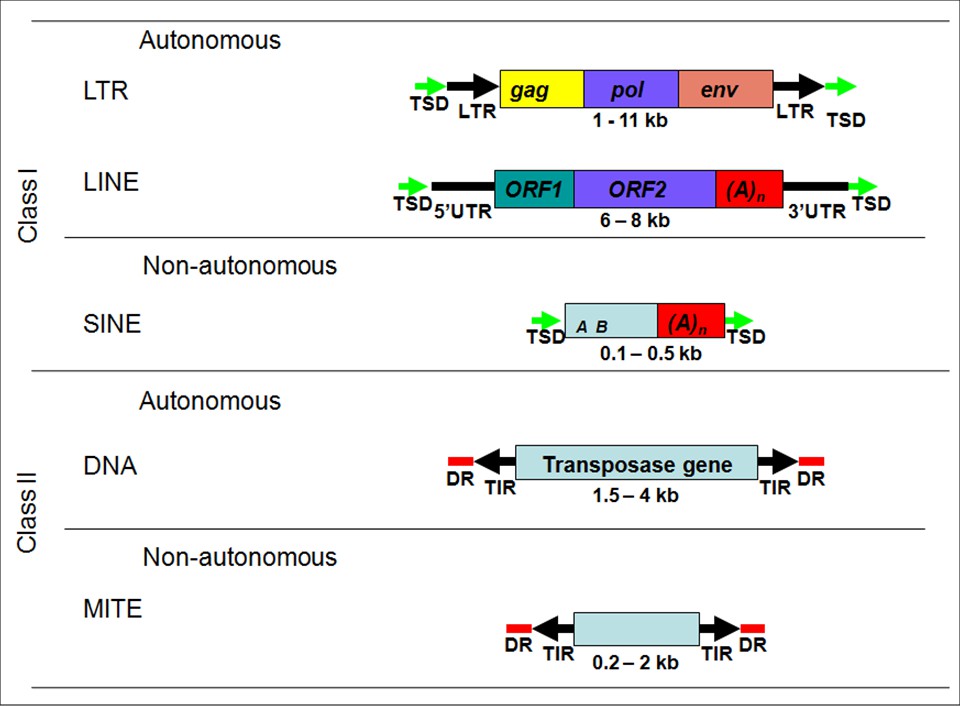
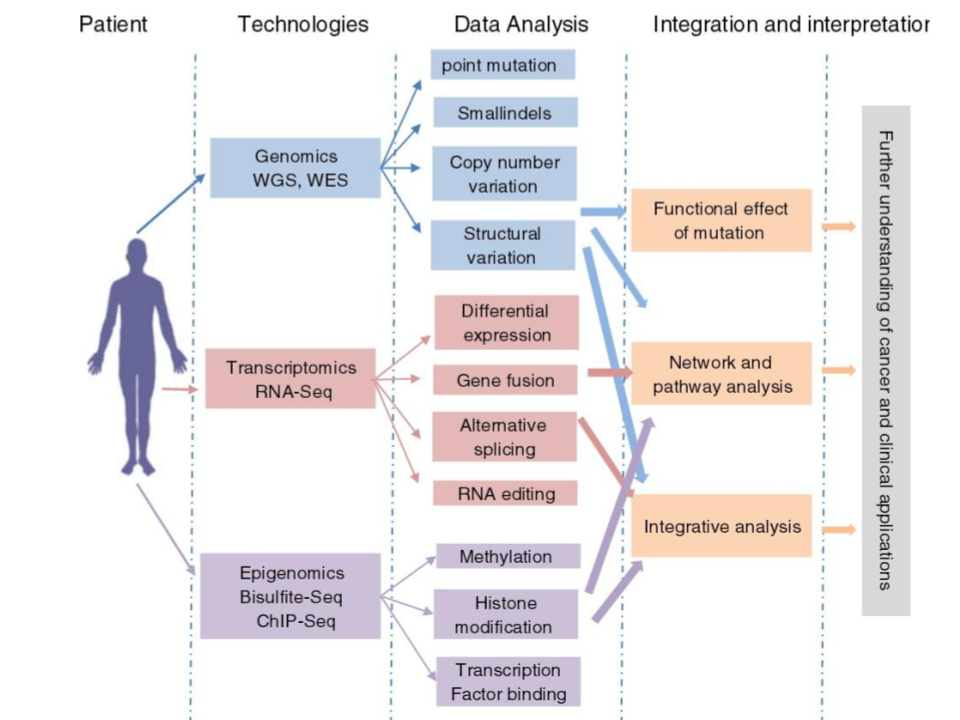
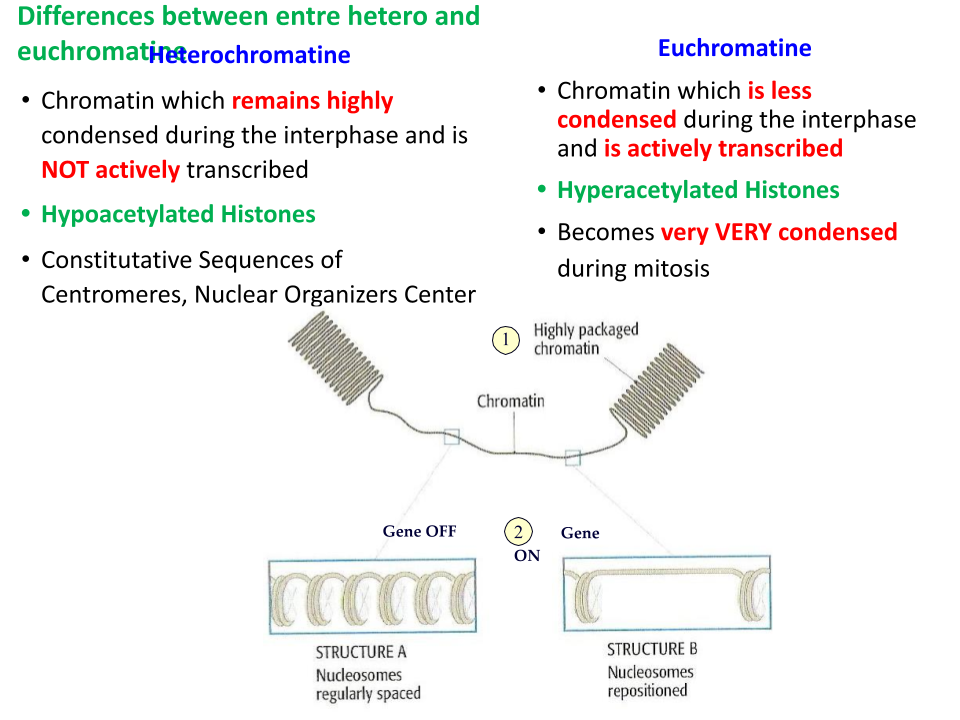
# Eukaryotes in Bioinformatics

* Using eukaryotic genomes as source info (reminder):
  + Prepare sample, extract DNA, maybe truncate sequence to look at specific genes, amplify DNA-> PCR multiplication.
  + Sequence at least 95% of the genome.
  + Assemble sequence after sequencing, different methods;
  + Annotate sequence, especially if it's a genomic sequence.
* The size of eukaryotic DNA used to be a disadvantage for sequencing, new technology is fine with it (NaNoPore). However, the **re-assembly and annotation steps** are accompanied with a **large computational burden**.
* The eukaryotic genomic **sequence size is not related to its complexity**, a large genome does not equal to a large or complex organism.
* A lot of the **DNA is non-coding (regulatory/life cycle related RNA), repetitive, redundant**. There are large areas-intergenic areas.
* Splicing and alternative splicing during transcription/translation exponentially multiply the possible proteins that can be produced.
* Genetic material is connected to histones, complicating extraction and purification steps.
* **Transposons** complicate the entire process.
* Chromosomes contain the eukaryotic DNA obvs, **centromeres and telomeres** **complicate the sequencing process**.
* **Introns** complicate the process as it makes it difficult to identify the start and the end, especially where **Introns and Exons** collide.
* Proximity of promoters to genes is often easy to understand, however there are cases of “**distant promoters**” **kilobases** away from the actual gene sequence.
* While re-assembling eukaryotic DNA you have to also account for organelle DNA, avoid contamination during cell lysis. Contamination from a prokaryote can also be confused for genes that have been transferred from the mitochondrion or chloroplast to the nucleus and are no longer encoded by the organelles/former prokaryotes themselves.

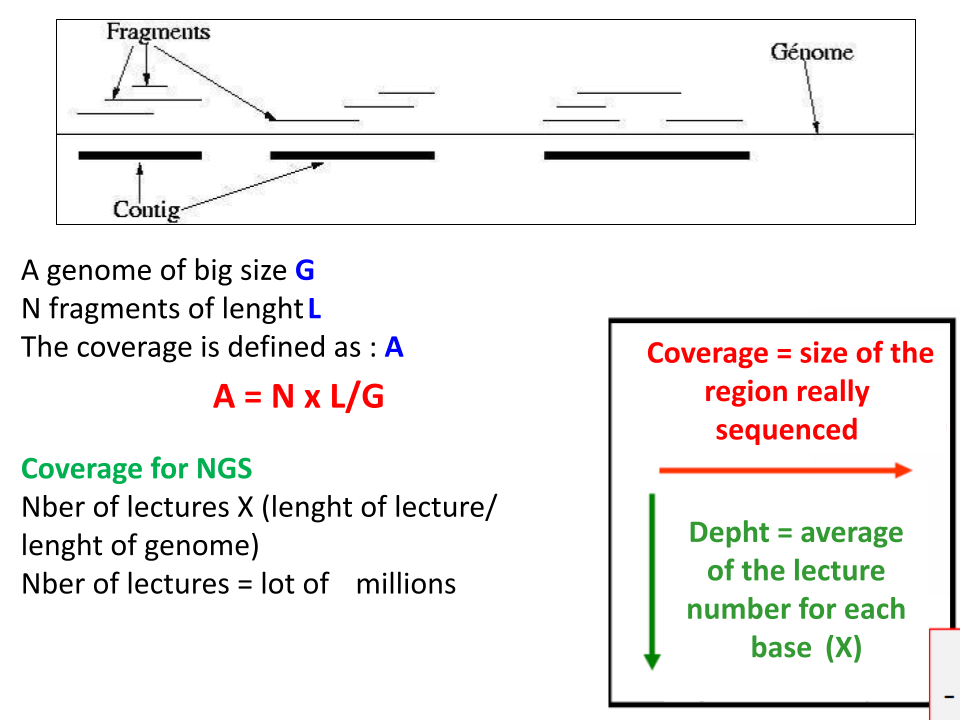


* **Gene families** are discerned from their ancestral genetic tie, their sequence similarity and often very similar function.
* As the genomic material is duplicated to produce redundant sequences, these redundant sequences can over evolution undergo mutation themselves; ergo, gene families, **neo-functionalization**.
* For example, in wheat, which can have multiple copies of its genome, it is paramount to identify which chromosomes belong to which genome-copy. As different copies will have evolved differently over time, owed to random mutational events. This is a prime problem with **polyploid** organisms.
* The **status of the chromatin can be altered** if necessitated by absences in the proteomic environment. **Epigenetic** changes can affect how easy it is to transcribe the DNA.
* Once genes are identified following the re-assembly of a genome, **transcriptomic methods are used to determine their function with a high level of confidence**. The theoretical function can be very different from reality.

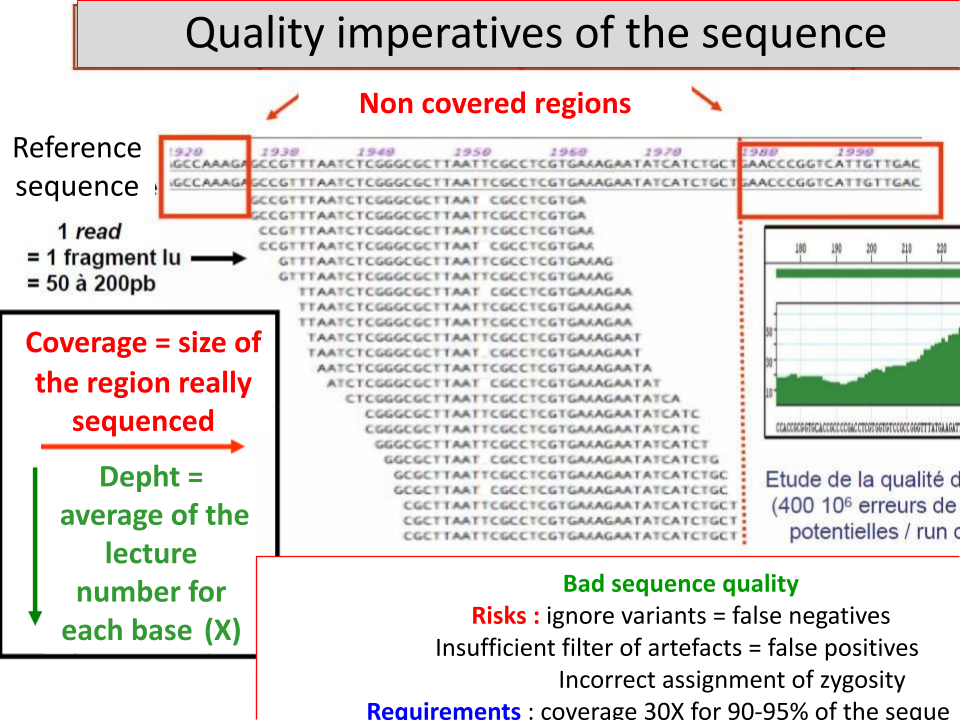
# Prokaryotes

* Way more compact genetic material, circular DNA with little buffer-intergenic-bases. Therefore easier to sequence and annotate/identify genes.

# Sequencing General Information

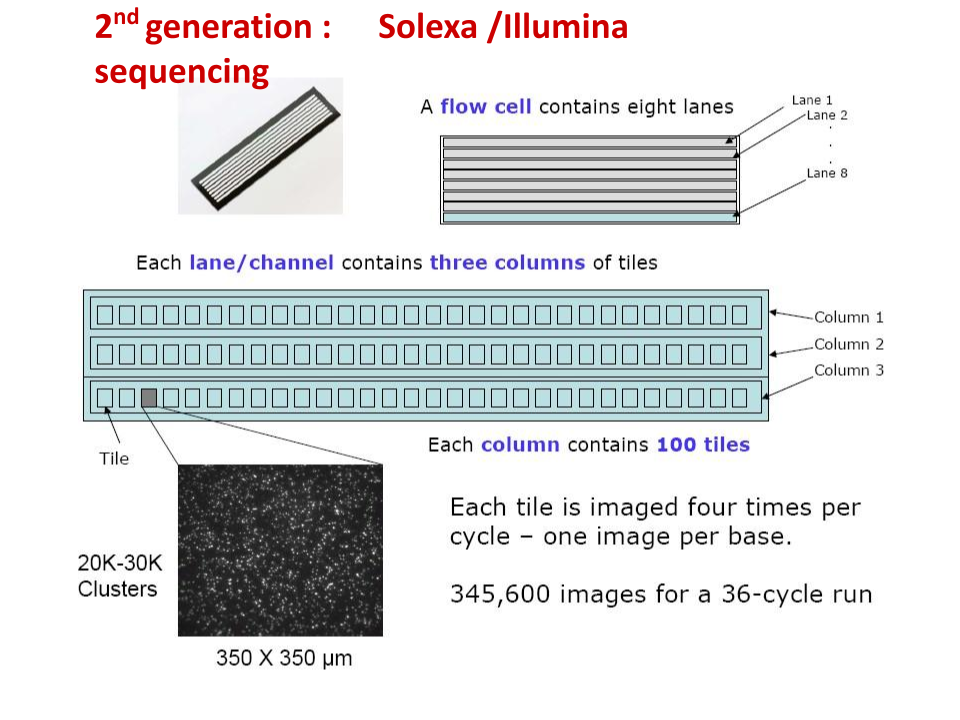


* When one says that they should sequence 95% of the genome for good practice, they imply that within that **5%, can be errors, but also centromeres and telomeres, which cannot be unwrapped**.
* The **depth** of the reads pertains to **how many times** a **nucleotide** has been read by the machine.



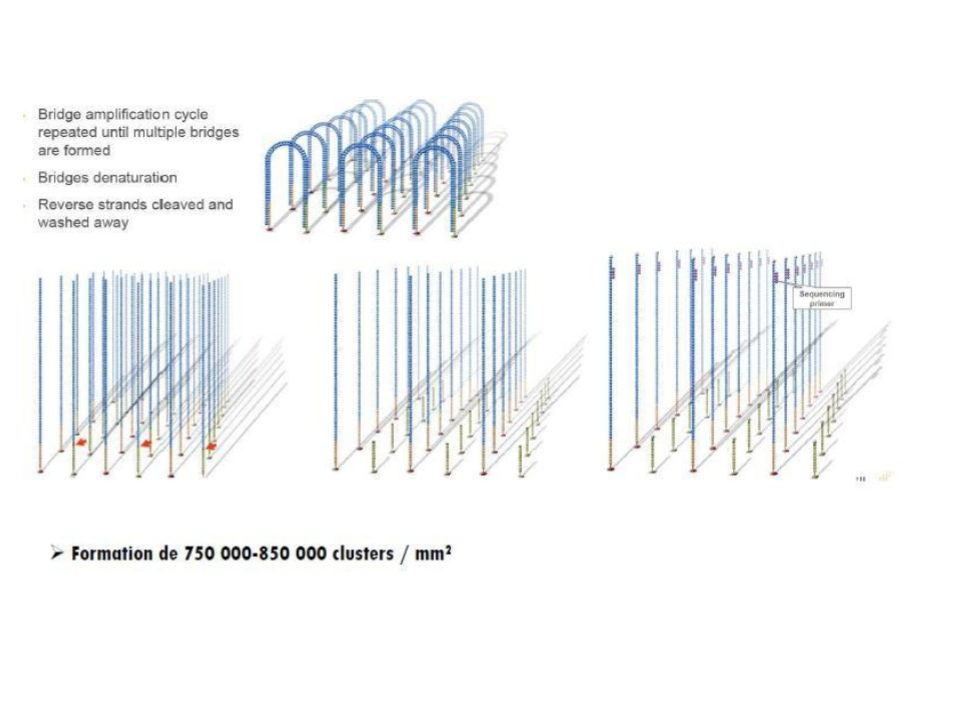
* 90% coverage at 100x.
  + Average nucleotide read rate is **100 times (the depth)**, over **90%** of the **genome**.
  + Nucleotides at the **ends of reads** are “sequenced” less often. A method to ascertain the confidence of a read is by **the average number of “lectures”.**
  + **Nucleotides that have been read (lectured)** **many times**, imply that there should be a **low chance of error**.
  + **Sequencing errors** can be determined by the times a nucleotide has been read; that number of reads can also be used to **ascertain Single Nucleotide Polymorphisms**.
* **Contigs**: An assembly of **overlapping short reads**, creating a **contiguous sequence**.
* **Chromosome Zero**: A collection of sequences that couldn’t be assembled together during sequence re-assembly.

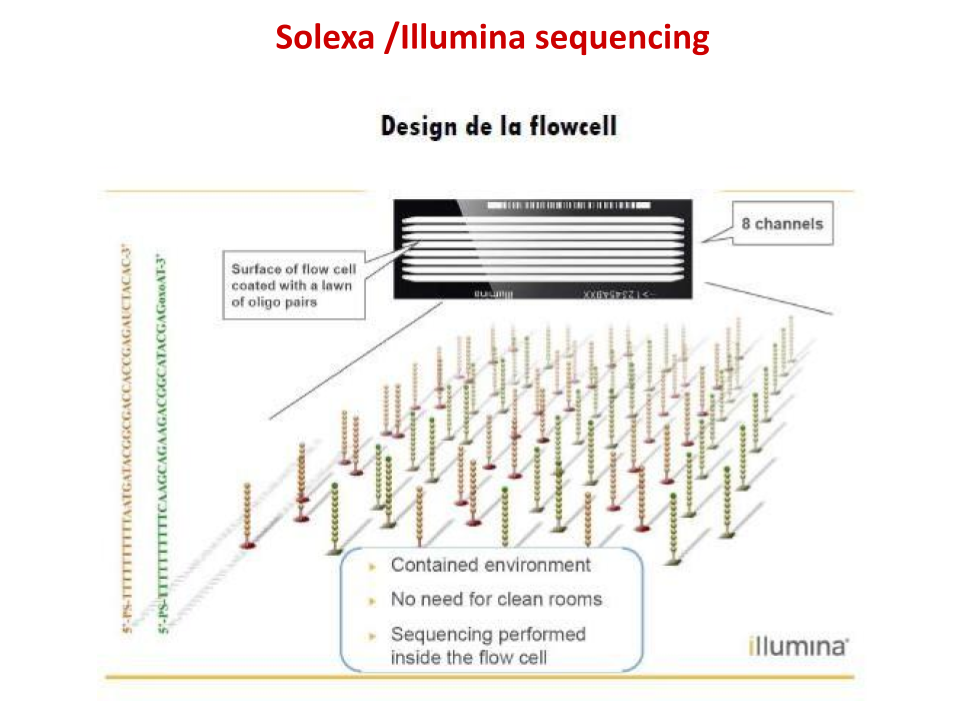
# Illumina Sequencing

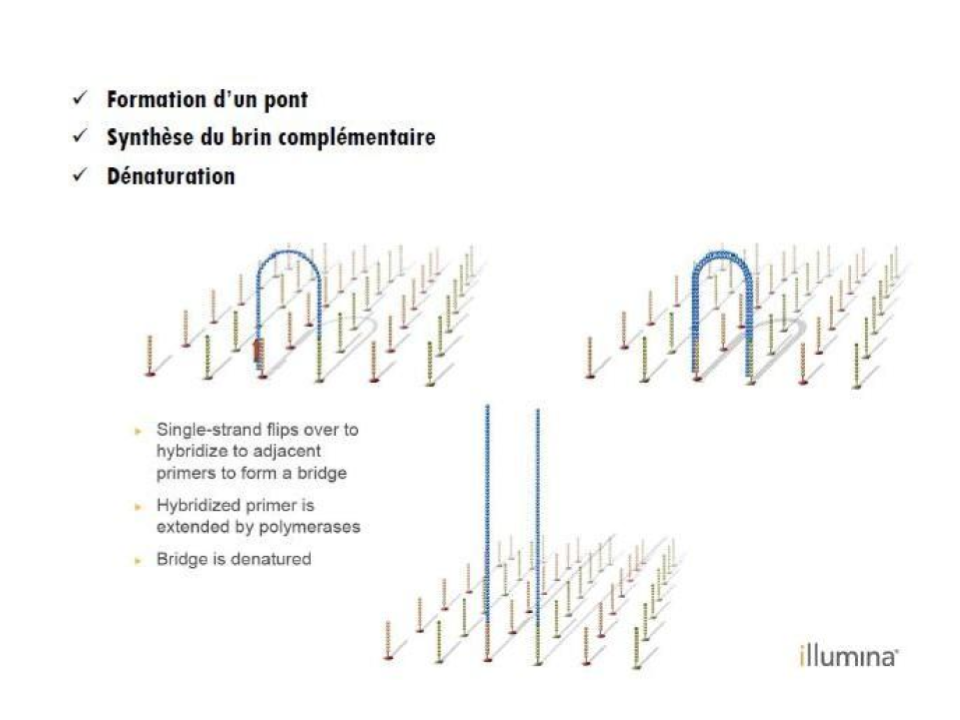


* Sheets of classified material, will contain 8 eight lanes, of which each lane has three columns, each column has 100 cells/tiles, which can **house DNA** to be sequenced.
* Each **cell/tile** is **lined** with small **segments** ~20bp, which attach to **DNA** inserted; specifically, they will attach to the **complementary** **adaptor segments** attached to **DNA molecules** that we want to test.
* **DNA is denatured,** double strand is **separated** moments before reading.
* The connecting forces between the **known segments, adaptors and DNA**, are designated as **hydrogen bonds**.
* During **PCR amplification**, the adaptor, cell liners and DNA will obtain a **new complementary strand**, as the other one has been removed, therefore, the complementary strand will be **preferred as all of these constituent parts are not just joined by hydrogen bonds, but covalent bonds between the nucleotide carbons**. -> The DNA **won’t be expelled by incoming buffer**.
* While this **PCR Amplification** occurs, the new strand **will attach** to a neighbouring **cell liner segment**, creating an **arch**, of **double stranded DNA**. Later, this **DS will be separated**, meaning that we now have two **complementary strands**, on different cell liners, **which have come from the same original sequence**, and can now be sequenced **separately**.
* TLDR: You use **known sequences**, with an active buffer, to **attach DNA, unwrap it, and read it**.

# IMAGES FOR YAPPING

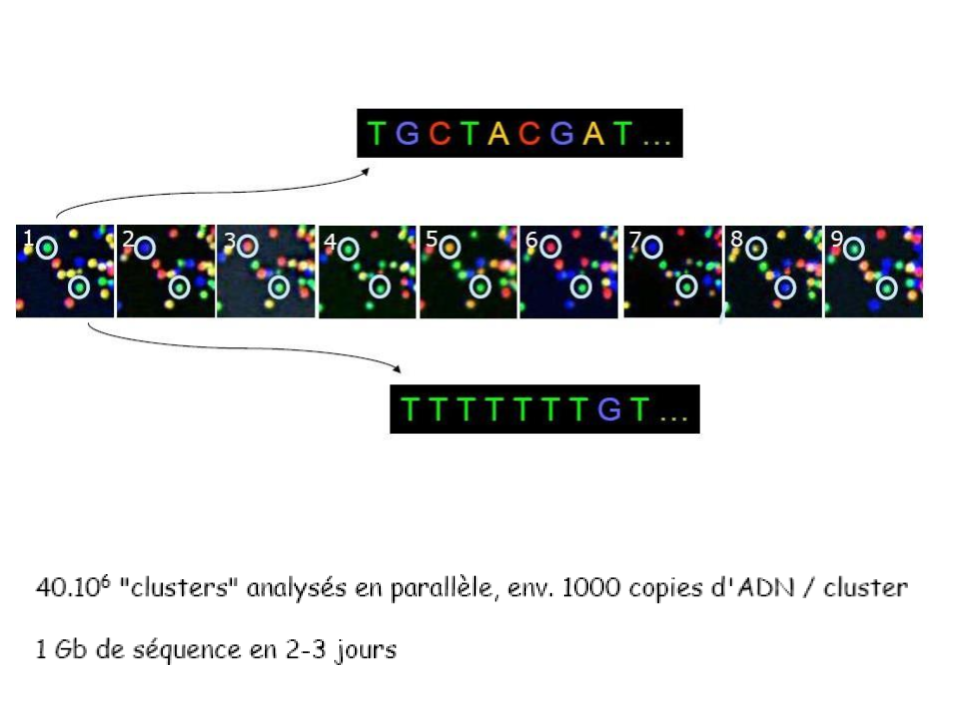






# Sequencing, how does it work

* Nucleotides are **tagged with nucleotide specific fluorophores,** which are **agitated** in order to emit light.
* Therefore, while the **DNA is being assembled on the floppy single strands in four nucleotide segments, yet again**, **tagged nucleotides** will be identified by the fluorophore that is added. After, the **fluorophores** on the 4 nucleotide segment **are cleaved from the nucleotides**, to prevent errors.



* The advantage of this method is that colours are easily discernible. Code will be read in clusters.
* **Multiplexing:** The use of multiple lanes in Illumina to obtain **sequencing data** on multiple different samples.
  + **Barcoding**: A known sequence which is used to **tag different samples** being sequenced within an **Illumina lane**.