CHAPTER 12C – DNA SEQUENCING

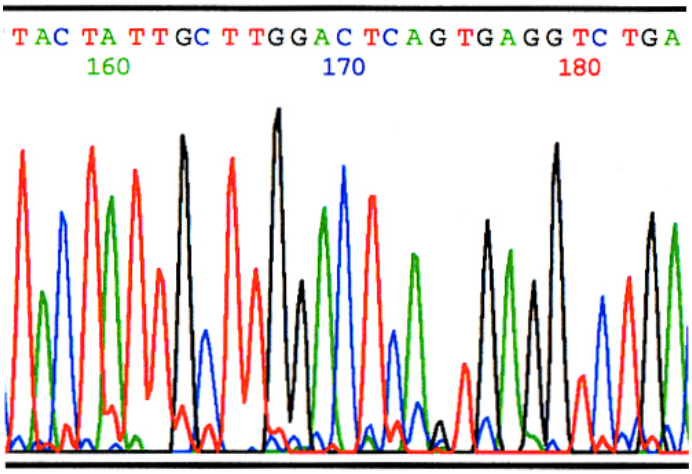

Figure 1.
Output from an automated Sanger DNA sequencer.
(Original-Harrington- CC BY-NC 3.0)

INTRODUCTION

DNA sequencing determines the order of nucleotide bases for a DNA molecule. These DNA molecules could be as small as a single restriction fragment, an entire gene, or as large as an organism's entire genome. Most DNA sequencing at the University of Alberta is done by the Molecular Biology Service Unit (MBSU). They use three machines: (1) an Applied Biosystems ABI 3730, (2) an Illumina MiSeq, and (3) an Illumina NextSeq 500, each has its own advantages and purposes. The 3730 uses an older technology called automated Sanger sequencing, while the Illumina machines perform next-generation DNA sequencing. This chapter will cover how these machines work and what they are used for.

1. AUTOMATED SANGER DNA SEQUENCING

1.1. HISTORICAL CONTEXT

DNA sequencing has had a long history. Beginning in the 1970s there have been many methods and improvements. Some dates that stand out are:

- 1977 - Frederick Sanger invents a popular method, later called manual Sanger sequencing.
- 1986 - Leroy Hood improves upon this method to invent **automated Sanger sequencing**.
- 1987 - **Applied Biosystems** begins selling a machine to perform automated Sanger sequencing, their ABI 370.
- 1995 to 2003 - Using ABI 370s, ABI 377s, and similar machines scientists in the US, UK, and other countries sequenced the human genome.
- 2002 - Applied Biosystems begins selling the **ABI 3730** (**Figure 2**) which became the most popular way to do automated Sanger sequencing and remains so to this day.

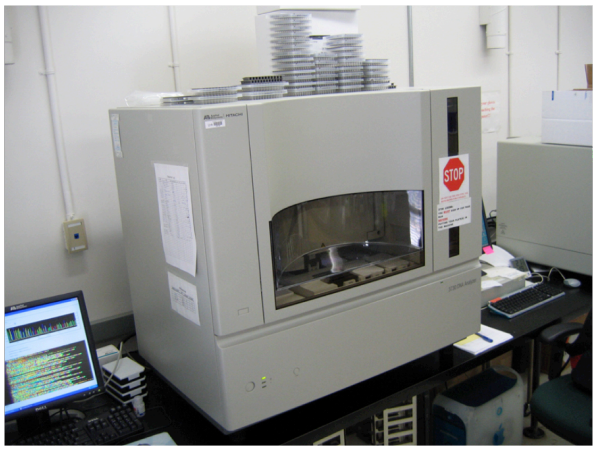

Figure 2.
The Applied Biosytems ABI 3730 in the MBSU (Molecular Biology Service Unit, Biological Sciences Department, U. of Alberta).
(Original-Harrington- CC BY-NC 3.0)

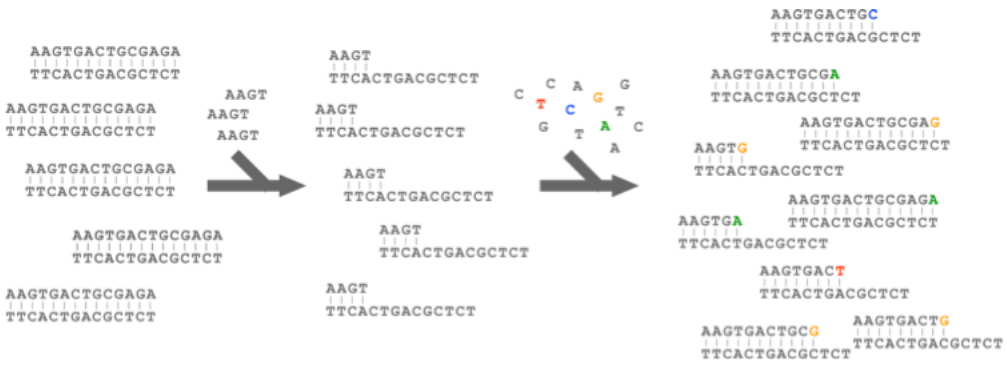

Figure 3.
An automated Sanger sequencing reaction. The regular dNTPs are shown here as black while the ddNTPs are in colour.
(Original-Deyholos- CC BY-NC 3.0)

1.2. HOW AUTOMATED SANGER DNA SEQUENCING WORKS

Recall that **DNA Polymerases** incorporate nucleotides (**dNTPs**) into a growing strand of DNA, based on the sequence of a template strand. DNA Polymerases add a new base only to the 3'-OH group of an existing strand of

DNA; this is why primers are required in natural DNA synthesis and in techniques such as PCR. Automated Sanger sequencing relies on the random incorporation of modified nucleotides called **dideoxy nucleotides (ddNTPs**, **Figure 3**).

These lack a 3'-OH group and therefore cannot serve as an attachment site for the addition of the next nucleotide. After a ddNTP is incorporated into a strand of DNA, no further elongation can occur. The ddNTPs are labelled with one of four fluorescent dyes, each specific for one the four nucleotide bases (**Figure 4**).

To sequence a DNA fragment, you need many copies of that fragment (**Figure 5**). Unlike PCR, Automated Sanger sequencing does not amplify the target sequence and only one **primer** is used. This primer is hybridized to the denatured template DNA, and determines where on the template strand the sequencing reaction will begin. A mixture of **regular dNTPs**, **fluorescently-labelled ddNTPs**, and DNA Polymerase is added to a tube containing the primer-template hybrid. The DNA Polymerase will then synthesize a new strand of DNA until a fluorescently-labelled ddNTP nucleotide is incorporated, at which point extension is terminated. Because the reaction contains millions of template molecules,

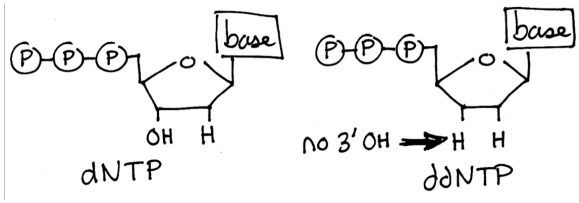

Figure 4.
ddNTPs are synthetic DNA nucleotides that lack a 3' hydroxyl group. If a DNA Polymerase uses one it can't continue.
(Original-Harrington- CC BY-NC 3.0)

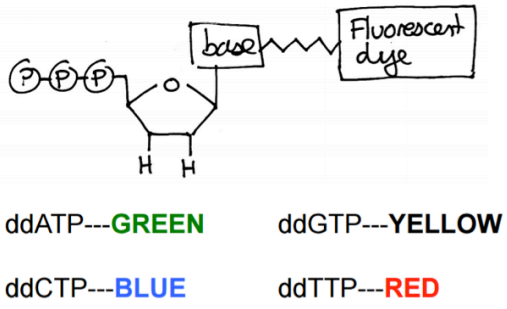

Figure 5.
Each ddNTP used in a sequencing reaction is attached to a different fluorescent molecule using a long chain of carbons. Note that black ink is used to represent yellow fluorescence.
(Original-Harrington- CC BY-NC 3.0)

a sufficient number of shorter molecules is synthesized, each ending in a fluorescent label that corresponds to the last base incorporated. The newly synthesized strands can be denatured from the template, and then separated electrophoretically based on their length (number of bases). The ABI machine is used for this electrophoresis step. While the original, old ABI 370 used a slab gel similar to the ones used in

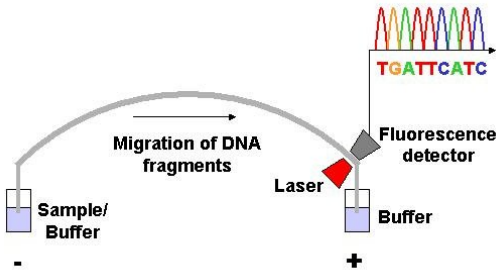

Figure 6.
Fluorescently labeled products can be separated by capillary electrophoresis, generating a chromatogram from which the sequence can be read.
(Wikimedia-Abizar-PD)

undergraduate labs, the newer ABI 3730 uses **capillary tube electrophoresis** (**Figure 6**). In this machine each sample travels through its own tube. Near the end of the tube is a laser, which excites any fluorescent dyes moving past and a detector that collects any emitted light. As each DNA molecule moves past the laser/detector it emits a specific colour. Because there will be several molecules with the same length and same colour the result appears as a peak of colour. A computer monitoring the results can add the sequence information to the colours since red = T and so on. In this way the DNA sequence can be read simply from the order of the colors in successive peaks.

The results from a sequencing reaction are presented as a **chromatograph**. While **Figure 6** only shows 9 peaks, a successful sequencing reaction will generate about 700 nucleotides worth of data. The figure shows the results from a single tube but in fact there can be 48 or 96 tubes in total. Thus in a single 'run' an ABI 3730 machine can sequence up to 67 000 bp of DNA.

1.3. USING AUTOMATED SANGER DNA SEQUENCING TO SEQUENCE A PLASMID

Making a new recombinant plasmid takes time and money. You will want to confirm that it has the DNA sequence it should before you use it for important experiments. A simple way to find out is to sequence it. Let's say you have put a 3.0 kb insert into a pBluescript II plasmid and right now the recombinant plasmid is in *E. coli* cells (**Figure 7**). The first step is to isolate plasmid DNA from some of the cells with a mini-prep protocol. This will be the template DNA. The primer will be oligonucleotides complementary to the pBluescript II vector adjacent to the insert. The sequencing reaction will tell you the sequence of the insert DNA within the plasmid.

1.4. USING AUTOMATED SANGER DNA SEQUENCING TO SEQUENCE A GENE

If you suspect that an organism has a mutation in a specific gene you can use automated Sanger sequencing to find out (**Figure 8**). Let's say you have a mouse strain and you think it has a mutation in a gene you are studying. As before, the first step is to isolate DNA. However, we can't sequence this DNA directly. Amongst all of the genomic DNA there just aren't enough copies of the gene to serve as the template DNA. To overcome this limitation a PCR reaction is used to amplify the gene sequence in question. Then we sequence the PCR product. Depending upon how large the gene is it may take several PCR products and several sequencing reactions to get the whole sequence.

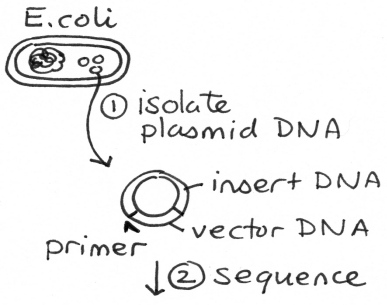

Figure 7
Using automated Sanger sequencing to sequence a plasmid.
(Original – Harrington- CC BY-NC 3.0)

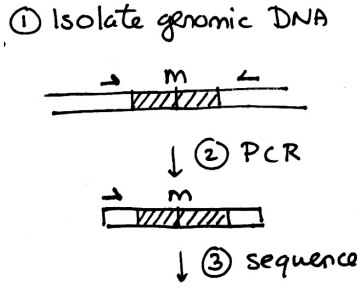

Figure 8.
Using automated Sanger sequencing to sequence a gene. The site of a point mutation is shown here as 'm'.
(Original –Harrington- CC BY-NC 3.0)

2. NEXT-GENERATION DNA SEQUENCING

2.1. HISTORICAL CONTEXT

Sequencing a single gene or plasmid with an ABI 3730 is quick and inexpensive. But sequencing a whole genome this way would be very slow and very expensive. There are two reasons for this.

The first is that automated Sanger sequencing requires many copies of the template DNA. A sample of purified plasmid DNA or purified PCR product has millions of copies of the target region. But a sample of genomic DNA has only a few copies of a specific target region. For many years the only way to sequence an organism was to isolate its genomic DNA, break the DNA into large pieces, and then clone these pieces into BAC (bacteria artificial chromosome) vectors. The BAC clones would then have to be sequenced one by one. Most of the 13 years and millions of dollars it took to sequence the human genome was spent making and organizing these BAC clones.

The second limitation of automated Sanger sequencing is that each reaction can only generate 700 nucleotides worth of data. It took literally millions of independent sequencing reactions to sequence the human genome.

Beginning in the late 1990s scientists realized that there was a need for a machine that could sequence genomic DNA directly and with a single reaction. In several instances a technology was invented in a university lab, developed in a small biotechnology company, and then purchased by a larger biotech company. An example of this is:

- 1996 - Swedish scientists invent a completely new way to sequence DNA called pyrosequencing. It is clever but very labour intensive.
- 2000 - An American inventor and entrepreneur, Jonathan Rothberg, refines their technique into automated pyrosequencing.
- 2004 - His company, 454 Life Sciences, markets the first so called **next-generation sequencing** machine.
- 2007 - The largest biotech company in the world, Roche, buys 454 Life Sciences.
- 2015 - 454, now a subsidiary of Roche, continues to develop and sell next-generation machines.

In 2015 there are several choices for next-generation sequencing. For a few hundred thousand dollars you can purchase a GS FLX (made by 454/Roche), an ABI 5500 (Applied Biosystems), or an Ion Proton (Life Technologies). Each is a fancy looking machine that uses a unique and proprietary technology.

2.2. HOW NEXT GENERATION DNA SEQUENCING WORKS

As mentioned in the introduction to this chapter, the MBSU recently purchased two next generation machines: an **Illumina MiSeq** and an **Illumina NextSeq 500** (**Figure 9**).

Both use a similar workflow (**Figure 10**). The scientist has to isolate genomic DNA from an organism (step 1) and then use a kit to break it into small fragments (step 2). The scientist then loads the fragments into the machine and turns it on. Once inside, the DNA fragments are isolated from each other (step 3), amplified in place (step 4), and finally sequenced (step 5). The technology is called **sequencing by synthesis**. Illumina has made animated movies of what happens within their machines:

www.youtube.com/watch?v=HMyCqWhwB8E

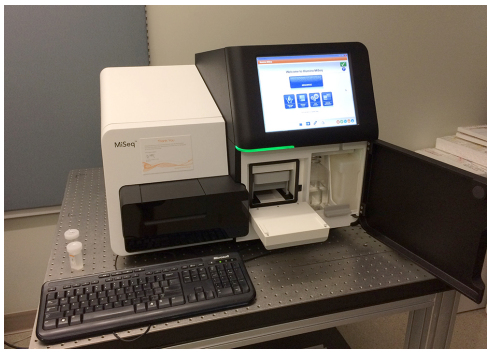

Figure 9.
The Illumina MiSeq in the MBSU (Biological Sciences Department, U. of Alberta). The door has been opened to show where the DNA sample and reaction mixtures are loaded. (Original-Harrington- CC BY-NC 3.0)

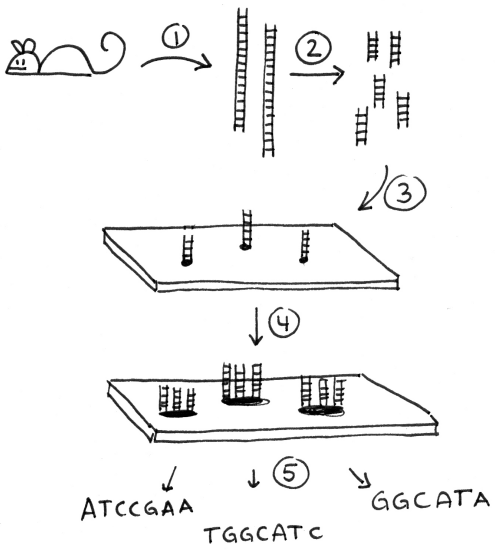

Figure 10.
Next generation DNA sequencing using Illumina's sequencing by synthesis technology. Steps 1 and 2 are done by the researcher and steps 3, 4, and 5 occur within the machine. (Original –Harrington- CC BY-NC 3.0)

Table 1. Comparison between different sequencing machines.

Machine	ABI 3730	Illumina MiSeq	Illumina NextSeq 500
DNA	plasmid or PCR product	genomic DNA	genomic DNA
Technology	automated Sanger	sequencing by synthesis	sequencing by synthesis
Data generated	700 bp	540 Mb to 15 Gb	16 Gb to 120 Gb
Price	$4.75	$1,250 - $1,850	$2,050 - $5,150

The output is just raw sequence data, there are no chromatograms. Powerful software is needed for **sequence assembly**, the process of joining these small pieces of sequence data into a continuous sequence (**Figure 11**). Ultimately there will be one sequence for each of the organism's chromosomes.

2.3. COMPARISON BETWEEN DNA SEQUENCING METHODS

Scientists all over the world now have a choice between automated Sanger sequencing and next-generation sequencing. For example, at the University of Alberta your choices at the MBSU are shown in **Table 1**.

Recall that DNA is measured in base pairs where:

1 kilobase (kb) = 1 000 base pairs (bp)
1 Megabase (Mb) = 1 000 000 bp
1 Gigabase (Gb) = 1 000 000 000 bp

Let's say you wanted to sequence a 2,000 bp long PCR product. You could do this with three sequencing reactions in the ABI 3730 or a single run in the Illumina MiSeq. The first method would cost $15 and the second would cost $1,250. Even though one machine is a decade older it is still the way to go! If you did use the MiSeq you'd end up sequencing the same PCR product over and over. It wouldn't produce any more data.

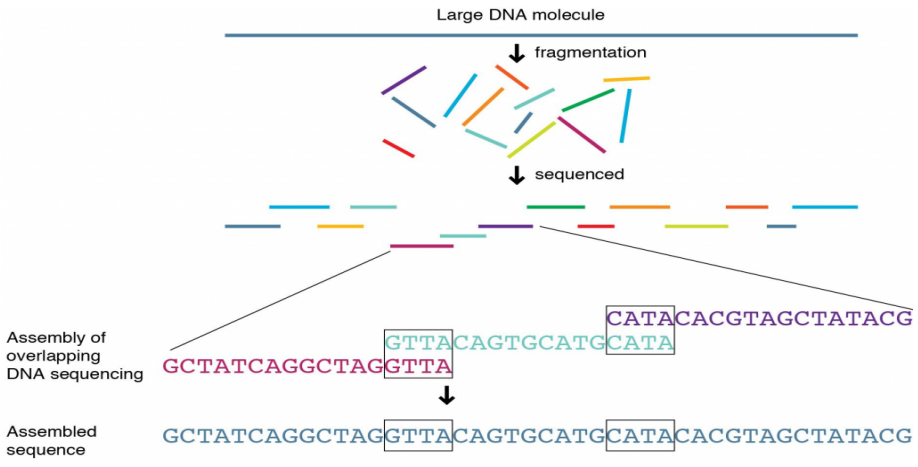

Figure 11.
Assembling the DNA sequence of a chromosome from many smaller sequences.
(NHGRI-Darryl Leja-PD)

On the other hand let's say you wanted to sequence your own DNA. Even if you don't consider the time and cost of making the BAC clones it would still cost millions of dollars to do all of the sequencing reactions in the ABI 3730. Conversely the MBSU could use their NextSeq 500 and have everything done in two days for about $4 000. Each of your 46 chromosomes would be sequenced about 30 times each. A more expensive machine, the Illumina HiSeq, can sequence human DNA for about $1 000 a person.

2.4. USING NEXT-GENERATION DNA SEQUENCING TO SEQUENCE HUMANS

Even though we know the *average* human DNA sequence, each of us is unique. There are two reasons why human DNA continues to be sequenced. (**Table 2**)

2.5. USING NEXT-GENERATION DNA SEQUENCING TO SEQUENCE OTHER ORGANISMS

Next-generation sequencing has made it feasible to sequence anything. Here are just a few examples. (**Table 3**)

Table 2. Using next-generation sequencing to sequence humans.

Use of next-generation sequencing	Description
Personalized genomics	If we sequence a person's DNA it can reveal information about their susceptibility to disease and their response to various medical treatments.
Tumour cell sequencing	If a person has cancer it is now possible to sequence individual cancer cells. This has revolutionized how physicians help their patients. Instead of treatments based upon the location of tumours, treatments can now be designed around the genetic defects that lead to the cells becoming cancerous in the first place.

Table 3. Using next-generation sequencing to sequence other organisms.

Use of next-generation sequencing	Description
***De novo* sequencing**	This is when an organism is sequenced for the first time. For example in 2014 researchers in Sierra Leone sequenced 99 Ebola virus genomes from 78 patients. They identified changes in the virus that caused the recent outbreak.
Metagenomics	This is when the entire collection of DNA in an environment is sequenced to determine which species are present. This technique has been used to show that a person's gut microbes vary with their diet.
RNA Seq	This is when RNAs from a tissue or organ are isolated, copied into DNA molecules, and then sequenced. This reveals which genes were active in the cell, tissue, or organ.

SUMMARY:

- Automated Sanger sequencing became commonplace in 1987 and is still used today. It is used to sequence plasmids and PCR products. The most popular machines are Applied Biosystem's ABI 3730s.
- Next-generation sequencing began in 2004 as a better way to sequence whole genomes. There are several competing technologies, for example Illumina's MiSeq machine and its sequencing by synthesis technology.
- Sequencing centres offer both types of sequencing today.
- Sequencing any DNA molecules, large or small, is now fast and inexpensive.

KEY TERMS:

automated Sanger sequencing
Applied Biosystems ABI 3730
DNA Polymerases
primer
regular dNTPs
fluorescently-labelled ddNTPs
capillary tube electrophoresis
chromatogram
next-generation sequencing
Ilumina MiSeq
Illumina NextSeq 500
sequencing by synthesis
sequence assembly
personalized genomics
tumour cell sequencing
de novo sequencing
metagenomics
RNA Seq

STUDY QUESTIONS:

1) What would the chromatogram look like if you set up an automated Sanger sequencing reaction with only template, primers, polymerase, and fluorescent ddNTPs?

2) How could you use DNA sequencing to identify new species of marine microorganisms?

3) An alternative name for automated Sanger sequencing is dye-terminator sequencing. Why is this term appropriate?

4) Ten years ago it would have cost $100,000,000 to sequence your DNA. Today it would cost as little as $1,000. Why did the cost go down so much?

5) Why haven't next-generation machines completely replaced the first generation of automated DNA sequencers?

6) True or false: Automated pyrosequencing and sequencing by synthesis are both considered next-generation DNA sequencing technologies.

Notes: