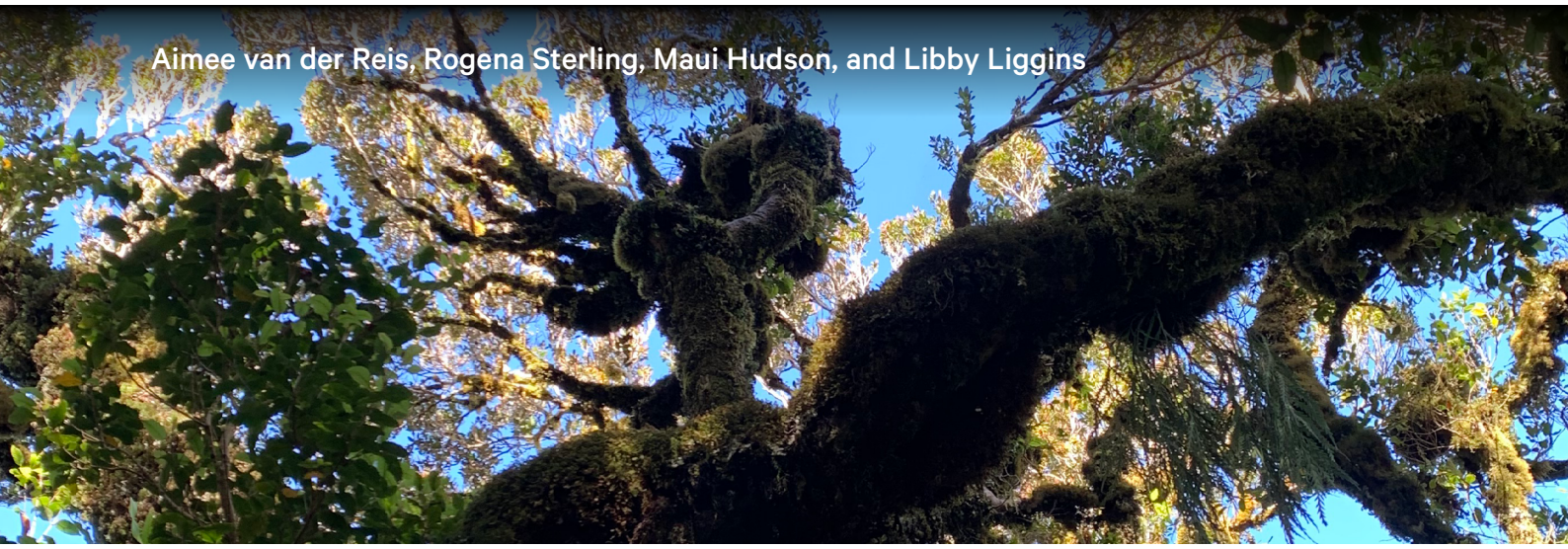
INFOSHEET – JULY 2024

# DNA Information Sheet

Te Kotahi Research Institute
THE UNIVERSITY OF WAIKATO



Aimee van der Reis, Rogena Sterling, Maui Hudson, and Libby Liggins

## What is DNA?

Deoxyribonucleic acid (DNA) is the molecule that carries the genetic information for the development and functioning of an organism.

## What is a DNA sequence?

DNA sequences are long stretches of nucleotide bases - adenine (A), cytosine (C), guanine (G) and thymine (T). DNA sequences can be very short (tens of base pairs long) and provide only a partial gene region, or very long (billions of bases long) that include many genes. The order and number of bases in the DNA sequence of a gene region, or a genome, is often unique to a species, so can help to distinguish and identify species. Similarities in these DNA sequences among species help us identify which species are closely related (i.e., share common ancestors), or have similar function (e.g., some DNA sequences provide the code for making proteins).

## Advances in DNA sequencing technology

Rapid development in the field of genetics since DNA discovery means we are increasingly able to sequence DNA from any part of the genome, for any species. Advances in sequencing technologies and decreasing costs in their application mean we can now sequence much longer DNA sequences. In some cases, the whole genome may be sequenced, but in many cases shorter DNA sequences are sufficient for the purpose of the study. Each data type has their own application(s). Here, we will define five frequently used data types.

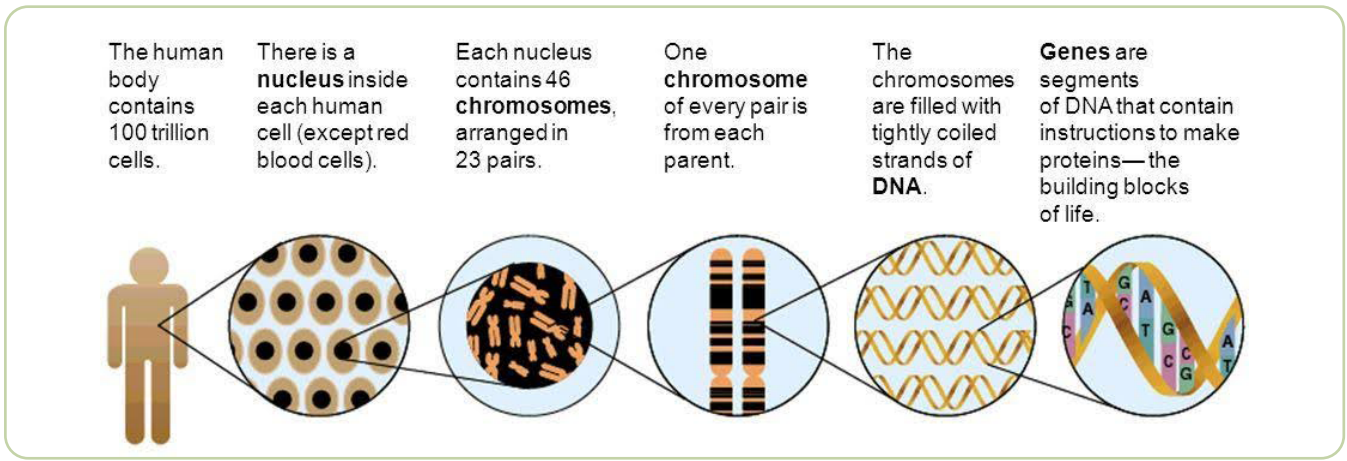



**Figure 1.** An overview of how and where DNA is arranged in human cells. Source: https://slideplayer.com/slide/8041855/

# DNA data types

**DNA barcode:** DNA barcoding is a system for species identification focused on the use of a short gene region acting as a 'barcode'. This is like scanning the barcode of grocery items at a self-checkout at a supermarket – a unique barcode will identify the grocery item scanned in the same way a DNA barcode will identify the species.

**What percentage of the genome is used as a DNA barcode?**
The commonly used mitochondrial cytochrome oxidase I (cox1 or COI) DNA barcode region used for species identification is 313–650 base pairs long. This is less than 0.0000002% of the common house mouse's (Mus musculus) genome which is 2.7 x 109 base pairs. This small gene section also allows for low quality DNA to be used, which is ideal for environmental DNA studies.

**DNA megabarcode:** A DNA megabarcode is a sequence of around 5,000 bases used for species identification. In this method, several DNA barcodes can be targeted at once because they lie relatively close to one another in the genome. With the rapid advances in sequencing technology this approach is increasingly becoming more common as it is time- and cost-effective for species identification, and is efficient when there is a limited amount of DNA or tissue available from a vouchered specimen.

**What percentage of the genome is used as a DNA megabarcode?**
For example, the cox1-3 DNA barcode region of the common house mouse is approximately 4,000 base pairs. This is less than 0.000001% of its genome.

**Genome:** A genome is all the DNA within each chromosome in an organism (Figure 1), it may be billions of bases in length. The genome provides all the information the organism requires to function (i.e., the organism's blueprint), and can be used to understand how an organism functions and what genes are responsible for the various functions.

**How big is a genome?**
The common house mouse has a total of 2.7 x 109 base pairs in its genome, but genome size varies among species. In all cases, to assemble a genome the DNA from all chromosomes (see Figure 1) would be sequenced. Very high-quality DNA is needed to ensure the entire genome is accounted for when sequencing.

**Genome skimming:** Genome skimming is the middle ground between a DNA (mega)barcode and a genome. The DNA is broken into random pieces and sequenced. This is helpful because it is likely to capture many DNA barcodes to identify species, with less time and cost than sequencing DNA (mega)barcodes individually. However, it also captures other areas of the genome that are not useful for species identification which are often seen as uninformative and are discarded.

**What percentage of the genome is used in genome skimming?** Genome skimming, also called low pass or low coverage sequencing, is defined as shallow sequencing down to 0.05× coverage of a genome. Typically, the low coverage does not allow the entire genome to be sequenced. For example, if 0.05× coverage is used this equates to 0.05% of the genome being randomly sequenced. The percentage of the genome sequenced depends on the coverage wanted.

An analogy to better understand genome skimming: All the pieces of a 1,000-piece puzzle are emptied into a large bowl and then all the pieces from another 19 identical puzzles. You want to get an idea of the puzzle's picture, so you select 500 puzzle pieces from the bowl. If you are lucky, all 500 pieces are unique, but the odds are that within those 500 pieces there will be duplicate pieces. The same concept applies to genome skimming, the genome is only partially sequenced with some parts sequenced more than once.

**Reduced-representation sequencing:** Short DNA regions, that are not necessarily barcodes, can also be used to tell us a bit more about individuals of a certain species, or closely related species. For example, how do individuals of a certain species differ between the North and South Island? These are the questions important to conservation and agriculture/aquaculture breeding programmes. To generate these data, the genome is fragmented into short DNA sequences (i.e., less than 100 base pairs) and the variation is compared at specific DNA bases that differ among individuals (aka single nucleotide polymorphisms or SNPs). Due to the reduced nature of this data, it is only informative for the species being studied and for the purposes of the study.

**What percentage of the genome is used in reduced genome studies?**
This answer is dependent on many variables, for species that we don't have much prior genomic information, we can expect to get a few thousand, up to 20,000 bases of variation, but for other organisms we can target specific SNPs, or SNPs at set intervals throughout the genome. For instance, a SNP technology developed by Illumina (a leading developer and manufacturer in DNA sequencing) can identify 143,259 SNPs for the house mouse, which is less than 0.03% of its genome.

| | Bases | Species ID | DNA quality needed | Unexplored functions | Technical difficulty | Cost (low-high; 1-5) |
|---|---|---|---|---|---|---|
| Barcode | 313-650 | Yes | Low | No | Low | 1 |
| Megabarcode | 4,000 | Yes | Low-medium | Unlikely | Low-medium | 2 |
| Reduced genome | $7.3 \times 10^5$ | Unlikely | Medium-high | Possibly | Medium | 3 |
| Genome skimming | $1.35 \times 10^8$ | Likely | High | Yes | Medium-high | 4 |
| Genome | $2.7 \times 10^9$ | Yes | High | Yes | High | 5 |

**Table 1.** A summary of the different DNA dataset types commonly used and their features. The mouse genome is used as an example.

**Funded by:** Ministry for the Environment
**Published by:** Te Mata Punenga o Te Kotahi | Te Kotahi Research Institute, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand. Email: rangahau@waikato.ac.nz
**DOI:** 10.15663/i56.28919
**Citation:** van der Reis AL, Sterling R, Hudson M, and Liggins L. 2024. DNA Information Sheet. Te Kotahi Research Institute.

Te Kotahi Research Institute
THE UNIVERSITY OF WAIKATO

IDIA
Indigenous Design & Innovation Aotearoa