DNA Sequencing

- The process of determining the order of bases adenine (A), thymine (T), cytosine (C), and guanine (G) along a DNA strand.

ATCGTACGAGAGAGAGAGAGAGATCAATTAGTACGTACTCAGTGG

- All the information required for the growth and development of an organism is encoded in the DNA of its genome. So, DNA sequencing is fundamental to genome analysis and understanding the biological processes in general.

Classical method of Sequencing:

1. Sanger Sequencing
2. Maxam-Gilbert sequencing

Sanger Sequencing

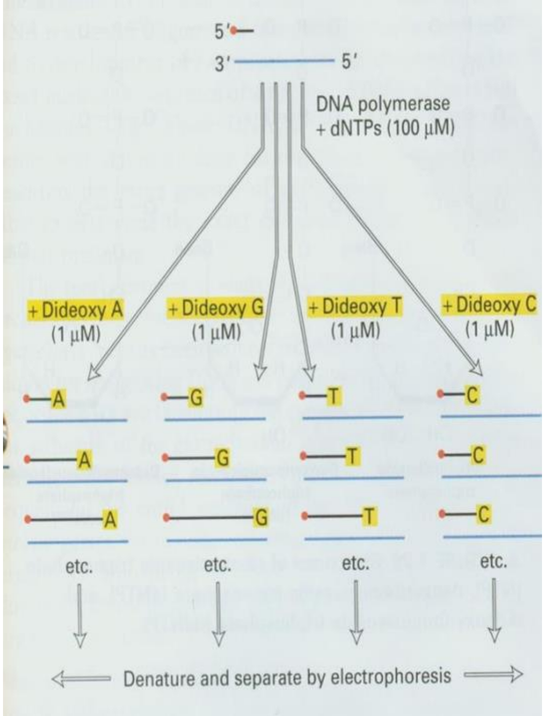
- Also known as dideoxy sequencing method or chain termination method because it involves the use of analogue of normal nucleotide 2’,3’-dideoxynucleoside triphosphates (ddNTPs). These are chain terminating nucleotides lacking 3’-OH ends. This method is based upon the incorporation of ddNTPs into a growing DNA strand to stop chain elongation.
- This method uses single-stranded DNA.

Steps:

- The DNA to be sequenced is called the template DNA. It is prepared as a single-stranded DNA after being spliced into M13 vector DNA. Infected *E. coli* host cells release phage particles which contains single-stranded recombinant DNA that includes the sample DNA. This DNA sample is then extracted from phage for sequencing purpose.
- A synthetic 5’-end-labeled oligodeoxynucleotide is used as the primer.
- The template DNA is hybridized to the primer.
- The primer elongation is performed in four separate polymerization reaction mixtures. Each mixture contains

  - 4 normal deoxynucleotides (dNTPs) in higher concentration and
  - a low concentration of the each of the 4 ddNTPs.
- There is initiation of DNA synthesis by adding enzyme DNA polymerase since the enzyme cannot distinguish between the normal nucleotides and their analogues.

- The strand synthesis continues until a ddNTP is added. The chain elongation ceases on the incorporation of a ddNTP because it lacks a 3’-OH group which prevents addition of the next nucleotide.
- There is a result of mixture of terminated fragments, all of different lengths.
- Denature DNA fragments.
- Each of the four mixtures are run together on a polyacrylamide gel for electrophoresis.
- The separated fragments are then visualized by autography.
- From the position of the bands of the resulting autoradiogram, the sequence of the original DNA template strand can be read directly.



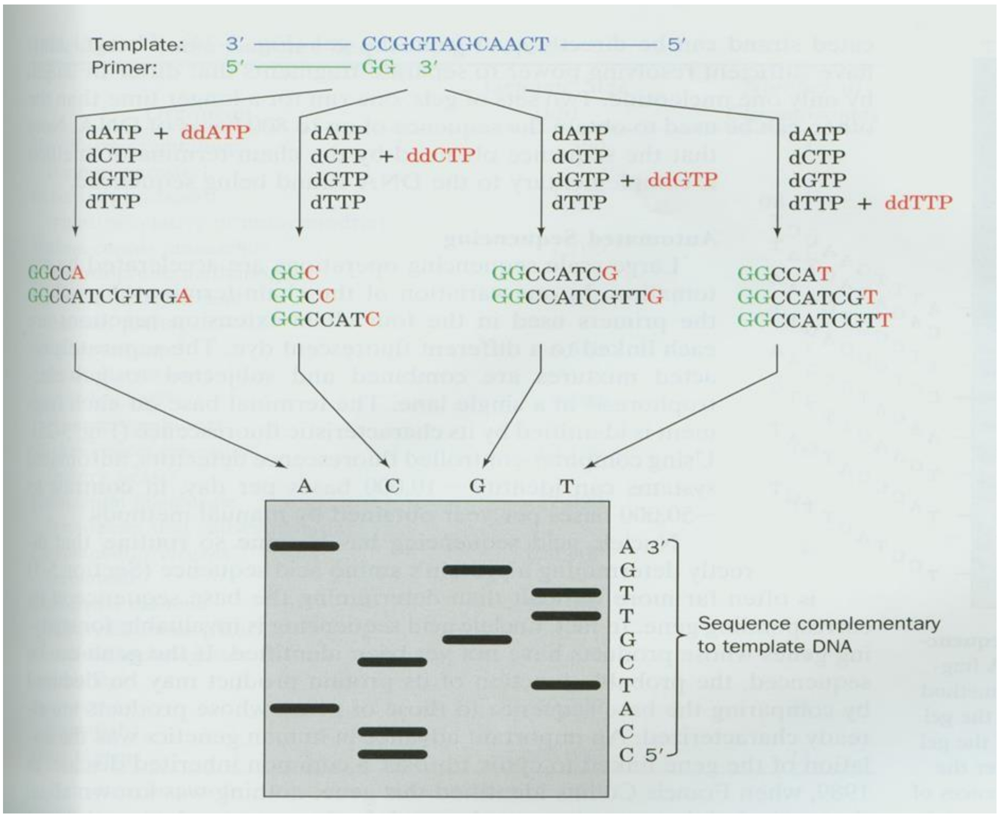


Figure: Sanger method

Advantages:

- Most popular method.
- Simpler and quicker allowing large output. Within an hour the primer-annealing and sequencing reaction can be completed.

Disadvantages:

- Yielding of poor results owing to secondary structure in the DNA as sometimes DNA polymerases terminate chain elongation prematurely.
- The sequence is obtained not from the original DNA molecule but from an enzymatic copy. So, there is a chance of incorporation of wrong bases.

Maxam-Gilbert Sequencing

- Also known as Chemical Cleavage Method
- This method uses double-stranded DNA samples.
- Involves modification of the bases in DNA followed by chemical base-specific cleavage.
- Sequences DNA fragments containing upto ~500 nucleotides in length.

Steps:

- The double-stranded fragment to be sequenced is isolated and radioactively labeled at the 5'-ends with ^{32}P.
- The fragment is then cut with restriction enzyme and thus the label is removed from one end.
- The fragment of DNA with one end labeled is denatured.
- Four identical samples of these end-labeled DNA restriction fragments are subjected to chemical cleavage at different chemical nucleotides.
- There are four specific sets of chemical reactions that selectively cut the DNA backbone at G, A+G, C+T, or C residues.

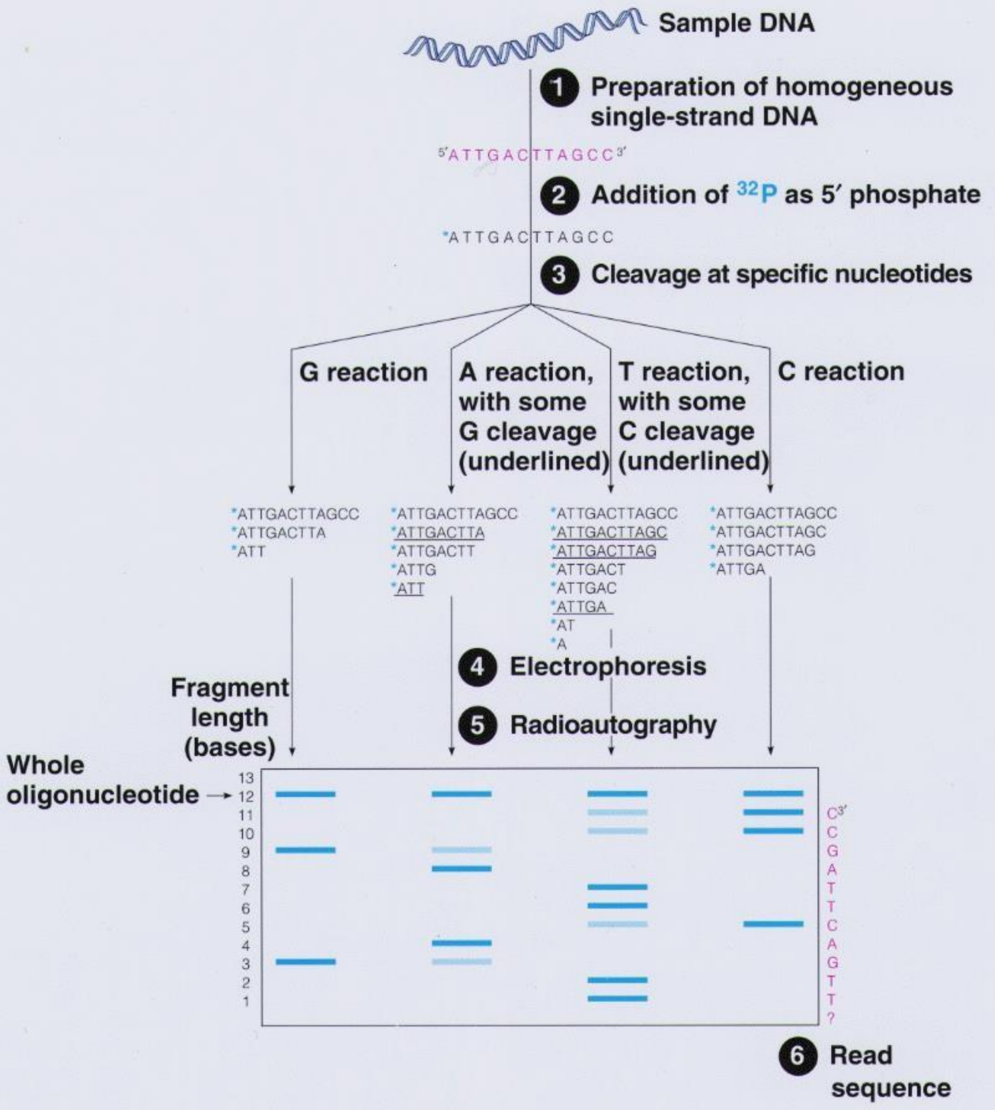
  G only: Dimethyl sulphate (DMS)and piperidine

  A+G: DMS, piperidine

  C+T: Hydrazine, piperidine

  C only: Hydrazine, alkali, piperidine
- For each labeled chain to be broken only once, the reactions are controlled.
- The labeled sub fragments created by the four reactions have the ^{32}P label at one end and the chemical cleavage point at the other end.
- The reaction products are separated by polyacrylamide gel electrophoresis which is based on size. Smallest fragment goes fastest.

Figure 4A.4 Sequencing an oligonucleotide by the Maxam-Gilbert method



From Mathews and van Holde: *Biochemistry 2/e.* © The Benjamin/Cummings Publishing Co., Inc.

Advantages:

- No premature termination due to DNA sequencing. So, no problem with polymerase to synthesize DNA.
- Stretches of DNA can be sequenced which cannot be done with enzymatic method.

Disadvantages:

- Not widely used.
- Use of radioactivity and toxic chemicals.

Next-generation sequencing

DNA sequencing is the process of determining the sequence of nucleotides in a section of DNA. The first commercialised method of DNA sequencing was Sanger sequencing. Next-generation sequencing (NGS) is a high-throughput methodology that enables rapid and cheaply sequencing of the base pairs in DNA or RNA samples than the previously used Sanger sequencing and as such revolutionised the study of genomics and molecular biology (Varshney et al. 2009).

Differences between NGS and Sanger Sequencing:

The concepts behind Sanger vs. next-generation sequencing (NGS) technologies are similar. In NGS and Sanger sequencing, DNA polymerase adds fluorescent nucleotides one by one onto a growing DNA template strand. Each incorporated nucleotide is identified by its fluorescent tag. The critical difference between Sanger sequencing and NGS is sequencing volume. While the Sanger method only sequences a single DNA fragment at a time, NGS is massively parallel, sequencing millions of fragments simultaneously per run. This high-throughput process translates into sequencing hundreds to thousands of genes at one time. NGS also offers greater discovery power to detect novel or rare variants with deep sequencing.

When to use NGS vs. Sanger Sequencing?

Sanger sequencing can be a good choice when interrogating a small region of DNA on a limited number of samples or genomic targets (~20 or fewer). NGS allows you to screen more samples cost-effectively and detect multiple variants across targeted areas of the genome an approach that would be costly and time-consuming using Sanger sequencing (Prober et al. 1987).

The four main advantages of NGS over classical Sanger sequencing are:

Sample size

NGS is significantly cheaper, quicker, needs significantly less DNA and is more accurate and reliable than Sanger sequencing. For Sanger sequencing, a large amount of template DNA is needed for each read. Several strands of template DNA are needed for each base being sequenced (i.e. for a 100bp sequence we need many hundreds of copies, for a 1000bp sequence we need many thousands of copies), as a strand that terminates on each base is needed to construct a full sequence. In NGS, a sequence can be obtained from a single strand. In both kinds of sequencing multiple staggered copies are taken for contig construction and sequence validation.

Speed

NGS is quicker than Sanger sequencing in two ways. Firstly, the chemical reaction may be combined with the signal detection in some versions of NGS, whereas in Sanger sequencing these are two separate processes. Secondly and more significantly, only one read (maximum ~1kb) can be taken at a time in Sanger sequencing, whereas NGS is massively parallel, allowing 300Gb of DNA to be read on a single run on a single chip.

Cost

The reduced time, manpower and reagents in NGS mean that the costs are much lower. The first human genome sequence cost in the region of £300M. Using modern Sanger sequencing methods, aided by data from the known sequence, a full human genome would still cost £6M. Sequencing a human genome with Illumina today would cost less than £1,000 (Smailus et al. 2005; Service et al. 2006).

Accuracy

Repeats are intrinsic to NGS, as each read is amplified before sequencing, and because it relies on many short overlapping reads, so each section of DNA or RNA is sequenced multiple times. Also, because it is so much quicker and cheaper, it is possible to do more repeats than with Sanger sequencing. More repeats means greater coverage, which leads to a more accurate and reliable sequence, even if individual reads are less accurate for NGS.

Sanger sequencing can be used to give much longer sequence reads. However, the parallel nature of NGS means that longer reads can be constructed from many contiguous short reads.

Next generation methods of DNA sequencing have three general steps:

Library preparation: libraries are created using random fragmentation of DNA, followed by ligation with custom linkers

Amplification: the library is amplified using clonal amplification methods and PCR

Sequencing: DNA is sequenced using one of several different approaches

Library Preparation

Firstly, DNA is fragmented either enzymatically or by sonication (excitation using ultrasound) to create smaller strands. Adaptors (short, double-stranded pieces of synthetic DNA) are then ligated to these fragments with the help of DNA ligase, an enzyme that joins DNA strands. The adaptors enable the sequence to become bound to a complementary counterpart.

Adaptors are synthesised so that one end is 'sticky' whilst the other is 'blunt' (non-cohesive) with the view to joining the blunt end to the blunt ended DNA. This could lead to the potential problem of base pairing between molecules and therefore dimer formation. To prevent this, the chemical structure of DNA is utilised, since ligation takes place between the 3′-OH and 5′-P ends. By removing the phosphate from the sticky end of the adaptor and therefore creating a 5′-OH end instead, the DNA ligase is unable to form a bridge between the two termini.

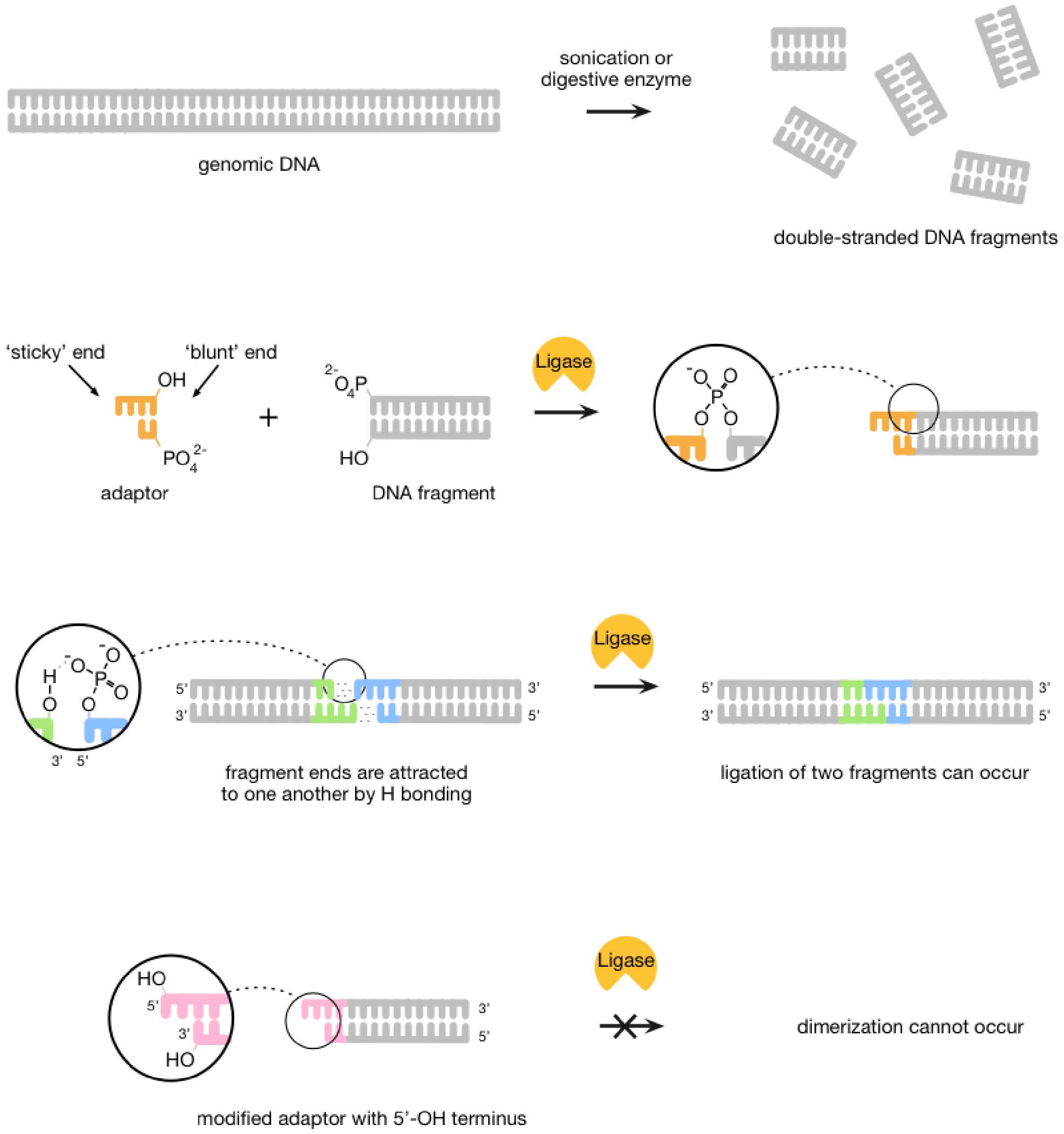

sonication or
digestive enzyme
genomic DNA
double-stranded DNA fragments
'sticky' end
'blunt' end
OH
PO_4^{2-}
adaptor
+
$^{2-}O_4P$
HO
DNA fragment
Ligase
^{-}O O P O O
H O ^{-}O P ^{-}O O O
3' 5'
5'
3'
fragment ends are attracted
to one another by H bonding
Ligase
ligation of two fragments can occur
HO
5'
3'
HO
modified adaptor with 5'-OH terminus
Ligase
dimerization cannot occur

Figure 1: **Library preparation of Next-generation sequencing**

In order for sequencing to be successful, the library fragments need to be spatially clustered in PCR colonies or 'polonies' as they are conventionally known, which consist of many copies of a particular library fragment. Since these polonies are attached in a planar fashion, the features of the array can be manipulated enzymatically in parallel. This method of library construction is much faster than the previous labour-intensive procedure of colony picking and *E. coli* cloning used to isolate and amplify DNA for Sanger sequencing, however, this is at the expense of read length of the fragments.

Amplification

Library amplification is required so that the received signal from the sequencer is strong enough to be detected accurately. With enzymatic amplification, phenomena such as biasing and duplication can occur leading to preferential amplification of certain library fragments. Instead, there are several types of amplification process which use PCR to create large numbers of DNA clusters.

Emulsion PCR

Emulsion oil, beads, PCR mix and the library DNA are mixed to form an emulsion which leads to the formation of micro wells. In order for the sequencing process to be successful, each micro well should contain one bead with one strand of DNA (approximately 15% of micro wells are of this composition). The PCR then denatures the library fragment leading two separate strands, one of which (the reverse strand) anneals to the bead. The annealed DNA is amplified by polymerase starting from the bead towards the primer site. The original reverse strand then denatures and is released from the bead only to re-anneal to the bead to give two separate strands. These are both amplified to give two DNA strands attached to the bead. The process is then repeated over 30-60 cycles leading to clusters of DNA. This technique has been criticised for its time consuming nature, since it requires many steps (forming and breaking the emulsion, PCR amplification, enrichment etc) despite its extensive use in many of the NGS platforms. It is also relatively inefficient since only around two thirds of the emulsion micro reactors will actually contain one bead. Therefore an extra step is required to separate empty systems leading to more potential inaccuracies.

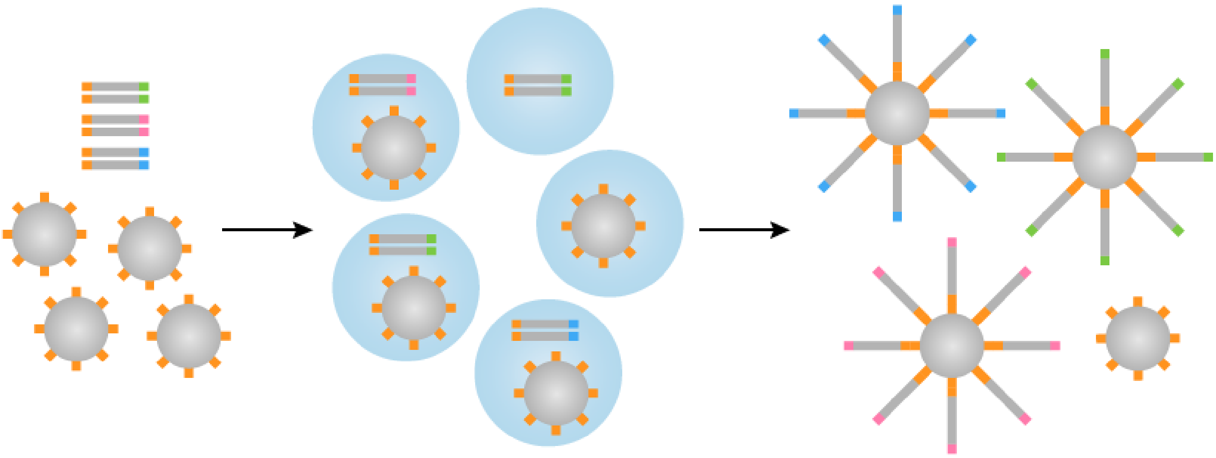


Figure 2: **Emulsion PCR**

Bridge PCR

The surface of the flow cell is densely coated with primers that are complementary to the primers attached to the DNA library fragments (Figure 3). The DNA is then attached to the surface of the cell at random where it is exposed to reagents for polymerase based extension. On addition of nucleotides and enzymes, the free ends of the single strands of DNA attach themselves to the surface of the cell via complementary primers, creating bridged structures. Enzymes then interact with the bridges to make them double stranded, so that when the denaturation occurs, two single stranded DNA fragments are attached to the surface in close proximity. Repetition of this process leads to clonal clusters of localised identical strands. In order to optimise cluster density, concentrations of reagents must be monitored very closely to avoid overcrowding.

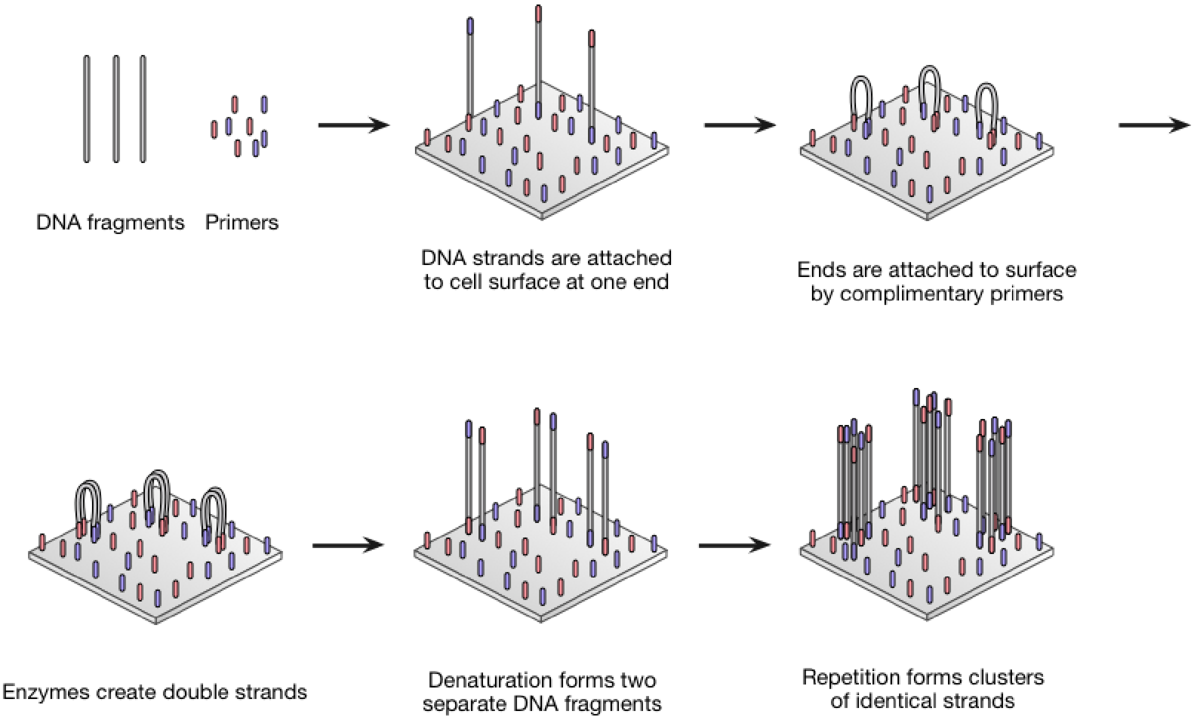

DNA fragments
Primers
DNA strands are attached to cell surface at one end
Ends are attached to surface by complimentary primers
Enzymes create double strands
Denaturation forms two separate DNA fragments
Repetition forms clusters of identical strands

Figure 3: **Bridging PCR**

Sequencing

Several competing methods of Next Generation Sequencing have been developed by different companies.

454 Pyrosequencing

Pyrosequencing is based on the 'sequencing by synthesis' principle, where a complementary strand is synthesised in the presence of polymerase enzyme (Figure 4). In contrast to using dideoxynucleotides to terminate chain amplification (as in Sanger sequencing), pyrosequencing instead detects the release of pyrophosphate when nucleotides are added to the DNA chain. It initially uses the emulsion PCR technique to construct the polonies required for sequencing and removes the complementary strand. Next, a ssDNA sequencing primer hybridizes to the end of the strand (primer-binding region), then the four different dNTPs are then sequentially made to flow in and out of the wells over the polonies. When the correct dNTP is enzymatically incorporated into the strand, it causes release of pyrophosphate. In the presence of ATP sulfurylase and adenosine, the pyrophosphate is converted into ATP. This ATP molecule is used for luciferase-catalysed conversion of luciferin to oxyluciferin, which produces light that can be detected with a camera. The relative intensity of light is proportional to the amount of base added (i.e. a peak of twice the intensity indicates two identical bases have been added in succession).

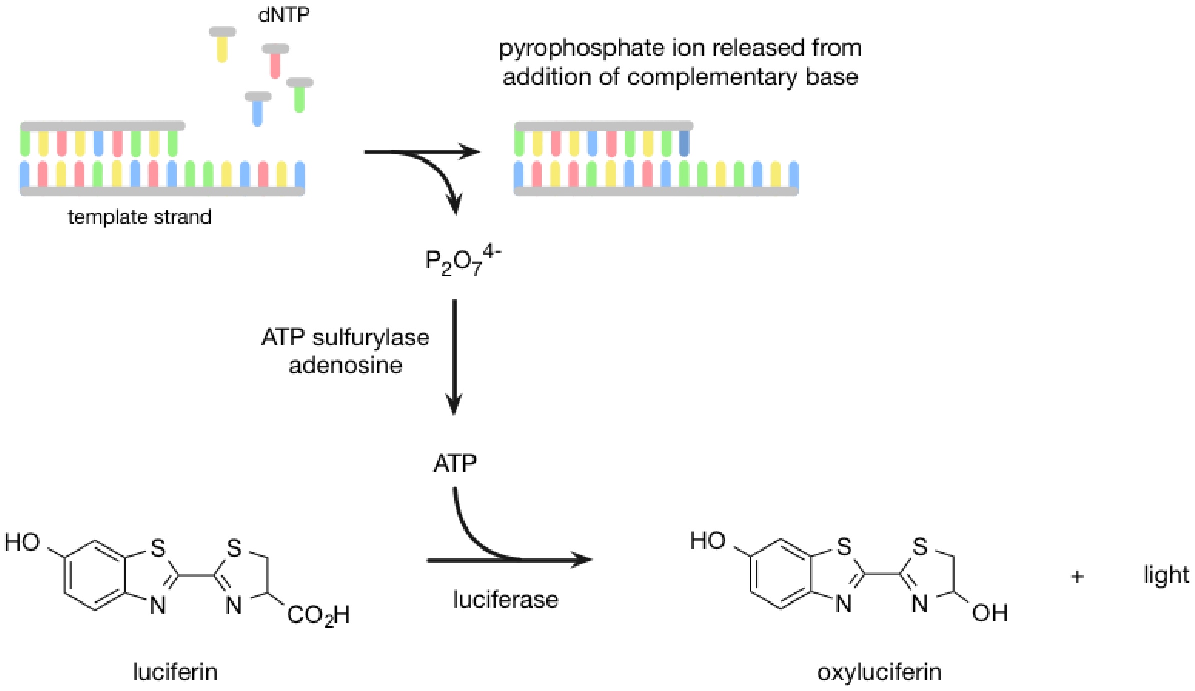


Figure 4: **454 Pyrosequencing**

Pyrosequencing, developed by 454 Life Sciences, was one of the early successes of Next-generation sequencing; indeed, 454 Life Sciences produced the first commercially available Next-generation sequencer. However, the method was eclipsed by other technologies and, in 2013, new owners Roche announced the closure of 454 Life Sciences and the discontinuation of the 454 pyrosequencing platform.

Ion torrent semiconductor sequencing

Ion torrent sequencing uses a "sequencing by synthesis" approach, in which a new DNA strand, complementary to the target strand, is synthesized one base at a time. A semiconductor chip detects the hydrogen ions produced during DNA polymerization (Figure

5). Following polony formation using emulsion PCR, the DNA library fragment is flooded sequentially with each nucleoside triphosphate (dNTP), as in pyrosequencing. The dNTP is then incorporated into the new strand if complementary to the nucleotide on the target strand. Each time a nucleotide is successfully added, a hydrogen ion is released, and it detected by the sequencer's pH sensor. As in the pyrosequencing method, if more than one of the same nucleotide is added, the change in pH/signal intensity is correspondingly larger.

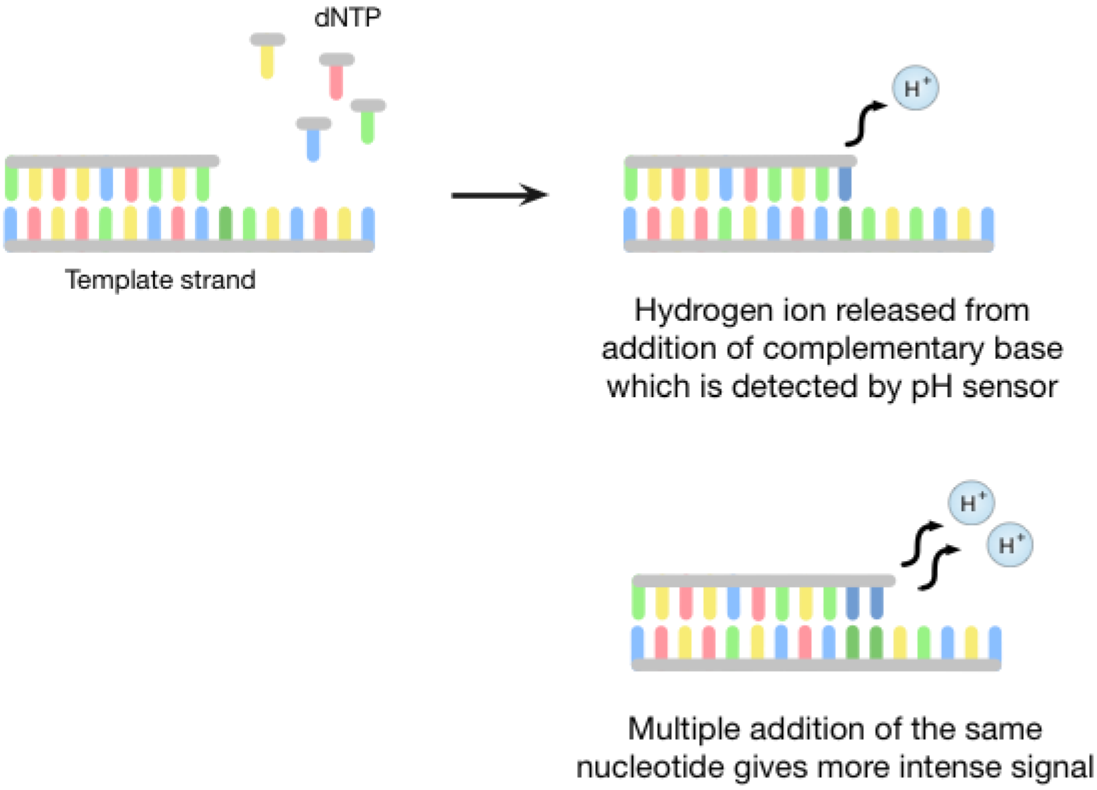

dNTP
Template strand
H^+
Hydrogen ion released from addition of complementary base which is detected by pH sensor
H^+
H^+
Multiple addition of the same nucleotide gives more intense signal

Figure 5: **Ion Torrent semiconductor sequencing**

Ion torrent sequencing is the first commercial technique not to use fluorescence and camera scanning; it is therefore faster and cheaper than many of the other methods. Unfortunately, it can be difficult to enumerate the number of identical bases added consecutively. For example, it may be difficult to differentiate the pH change for a homorepeat of length 9 to one of length 10, making it difficult to decode repetitive sequences. **Sequencing by ligation (SOLiD)**

SOLiD is an enzymatic method of sequencing that uses DNA ligase, an enzyme used widely in biotechnology for its ability to ligate double-stranded DNA strands (Figure 6). Emulsion PCR is used to immobilise/amplify a ssDNA primer-binding region (known as an adapter) which has been conjugated to the target sequence (i.e. the sequence that is to be sequenced) on a bead. These beads are then deposited onto a glass surface − a high density of beads can be achieved which which in turn, increases the throughput of the technique.

Once bead deposition has occurred, a primer of length N is hybridized to the adapter, then the beads are exposed to a library of 8-mer probes which have different fluorescent dye at the 5' end and a hydroxyl group at the 3' end. Bases 1 and 2 are complementary to the nucleotides to be sequenced whilst bases 3-5 are degenerate and bases 6-8 are inosine bases. Only a complementary probe will hybridize to the target sequence, adjacent to the primer. DNA ligase is then uses to join the 8-mer probe to the primer. A phosphorothioate linkage between bases 5 and 6 allows the fluorescent dye to be cleaved from the fragment using silver ions. This cleavage allows fluorescence to be measured (four different fluorescent dyes are used, all of which have different emission spectra) and also generates a 5'-phosphate group which can undergo further ligation. Once the first round of sequencing is completed, the extension product is melted off and then a second round of sequencing is perfomed with a primer of length N−1. Many rounds of sequencing using shorter primers each time (i.e. N−2, N−3 etc) and measuring the fluorescence ensures that the target is sequenced. Due to the two- base sequencing method (since each base is effectively sequenced twice), the SOLiD technique is highly accurate (at 99.999% with a sixth primer, it is the most accurate of the second-generation platforms) and also inexpensive. It can complete a single run in 7 days and in that time can produce 30 Gb of data. Unfortunately, its main disadvantage is that read lengths are short, making it unsuitable for many applications.

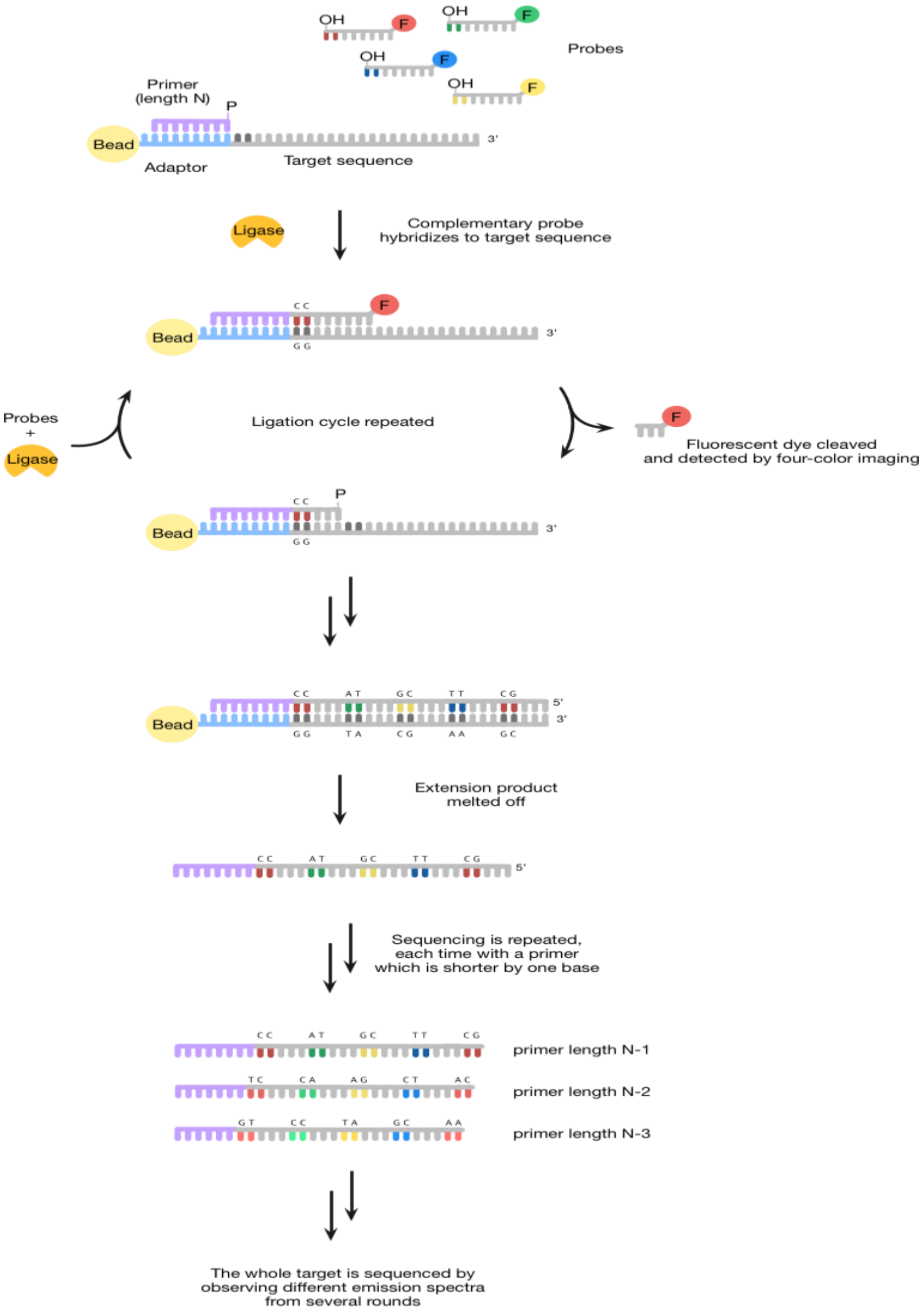


Figure 6: **Sequencing by ligation**

Illumina Sequencing

In this method, the sample DNA is fragmented, and two different adapters are ligated to their 50 and 30 ends. The fragments are attached to an especially prepared substrate on a flow cell, which contains a dense lawn of primers to be used in the next step of solid phase PCR. Fold-back PCR or bridge PCR produces up to 1,000 identical copies of each DNA fragment. All the copies of one fragment form an isolated cluster of molecules on the flow cell. All the clusters formed on a flow cell together represent the in vitro library. The sequencing primer is now attached to the free ends of the fragments. The four dNTPs used for DNA synthesis have fluorophores linked to them; these fluorophores also serve as chain terminators. The dNTPs are added one at a time, and a CCD camera records their incorporation at the 30 end of the sequencing primer/growing chain as fluorescence from the fluorophores attached to them. The fluorophore terminator is removed from the dNTP that has just been added to the primer/growing chain, making this nucleotide available for further DNA synthesis. A new dNTP is now added to the reaction mixture, it is incorporated at the ends of the growing chains, the fluorescence is recorded, and then the fluorophore is removed. In this way, the sequence of each DNA fragment is determined. The use of fluorophore chain terminators linked to the dNTPs eliminates the error in base sequence determination when the same base is present at two or more consecutive positions in the template strand.

This technique was pioneered by Illumina, with their HiSeq and MiSeq platforms. HiSeq is the cheapest of the second-generation sequencers with a cost of $0.02 per million bases. It also has a high data output of 600 Gb per run which takes around 8 days to complete.

Third Generation Sequencing

A new cohort of techniques has since been developed using single molecule sequencing and single real time sequencing, removing the need for clonal amplification. This reduces errors caused by PCR, simplifies library preparation and, most importantly, gives a much higher read length using higher throughput platforms. Examples include Pacific Biosciences' platform which uses SMRT (single molecule real time) sequencing to give read lengths of around one thousand bases and Helicos Biosciences which utilises single molecule sequencing and therefore does not require amplification prior to sequencing. Oxford Nanopore Technologies are currently developing silicon-based nanopores which are subjected to a current that changes as DNA passes through the pore. This is anticipated to be a high-throughput rapid method of DNA sequencing, although problems such as slowing transportation through the pore must first be addressed.

Applications of Next-Generation Sequencing

Next generation sequencing has enabled researchers to collect vast quantities of genomic sequencing data. This technology has a plethora of applications, such as: wholegenome sequencing, analysis of epigenetic modifications, mitochondrial sequencing, transcriptome sequencing − understanding how altered expression of genetic variants affects an organism and exome sequencing.

As the cost of DNA sequencing goes down, sequencing produces huge volumes of data, and there are many computational challenges associated with processing and storing the data.

References

Prober, J.M. et al. (1987) A system for rapid DNA sequencing with fluorescent chain terminating dideoxynucleotides. Science 238, 336–341

Smailus, D.E. et al. (2005) Simple, robust methods for high-throughput nanoliter-scale DNA sequencing. Genome Res. 15, 1447–1450

Service, R.F. (2006) The race for the $1000 genome. Science 311, 1544–1546

Varshney, R. et al. (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends in Biotechnology 27(9), 523-530