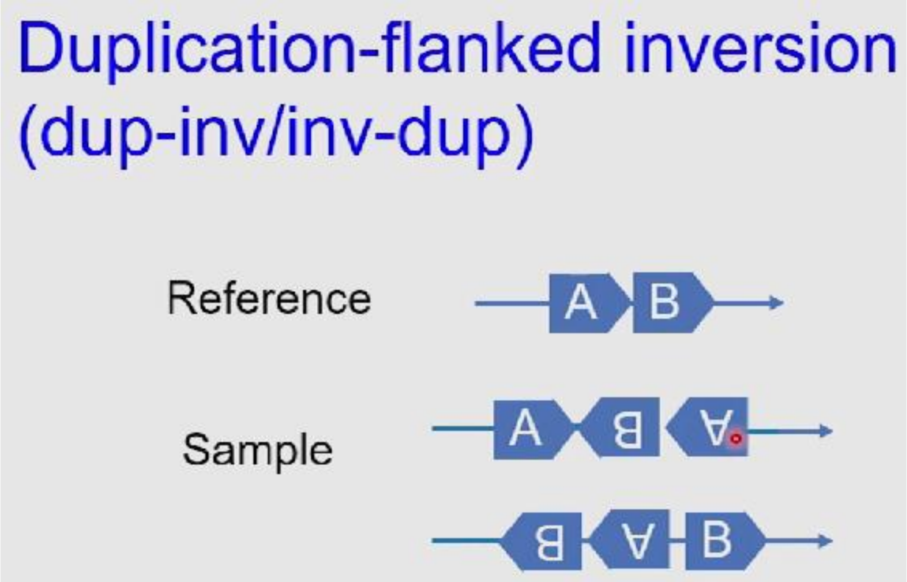
So, this inversion event is there. So, that is why it is called paired deletion duplication inversion. So, you have deletion as well as duplication and inversion, right? So, you have three events occurring together. You can have other types of complex SV. So, this one is called deletion-flanked inversion right? So, del inversion or in del. So, if you look at the reference sequence, is the A B right these two regions, and what you see is that one region is deleted and the other region is inverted, ok? So, you have one deletion event and one inversion event, ok? So, you see deletion of gene A here or region A here and inversion of region B, and in this case, you have



Deletion-flanked inversion (del-inv/inv-del) inversion of region A and deletion of region B. So, this is another type of complex SV.

So, we can look at the other one right, which is the duplication flanked inversion right.

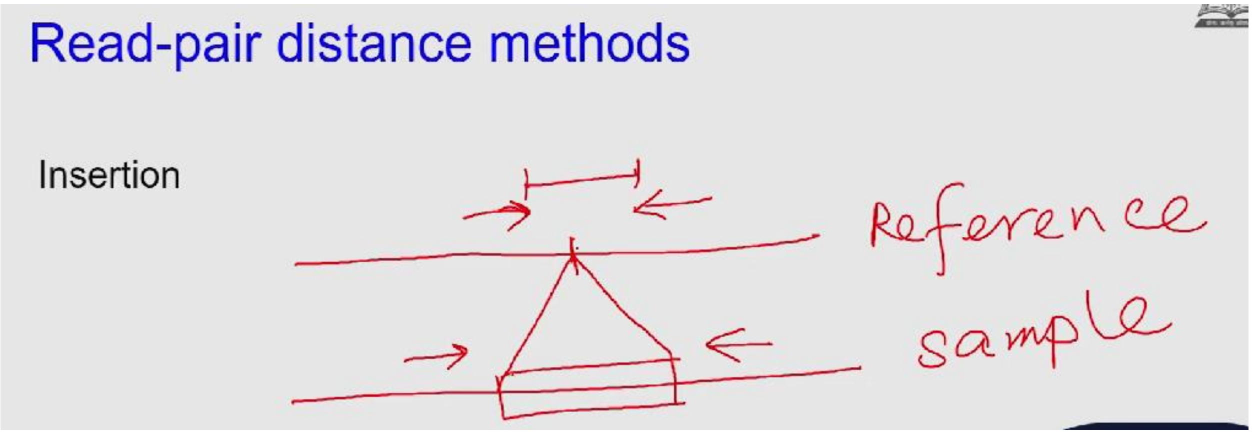
So, we talked about deletion-flanked inversion. Now, we have duplication-flanked inversion, and you can probably guess what it would look like. So, you have this A-B reference sequence, and in the sample, you have a duplication event as well as an inversion event, right? So, in this case, the region A is duplicated and the A-B region is inverted. In this lower case, you have region B amplified or duplicated, and this A-B region is inverted, ok?


Duplication-flanked inversion (dup-inv/inv-dup)

So, you have two events now: duplication and inversion, all right? So, there are different types of S V's, and you can probably find other types of S V's as well if we look for them in the human genomes. So, how do we identify them from the genome data? So, this is one very important area right? So, we have now talked about different types of S-V's that could be identified in the human genome. Now, the question is: how do we actually identify these structural variances, or S's? So, we will talk about some methods that allow us to identify these structural variances. So, before we go there, we should remember that this is a comparative analysis. So, we are working with read data from two sets of samples, and this would be set on versus set to samples, and we want to identify whether there are structural variations in, for example, set to samples compared to set 1 right. So, we need to read data from two sets of samples, and this is something that is really important. So, we will talk about some of the methods that can be applied to call these structural variances. So, the first method is called the read-pair distance method, right? So, the idea is actually quite simple. So, we are working with paired end sequencing data. We have talked about single end sequencing and paired end sequencing. So, here we are working with paired end sequencing data, and what happens is that the distance between the pairs actually changes because of the presence of structural variance. So, it is not just the distance, but sometimes, in the case of other complexes, we can also have a change in orientation.

So, in this case, if you have a change in the distance, that might actually tell us that there might be a structural variance present in the AH sample. So, what we do is compare the actual distance with the expected distance. So, the expected distance we derived from the reference is okay, and if you have a significant deviation right. So, we can identify ok; there are probably SVs that are ah present ok. So, let us take a couple of examples and see how this ah method will work right.

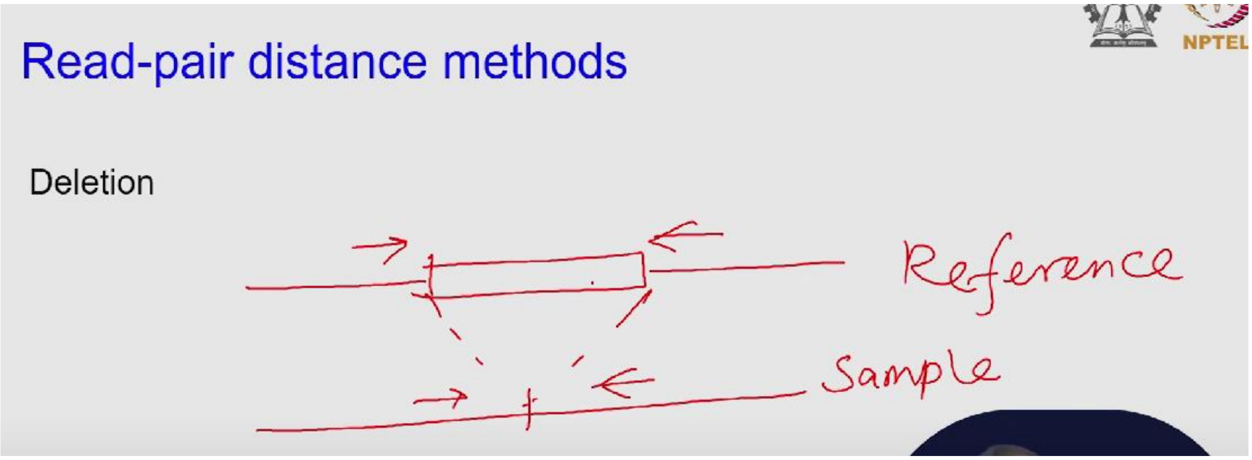
So, let us imagine, right? This is our reference sequence, right? So, we have this paired end that reads right, and this is the expected distance, ok? So, this is the reference, and then we have the samples you know, right? So, this is our sample that we are working with, and now imagine that at this position we have an insertion. Okay, we get this insertion here in this position in the sample. So, what will happen is that for the reads that are mapping here the distance between these pairs will increase. So, the reads that will come from here right this distance will increase because of this presence of these AH insertions around here, ok? So, this is something that we can see, right, and then we will see ok, there is probably an insertion here, ok. So, the expected distance is something here like this, but the actual distance will be something like this, ok?



Read-pair distance methods

Insertion

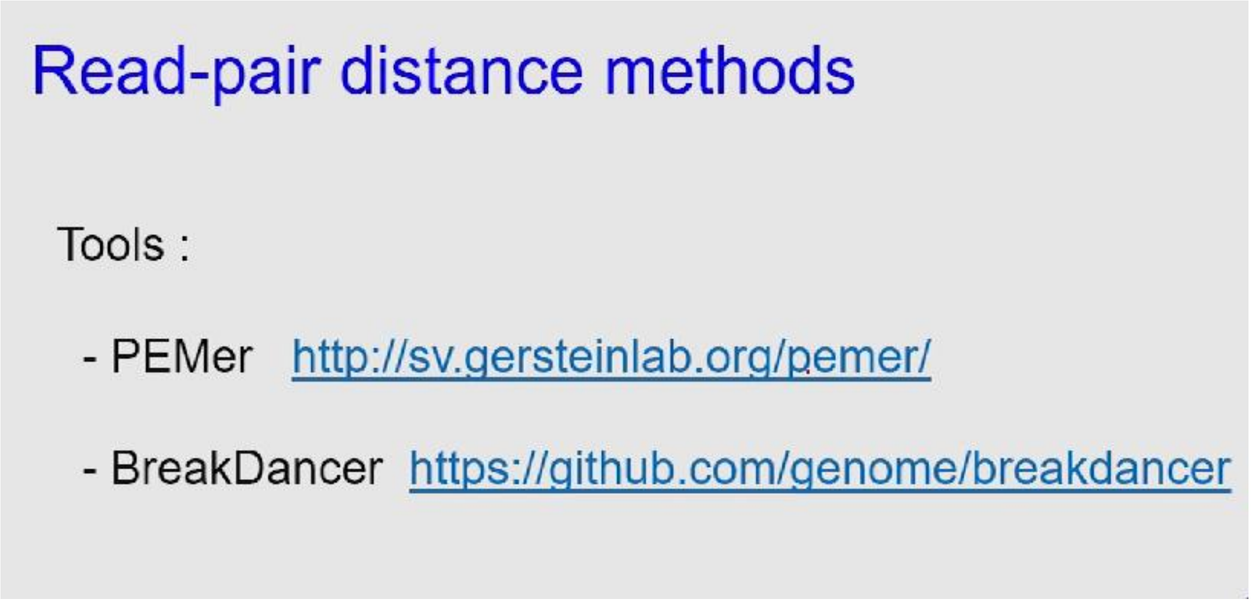
Now let us think about the deletion case. So, what will happen in the case of deletion? OK. So, imagine that this is the reference genome, and imagine that this region is deleted in the sample sequence. So, what will happen is that if you are mapping these reads, the distance between these reads will change and will be reduced because this region is now deleted. So, if we observe this kind of change, we can say, OK, there is this ah difference in the distance, and that might tell us, OK, there is probably an sp that is present in there, ok, alright.



Read-pair distance methods

Deletion

So, now what will happen? Ah, so here are some tools that are able to utilize this kind of method. So, paired and ah, the distance method is right, so, paired and reads ah.

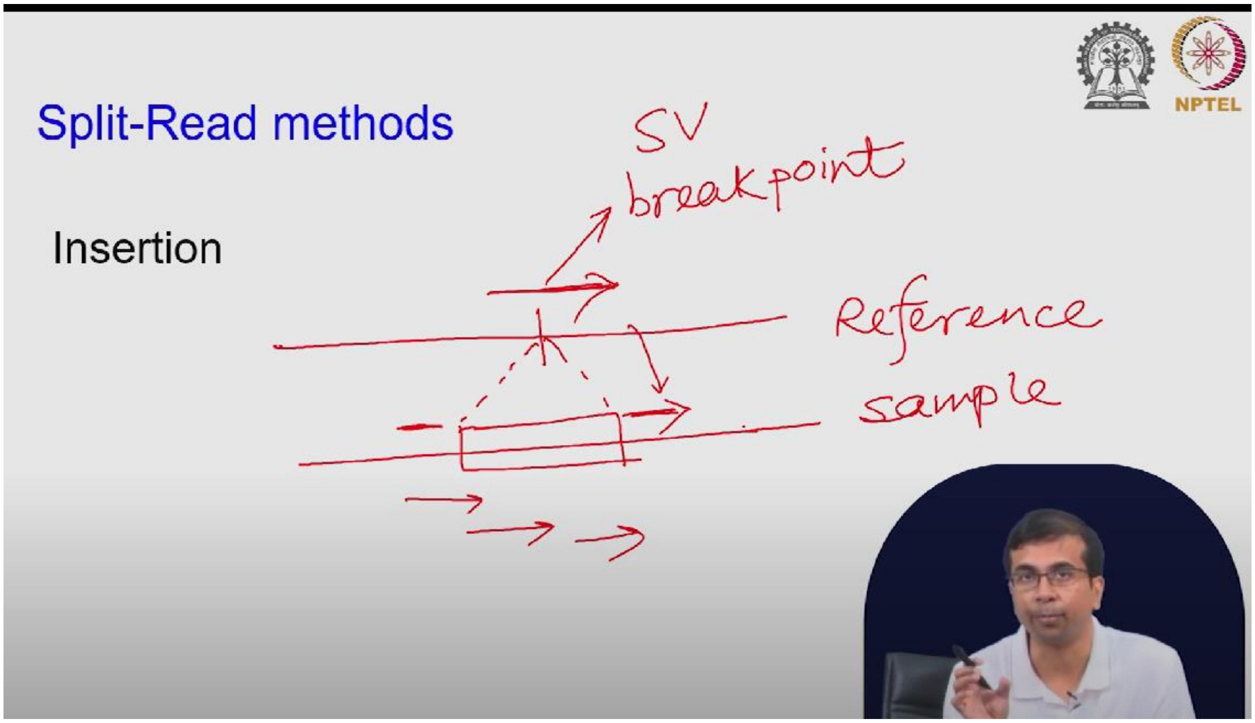


Read-pair distance methods

Tools :

- PEMer http://sv.gersteinlab.org/pemer/
- BreakDancer https://github.com/genome/breakdancer

So, I have given you the names and the links. So, you have this is ah is p ah detection tools, and of course, there are more tools out there and you can explore them ah to understand how they actually work. There are some differences, but the general principle is the same. Then you have a second type of method, which is called the split read method. So, we will talk about them in a moment, ok? So, how does this work? So, what happens is that these are the reads that actually cover the structural variant break point. So, break point is the region where you have this ah in ah structural very ah events happening ok, and we can again look at this in case of insertion right? We are taking very simple examples. So, here is the reference, and here is the sample with an insertion right. So, this is the sample, and imagine that this is the break point, right? This is where this insertion event is happening. So, this is the break point, right? This is the SV break point, ok? Now, imagine there was a kind of read right that actually maps across this position. So, you can imagine you have a read here map across this position. So, what will happen in the case of this insertion? So, part of the read will come from here, and part of the read will come from here, ok? So, because there is an insertion, part of the read will map to this position, and part of the read will map to this position, right?

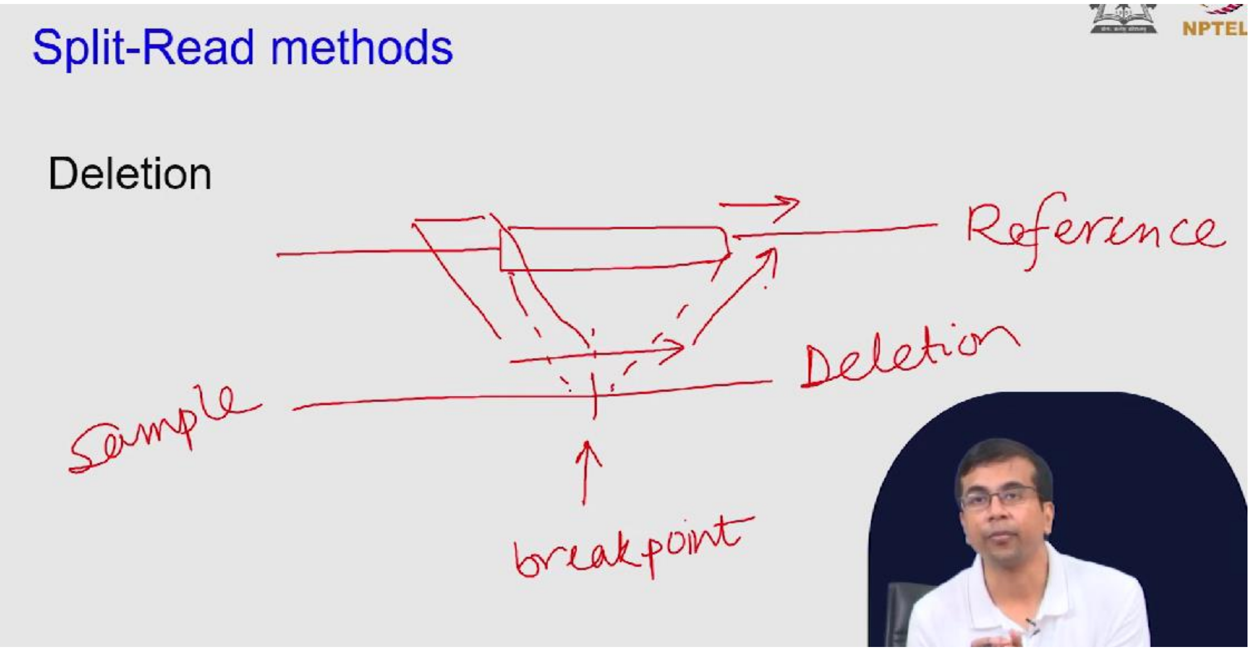


Split-Read methods

Insertion

This is what will happen. Now, if you are generating this sample and mapping back to the reference data, what you will see is that you will not find any reads that will actually map across this s p break point, because the reads from the sample sequence, when they are generated right, will give you continuous reads like this. So, you will not get any reads that actually map across this break point, right? So, this is something that you can identify as right and say, OK, this is expected, but we do not see this kind of reading as right. So, this kind of shows that probably this is where the insertion is happening, and of course, you can also identify the insertion that is happening. Now, for deletion, we can again think about this using this example of a reference genome, right? So, you have a reference, and this region is deleted. So, this is the SV break point, ok, and here you will find some reads right when you are sequencing. Ah, this sample, sorry, this is sample right.

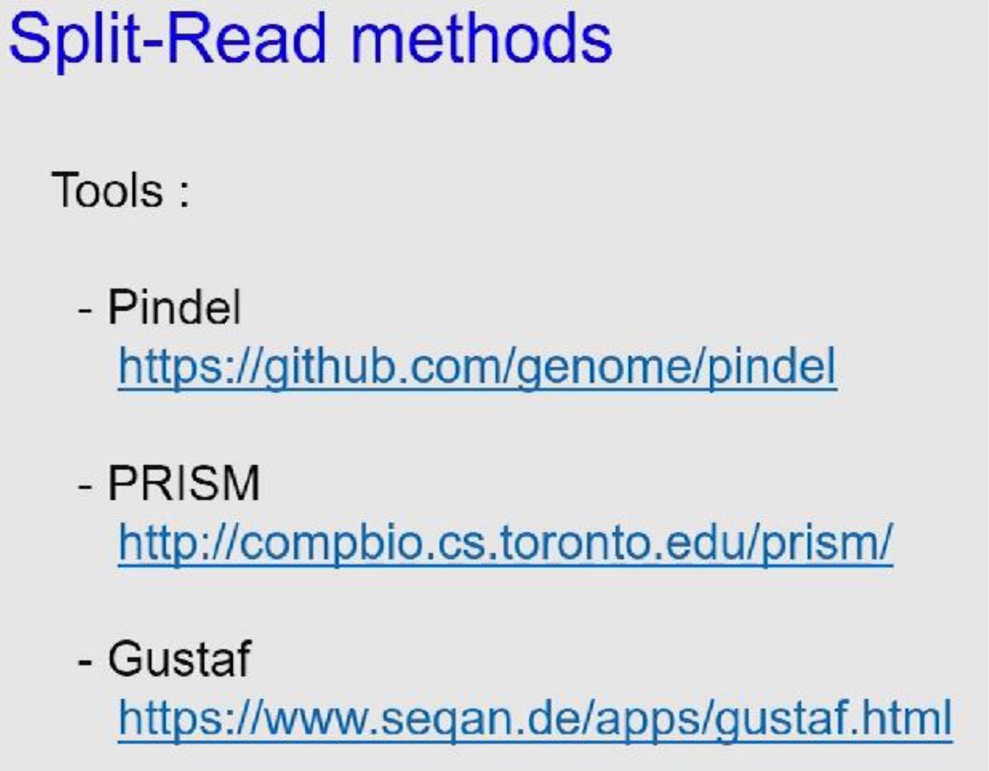
So, when you are getting some reads from this sample data and you try to map it against the reference, what you will see is that this part of the reads will map the two different locations. So, one part will map to this location, and the other part will map to this location, ok? So, this kind of mapping difference will arise, and based on these differences, you can then identify OK. There is probably a SV ah break point here, right?



Split-Read methods

Deletion

This actually has ah in it, which will indicate that there is a structural variant that is present in the sample. So, I think one idea you probably get now in general is that we can use this distance between reads or this split mapping of reads to identify the occurrence of certain fields, and then, of course, we can dig deeper and identify the actual SV that has happened, whether it is insertion, division, etcetera, depending on the different cases or scenarios. In some cases, we can have very complex cases, and then probably again, we have to look carefully at the data. So, here are some tools that actually use this split read method, ah, again, the links I have given.

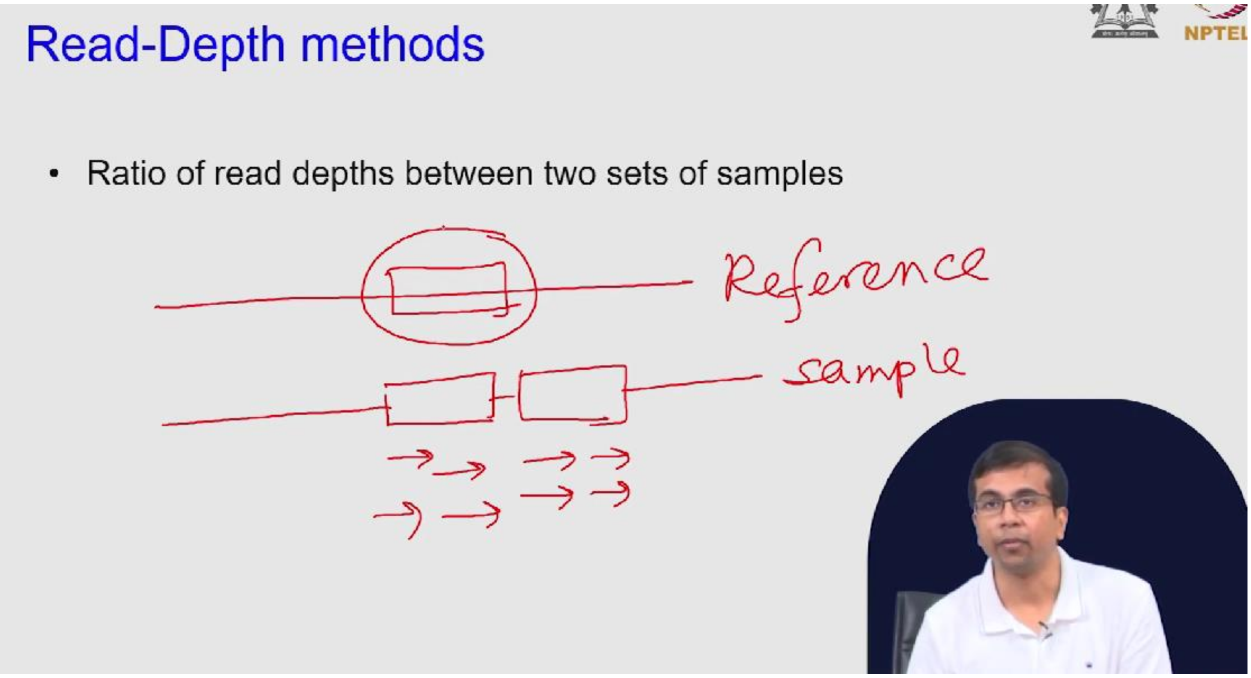


Split-Read methods

Tools :

- Pindel
  https://github.com/genome/pindel
- PRISM
  http://compbio.cs.toronto.edu/prism/
- Gustaf
  https://www.seqan.de/apps/gustaf.html

So, you can look up and see how these methods are implementing these split-read methods. Of course, there are more tools out there, and their implementations are slightly different, but the general principle is the same. The other method the third method is called the read-depth method, and this method is mostly employed for looking at the copy number variations. So, we have talked about copy number variations, and we can use this read-depth method for copy number variation. So, what we do here in this method is look at the ratio of read depths between two sets of samples, and depending on that, we can say that there is a change in copy number.

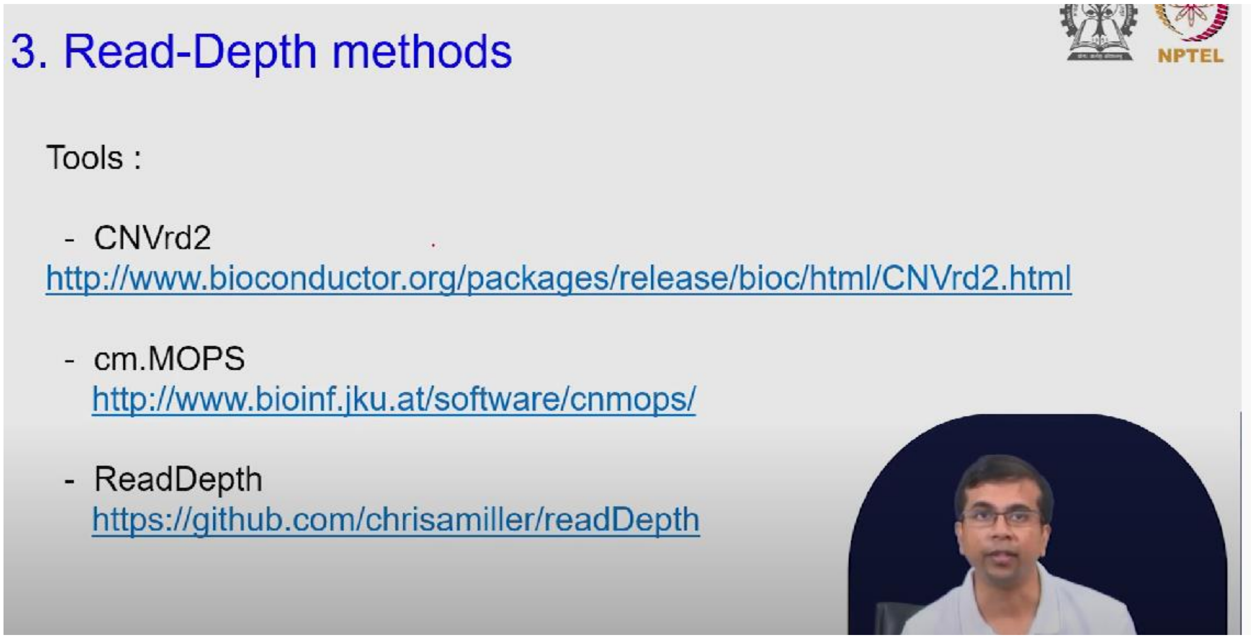


Read-Depth methods

- Ratio of read depths between two sets of samples

So, what you will expect is that if you have, let us say, a gain, amplification, or duplication, you will have more reads coming from that region. So, which means, in turn, you will get a lot more depth in that region, ok? So, you can take a very simple example of this right. So, let us say we have this region in the reference, and this region is duplicated OK in the sample right.

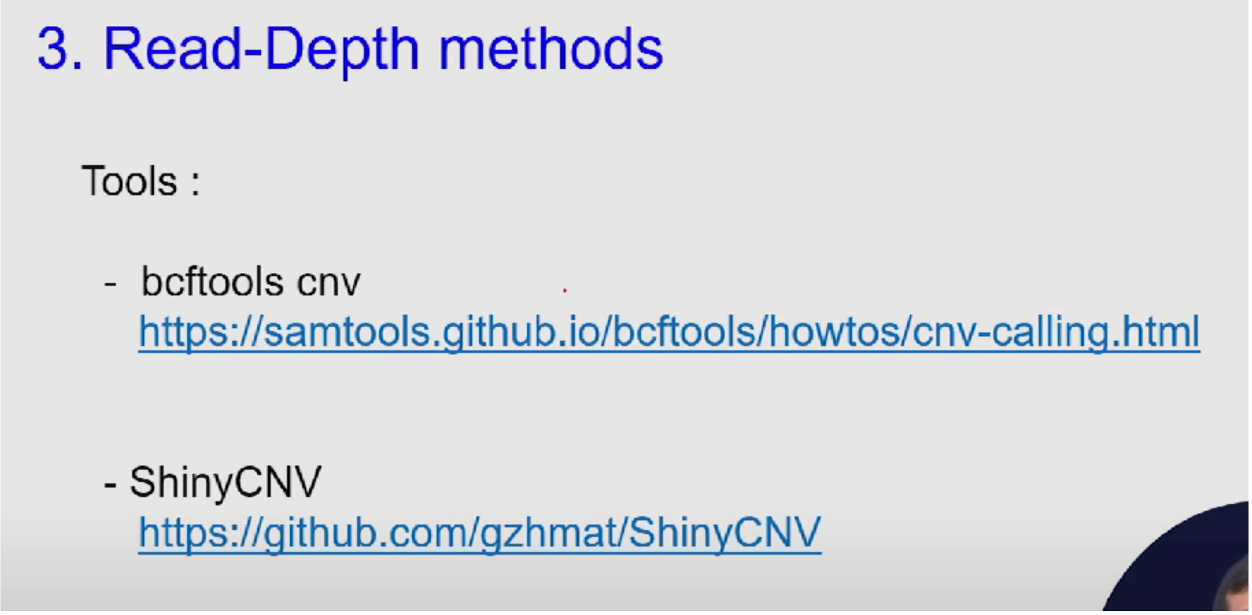
So, we have two copies of this region in the sample. Now, when you sequence this sample, we will get these reads that will be mapping or coming from these regions, and when you do the mapping, all these reads together will map against this location here, right in the reference genome. They will all map to this location, ok? And then, if you now compare the read depth in the reference genome versus the sample you will see that in the sample, you will have more read depth because you have two regions that are generating these reads that map to the same location in the reference genome. So, by identifying this change in read depth, you can then say, OK, there is an increase in copy number or a decrease in copy number. So, in the case of a deletion or decrease in copy number, you will see a reduction in read depth because you will have lost some regions of this right. So, if there are two copies of the region in the reference and one copy is lost, then you will see a reduction in depth, but if you have amplification or duplication, you will see an increase in read depth. So, based on this, you can then identify OK. There is a copy number variation on this, of course, and there are a lot of statistics as well, but this is the very simple idea behind this kind of call OK. So, as I have said, a higher read depth may indicate gain or duplication, whereas a reduction in read depth might indicate loss or deletion. So, this method is mostly used for the detection of CNVs. So, you cannot use this for other types of SVs, etcetera. So, you have illustrated these read-depth methods, and here are some tools that are out there that you can use with these read-depth methods.



3. Read-Depth methods

Tools :

- CNVrd2
http://www.bioconductor.org/packages/release/bioc/html/CNVrd2.html
- cm.MOPS
http://www.bioinf.jku.at/software/cnmops/
- ReadDepth
https://github.com/chrisamiller/readDepth

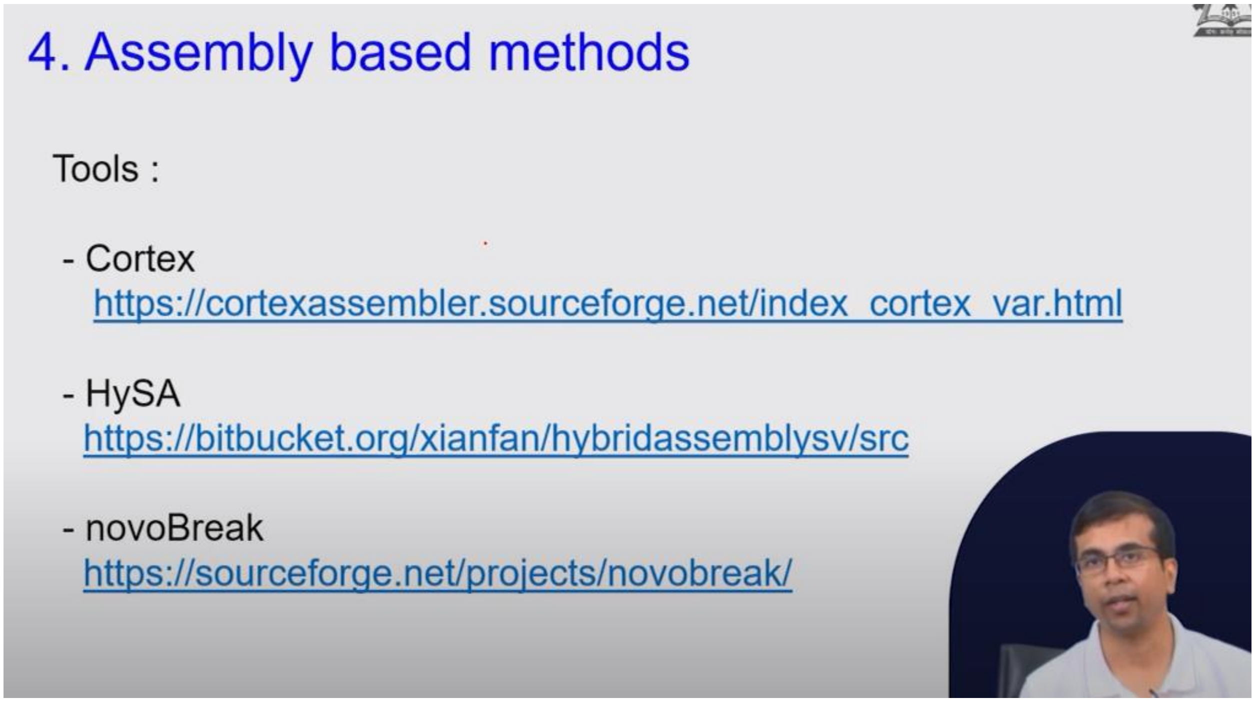


3. Read-Depth methods

Tools :

- bcftools cnv
https://samtools.github.io/bcftools/howtos/cnv-calling.html
- ShinyCNV
https://github.com/gzhmat/ShinyCNV

Of course, there are many more, and again, I have given you the links so you can actually go and read about them. You can see the implementations if you are interested. For a few more tools, right, for example, the bcftools CNV, this can actually be very useful when you are doing the AH CNV calling, as we will see. The fourth method relies on the assembly-based method, right? So, the idea is actually quite simple here: we do this genome assembly of reads from two sets of samples, right? We will talk about the algorithms later on and how we actually do the assembly from read ah to the genome, and if you see any difference in assembly between the two samples, this might indicate the presence of SVs. So, if the outcome is different for two sample trainings, we can probably say that there is some SV present, which is why we are getting different assemblies. Now, one of the limitations of this method is because of the computational requirements. So, genome assembly is a very complex process that requires a lot of computational time and resources. So, maybe this is not the best way to go about it, right? Here are some tools that actually implement these assembly-based methods for SV detection, and again, you can look them up.

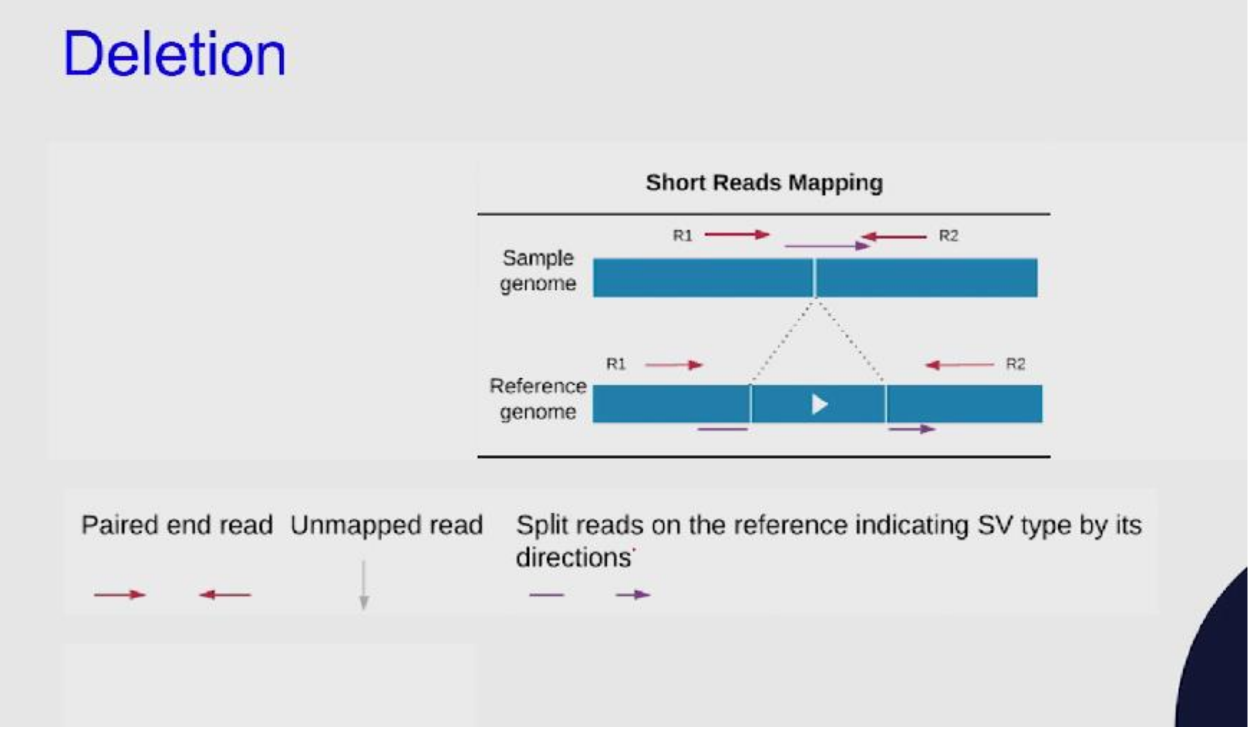


4. Assembly based methods

Tools :

- Cortex
  https://cortexassembler.sourceforge.net/index_cortex_var.html
- HySA
  https://bitbucket.org/xianfan/hybridassemblysv/src
- novoBreak
  https://sourceforge.net/projects/novobreak/

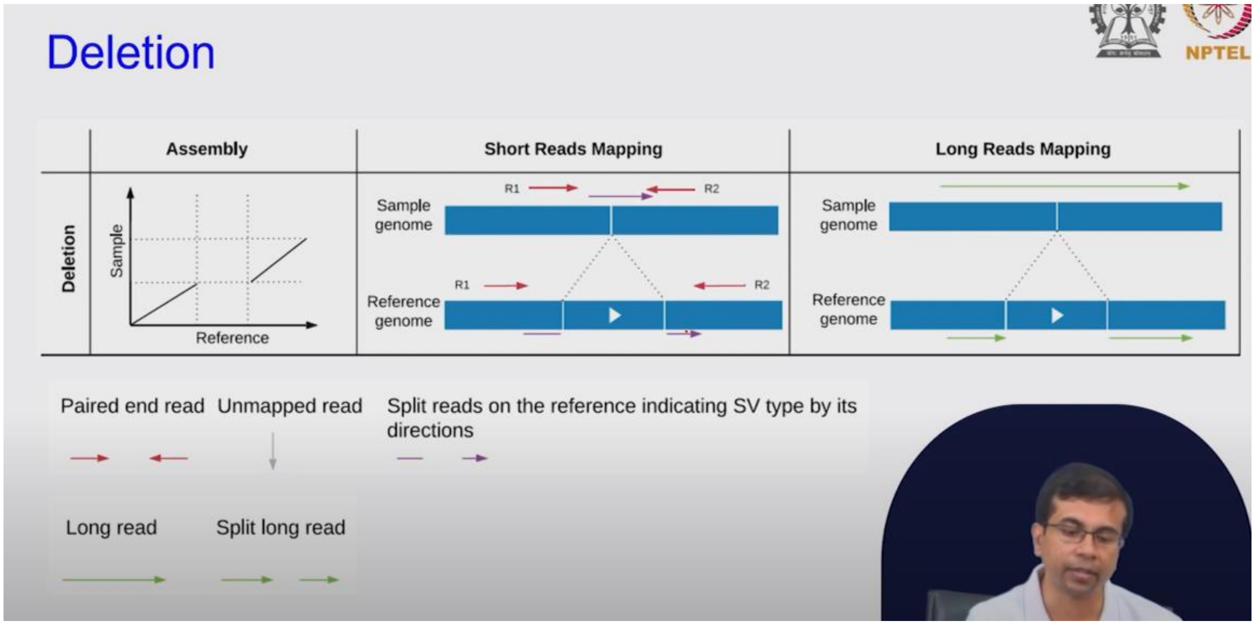
Now, we can use these methods, and you can have some sort of combined observation to actually identify different cases and different scenarios. So, we will discuss them briefly. So, we will start with deletion, right? So, let us say we are working with the short-read data. So, ah, we have the short read mappings, and you can see we are looking at paired end reads, we are looking at split reads, and we have talked about these methods. Again, what are the scenarios that we will see? So, in the case of deletion right, you can see that we have the reference genome now here, and then we have the sample genome. Now, what will happen is that in the case of the reference genome, right r 1 r 2 right, if this region is deleted, the distance between them in the sample genome will be reduced. So, there will be a reduction in the expected AH distance, and similarly, if you have read that AH is coming from the sample genome, what will happen is that they will map AH across this deleted region.



Deletion

Paired end read Unmapped read Split reads on the reference indicating SV type by its directions

So, there will be split mapping, okay? Again, this is something we have discussed already. If you take the long reads, you will have very similar observations. So, long reads might cover this entire region. Now, when you map this long read against the reference genome, you will see this split mapping OK. This is what we are going to see. What about the assembly-based method? So, if you have assembly in case of deletion, you will see this gap in assembly right between sample and reference. So, this region will be absent. So, we will not be able to find any corresponding region in the sample, ok? So, we will see these differences in the assembly process.



Deletion

Paired end read

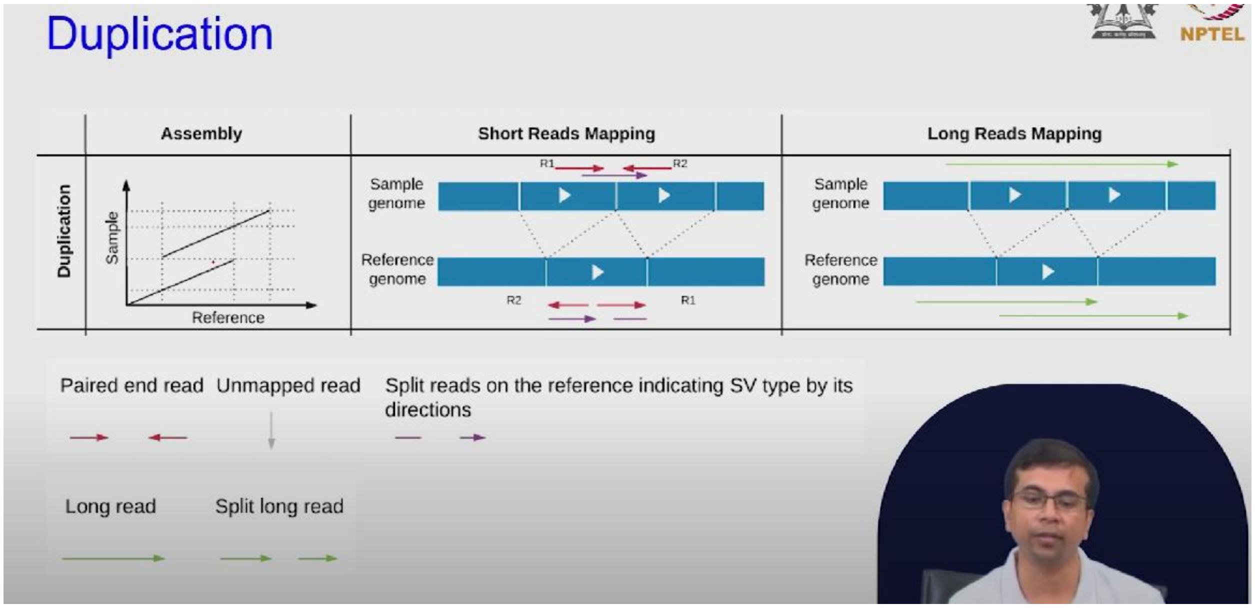
Unmapped read

Split reads on the reference indicating SV type by its directions

Long read

Split long read

So, we can then combine these methods, and perhaps we can come up with different tools. So, continuing this discussion against duplication is right for duplication events. So, again, we have the short read mapping data, the long read mapping data, and the assembly data. So, in the case of duplication, what you will see again in the reference genome is that this is duplicated. Now, if you look at this paired end data, right? So, this arrows in red color right their orientations ah will change right. We will see this kind of orientation change, ok? And what you mean is that we will probably see this coordinate mapping in the AH results, okay? And then again, we have this read if you come across this AH break point, right? This one is in violet color. This will show a split mapping in the case of the reference genome. When you are mapping the data against the reference genome, we will see this kind of split mapping.



Duplication

Paired end read

Unmapped read

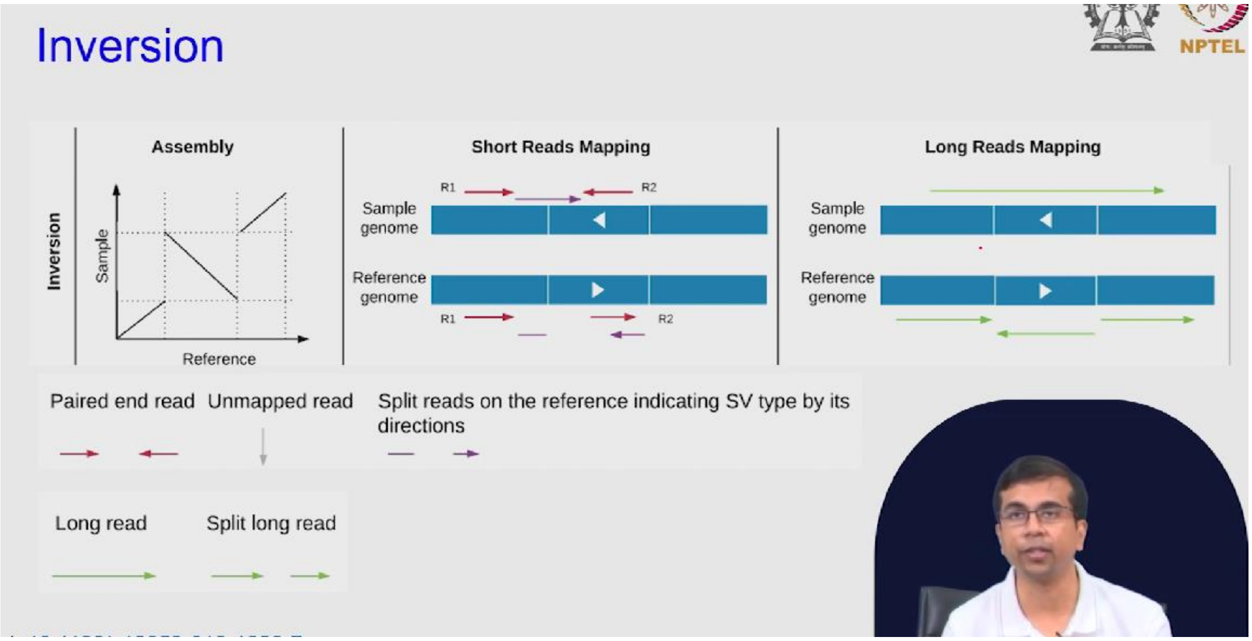
Split reads on the reference indicating SV type by its directions

Long read

Split long read

For long reads, we will see this kind of ah again overlap ah in the mapping right because one region has been sequenced twice. This is one region in the reference. So, this has been sequenced twice. So we will see this kind of overlap in the mapping. And again, if you look at the assembly part, you will see this kind of overlap in the assembly data, right? So, this will mean, ok, there is some duplication event that has happened in the sample genome.

Similarly, you can extend these two, for example, inversion events. Right again, there are different scenarios without going into a lot of details; you can actually look up and think about them. So, again, with inversion events, you will have different kinds of scenarios, but again, depending on the distance orientation of the reads split read mapping, these things will change. So, similarly, if you look at the long reads, you will see this kind of change in the orientation mapping process, as well as if you look at the assembly again, the orientation, etcetera, will change because of the inversion event.



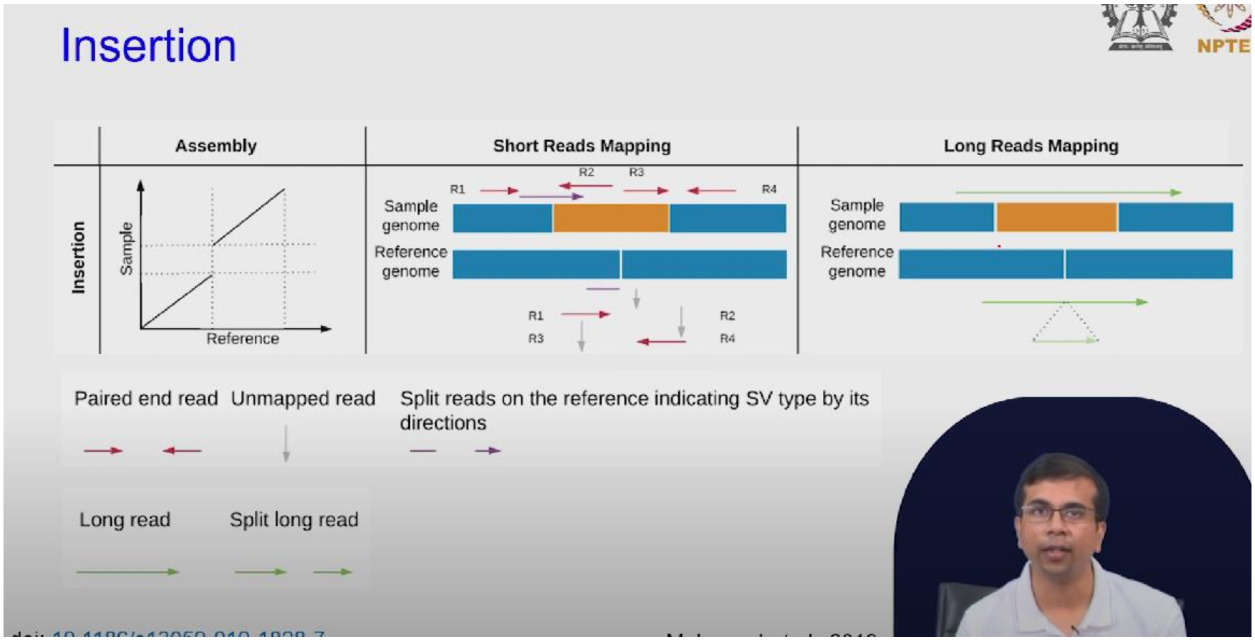
Inversion

Paired end read Unmapped read Split reads on the reference indicating SV type by its directions

Long read Split long read

So, what I will do is encourage you to actually think about these scenarios—what will happen if you have inversion, deletion, etcetera—very carefully, and maybe take a piece of pen and paper and think about these situations and draw the outcomes yourself.

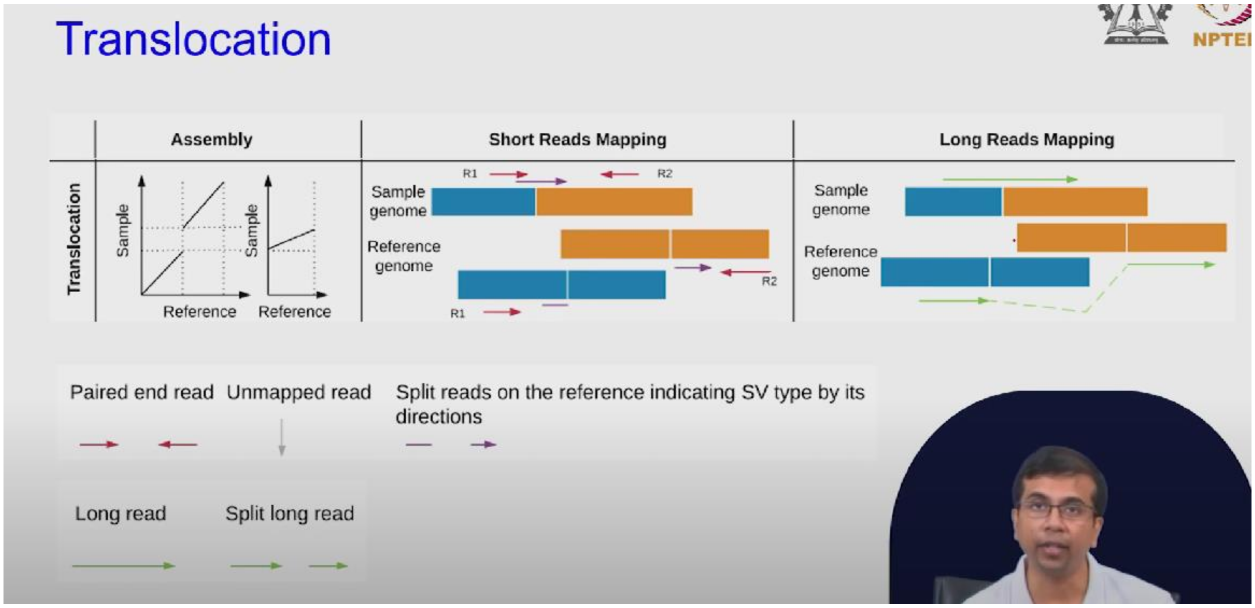
So, that will actually help in better understanding. So, you can have insertions of extra regions right in the reference sequence, which is denoted by this orange region. Here, this region is inserted, and what will happen is that you will generate some leads from this sample that will not be able to map against the reference genome because this sequence is not present in the reference genome, and similarly, you have the same issue with long-read data. So, again, you have this read that will map the inserted region, but part of this read will not be able to map in the reference genome, and similarly, you can think about it the same way in the assembly process. If you are comparing the sample assembly and the reference assembly, you will see in the sample that some sequences are extra ah and that they are not present in the reference assembly.



Insertion

doi: 10.1186/s13059-019-1828-7 Mahmoud et al., 2019

So, ah, these kinds of situations will describe the type of SV that is happening, and similarly, you can think of translocation. So, translocation means you will have a change in some kind of combination, right? So, again, you have some regions that were not present, and again, you can think about all these, ah, what will actually happen if you are generating certain reads, ah, some reads in the translocation. So, two, let us say, two chromosome parts of two chromosomes have fused together, right? So, they have been translocated; they are fused together, and you are generating these reads R 1 and R 2 right from the sample, and when you are mapping them back to the reference genome, they will map to two different chromosomes. So, this means we will get this discordant mapping right, and those results will then become important for this kind of scenario. You can think about the split read case, etcetera, and what will happen. In the case of long reads again, part of the read will map to one chromosome, and the other part will map to another chromosome, and again, this discordant mapping will help us identify this ah SV and also will help us understand the type of SV that has happened. So, what we will talk about is that there are an array of methods that actually take this combined approach right; they will actually combine all these different methods that we have discussed, such as read-depth assembly, etcetera.



Translocation

Paired end read Unmapped read Split reads on the reference indicating SV type by its directions

Long read Split long read

They can combine some of these methods and generate much more accurate results. So, ah, they will combine maybe two or three of these, right? So, paired and read plus split read and read depth ah again, assembly is not really favored because of the computational resource requirements, etcetera, and it can. So, these methods can take any two of these or all three of these and can help generate a very accurate picture of the SVs. So, when we combine these methods, we get the benefits of all these approaches, and we also get lower false positives. Here are some tools that actually implement this. They kind of come up with this combined approach, and they get much more accurate results. So, here are the references that we have used for this class, and in this class, to summarize, we have seen the different types of SVs we have discussed and which are observed in ah genomes. These are not just made up or thought of, but we actually identify these different types of SVs in the genomes, and we have discussed different methods and tools that ah can help us identify these structural variants, and what we have also seen that the combined approach brings together the strengths of each of these methods. So, we get a much more accurate view of these structural variants that are out there, we get the best results, and we reduce the number of false positive calls. Thank you.