DNA sequencing and analysis: How it's done

David Glenny[1], Steve Wagstaff[2], Murray Dawson[3]

DNA sequencing is now a very established technique, practiced in many labs throughout New Zealand and the world. The approaches are similar for plant and animal tissue, only the DNA regions that are studied differ. The steps are:

Extraction

DNA is extracted from fresh plant material by grinding it in a mortar and pestle with chemicals that block the action of enzymes that would otherwise destroy the DNA. The extract is washed and centrifuged to clean and concentrate the DNA.

Amplification

Small segments (gene regions) of DNA from the extract are copied repeatedly ('amplified'), millions of times, using the now famous PCR (polymerase chain reaction) technique (Fig. 1 A–D). PCR has been one of the most important advances made in molecular biology, and the technique revolutionised the study of DNA to such an extent that its creator (Dr Kary Mullis) was awarded a Nobel Prize in Chemistry in 1993.

PCR depends on an enzyme called DNA polymerase, originally found in a thermophilic (heat loving) bacterium, that copies DNA and operates at higher temperatures than normal. DNA polymerase was sold for about NZ$500 per millilitre, but the patent has expired so costs are now decreasing. The reactions are done in small plastic test-tubes (0.2–0.5 ml) placed in wells in a metal block that is alternately heated and cooled, under the control of a microprocessor chip similar to that found in a bread-maker (Fig. 2 A–B).

Heating the samples separates the two halves of each DNA strand, a cocktail including DNA polymerase copies each half, and cooling the sample then re-anneals the various half-strands back into double strands.

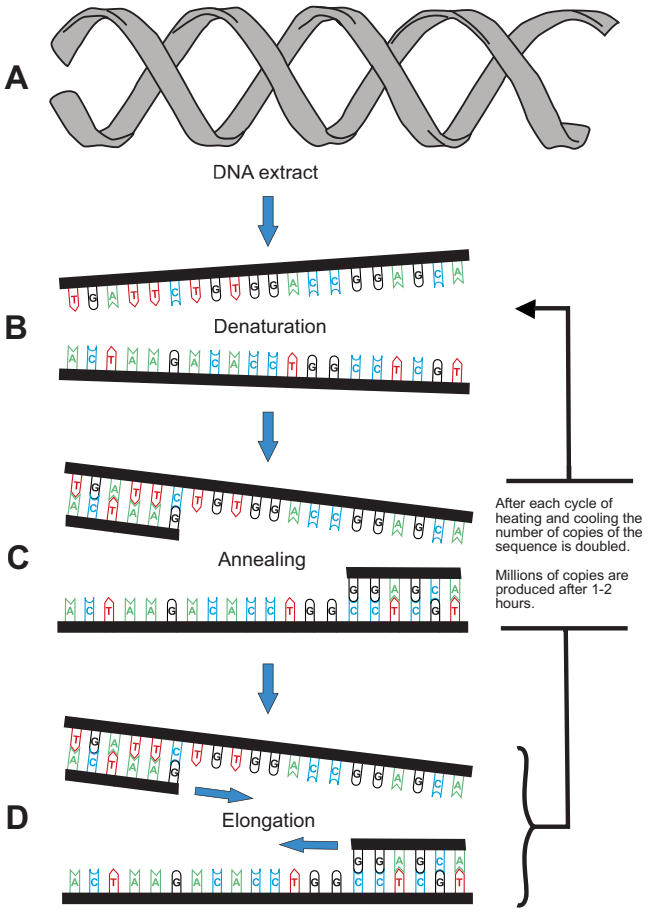

Fig. 1 Amplification of DNA using the polymerase chain reaction (PCR): **A**, double stranded DNA extract; **B**, denaturation of DNA strands at 95°C; **C**, annealing (binding of primers, e.g., at 55°C); **D**, elongation (replication of DNA starting from the primers, 72°C). The two resulting strands make up the template DNA for the next cycle, thus doubling the amount of DNA duplicated for each cycle. Each cycle is repeated many times.

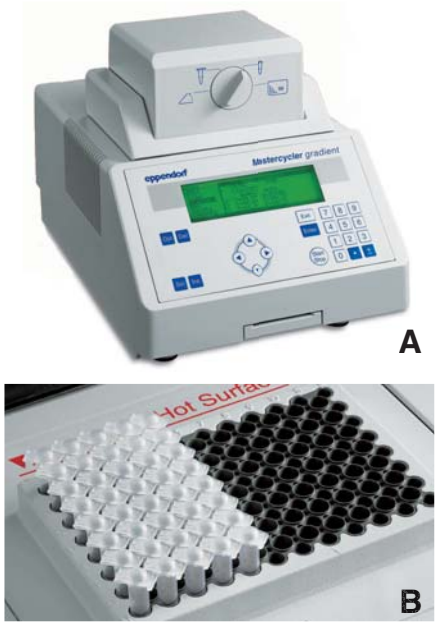

Fig. 2 **A**, PCR machine (also known as a Thermal Cycler); **B**, under the lid, there is a heating block loaded with small tubes. Each tube contains the four nucleotide building blocks for a DNA molecule, heat stable DNA polymerase, and a DNA extraction from one species that serves as a template for the synthesis. Images courtesy of Eppendorf*[1].

Sequencing

A dye marker is attached to the end of each DNA molecule, one colour for each of the four terminating DNA bases, A, C, G, T (Adenine, Cytosine, Guanine, and Thiamine). The amplified DNA is then loaded into a DNA sequencing machine. This machine has a row of about 100 micro-capillary tubes with an electric potential across the tube ends and processes a batch of about 25 samples at a time, with 4 tubes per sample, one for each A, C, G, and T base. Since DNA molecules are slightly negatively charged they move through the capillary tube, attracted to the positive terminal at the far end of the tube, at rates according to their size. So it is

[1] Landcare Research, PO Box 40, Lincoln 7640; glennyd@landcareresearch.co.nz;
[2] wagstaffs@landcareresearch.co.nz;
[3] dawsonm@landcareresearch.co.nz

[1] Eppendorf* is a registered trademark.

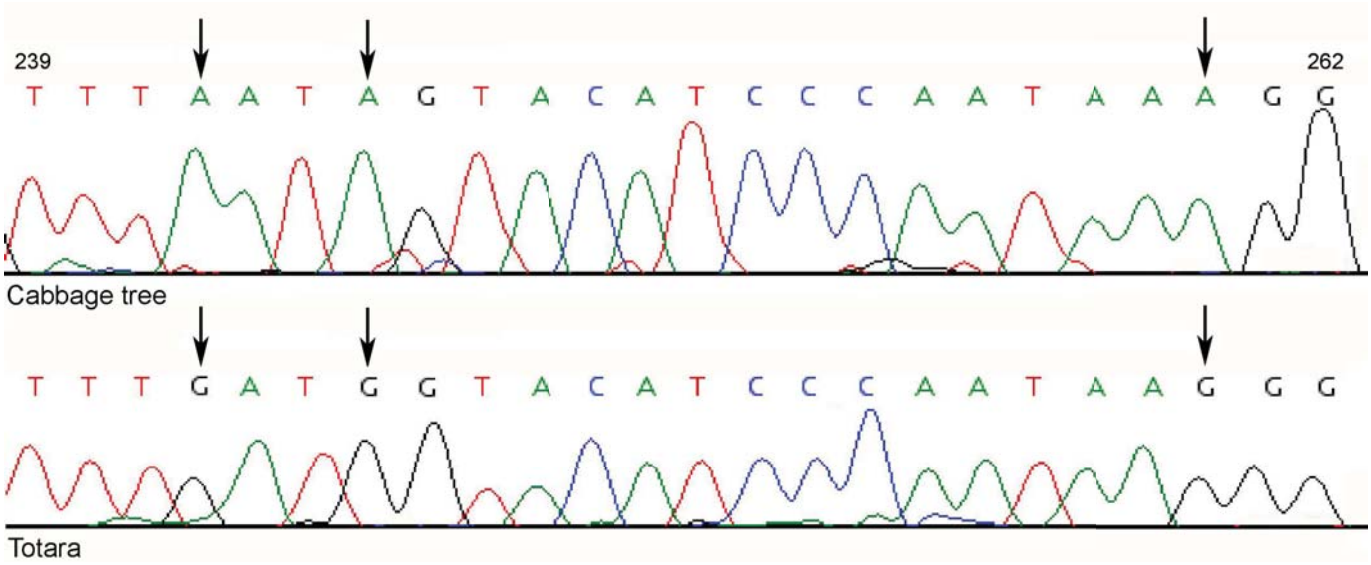

Fig. 3 Small portion of a printout from the results of DNA sequencing. Sequences can be hundreds of bases long; only 23 bases are shown here. The sequencing machine gives the sequence of bases, but also plots the intensity of the dye as read by the laser so that the sequence can be visually interpreted. Each base corresponds to a different coloured dye peak. Arrows have been added to show three base differences between these two sequence fragments (top row is a portion of a cabbage tree sequence and the bottom row is of totara).

Taxa \ Characters		239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262
1	Cabbage tree	T	T	T	A	A	T	A	G	T	A	C	A	T	C	C	C	A	A	T	A	A	A	G	G
2	Totara	T	T	T	G	A	T	G	G	T	A	C	A	T	C	C	C	A	A	T	A	A	G	G	G
3	Rata	T	T	T	A	A	T	A	G	T	A	C	A	T	C	C	C	A	A	T	A	G	G	G	G
4	Hebe	T	T	T	A	A	T	A	G	T	A	C	A	T	C	C	C	A	A	C	A	G	A	G	G
5	Karaka	T	T	T	A	A	T	A	G	T	A	C	A	T	C	C	C	A	A	T	A	G	A	G	G

Fig. 4 Small portion of alignment of sequences using the computer program MacClade. Each row represents a sequence from the DNA of an individual plant (in this simple demonstration, only five sequences were aligned – from a cabbage tree, totara, rata, hebe, and karaka). Each column represents the nucleotide (A, C, G, T) position in the DNA molecule.

essentially a sorting process with small fragments migrating further than large fragments. A laser then scans the capillary tubes, detecting the DNA fragments that are labelled with the four different dyes. The results of the scans are provided as a chart and a sequence of bases (Fig. 3).

DNA sequence analysis

Alignment

The sequences of each sample are aligned by columns using a computer program that recognises the pattern shared by each, using the fact that in sequences of related species, about 95–99% of the bases will be the same (Fig. 4).

Tree-building

Once aligned, the sequences are compared to detect differences between lineages caused by mutations at single base sites, again using a computer program.

The differences of interest are those shared by two or more species as these suggest a relationship between those species – a mutation inherited from a common ancestor. A mutation occurring independently two or more times by coincidence can't be ruled out, so the assertion that two samples are related because they share a mutation is a statement of probability.

Using the data provided by the shared mutations, the same computer program draws the tree diagram (Fig. 5). This tree is a statistical inference about the true phylogenetic (evolutionary) tree. The true phylogenetic tree can never be known with certainty, but at their best, trees based on DNA sequence results have a high probability of being correct. The trees are often presented with an estimate of probability on each branch in the tree.

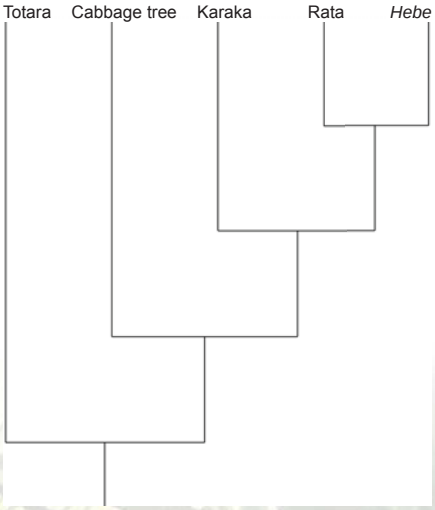

Fig. 5 A tree diagram that attempts to represent the evolutionary relationships within a group of related species.

These phylogenetic reconstructions are powerful tools for investigating the evolutionary and taxonomic relationships of living organisms. The huge impact that DNA sequencing has made on our understanding of the New Zealand flora is discussed in the following articles.