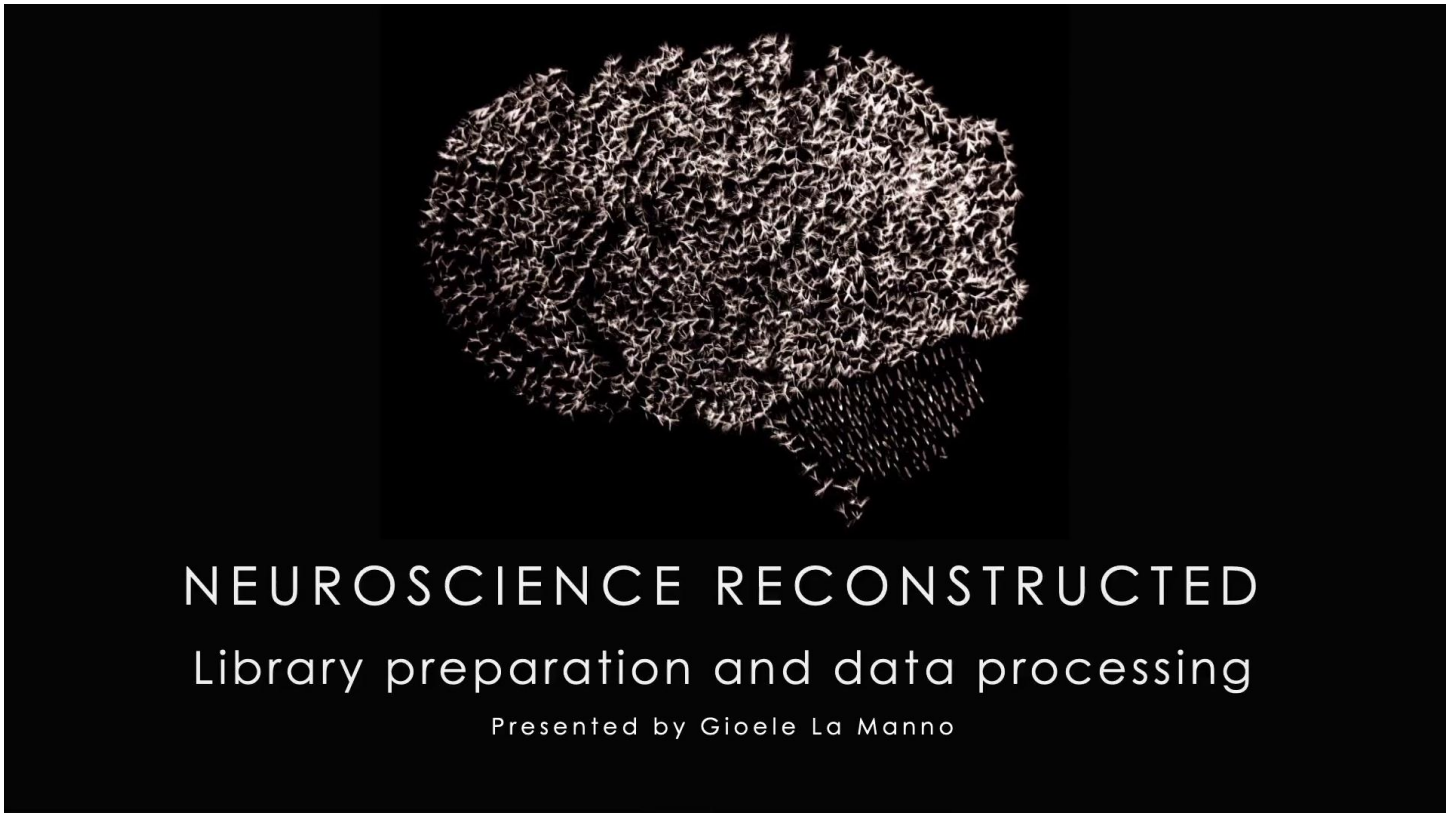
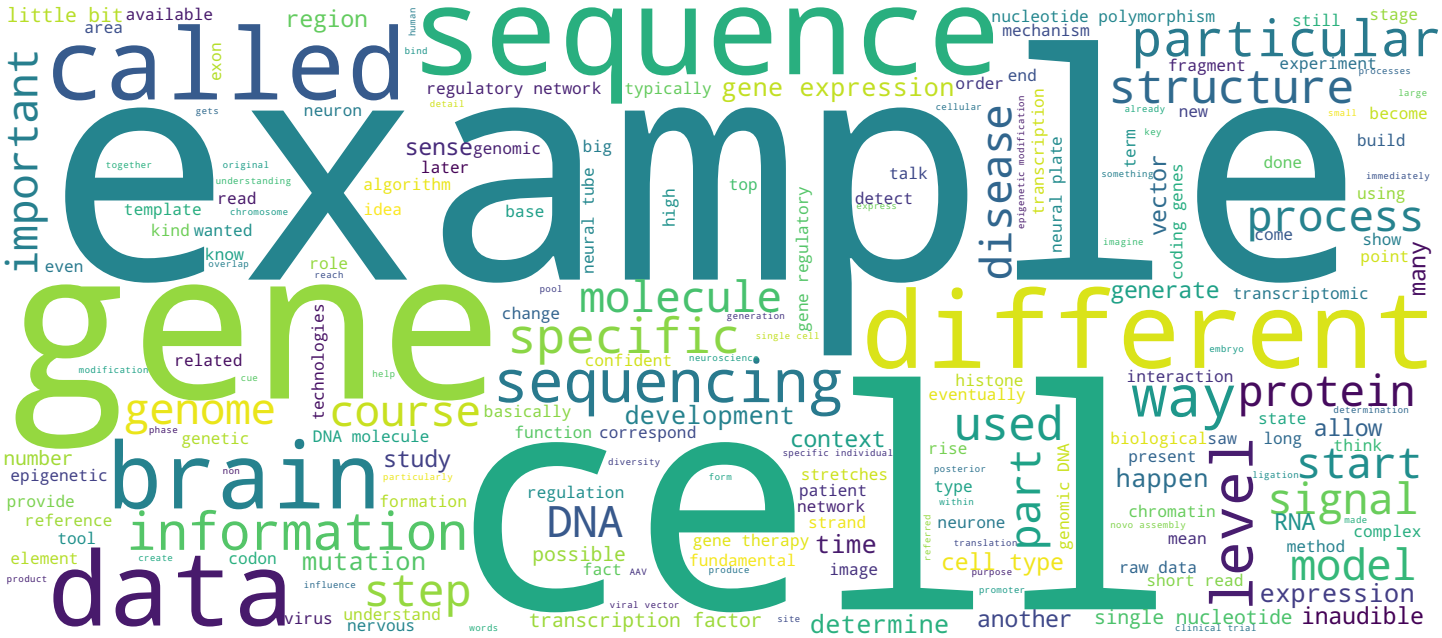




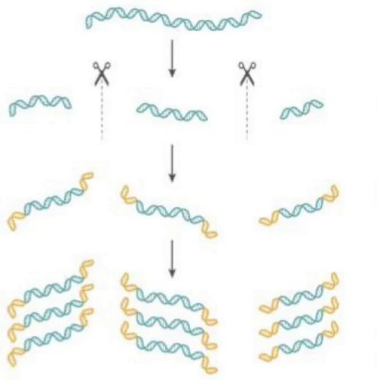
Search MOOC

Video

EPFL

From molecule to sequence

- Fragmentation: the long DNA strands must be cut into smaller pieces
- Ligation: The DNA fragments must be ligated to adaptors in order to be bound to surface or a bead
- Amplification: The DNA must be amplified in order to obtain clusters of clonal DNA



Now I want to discuss about some general principle about what is called library preparation. Those are a set of steps that are used to process an original pool of DNA molecules, what we can still call a library of DNA molecules, into our sequencing libraries. In other words, a pool of molecules that corresponds to the original one we wanted to get information about the sequence, but it's been now processed in such a way that it's compatible to the specific sequencing method we want to use. This is example a little bit described by thinking specifically of luminal sequencing, but most of these steps should apply to many of the technologies. We have the step of fragmentation is one of the earliest step. You typically have long molecules and you can only sequence stretches of 200, 300 base pairs, typically. You want to fragment the DNA, cutting it in smaller pieces. This can be achieved enzymatically or more mechanical [inaudible 00:01:13]. Then there is the step of ligation where one typically ligates adaptors to the two extremities of the double-strain DNA molecule. Those adaptors have a known sequence that can be used to further amplify the library to reach now higher concentration for each of the templates so that it will be easier to detect and determine the sequence.

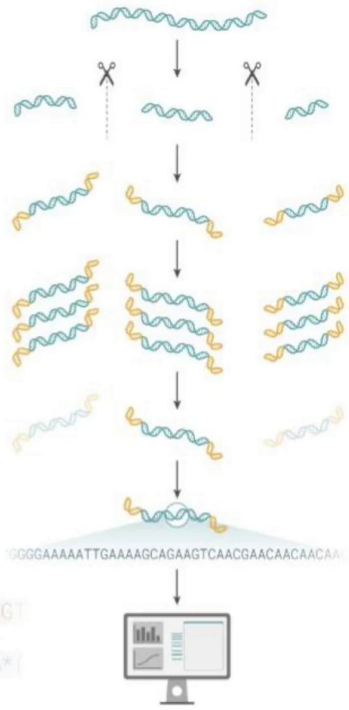
Notes

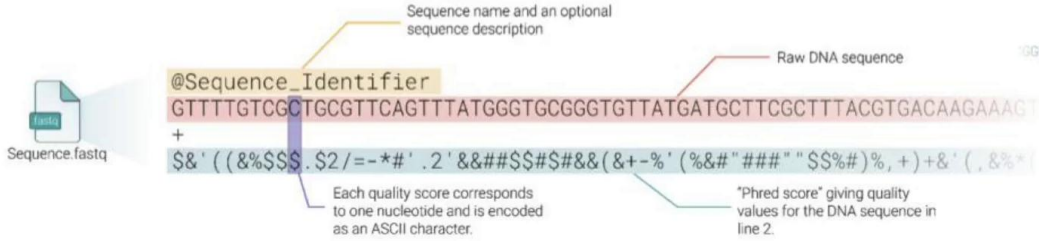
Summary

0m 05s

From molecule to sequence

- Fragmentation: the long DNA strands must be cut into smaller pieces
- Ligation: The DNA fragments must be ligated to adaptors in order to be bound to surface or a bead
- Amplification: The DNA must be amplified in order to obtain clusters of clonal DNA
- Size selection of the amplified fragments
- Sequencing
- Data processing and sequence alignment



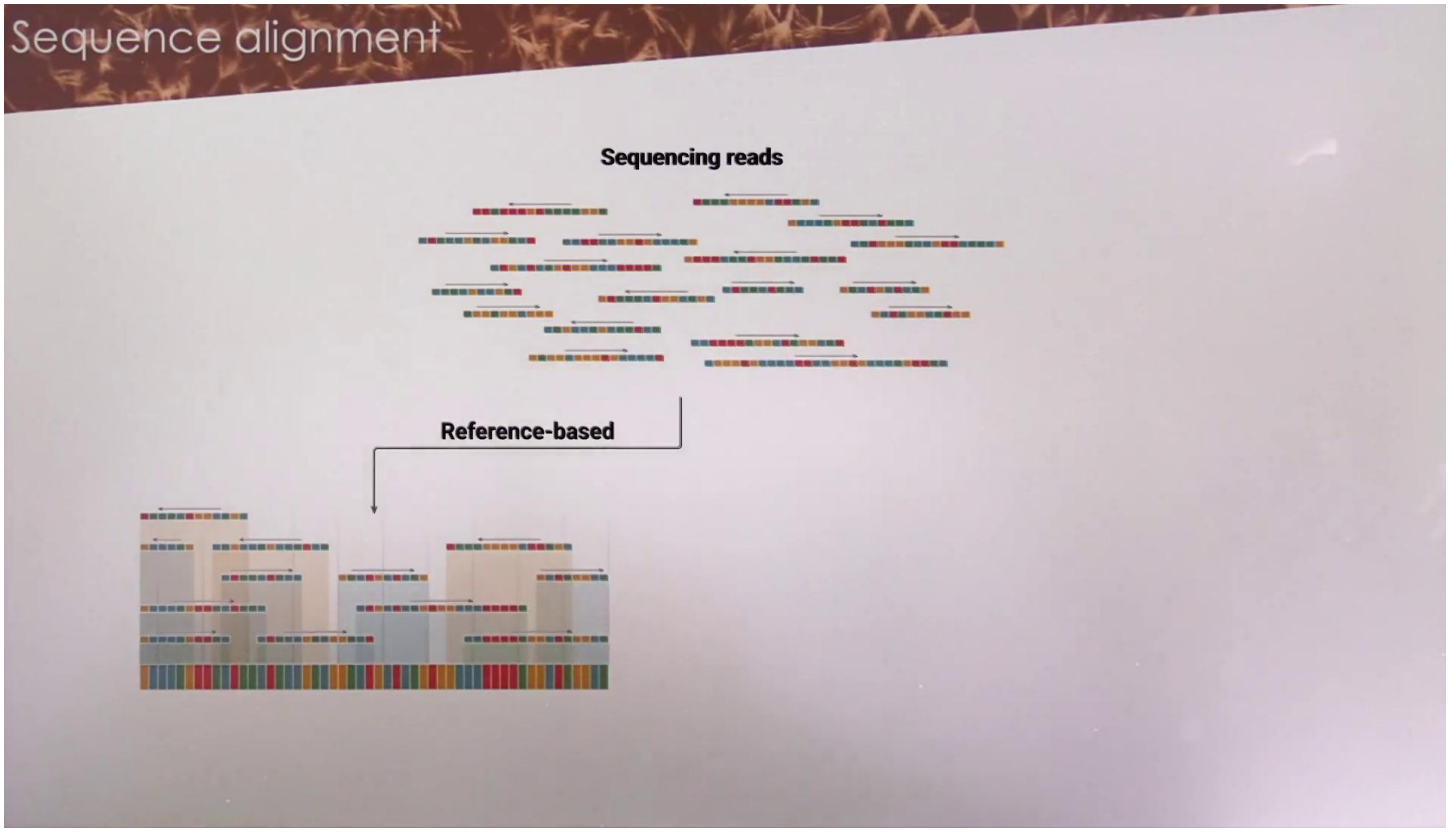


Then this amplified library usually can be, for example, size selected to even become more precise using, for example, cutting a gel or using some specific size selection kit can be selected to have a particular molecular weight range and then can be loaded for sequencing. Each sequencer is typically gathering different kind of raw data that get typically digested in what is referred to as raw data in standardized format. It is what is called the FASQ format. Typically, this is really a simple text-based file that contains something like a sequence identifier, eventually the raw DNA sequence, and a compounding by it's quality score that correspond how confident the algorithms that are used to processing the really raw data, corresponding, for example, to images or the level of intensity of different fluorophores, how confident the algorithm is that the determination of that base is coherent.

Notes

Summary

1m 37s



In the context of genomics, the main aim of sequencing is determining actually the sequence, for example, DNA or genomic DNA. This might sound obvious, but you're going to see immediately after how you can use sequencing for the purpose of quantifying the number of molecules in a specimen like in transcriptomics. But in the context of, for example, genomics, you want to determine, for example, the sequence of genomic DNA. To do this, you need now to make a little bit of inference because of the fragmentation step that you need to perform, and because of the short read length that is available for most of the attribute sequencer. At the end of sequencing, you have a set of stretches of sequencing reads that you need to make sense of. The most common way to make sense of those stretches is to use a reference-based analysis, meaning that one would use specific mappers, software such as Bowtie, which is a very popular one, for example, or Star, that will align your sequence to a reference. For example, a reference genome of a specific individual of a species.

Notes

Summary

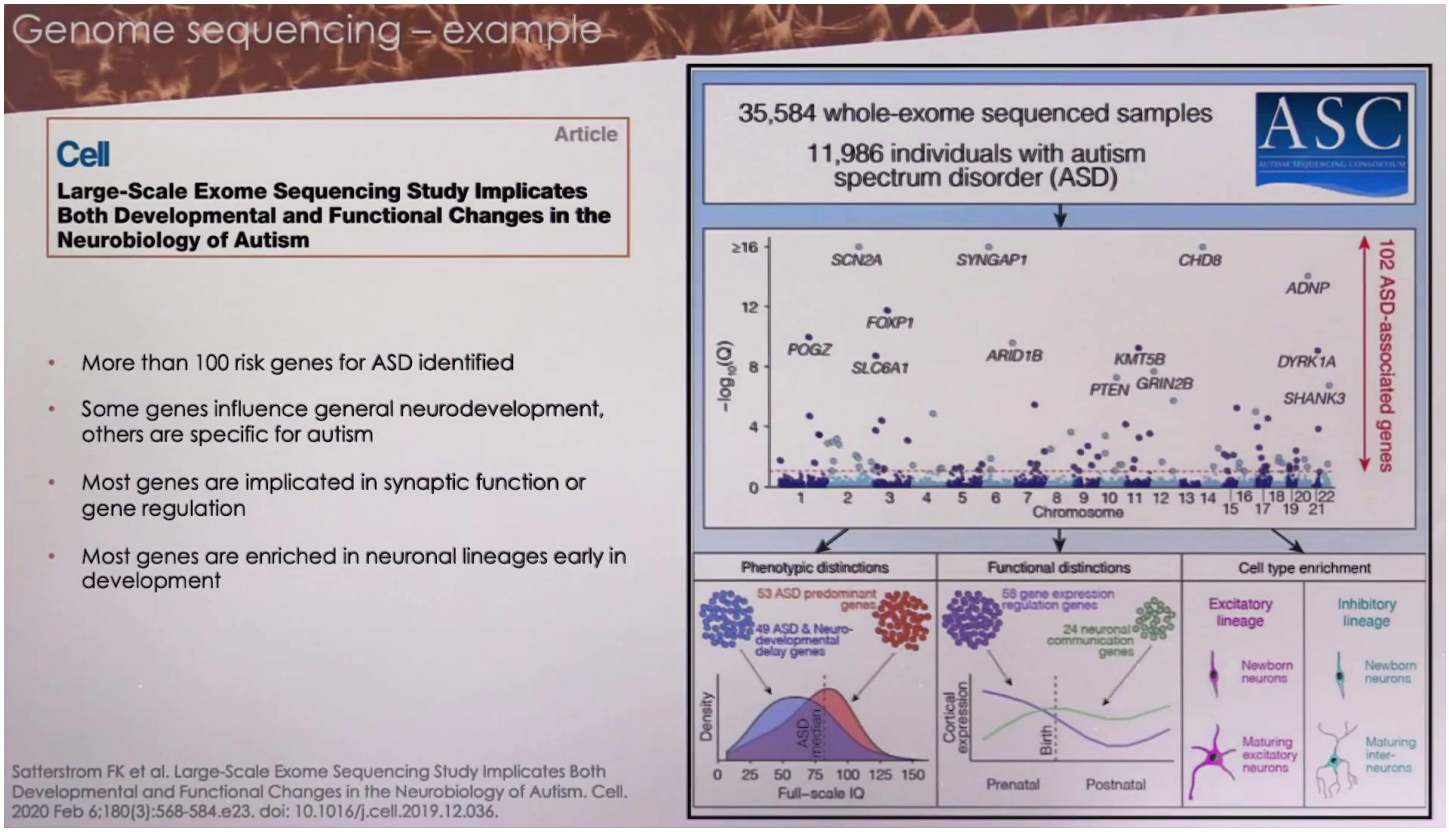
2m 46s

Eventually, this would allow you to broadly map its read and eventually found that there were few, for example, mismatches, so-called single-nucleotide polymorphism, or other kinds of polymorphisms that made the specific individual genetically different from the reference. The other way to go about this, and this is done, for example, if this is the first time the DNA or the genome of a particular species is being sequenced, is to do De novo assembly. De novo assembly involves the use of different algorithms that try to build what is called a [inaudible 00:04:43], try to get information on the overlap of the different short reads, and to actually build a long single stretch that encompass the whole sequence of a chromosome.

Notes

Summary

4m 10s



Here, I wanted to just show an example of a large-scale exome sequencing that's been used to study specific aspects of developmental neuroscience here in the context of human biology, and in particular of the neurobiology vertice. In this study, for example, the scientists here have been getting information on the single-nucleotide polymorphism associated close by to different coding genes. They're doing what is called an exome, in other words, sequencing only the part of the genome that associate with exomes of protein-coding genes. In this way, they have identified single-nucleotide polymorphism that are associated and can be recognized as increasing the odd ratio of the carrier to actually develop acute spectrum disorder. Another aspect I wanted to provide about this study is the fact that one can then further dig on the variability of this information. In the study, for example, they go on and look at further structure and correlation in the stratification of this information using different covariate they're collecting in this score and eventually, enable the others to make hypotheses on the specific involvement of specific cell types in their own system into the disease. This is particularly, I think, interesting as an example of how one can use the information that comes out of the sequencing data to draw interesting neuroscience and biology conclusion.

Notes

Summary

4m 58s