Review

New developments in ancient genomics

Craig D. Millar[1], Leon Huynen[2], Sankar Subramanian[2], Elmira Mohandesan[2] and David M. Lambert[2]

[1] Allan Wilson Centre for Molecular Ecology and Evolution, School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland 1010, New Zealand
[2] Allan Wilson Centre for Molecular Ecology and Evolution, Institute of Molecular Biosciences, Massey University, Private Bag 102904, Auckland 0632, New Zealand

Ancient DNA research is on the crest of a 'third wave' of progress due to the introduction of a new generation of DNA sequencing technologies. Here we review the advantages and disadvantages of the four new DNA sequencers that are becoming available to researchers. These machines now allow the recovery of orders of magnitude more DNA sequence data, albeit as short sequence reads. Hence, the potential reassembly of complete ancient genomes seems imminent, and when used to screen libraries of ancient sequences, these methods are cost effective. This new wealth of data is also likely to herald investigations into the functional properties of extinct genes and gene complexes and will improve our understanding of the biological basis of extinct phenotypes.

'Waves' of ancient DNA research

Ancient DNA (aDNA) is that recovered from any *post mortem* material such as archaeological or historical specimens. Ancient DNA has, for example, been successfully recovered from mummified specimens [1], archival collections of biological material and preserved plant remains [2,3], as well as ice and permafrost material [4,5]. Developments in ancient DNA research have enabled more precise estimates of rates and patterns of microevolution, and an improved understanding of the relationships of extinct species to modern ones. The study of ancient DNA has fascinated the public and scientists alike [6]. In 1984, among the first 'wave' of ancient DNA studies, researchers were successful in isolating and sequencing DNA from *post mortem* material. Using well-established molecular biology techniques, researchers sequenced 229 base pairs (bp) of mitochondrial DNA from the extinct quagga (*Equus quagga*) [7]. The recovered sequences, supported by a phylogenetic analysis, allowed the authors to establish the relationship between the quagga and its close relative the zebra.

The invention of the polymerase chain reaction (PCR) [8] resulted in a 'second wave' of progress in the field. PCR resolved a central problem in ancient DNA research, that of the low quantity of target DNA inherent in subfossils (partially mineralised remains) and historical samples. This low quantity is due to the degradation of DNA, although short fragments typically remain. During this second wave, most ancient DNA research used PCR to focus on obtaining data from mitochondrial or plastid DNAs that are found in hundreds to thousands of copies per cell [9]. During this period the field had a somewhat checkered history. The results of some high-profile studies published during that period are now generally accepted to be the result of contamination [10–13].

Recent ancient DNA successes include the analysis of the population dynamics of extinct bear and bison [14–16], the detection of changes in genes responsible for desirable traits in maize as a result of cultivation [17], the sequencing of the entire mitochondrial genomes of extinct moa of New Zealand [18,19] and the advent of molecular sexing methods that enable the amplification of single-copy nuclear DNA [20,21]. Despite these successes, some authors have argued that it is impossible to sequence the entire nuclear genome of extinct animals and that this is likely to remain the case. For example, it has been suggested that due to the degraded nature of aDNA (Box 1), it is currently impossible to correctly assemble the large number of repeats found in complex genomes [22]. Although the recent development of nested multiplex (from multiple loci) PCR (Figure 1) now allows the rapid recovery of large numbers of complete mitochondrial genomes, the recovery of complete nuclear genomes is a quite different matter [23]. As a result, no complete ancient nuclear genomes have been recovered to date. However, the recent recovery of millions of bases of nuclear DNA from Neanderthals and woolly mammoths has paved the way for large-scale sequencing and alignment of ancient nuclear genomes [24,25]. These large-scale studies signal the beginning of a 'third wave' of progress in ancient DNA

Glossary

Adapter ligation: ligation of oligos of known sequence to the ends of DNA fragments.
Cluster sequencing: sequencing of clusters of DNA, each cluster being produced, by PCR, from a single DNA molecule.
Oligo: oligonucleotides are short sequences of nucleotides (RNA or DNA).
Paired-end reads: sequence reads from both ends of a DNA fragment.
Poly-A tailing: the enzymatic addition of dATPs to the 3′ terminus of DNA.
Shotgun sequencing: in shotgun sequencing, DNA is broken up randomly into numerous small overlapping fragments, cloned, sequenced and the sequences reassembled.
Slide-anchored oligo dT primers: DNA oligos consisting of multiple dTTPs are bound by their 5′ terminus to glass.

Corresponding author: Lambert, D.M. (d.m.lambert@massey.ac.nz).

0169-5347/$ – see front matter doi:10.1016/j.tree.2008.04.002 Available online 22 May 2008

Box 1. Characteristics of aDNA

Three major characteristics of aDNA are important for the recovery of these sequences.

First, the concentration of aDNA in biological material is generally low in comparison to modern samples [53]. In fact, using PCR amplification conditions capable of detecting single molecules, aDNA is not detectable in many ancient biological samples. For example, in a recent study of the tuatara (*Sphenodon punctatus*), 27% of ancient samples, ranging in age from 649 to 8748 years BP, yielded no DNA sequences [54]. The amount of aDNA capable of being retrieved from tissues is usually of the order of 1 ng per 1 mg of tissue, a few orders of magnitude lower than what can be extracted from fresh tissue.

Second, DNA in ancient samples is typically degraded into very short fragments. Nuclear DNA is rarely longer than 150 bp, whereas mitochondrial fragments are typically less than ~400 bp [55]. Interestingly, the size of these fragments varies little with the age of ancient samples, suggesting that most damage occurs relatively soon *post mortem* [56,57]. It is well known that temperature and the presence of water have major effects on the degradation of DNA [2,3]. In cold environments such as Antarctica, DNA is relatively stable and consequently sequences recovered from such permafrost conditions are of generally high quality. For example, up to 80% of the ancient samples from Adélie penguins (*Pygoscelis adeliae*) from Antarctica have been amplified successfully [4,58]. By contrast, only 2–4% of samples recovered from cattle (*Bos*) from arid hot environments in Europe and Africa yielded aDNA [59].

Third, aDNA is usually modified as a result of two major types of processes. These can be summarised as those caused by oxidative damage and those resulting from hydrolytic processes [60]. Common damage found in aDNA includes the oxidation of the purines adenine and guanine and the formation of hydantoin derivatives of the pyrimidines cytosine and thymine [61]. Hydrolytic damage mostly results in the deamination of cytosine and guanine nucleotides, to give uracil and xanthine, respectively. This results in C/G to T/A transitions during subsequent amplification. Unique to aDNA is the observation that transitions from C to T are at least three times more common than those from G to A [62,63]. Recently it has been suggested that C to T transitions represent 'the overwhelming majority of misincorporations' [4], whereas others have even argued that this represents the sole type of ancient damage [64]. Damage levels vary between samples but can be up to 1 in every 200 nucleotides [24]. Studies using hair shafts have shown DNA to be very well preserved in this tissue [65]. This is thought to be a result of the protective properties of the keratin shaft. However, although aDNA is typically damaged and found in low concentrations [27,66], these problems are not insurmountable, given the new technologies at hand.

research. New massively parallel DNA sequencing technologies, capable of producing over 1 gigabase (Gb) (one-third of the human genome) of sequence in a single run, will inevitably propel ancient DNA research forward and, in particular, should make large-scale studies of nuclear genomes feasible [6]. These new sequencing methods are particularly suited to ancient DNA analyses as they sequence fragments up to ~250 bp in length, a size comparable to that found in most degraded ancient genomes.

Here we review the new-generation DNA sequencing technologies, highlighting the advantages and disadvantages of the different systems. We will also detail their likely future uses as they relate to the study of ancient genomes.

New massively parallel DNA sequencing technologies

Prior to the development of the new-generation DNA sequencers [26,27], sequencing relied on cloning or PCR amplification to generate large amounts of template. These templates are then sequenced individually and the products separated by capillary electrophoresis. The most advanced conventional sequencers produce at best ~70 kb of sequence per run. The new sequencers do not rely on capillary electrophoresis and are capable of detecting sequences as they are generated by 'sequencing by synthesis' (SBS) in hundreds of thousands of minute DNA clusters (cluster sequencing; see Glossary). Two of these new DNA sequencers rely on emulsion-based technology: the FLX system from Roche and the SOLiD system manufactured by Applied Biosystems (Table 1). Both machines use primer-coated beads and amplify templates for DNA sequencing within an oil emulsion. By contrast, the Solexa system, marketed by Illumina, uses primers bound to a silica matrix together with a process of 'bridge amplification.' In bridge amplification, DNA fragments are hybridised to bound primers, copied by a polymerase enzyme and denatured; the freshly copied strand hybridises to the 'reverse' primer that is also bound to the silica matrix. Both oil emulsion and silica matrices are very efficient generators of clonal DNA clusters suitable for sequencing. The HeliScope true single-molecule sequencing (tSMS) technology from Helicos BioSciences also relies on SBS but does not require amplified DNA; instead, using high-definition optics, this technology is capable of detecting single-base additions to single DNA or RNA strands [28].

The capacity of the machines varies significantly as a result of the different methods of DNA bead, cluster or strand deposition. The FLX sequencer relies on the successful deposition of single beads into a synthesised array of wells. By contrast, the SOLiD, Solexa and HeliScope machines read from high concentrations of randomly deposited DNA beads, clusters or strands resulting in 100-fold greater amounts of DNA sequence. However, all four machines are capable of sequencing only short regions of DNA in comparison to what can be achieved using current Sanger dideoxy chemistry [29]. The SOLiD, Solexa and HeliScope tSMS systems can sequence very short targets, typically in the range of 20–30 bp, whereas the FLX system is able to sequence longer regions, but is still limited to only ~250 bp. The strength of these machines, however, is their ability to sequence hundreds of thousands to millions of templates simultaneously. Table 1 summarises the characteristics of these machines including their advantages and disadvantages. Because these machines sequence all available templates in a sample, they are particularly suited to the recovery and analysis of complex DNA samples such as those from ancient materials.

Metagenomic analyses

Metagenomics is the recovery of DNA sequences from biological samples that contain DNA from many species. For example, gut tissue will contain not only DNA from the host but also DNA from microbes and food material. Similarly, soil and water samples will inevitably contain cells from a diverse array of organisms [5,30]. Historically, metagenomic analyses involved the amplification of 'signature' regions of DNA such as rRNA genes, to enable the identification of major groups of organisms within a

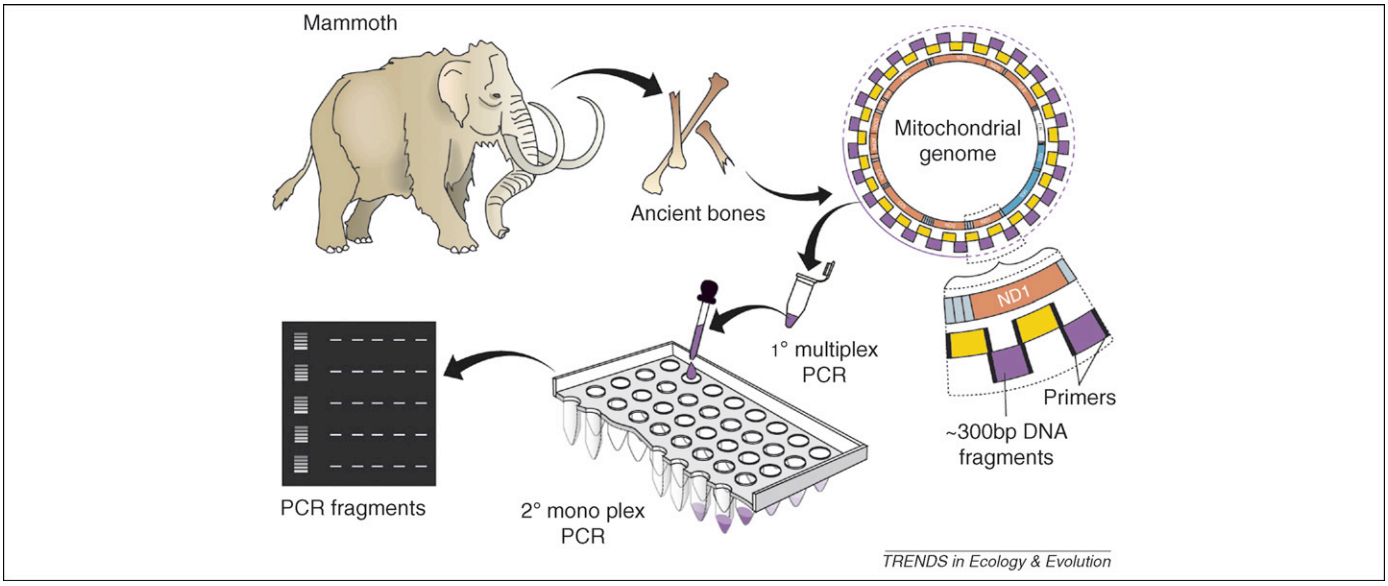

Figure 1. Multiplex PCR of mitochondrial genomes. Ancient DNA is extracted from biological remains such as bones. Following this, the multiplex PCR method uses inner (yellow) and outer (purple) PCR primer sets to amplify the complete mitochondrial genome of the target species. This is completed in two rounds of PCR: the first or 1° multiplex round is performed in a single tube and the second monoplex (from a single locus) PCR is conducted in a multiwell plate.

sample, such as bacteria. Using 'whole-genome shotgun sequencing' and conventional Sanger sequencing methods, >1 billion bp of sequence was recovered from samples taken from the Sargasso Sea [31]. Analysis of these data revealed nearly 150 new bacterial 'phylotypes' [31], and habitat-specific 'DNA fingerprints' were detected by others [32]. Similar projects using massively parallel DNA sequencers will be more efficient, provide more data and be very cost effective [33,34]. A potentially fruitful area of research will be the metagenomics of coprolites (subfossil excrement, faeces or droppings of ancient animals). Previous studies have shown that both mitochondrial and single-copy nuclear DNA sequences can be recovered from coprolites [35], and the new DNA sequencing technologies will massively increase the power to sequence, with a very high level of coverage, all DNA templates in such remains.

Ancient DNA analyses will also benefit greatly from such metagenomic approaches using the new sequencers, taking advantage of what has been, in the past, a disadvantage of ancient samples, namely that they are composed of a 'soup' of DNAs from various species. In particular, studies of animal, plant and microbial samples from permafrost environments [36–38] are likely to contain well-preserved DNA templates. Furthermore, the quantitative nature of the results obtained from parallel sequencers will allow us to not only detect the types of organisms that once existed in these environments but also to document any subtle changes in their population dynamics over time. Although such metagenomic studies will continue to be high profile and will represent a growing research field, genomic approaches where ancient DNA sequences are selected using homologous probes from closely related species might be preferred.

Selected genomics: 'sorting the wheat from the chaff'

Massively parallel sequencing of ancient DNA samples is not specific, in comparison to PCR-based methods that employ sequence-specific primers to amplify particular targets. The new technologies are designed to recover large amounts of sequence and possibly whole genomes by massive-scale random sequencing of all available DNA templates. Repeated rounds of DNA sequencing will progressively result in a larger number of independent sequences of the genome. One of the difficulties with ancient material is that the ratio of endogenous to contaminating DNA can be low. Hence, sequences from other genomes such as microbial flora and fauna will also be sequenced using these new technologies. For example, despite extensive screening of bones for their level of DNA preservation, in a recent Neanderthal study [24] only 6% of sequence reads matched those of primates and were therefore regarded as authentic (Figure 2). Moreover, the vast majority of the recovered sequences (79%) had no match in the public sequence databases, suggesting contamination by unknown (or as yet unsequenced) microorganisms [24]. Furthermore, contamination with modern human DNA is common in Neanderthal samples, as is the case for any samples that have been regularly handled. Such contamination is difficult to detect in the case of Neanderthals because the two species are so closely related. A recent reanalysis of published Neanderthal genome sequences suggests that human contamination constituted as much as 80% of the nucleotides that were reported to be of Neanderthal origin [39,40]. By contrast, well-preserved mammoth samples from the permafrost of Siberia were shown to have a relatively low level of contamination, as approximately half of the sequence reads from the permafrost material could be aligned to the elephant genome [25]. Not surprisingly, therefore, samples containing a high proportion of target DNA are typically chosen for ancient DNA analyses [24]. However, given the capacity of the new sequencers, large-scale random sequencing will still result in the recovery of a significant amount of target DNA sequences (e.g. >1 million base pairs), even if the samples are highly contaminated and

Table 1. Comparison of massively parallel sequencing technologies

	FLX (Roche)	Solexa (Illumina)	SOLiD (Applied Biosystems)	HeliScope (Helicos BioSciences)
Website addresses	http://www.roche-applied-science.com	http://www.illumina.com	http://www.appliedbiosystems.com/index.cfm	http://www.helicosbio.com
Chemistry	Emulsion PCR [67] of bead-anchored oligos [68]. Clonal plate amplification. Pyrosequencing using light emission and detection [69,70].	Solid-phase-anchored oligo bridge amplification [71]. Cluster sequencing using reversible fluorescent dNTP terminators [72].	Paired-end oligo cloning. Emulsion PCR of bead-anchored oligos. Fluorescent oligo ligation and detection [73,74].	True single-molecule sequencing (tSMS). Polymerase-mediated addition of reversible fluorescent dNTP terminators to polyA-tailed single RNA or DNA molecules hybridised to slide-anchored oligo dT primers [75].
Read length	~250 bases	~35 bases	~25 bases	25–35 bases (45% of reads >30 bases, 78% of reads >20 bases)
Machine costs	US$500 000	US$520 000	US$600 000	US$1 350 000
Costs/run (reagents)	Not currently available	US$3000	US$3000	Not currently available
Availability	Immediate	Immediate	Immediate	Not currently available
Capacity/run	0.1 Gb	1 Gb	1 Gb	3 Gb (originally reported to be 100 Gb). Fast processing up to 0.1 Gb per hour to give over 2 Gb per day.
Run duration	7.5 h	67–91 h	4 days for fragment library, 8 days for paired library	1.5 days
Advantages	Relatively long read length	Possible to sequence through single-repeat DNA nucleotide stretches	High base-calling accuracy (bases are read twice). Homopolymer sequencing.	No amplification of template required before sequencing. Similarly, preparation-free paired-end reads. Relatively high throughput. Reversible dNTP terminators allow sequencing through homopolymer DNA.
Disadvantages	Inability to determine length of large repeats, including length of repeat homopolymers >8 bases long	Read length is limiting. Inability to determine length of large repeats.	Read length is limiting. Sequence identification is difficult. Inability to determine length of large repeats.	Reduced accuracy (extensive coverage required)
Machine	Genome Sequencer FLX	Illumina Genome Analyzer (Solexa) 1G Genetic Analyzer	SOLiD (Supported Oligonucleotide Ligation and Detection) gene sequencer	HeliScope true Single Molecule Sequencer
Accuracy	99.5%	99%	99.9%	~90% (constant over 10–45 base reads). Accuracy is increased by multipass sequencing.
Paired-end reads	Yes	Yes	Yes	Yes
Sample size required	0.5–5 μg	0.1–1 μg	10–30 μg	Not available
Bench work	(i) DNA fragmentation for paired-end library construction or adaptor ligation (ii) Clonal emulsion PCR of fragments on beads (iii) Sequencing by synthesis of DNA beads in wells (iv) Detection of fluorescence by laser excitation	(i) DNA fragmentation for paired-end library construction or adaptor ligation (ii) Attachment to solid-phase-anchored oligo (iii) Cluster PCR (iv) Sequencing by synthesis of clustered DNA (v) Detection of fluorescence by laser excitation	(i) DNA fragmentation for paired-end library construction or adaptor ligation (ii) Clonal emulsion PCR of fragments on DNA beads (iii) Sequencing by oligo hybridisation of beads in wells (iv) Detection of fluorescence by laser excitation	(i) Shear DNA (ii) Ligate adaptor to 5′ terminus (iii) Add polyA to 3′ terminus (iv) Hybridise to polyT-covered slide (v) Wash with polymerase and a fluorescent dNTP (vi) Detect fluorescence using precision optics
Future developments	500 base reads currently being developed; 1 Gb per run in 2008	>6 Gb per run in 2008	50 bp single reads; ~6 Gb per run in 2008	Company predicts US$5000 human-size genome in 3 days

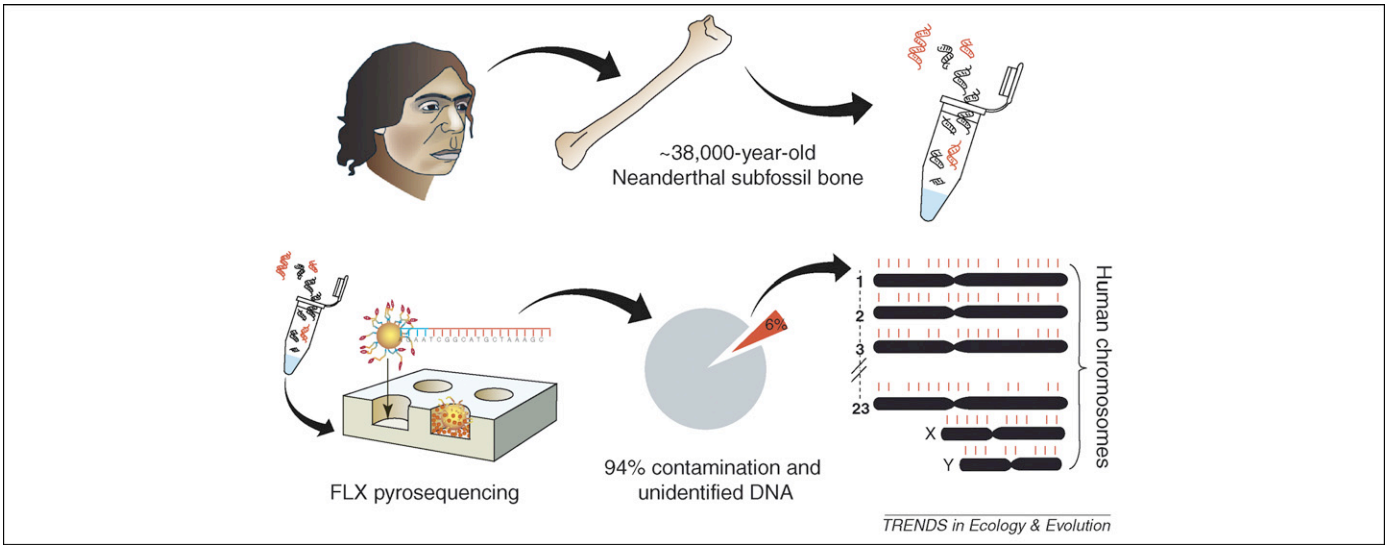

Figure 2. Sequencing of nuclear and mitochondrial DNA from Neanderthals using pyrosequencing technology. This approach results in random sequencing of all available DNA templates extracted from subfossil bones. DNA sequencing takes place on beads in an ordered array of wells. Approximately 94% of the sequences recovered were derived from contaminating microorganisms and DNA sequences not present in the databases. The remaining ~6% primate-specific sequences were mapped against a 'scaffold' of the human genome [24].

contain only a small proportion of target DNA. Contaminated sequences can often be identified by reference to database sequences.

A strategy to ensure high coverage of ancient genomes at an affordable cost is to target specific regions, as opposed to sequencing the entire genome. A selected genomic approach involves enriching for particular loci (Figure 3). By using known genes as probes, it is possible to isolate homologues from an ancient species by cross-hybridisation. For example, using a selected metagenomics approach [41], fragments of 29 of 35 Neanderthal genes targeted were recovered by screening a metagenomic library constructed from a Neanderthal bone using human genes as probes. In the future, this will allow a precise comparison of evolutionary differences between known genes in humans and Neanderthals, for example those involved in brain development and speech.

The importance of a 'scaffold'

Whereas the FLX sequencer generates relatively long sequences (~250 bp), those generated by the SOLiD, Solexa and HeliScope systems are comparatively short (25–35 bp; Table 1). Although the longer sequences generated by the FLX allow *de novo* assembly of sequences into contigs [42], the genome of a closely related species is still important in assembling complex genomes. A closely related species effectively acts as a 'scaffold' onto which it is possible to 'hang' new sequences (Figure 2). Recent ancient genomic studies on Neanderthals and woolly mammoths [25,41] have employed this approach owing to the availability of genomes of closely related species (human and elephant, respectively). In the absence of a close relative, it is difficult to identify and assemble many non-coding regions of the genome because these regions often evolve quickly and, consequently, similarity between

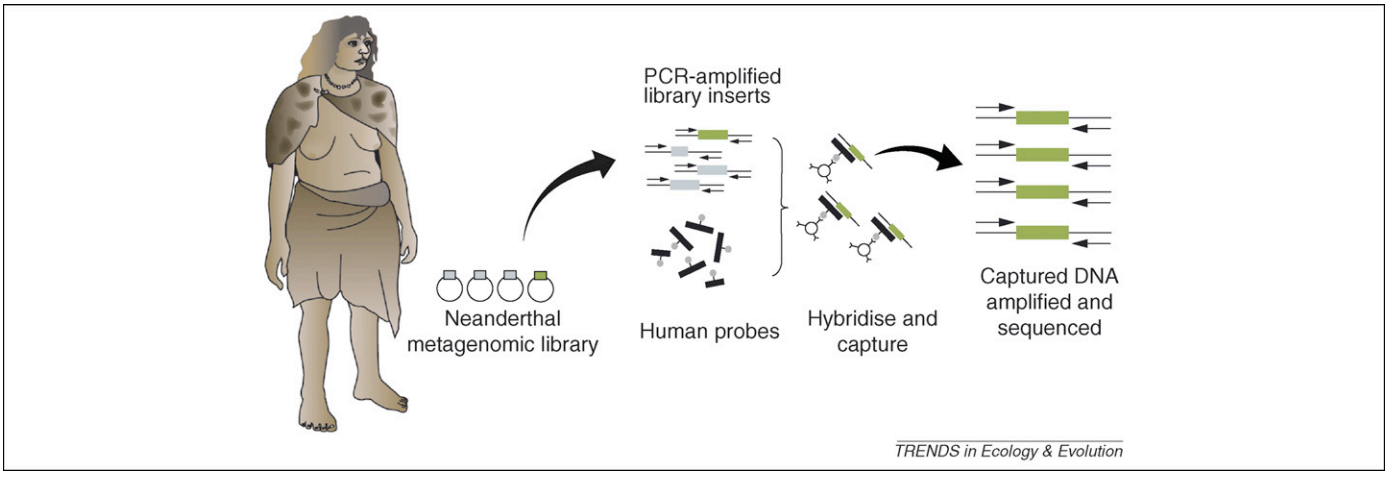

Figure 3. A selected genomic approach that targets specific regions of ancient genomes. One method of selected genomics involves the cloning of short ancient DNA sequences into a vector such as a plasmid to construct a metagenomic library. The library will inevitably comprise both target (in this case Neanderthal) and contaminating sequences. Clones are selected from the library by hybridisation to homologous probes (e.g. from humans) and amplified. These steps are followed by sequencing of the recovered regions using conventional Sanger methods or massively parallel technology.

distantly related species is low. Although as much as 40% of the human and mouse genomes are alignable, the location of the aligned sequences within the genome cannot be determined for a large proportion of these sequences [43]. Generally, then, the availability of a genome of a closely related species is a prerequisite for the correct identification and alignment of any ancient genome. However, short sequence reads (e.g. those from the SOLiD, Solexa and HeliScope systems) should not hinder the analysis or reconstruction of most of the coding sequences, as these are highly conserved, even between distantly related species. Notwithstanding these complexities, the recent developments in DNA sequencing technology provide reasons to be optimistic about the future recovery and assembly of many ancient genomes [44].

Recovering sequences from damaged ancient genomes

Despite the high levels of contamination characteristic of ancient DNA, the capacity of massively parallel sequencers is likely to result in the recovery of entire ancient genomes. However, there are difficulties. Ancient DNA is invariably damaged (Box 1), and therefore any sequence recovered will derive from a combination of both intact and damaged templates. To overcome this difficulty, a high level of coverage of the genome is required. This means sequencing each base multiple times, thereby enabling us to infer the real sequence from the consensus [45]. For the Neanderthal sample [24], the level of nucleotide damage was estimated to be ~0.4%, that is, 4 out of each 1000 bases [24]. Most of this damage is due to the deamination of cytosine resulting in C/T transitions (Box 1).

To determine whether any particular transition is real and not merely a result of DNA damage, extensive coverage of the genome is required. Using a binomial distribution and calculating to the 99% certainty level, the genome needs to be sequenced a minimum of eight times. This level of coverage will also allow us to identify with confidence any heterozygous sites within the genome. The number of bases that need to be sequenced to achieve an eightfold coverage of a complex genome can be calculated using the equations commonly applied to the estimation of genome coverage [46] and the probability of finding a clone in a genomic library [47]. To sequence the Neanderthal genome using, for example, the Solexa system and assuming an average read length of 30 bases, a total of 2.4×10^{10} bases of sequence will need to be generated to achieve the eightfold coverage. Taking into account that, from the best available sample, only 5% of sequences obtained will be Neanderthal, this raises the amount of sequence needed to 4.8×10^{11} bases, or just under 500 sequencing runs. The cost of this exercise, at US$3000 per run, will be US$1.5M. This amount compares well to the cost of sequencing the human genome using Sanger chemistry and uncontaminated DNA, at US$2.7B. However, parallel sequencers are being developed at a very fast rate, with both Applied Biosystems and Solexa claiming a capacity of >6 Gb per run, six times that available now (Table 1). This is likely to reduce the cost of sequencing further.

In summary, in relation to the recovery of ancient genomes, researchers will be faced with either the costs of a high level of coverage of an ancient genome or with a large number of sites that might be the result of damage; without a high level of coverage, damaged sites can be mistaken for real evolutionary differences.

Answering ancient mysteries?

Several previously unsolved problems can now be addressed using the large amounts of nuclear data that will inevitably be recovered from ancient organisms. For example, the question of whether Neanderthals interbred with modern humans remains contentious, despite extensive sequence data [48]. Current work by several research groups aimed at sequencing the entire Neanderthal nuclear genome is likely to finally resolve this issue [24].

The phylogenetic position of some extinct animals is open to debate. An example is the relationship of New Zealand's giant moa to the other ratites. Despite the availability of complete mitochondrial sequences for many ratites, no definitive tree can be constructed [18,19]. Here again, nuclear genome sequences from these birds should allow us to determine with confidence the relationships between these animals.

Many extinct species possessed unusual morphologies or phenotypes not seen today. The current wealth of knowledge of extant genes that influence body plan, size and morphology will enable comparisons with their extinct counterparts and might also uncover unexpected genes or gene families. For example, studies of the chicken genome have shown that although this animal has only a limited sense of smell, the species has retained a large number of olfactory genes [49]. Even more unusual is that the chicken appears to have retained all the genes required for dentition [49]. This illustrates a general principle of genomic evolution, namely that nature adds but rarely deletes. The genomes of organisms often contain a record of previous sequences and how, throughout evolution, these sequences have been modified. Massively parallel sequencers will now greatly enhance our ability to study these accumulated genomic 'signatures' in the case of extinct animals such as the mammoth, sabre-toothed tiger and giant ground sloth, which possessed unique phenotypes. Finally, the successful isolation of ancient animal DNA from soil will allow in-depth genetic analyses of extinct populations in response to climatic and geographical changes [50].

Sequencing ancient genomes

Despite the significant advances in DNA sequencing technology, the first successful recovery of an entire ancient genome will not be an insignificant task, to say the least. The correct assembly of even a single gene is difficult, as genes usually exist as variants even within an individual. With the average size of a gene's coding sequence at 2 kb and the average size of ancient nuclear DNA at ~100 bp, the chance of producing a chimaeric gene construct (a hybrid mixture of original sequences) is very high. Despite this, progress in the field has already been made with the successful reconstruction of a coat colour gene from mammoths [51]. Another potential problem is the correct assembly of the highly repetitive 'junk' DNA that is found between genes and which might play a role in gene regu-

lation [52]. However, the reconstruction of some of the 'smaller' ancient genes is likely to be possible.

Massively parallel sequencing should allow the construction of at least the coding sequences, and possibly a large segment of adjacent or flanking sequence, of genes from well-preserved extinct animals. Such flanking regions are likely to contain promoters and other regulatory elements essential for gene expression. Using available transgenic technology, it might then be possible to "reactivate" these genes, for example by syntenic replacement of their homologues in a modern species. It would be interesting, and no doubt subject to considerable ethical debate, to determine how many genes can be replaced in this manner. In any event, this work would provide, at the very least, a starting point for the analysis and for an improved understanding of the activity and function of ancient genomes.

Concluding remarks

Within a few years, a small number of relatively complete genomes of iconic ancient and extinct organisms are likely to be published, for example Neanderthals and woolly mammoths. For the foreseeable future, however, these developments are likely to come from major research groups and programs with large budgets. The vast majority of researchers with limited budgets will need to focus on more achievable projects. One way to do this is to recover specific sequences from DNA libraries of ancient organisms. The search for specific DNA loci important in population analyses, species identification or genes known to control particular characteristics such as coat colour in mammals or perhaps particular metabolic pathways such as the production of toxins will allow researchers to compare ancient and modern sequences from homologous genetic loci. This then provides a new comparative biology where instead of comparing different species at the same point in time (typically the present), we will be able to compare the same or closely related species at different points in time. Such ancient/modern genomic comparisons can provide insights into the biology of phenotypes (extinct and modern) and their evolution. Furthermore, because single-gene studies rarely represent a complete explanation for complex phenotypes, a systems approach is required where we recover entire ancient gene complexes. The new DNA sequencing technologies, in combination with metagenomics, are likely to propel such studies. In summary, we predict that the 'third wave' of ancient DNA progress will be characterised by developments such as an expansion of studies aimed at partial mitochondrial genomes to the recovery of large numbers of complete ones. In addition, there is likely to be a focus on the recovery of large regions of nuclear genomes, particularly protein-coding regions, and a diversification into genomes of organisms that are currently genetically unknown. Either way, the third wave will surely have a major impact on the science of ancient genomics.

Acknowledgements

We are grateful to Vivian Ward for graphic design. Our research is supported by funds from the Centres of Research Excellence to the Allan Wilson Centre for Molecular Ecology and Evolution, the Marsden Fund, Massey University and the University of Auckland and support from a James Cook Fellowship to D.M.L.

References

1 Donoghue, H.D. *et al.* (2005) Co-infection of *Mycobacterium tuberculosis* and *Mycobacterium leprae* in human archaeological samples – a possible explanation for the historical decline of leprosy. *Proc. Biol. Sci.* 272, 389–394
2 Hofreiter, M. *et al.* (2001) Ancient DNA. *Nat. Rev. Genet.* 2, 353–359
3 Willerslev, E. and Cooper, A. (2005) Ancient DNA. *Proc. Biol. Sci.* 272, 3–16
4 Lambert, D.M. *et al.* (2002) Rates of evolution in ancient DNA from Adélie penguins. *Science* 295, 2270–2273
5 Willerslev, E. *et al.* (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science* 317, 111–114
6 Lambert, D.M. and Millar, C.D. (2006) Evolutionary biology: ancient genomics is born. *Nature* 444, 275–276
7 Higuchi, R. *et al.* (1984) DNA sequences from the quagga, an extinct member of the horse family. *Nature* 312, 282–284
8 Mullis, K.B. and Faloona, F.A. (1987) Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. *Methods Enzymol.* 155, 335–350
9 Pääbo, S. *et al.* (2004) Genetic analyses from ancient DNA. *Annu. Rev. Genet.* 38, 645–679
10 Cano, R.J. *et al.* (1993) Amplification and sequencing of DNA from a 120–135-million-year-old weevil. *Nature* 363, 536–538
11 Desalle, R. *et al.* (1992) DNA sequences from a fossil termite in Oligomiocene amber and their phylogenetic implications. *Science* 257, 1933–1936
12 Sidow, A. *et al.* (1991) Bacterial DNA in *Clarkia* fossils. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 333, 429–432
13 Woodward, S.R. *et al.* (1994) DNA sequence from Cretaceous period bone fragments. *Science* 266, 1229–1232
14 Leonard, J.A. *et al.* (2000) Population genetics of ice age brown bears. *Proc. Natl. Acad. Sci. U. S. A.* 97, 1651–1654
15 Shapiro, B. *et al.* (2004) Rise and fall of the Beringian steppe bison. *Science* 306, 1561–1565
16 Noonan, J.P. *et al.* (2005) Genomic sequencing of Pleistocene cave bears. *Science* 309, 597–599
17 Jaenicke-Despres, V. *et al.* (2003) Early allelic selection in maize as revealed by ancient DNA. *Science* 302, 1206–1208
18 Haddrath, O. and Baker, A.J. (2001) Complete mitochondrial DNA genome sequences of extinct birds: ratite phylogenetics and the vicariance biogeography hypothesis. *Proc. Biol. Sci.* 268, 939–945
19 Cooper, A. *et al.* (2001) Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* 409, 704–707
20 Huynen, L. *et al.* (2003) Nuclear DNA sequences detect species limits in ancient moa. *Nature* 425, 175–178
21 Bunce, M. *et al.* (2003) Extreme reversed sexual size dimorphism in the extinct New Zealand moa *Dinornis*. *Nature* 425, 172–175
22 Cooper, A. (2006) The year of the mammoth. *PLoS Biol.* 4, e78
23 Krause, J. *et al.* (2006) Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature* 439, 724–727
24 Green, R.E. *et al.* (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* 444, 330–336
25 Poinar, H.N. *et al.* (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* 311, 392–394
26 Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141
27 Hansen, A. *et al.* (2001) Statistical evidence for miscoding lesions in ancient DNA templates. *Mol. Biol. Evol.* 18, 262–265
28 Blows, N. (2008) DNA sequencing: generation next-next. *Nat. Methods* 5, 267–274
29 Sanger, F. *et al.* (1977) Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature* 265, 687–695
30 Johnson, S.S. *et al.* (2007) Ancient bacteria show evidence of DNA repair. *Proc. Natl. Acad. Sci. U. S. A.* 104, 14401–14405
31 Venter, J.C. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74
32 Tringe, S.G. and Rubin, E.M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* 6, 805–814

33 Turnbaugh, P.J. *et al.* (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031
34 Gill, S.R. *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359
35 Poinar, H. *et al.* (2003) Nuclear gene sequences from a late pleistocene sloth coprolite. *Curr. Biol.* 13, 1150–1152
36 Willerslev, E. *et al.* (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* 300, 791–795
37 Lydolph, M.C. *et al.* (2005) Beringian paleoecology inferred from permafrost-preserved fungal DNA. *Appl. Environ. Microbiol.* 71, 1012–1017
38 Vishnivetskaya, T.A. *et al.* (2006) Bacterial community in ancient Siberian permafrost as characterized by culture and culture-independent methods. *Astrobiology* 6, 400–414
39 Wall, J.D. and Kim, S.K. (2007) Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genet.* 3, 1862–1866
40 Dalton, R. (2007) DNA probe finds hints of human. *Nature* 449, 7
41 Noonan, J.P. *et al.* (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science* 314, 1113–1118
42 Sundquist, A. *et al.* (2007) Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE* 2, e484
43 Waterston, R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
44 Pop, M. and Salzberg, S.L. (2008) Bioinformatic challenges of new sequencing technology. *Trends Genet.* 24, 142–149
45 Hofreiter, M. *et al.* (2001) DNA sequences from multiple amplifications reveal artefacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* 29, 4793–4799
46 Paterson, A.H. (1996) *The DNA Revolution.* Academic Press
47 Clarke, L. and Carbon, J. (1976) A colony bank containing synthetic ColE1 hybrid plasmids representative of the entire *E. coli* genome. *Cell* 9, 91–101
48 Serre, D. *et al.* (2004) No evidence of Neandertal mtDNA contribution to early modern humans. *PloS Biol.* 2, e57
49 Wallis, J.W. *et al.* (2004) A physical map of the chicken genome. *Nature* 432, 761–764
50 Shepherd, L.D. *et al.* (2005) Microevolution and mega-icebergs in the Antarctic. *Proc. Natl. Acad. Sci. U. S. A.* 102, 16717–16722
51 Römpler, H. *et al.* (2006) Nuclear gene indicates coat-color polymorphism in mammoths. *Science* 313, 62
52 Nóbrega, M.A. *et al.* (2004) Megabase deletions of gene deserts result in viable mice. *Nature* 431, 988–993
53 Geigl, E.M. (2002) On the circumstance surrounding the preservation and analysis of very old DNA. *Archaeometry* 44, 337–342
54 Hay, J.M. *et al.* (2008) Rapid molecular evolution in a living fossil. *Trends Genet.* 24, 106–109
55 Pääbo, S. (1989) Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc. Natl. Acad. Sci. U. S. A.* 86, 1939–1943
56 Wandeler, P. *et al.* (2003) Patterns of nuclear DNA degeneration over time – a case study in historic teeth samples. *Mol. Ecol.* 12, 1087–1093
57 Lindahl, T. (1993) Instability and decay of the primary structure of DNA. *Nature* 362, 709–715
58 Ritchie, P.A. *et al.* (2004) Ancient DNA enables timing of the Pleistocene origin and Holocene expansion of two Adélie penguin lineages in Antarctica. *Mol. Biol. Evol.* 21, 240–248
59 Edwards, C.J. *et al.* (2004) Ancient DNA analysis of 101 cattle remains: limits and prospects. *J. Archaeol. Sci.* 31, 695–710
60 Höss, M. *et al.* (1996) DNA damage and DNA sequence retrieval from ancient tissues. *Nucleic Acids Res.* 24, 1304–1307
61 Binladen, J. *et al.* (2006) Assessing the fidelity of ancient DNA sequences amplified from nuclear genes. *Genetics* 172, 733–741
62 Gilbert, M.T. *et al.* (2003) Characterization of genetic miscoding lesions caused by postmortem damage. *Am. J. Hum. Genet.* 72, 48–61
63 Briggs, A.W. *et al.* (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* 104, 14616–14621
64 Brotherton, P. *et al.* (2007) Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res.* 35, 5717–5728
65 Gilbert, M.T. *et al.* (2007) Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science* 317, 1927–1930
66 Stiller, M. *et al.* (2006) Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. *Proc. Natl. Acad. Sci. U. S. A.* 103, 13578–13584
67 Nakano, M. *et al.* (2003) Single-molecule PCR using water-in-oil emulsion. *J. Biotechnol.* 102, 117–124
68 Dressman, D. *et al.* (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. U. S. A.* 100, 8817–8822
69 Ronaghi, M. *et al.* (1998) A sequencing method based on real-time pyrophosphate. *Science* 281, 363–365
70 Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380
71 Adams, C.P., and Kron, S.J. Mosaic Technologies, Whitehead Institute for Biomedical Research, U. S. A. (1997) Method for performing amplification of nucleic acid with two primers bound to a single solid support, US Patent 5,641,658
72 Ju, J. *et al.* (2006) Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. U. S. A.* 103, 19635–19640
73 Shendure, J. *et al.* (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309, 1728–1732
74 Brenner, S. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18, 630–634
75 Mitchelson, K.R. (2007) *New High Throughput Technologies for DNA Sequencing and Genomics.* Elsevier