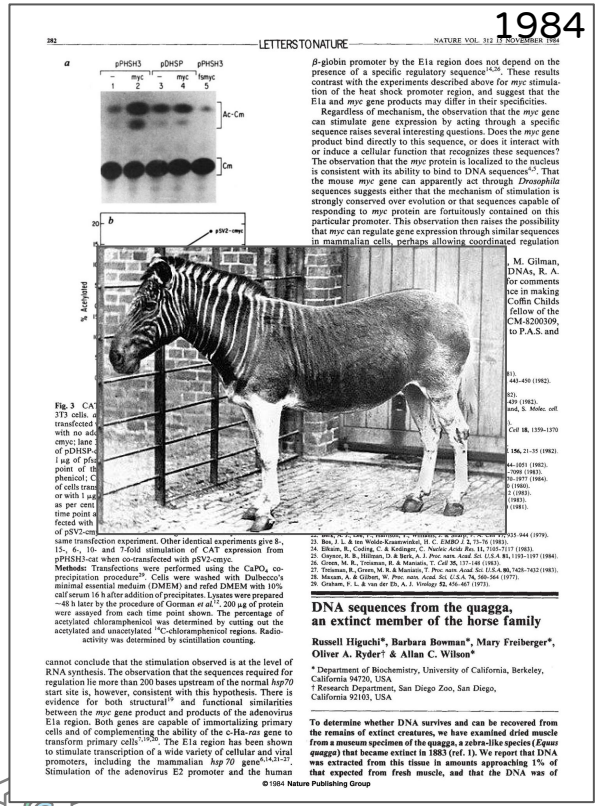
SPAAM

Standards, Precautions & Advances in Ancient Metagenomics

Lecture 2A: Introduction to Ancient DNA

Christina Warinner

We've come a long way



1984

282 LETTERS TO NATURE NATURE VOL. 312 15 NOVEMBER 1984

a

pPHSH3 pDHSP pPHSH3
– myc – myc fsmyc
1 2 3 4 5

Ac-Cm

Cm

b

pSV2-cmyc

% Acetylated

Fig. 3 CAT
3T3 cells. a
transfected
with no ada
cmyc; lane
of pDHSP-
1 μg of pfos
point of th
phenicol; C
of cells tran
or with 1 μg
as per cent
time point a
fected with
of pSV2-cm
same transfection experiment. Other identical experiments give 8-, 15-, 6-, 10- and 7-fold stimulation of CAT expression from pPHSH3-cat when co-transfected with pSV2-cmyc.

Methods: Transfections were performed using the $CaPO_4$ co-precipitation procedure[19]. Cells were washed with Dulbecco's minimal essential medium (DMEM) and refed DMEM with 10% calf serum 16 h after addition of precipitates. Lysates were prepared ~48 h later by the procedure of Gorman *et al.*[12]. 200 μg of protein were assayed from each time point shown. The percentage of acetylated chloramphenicol was determined by cutting out the acetylated and unacetylated ^{14}C-chloramphenicol regions. Radioactivity was determined by scintillation counting.

cannot conclude that the stimulation observed is at the level of RNA synthesis. The observation that the sequences required for regulation lie more than 200 bases upstream of the normal *hsp70* start site is, however, consistent with this hypothesis. There is evidence for both structural[19] and functional similarities between the *myc* gene product and products of the adenovirus E1a region. Both genes are capable of immortalizing primary cells and of complementing the ability of the c-Ha-*ras* gene to transform primary cells[1,19,20]. The E1a region has been shown to stimulate transcription of a wide variety of cellular and viral promoters, including the mammalian *hsp 70* gene[6,14,21–27]. Stimulation of the adenovirus E2 promoter and the human β-globin promoter by the E1a region does not depend on the presence of a specific regulatory sequence[18,20]. These results contrast with the experiments described above for *myc* stimulation of the heat shock promoter region, and suggest that the E1a and *myc* gene products may differ in their specificities.

Regardless of mechanism, the observation that the *myc* gene can stimulate gene expression by acting through a specific sequence raises several interesting questions. Does the *myc* gene product bind directly to this sequence, or does it interact with or induce a cellular function that recognizes these sequences? The observation that the *myc* protein is localized to the nucleus is consistent with its ability to bind to DNA sequences[1,3]. That the mouse *myc* gene can apparently act through *Drosophila* sequences suggests either that the mechanism of stimulation is strongly conserved over evolution or that sequences capable of responding to *myc* protein are fortuitously contained on this particular promoter. This observation then raises the possibility that *myc* can regulate gene expression through similar sequences in mammalian cells, perhaps allowing coordinated regulation

M. Gilman, DNAs, R. A. for comments ce in making Coffin Childs fellow of the CM-8200309, to P.A.S. and

81). 443–450 (1982).

82). 439 (1982). and, S. Molec. cell.

Cell 18, 1359–1370

156, 21–35 (1982).

44–1051 (1982). –7098 (1983). 78–1977 (1984). 20 (1980). 2 (1983). (1983). (1981).

22. Berk, A. J. & Sharp, P. A. *Proc. natn. Acad. Sci. U.S.A.* 76, 2910–2915–944 (1979).
23. Bos, J. L. & ten Wolde-Kuanwinkel, H. C. *EMBO J.* 2, 75–76 (1983).
24. Eliksim, R., Coding, C. & Kedinger, C. *Nucleic Acids Res.* 11, 7165–7177 (1983).
25. Gaynor, R. B., Hillman, D. & Berk, A. J. *Proc. natn. Acad. Sci. U.S.A.* 81, 1193–1197 (1984).
26. Green, M. R., Treisman, R. & Maniatis, T. *Cell* 35, 137–148 (1983).
27. Treisman, R., Green, M. R. & Maniatis, T. *Proc. natn. Acad. Sci. U.S.A.* 80, 7428–7432 (1983).
28. Maxam, A. & Gilbert, W. *Proc. natn. Acad. Sci. U.S.A.* 74, 560–564 (1977).
29. Graham, F. L. & van der Eb, A. J. *Virology* 52, 456–467 (1973).

DNA sequences from the quagga, an extinct member of the horse family

Russell Higuchi*, Barbara Bowman*, Mary Freiberger*, Oliver A. Ryder† & Allan C. Wilson*

* Department of Biochemistry, University of California, Berkeley, California 94720, USA
† Research Department, San Diego Zoo, San Diego, California 92103, USA

To determine whether DNA survives and can be recovered from the remains of extinct creatures, we have examined dried muscle from a museum specimen of the quagga, a zebra-like species (*Equus quagga*) that became extinct in 1883 (ref. 1). We report that DNA was extracted from this tissue in amounts approaching 1% of that expected from fresh muscle, and that the DNA was of

We've come a long way

1984

282 LETTERS TO NATURE NATURE VOL. 312 15 NOVEMBER 1984

Unidentified reading frame 1

```
Quagga  C CCA ATC CTG CTC GCC GTA GCA TTC CTC ACA CTA GTT GAA CGA AAA GTC TTA GGC TAC ATA CAA CTT CGT AAA GGA CCC AAC ATC GTA GGC CCC TAT GGC CTA CTA CAA CCC ATT AC
Zebra   ..... T ..... G ..... T ..... C ..... G*
```

Cytochrome oxidase I

```
Quagga  A GGA GGA TTC GTT CAC TGA TTC CCT CTA TTC TCA GGA TAC ACA CTC AAC CAA ACC TGA GCA AAA ATT CAC TTT ACA ATT ATA TTC GTA GGG GTC AAC ATA ATT TTC TTC CCA
Zebra   G ..... T ..... G ..... C ..... A ..... T ..... C* .....
```

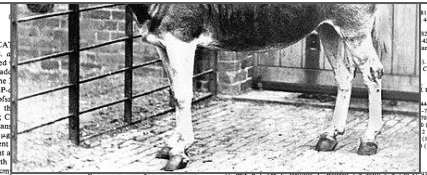


Fig. 3 CAT
3T3 cells. a
transfected
with no ada
emyc; lane
of pDHSP-
1 μg of pfos
point of th
phenicol; C
of cells tran
or with 1 μg
as per cent
time point a
fected with
of pSV2-cm

same transfection experiment. Other identical experiments give 8-, 15-, 6-, 10- and 7-fold stimulation of CAT expression from pPHSH3-cat when co-transfected with pSV2-cmyc.

Methods: Transfections were performed using the $CaPO_4$ co-precipitation procedure[19]. Cells were washed with Dulbecco's minimal essential medium (DMEM) and refed DMEM with 10% calf serum 16 h after addition of precipitates. Lysates were prepared ~48 h later by the procedure of Gorman *et al.*[12]. 200 μg of protein were assayed from each time point shown. The percentage of acetylated chloramphenicol was determined by cutting out the acetylated and unacetylated ^{14}C-chloramphenicol regions. Radioactivity was determined by scintillation counting.

cannot conclude that the stimulation observed is at the level of RNA synthesis. The observation that the sequences required for regulation lie more than 200 bases upstream of the normal *hsp70* start site is, however, consistent with this hypothesis. There is evidence for both structural[19] and functional similarities between the *myc* gene product and products of the adenovirus E1a region. Both genes are capable of immortalizing primary cells and of complementing the ability of the c-Ha-*ras* gene to transform primary cells[1,19,20]. The E1a region has been shown to stimulate transcription of a wide variety of cellular and viral promoters, including the mammalian *hsp 70* gene[6,14,21-27]. Stimulation of the adenovirus E2 promoter and the human

81).
443–450 (1982).
82).
439 (1982).
and, S. Molec. cell.
Cell 18, 1359–1370
156, 21–35 (1982).
44–1051 (1982).
–7098 (1983).
70–1977 (1984).
20 (1980).
2 (1983).
(1983).
(1981).
22. Berk, A. J. and Sharp, P. A. Nature 287, 479–482 (1980).
23. Bos, J. L. & ten Wolde-Kuanwinkel, H. C. EMBO J. 2, 75–78 (1983).
24. Eliasen, R., Coding, C. & Kedinger, C. Nucleic Acids Res. 11, 7105–7117 (1983).
25. Gaynor, R. B., Hillman, D. & Berk, A. J. Proc. natn. Acad. Sci. U.S.A. 81, 1193–1197 (1984).
26. Green, M. R., Treisman, R. & Maniatis, T. Cell 35, 137–148 (1983).
27. Treisman, R., Green, M. R. & Maniatis, T. Proc. natn. Acad. Sci. U.S.A. 80, 7428–7432 (1983).
28. Maxam, A. & Gilbert, W. Proc. natn. Acad. Sci. U.S.A. 74, 560–564 (1977).
29. Graham, F. L. & van der Eb, A. J. Virology 52, 456–467 (1973).

DNA sequences from the quagga, an extinct member of the horse family

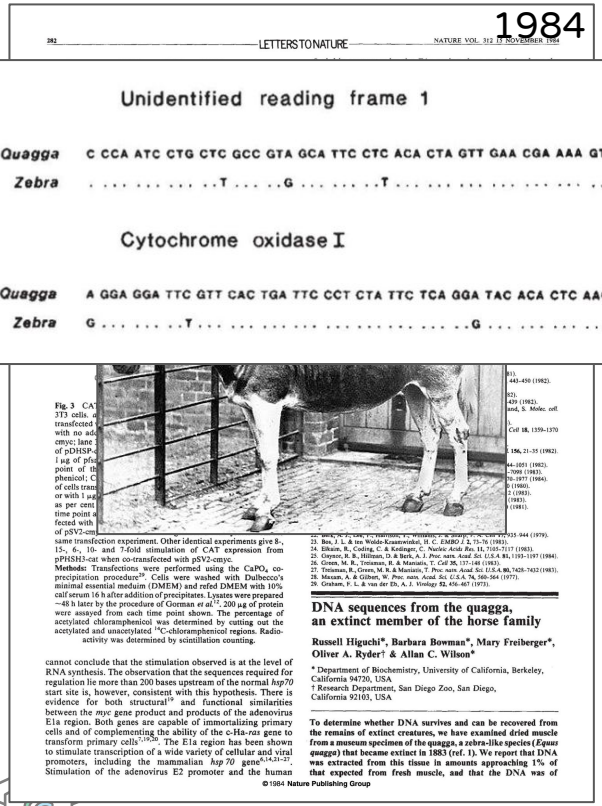
Russell Higuchi*, Barbara Bowman*, Mary Freiberger*, Oliver A. Ryder† & Allan C. Wilson*

* Department of Biochemistry, University of California, Berkeley, California 94720, USA
† Research Department, San Diego Zoo, San Diego, California 92103, USA

To determine whether DNA survives and can be recovered from the remains of extinct creatures, we have examined dried muscle from a museum specimen of the quagga, a zebra-like species (*Equus quagga*) that became extinct in 1883 (ref. 1). We report that DNA was extracted from this tissue in amounts approaching 1% of that expected from fresh muscle, and that the DNA was of

We've come a long way

1984



282 LETTERS TO NATURE NATURE VOL. 312 15 NOVEMBER 1984

Fig. 3 CAT
3T3 cells. a
transfected
with no ada
emyc; lane
of pDHSP-
1 μg of pfo
point of th
phenicol; C
of cells tran
or with 1 μg
as per cent
time point a
fected with
of pSV2-cm
same transfection experiment. Other identical experiments give 8-, 15-, 6-, 10- and 7-fold stimulation of CAT expression from pPHSH3-cat when co-transfected with pSV2-cat/yc.

Methods: Transfections were performed using the $CaPO_4$ co-precipitation procedure[29]. Cells were washed with Dulbecco's minimal essential medium (DMEM) and refed DMEM with 10% calf serum 16 h after addition of precipitates. Lysates were prepared ~48 h later by the procedure of Gorman *et al.*[12]. 200 μg of protein were assayed from each time point shown. The percentage of acetylated chloramphenicol was determined by cutting out the acetylated and unacetylated ^{14}C-chloramphenicol regions. Radioactivity was determined by scintillation counting.

cannot conclude that the stimulation observed is at the level of RNA synthesis. The observation that the sequences required for regulation lie more than 200 bases upstream of the normal *hsp70* start site is, however, consistent with this hypothesis. There is evidence for both structural[19] and functional similarities between the *myc* gene product and products of the adenovirus E1a region. Both genes are capable of immortalizing primary cells and of complementing the ability of the c-Ha-*ras* gene to transform primary cells[1,19,20]. The E1a region has been shown to stimulate transcription of a wide variety of cellular and viral promoters, including the mammalian *hsp 70* gene[6,14,21-27]. Stimulation of the adenovirus E2 promoter and the human

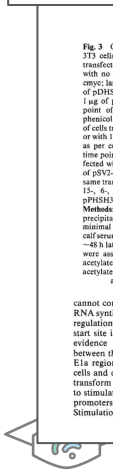
81).
443–450 (1982).
82).
439 (1982).
and, S. Molec. cell.
C.
Cell 18, 1359–1370
156, 21–35 (1982).
44–1051 (1982).
–7098 (1983).
76–1977 (1984).
20 (1980).
2 (1983).
(1983).
(1981).
23. Bos, J. L. & ten Wolde-Kuanwinkel, H. C. EMBO J. 2, 75–76 (1983).
24. Eliasen, R., Coding, C. & Kedinger, C. Nucleic Acids Res. 11, 7105–7117 (1983).
25. Gruysen, R. B., Hillman, D. & Berk, A. J. Proc. natn. Acad. Sci. U.S.A. 81, 1193–1197 (1984).
26. Green, M. R., Treisman, R. & Maniatis, T. Cell 35, 137–148 (1983).
27. Treisman, R., Green, M. R. & Maniatis, T. Proc. natn. Acad. Sci. U.S.A. 80, 7428–7432 (1983).
28. Maxam, A. & Gilbert, W. Proc. natn. Acad. Sci. U.S.A. 74, 560–564 (1977).
29. Graham, F. L. & van der Eb, A. J. Virology 52, 456–467 (1973).

DNA sequences from the quagga, an extinct member of the horse family

Russell Higuchi*, Barbara Bowman*, Mary Freiberger*, Oliver A. Ryder† & Allan C. Wilson*

* Department of Biochemistry, University of California, Berkeley, California 94720, USA
† Research Department, San Diego Zoo, San Diego, California 92103, USA

To determine whether DNA survives and can be recovered from the remains of extinct creatures, we have examined dried muscle from a museum specimen of the quagga, a zebra-like species (*Equus quagga*) that became extinct in 1883 (ref. 1). We report that DNA was extracted from this tissue in amounts approaching 1% of that expected from fresh muscle, and that the DNA was of

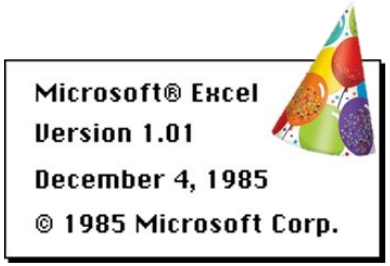


Unidentified reading frame 1

```
Quagga  C CCA ATC CTG CTC GCC GTA GCA TTC CTC ACA CTA GTT GAA CGA AAA GTC TTA GGC TAC ATA CAA CTT CGT AAA GGA CCC AAC ATC GTA GGC CCC TAT GGC CTA CTA CAA CCC ATT AC
Zebra   . . . . . T . . . . G . . . . . T . . . . . C . . . . . G*
```

Cytochrome oxidase I

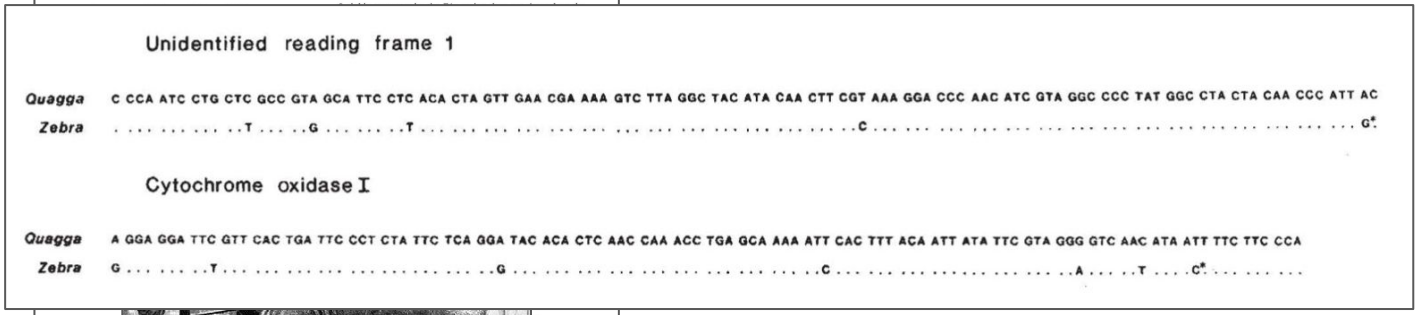
```
Quagga  A GGA GGA TTC GTT CAC TGA TTC CCT CTA TTC TCA GGA TAC ACA CTC AAC CAA ACC TGA GCA AAA ATT CAC TTT ACA ATT ATA TTC GTA GGG GTC AAC ATA ATT TTC TTC CCA
Zebra   G . . . . . T . . . . . G . . . . . C . . . . . A . . . . T . . . . C* . . . . . .
```



1985

We've come a long way

1984



Unidentified reading frame 1

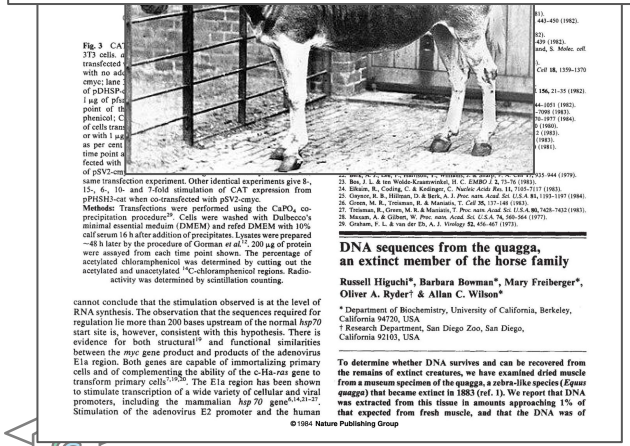
Quagga C CCA ATC CTG CTC GCC GTA GCA TTC CTC ACA CTA GTT GAA CGA AAA GTC TTA GGC TAC ATA CAA CTT CGT AAA GGA CCC AAC ATC GTA GGC CCC TAT GGC CTA CTA CAA CCC ATT AC

Zebra . . . . . T . . . . . G . . . . . T . . . . . C . . . . . G*.

Cytochrome oxidase I

Quagga A GGA GGA TTC GTT CAC TGA TTC CCT CTA TTC TCA GGA TAC ACA CTC AAC CAA ACC TGA GCA AAA ATT CAC TTT ACA ATT ATA TTC GTA GGG GTC AAC ATA ATT TTC TTC CCA

Zebra G . . . . . T . . . . . G . . . . . C . . . . . A . . . . T . . . . C*. . . . . .



LETTERS TO NATURE

NATURE VOL. 312 15 NOVEMBER 1984

Fig. 3 CAT

3T3 cells.

transfected

with no ad

emyc; lane

of pDHSP-

1 μg of pfo

point of th

phenicol; C

of cells tran

or with 1 μg

as per cent

time point a

fected with

of pSV2-cm

same transfection experiment. Other identical experiments give 8-, 15-, 6-, 10- and 7-fold stimulation of CAT expression from pPHSH3-cat when co-transfected with pSV2-cat/myc.

Methods: Transfections were performed using the $CaPO_4$ co-precipitation procedure[19]. Cells were washed with Dulbecco's minimal essential medium (DMEM) and refed DMEM with 10% calf serum 16 h after addition of precipitates. Lysates were prepared ~48 h later by the procedure of Gorman *et al.*[12]. 200 μg of protein were assayed from each time point shown. The percentage of acetylated chloramphenicol was determined by cutting out the acetylated and unacetylated ^{14}C-chloramphenicol regions. Radioactivity was determined by scintillation counting.

cannot conclude that the stimulation observed is at the level of RNA synthesis. The observation that the sequences required for regulation lie more than 200 bases upstream of the normal *hsp70* start site is, however, consistent with this hypothesis. There is evidence for both structural[19] and functional similarities between the *myc* gene product and products of the adenovirus E1a region. Both genes are capable of immortalizing primary cells and of complementing the ability of the c-Ha-*ras* gene to transform primary cells[1,19,20]. The E1a region has been shown to stimulate transcription of a wide variety of cellular and viral promoters, including the mammalian *hsp 70* gene[6,14,21-27]. Stimulation of the adenovirus E2 promoter and the human

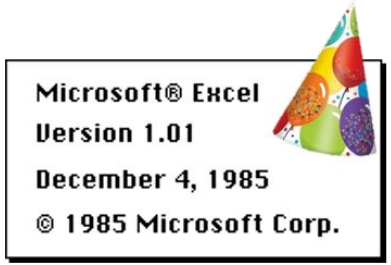
81).
443–450 (1982).
82).
439 (1982).
and, S. Molec. cell.
C.
Cell 18, 1359–1370
156, 21–35 (1982).
44–1051 (1982).
-7098 (1983).
76–1977 (1984).
0 (1980).
2 (1983).
(1983).
(1981).
23. Bos, J. L. & ten Wolde-Kuanwinkel, H. C. *EMBO J.* 2, 75–76 (1983).
24. Eliksim, R., Coding, C. & Rodinger, C. *Nucleic Acids Res.* 11, 7105–7117 (1983).
25. Gruysier, R. B., Hillman, D. & Berk, A. J. *Proc. nats. Acad. Sci. U.S.A.* 81, 1193–1197 (1984).
26. Green, M. R., Treisman, R. & Maniatis, T. *Cell* 35, 137–148 (1983).
27. Treisman, R., Green, M. R. & Maniatis, T. *Proc. nats. Acad. Sci. U.S.A.* 80, 7428–7432 (1983).
28. Maxam, A. & Gilbert, W. *Proc. nats. Acad. Sci. U.S.A.* 74, 560–564 (1977).
29. Graham, F. L. & van der Eb, A. J. *Virology* 52, 456–467 (1973).

DNA sequences from the quagga, an extinct member of the horse family

Russell Higuchi*, Barbara Bowman*, Mary Freiberger*, Oliver A. Ryder† & Allan C. Wilson*

* Department of Biochemistry, University of California, Berkeley, California 94720, USA
† Research Department, San Diego Zoo, San Diego, California 92103, USA

To determine whether DNA survives and can be recovered from the remains of extinct creatures, we have examined dried muscle from a museum specimen of the quagga, a zebra-like species (*Equus quagga*) that became extinct in 1883 (ref. 1). We report that DNA was extracted from this tissue in amounts approaching 1% of that expected from fresh muscle, and that the DNA was of

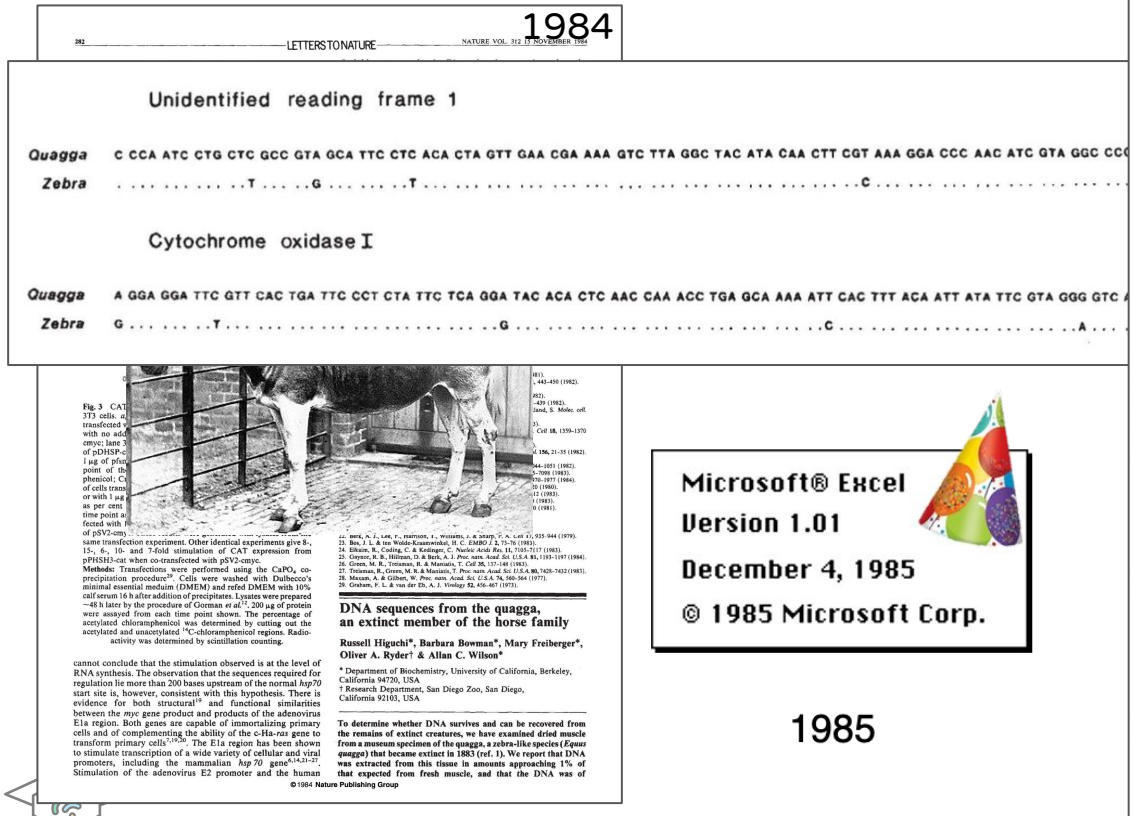


1985

"...quagga DNA sequences... [were sequenced] via the primed-synthesis, dideoxynucleoside chain-termination method of Sanger et al."

We've come a long way

1984



1985

1977


Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage φX174)

F. SANGER, S. NICKLEN, AND A. R. COULSON
Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977

ABSTRACT A new method for determining nucleotide sequences in DNA is described. It is similar to the "plus and minus" method [Sanger, F. & Coulson, A. R. (1975) J. Mol. Biol. 94, 441–448] but makes use of the 2′,3′-dideoxy and arabinonucleoside analogues of the normal deoxynucleoside triphosphates, which act as specific chain-terminating inhibitors of DNA polymerase. The technique has been applied to the DNA of bacteriophage φX174 and is more rapid and more accurate than either the plus or the minus method.

The "plus and minus" method (1) is a relatively rapid and simple technique that has made possible the determination of the sequence of the genome of bacteriophage φX174 (2). It depends on the use of DNA polymerase to transcribe specific regions of the DNA under controlled conditions. Although the method is considerably more rapid and simple than other available techniques, neither the "plus" nor the "minus" method is completely accurate, and in order to establish a sequence both must be used together, and sometimes confirmatory data are necessary. W. M. Barnes (J. Mol. Biol., in press) has recently developed a third method, involving ribo-substitution, which has certain advantages over the plus and minus method, but this has not yet been extensively exploited.

Another rapid and simple method that depends on specific chemical degradation of the DNA has recently been described by Maxam and Gilbert (3), and this has also been used extensively for DNA sequencing. It has the advantage over the plus and minus method that it can be applied to double-stranded DNA, but it requires a strand separation or equivalent fractionation of each restriction enzyme fragment studied, which makes it somewhat more laborious.

This paper describes a further method using DNA polymerase, which makes use of inhibitors that terminate the newly synthesized chains at specific residues.

Principle of the Method. Atkinson et al. (4) showed that the inhibitory activity of 2′,3′-dideoxythymidine triphosphate (ddTTP) on DNA polymerase I depends on its being incorporated into the growing oligonucleotide chain in the place of thymidylic acid (dT). Because the ddT contains no 3′-hydroxyl group, the chain cannot be extended further, so that termination occurs specifically at positions where dT should be incorporated. If a primer and template are incubated with DNA polymerase in the presence of a mixture of ddTTP and dTTP, as well as the other three deoxyribonucleoside triphosphates (one of which is labeled with ^{32}P), a mixture of fragments all having the same 5′ and with ddT residues at the 3′ ends is obtained. When this mixture is fractionated by electrophoresis on denaturing acrylamide gels the pattern of bands shows the distribution of dTs in the newly synthesized DNA. By using analogous terminators for the other nucleotides in separate incubations and running the samples in parallel on the gel, a pattern of bands is obtained from which the sequence can be read off as in the other rapid techniques mentioned above.

Two types of terminating triphosphates have been used—the dideoxy derivatives and the arabinonucleosides. Arabinose is a stereoisomer of ribose in which the 3′-hydroxyl group is oriented in trans position with respect to the 2′-hydroxyl group. The arabinosyl (ara) nucleotides act as chain terminating inhibitors of Escherichia coli DNA polymerase I in a manner comparable to ddT (4), although synthesized chains ending in 3′ araC can be further extended by some mammalian DNA polymerases (5). In order to obtain a suitable pattern of bands from which an extensive sequence can be read it is necessary to have a ratio of terminating triphosphate to normal triphosphate such that only partial incorporation of the terminator occurs. For the dideoxy derivatives this ratio is about 100, and for the arabinosyl derivatives about 5000.

METHODS

Preparation of the Triphosphate Analogues. The preparation of ddTTP has been described (6, 7), and the material is now commercially available. ddA has been prepared by McCarthy et al. (8). We essentially followed their procedure and used the methods of Tener (9) and of Hoard and Ott (10) to convert it to the triphosphate, which was then purified on DEAE-Sephadex, using a 0.1–1.0 M gradient of triethylamine carbonate at pH 8.4. The preparation of ddGTP and ddCTP has not been described previously; however we applied the same method as that used for ddATP and obtained solutions having the requisite terminating activities. The yields were very low and this can hardly be regarded as adequate chemical characterization. However, there can be little doubt that the activity was due to the dideoxy derivatives.

The starting material for the ddGTP was N-isobutyryl-5′-O-monomethoxytrityldeoxyguanosine prepared by F. E. Baralle (11). After tosylation of the 3′-OH group (12) the compound was converted to the 2′,3′-didehydro derivative with sodium methoxide (8). The isobutyryl group was partly removed during this treatment and removal was completed by incubation in NH_3 (specific gravity 0.88) overnight at 45°. The didehydro derivative was reduced to the dideoxy derivative (8) and converted to the triphosphate as for the ddATP. The monophosphate was purified by fractionation on a DEAE-Sephadex column using a triethylamine carbonate gradient (0.025–0.3 M) but the triphosphate was not purified.

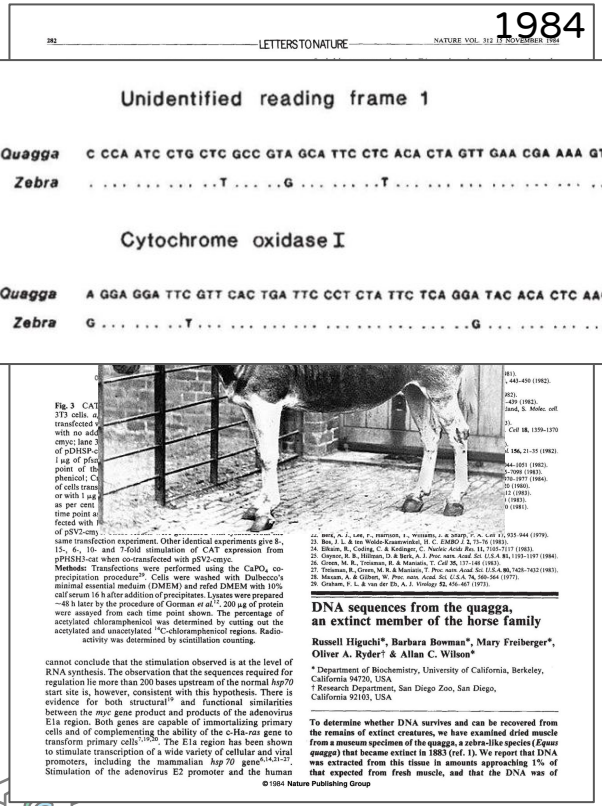
ddCTP was prepared from N-anisoyl-5′-O-monomethoxytrityldeoxycytidine (Collaborative Research Inc., Waltham, MA) by the above method but the final purification on DEAE-Sephadex was omitted because the yield was very low and the solution contained the required activity. The solution was used directly in the experiments described in this paper.

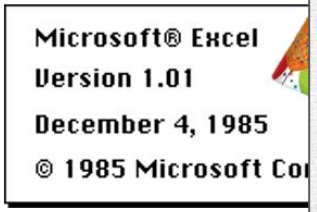
An attempt was made to prepare the triphosphate of the intermediate didehydrodideoxycytidine because Atkinson et

Abbreviations: The symbols C, T, A, and G are used for the deoxyribonucleotides in DNA sequences; the prefix dd is used for the 2′,3′-dideoxy derivatives (e.g., ddATP is 2′,3′-dideoxyadenosine 5′-triphosphate); the prefix ara is used for the arabinose analogues.

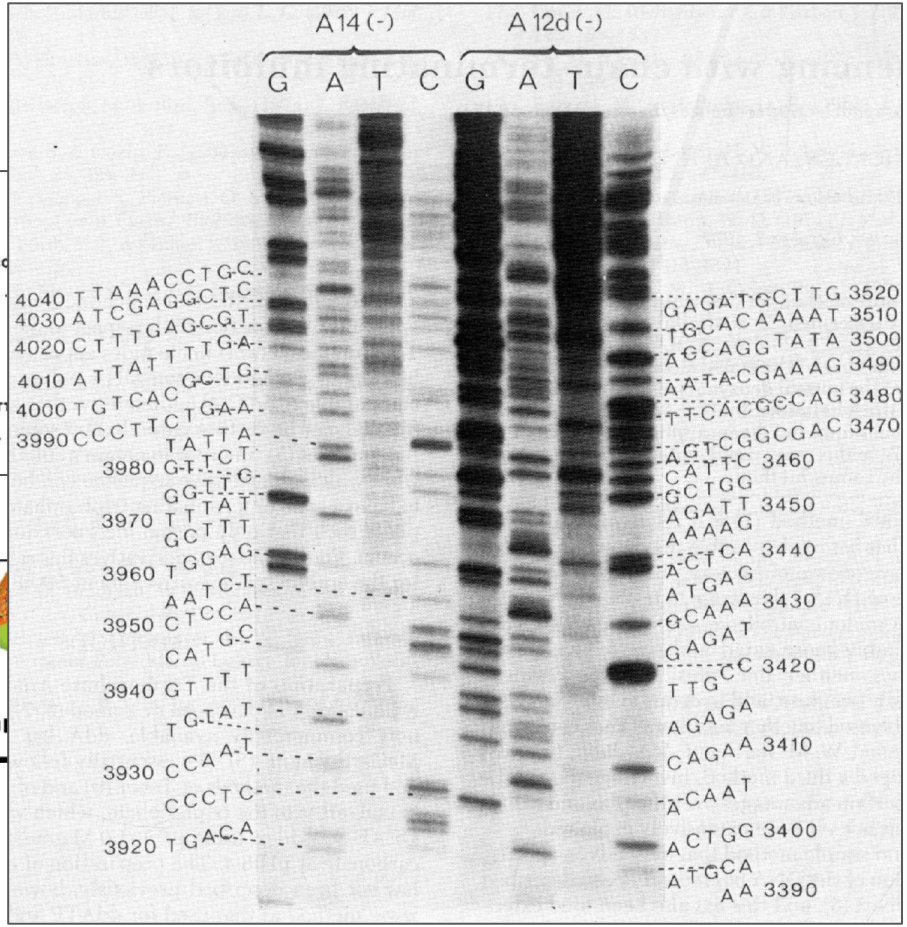
We've come a long way

1984

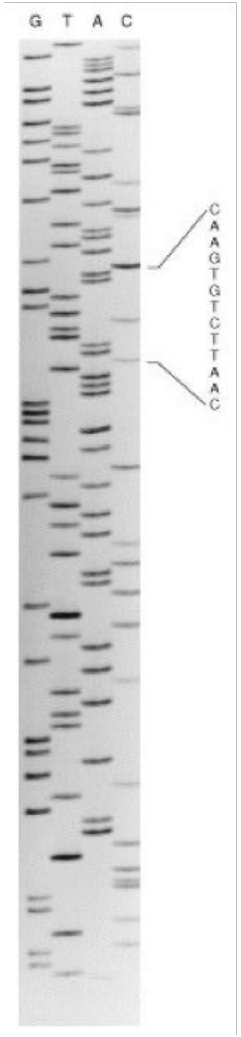




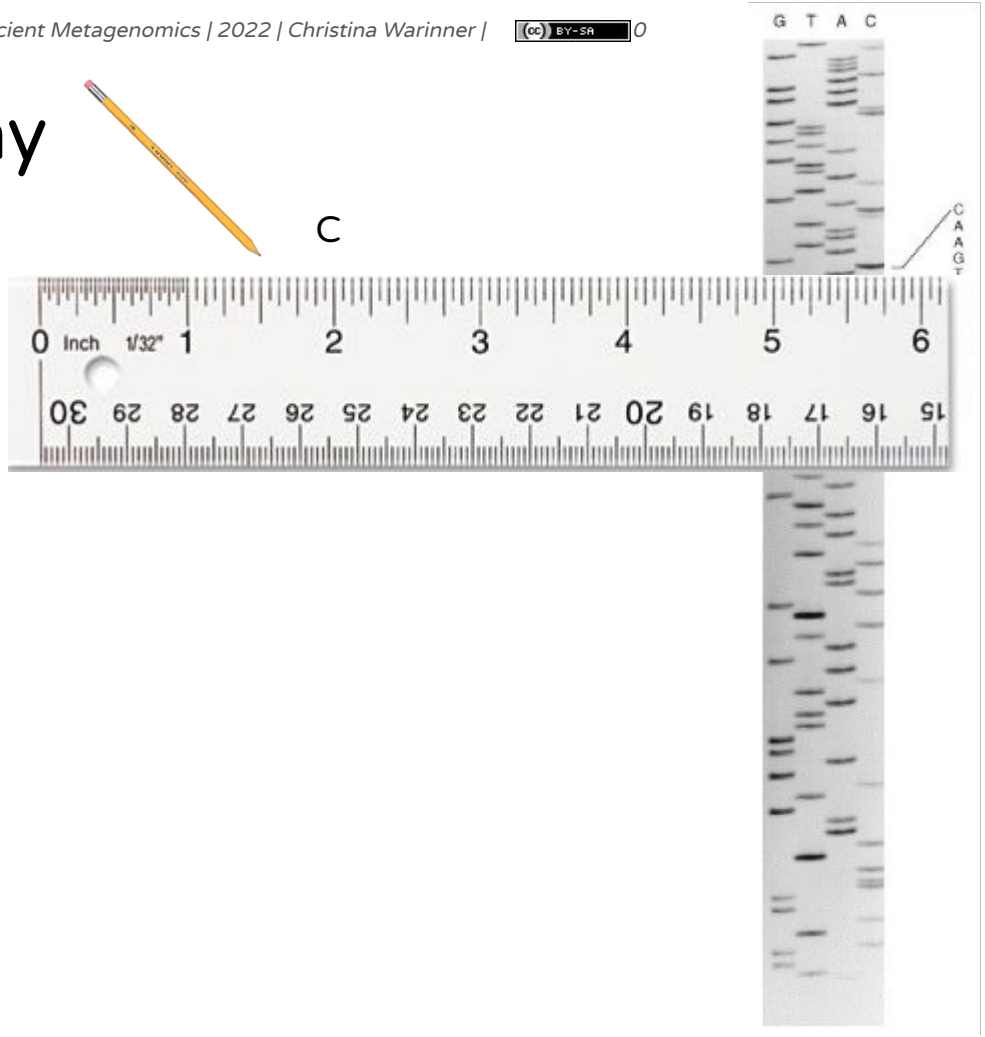
1985



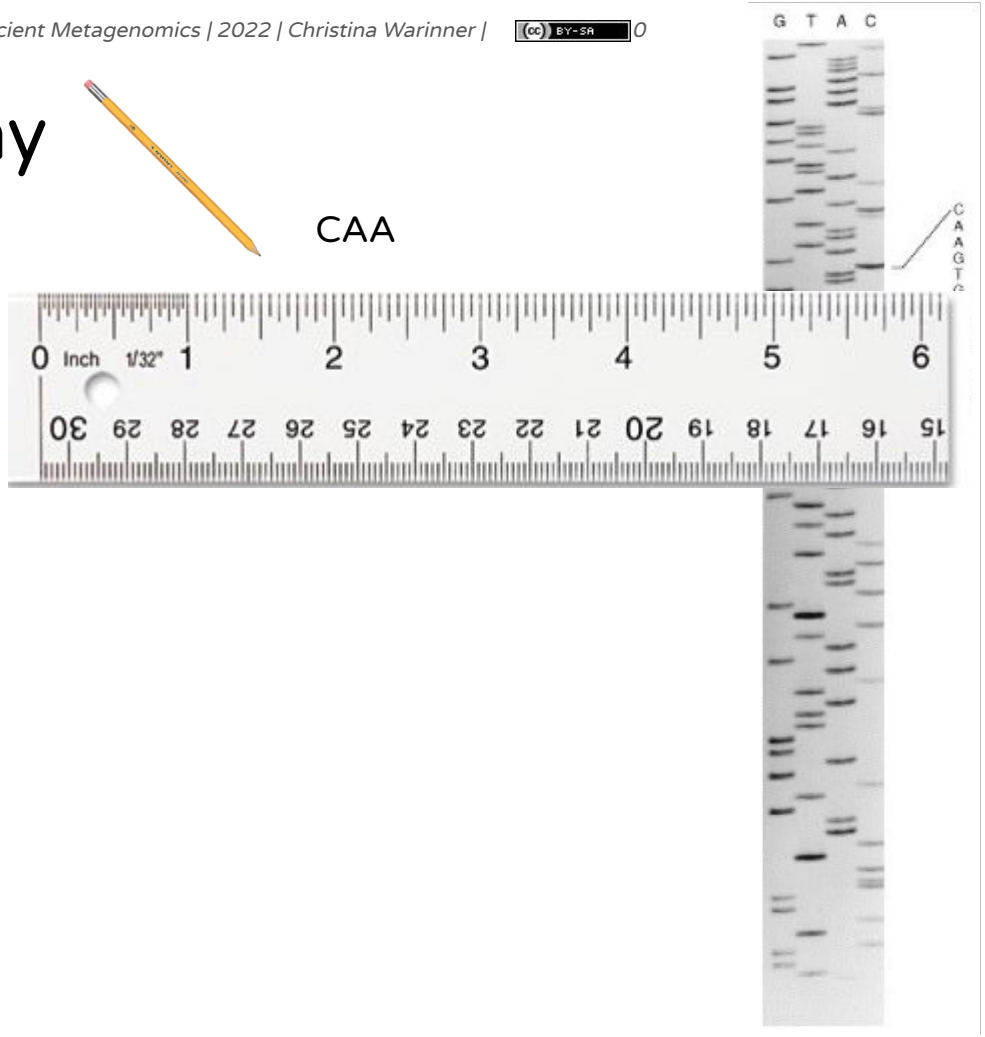
We've come a long way



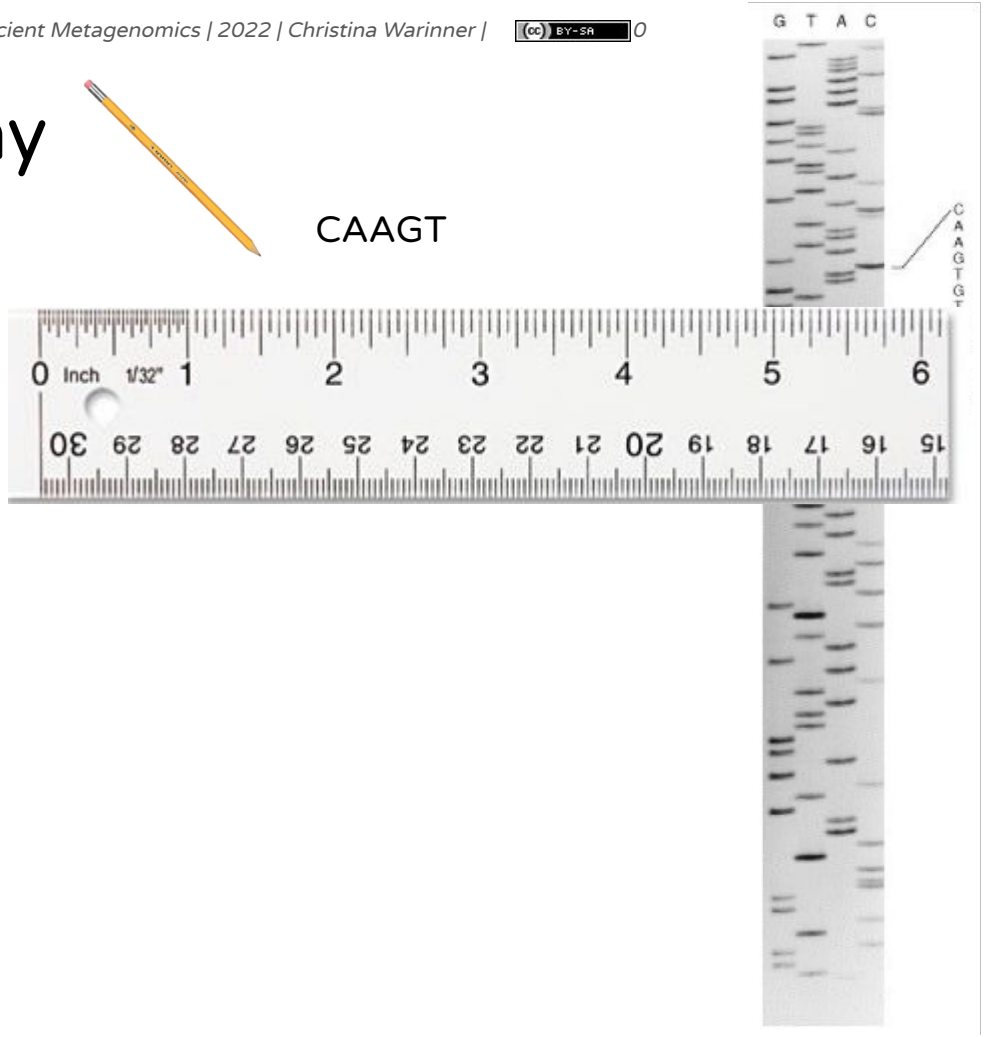
We've come a long way



We've come a long way

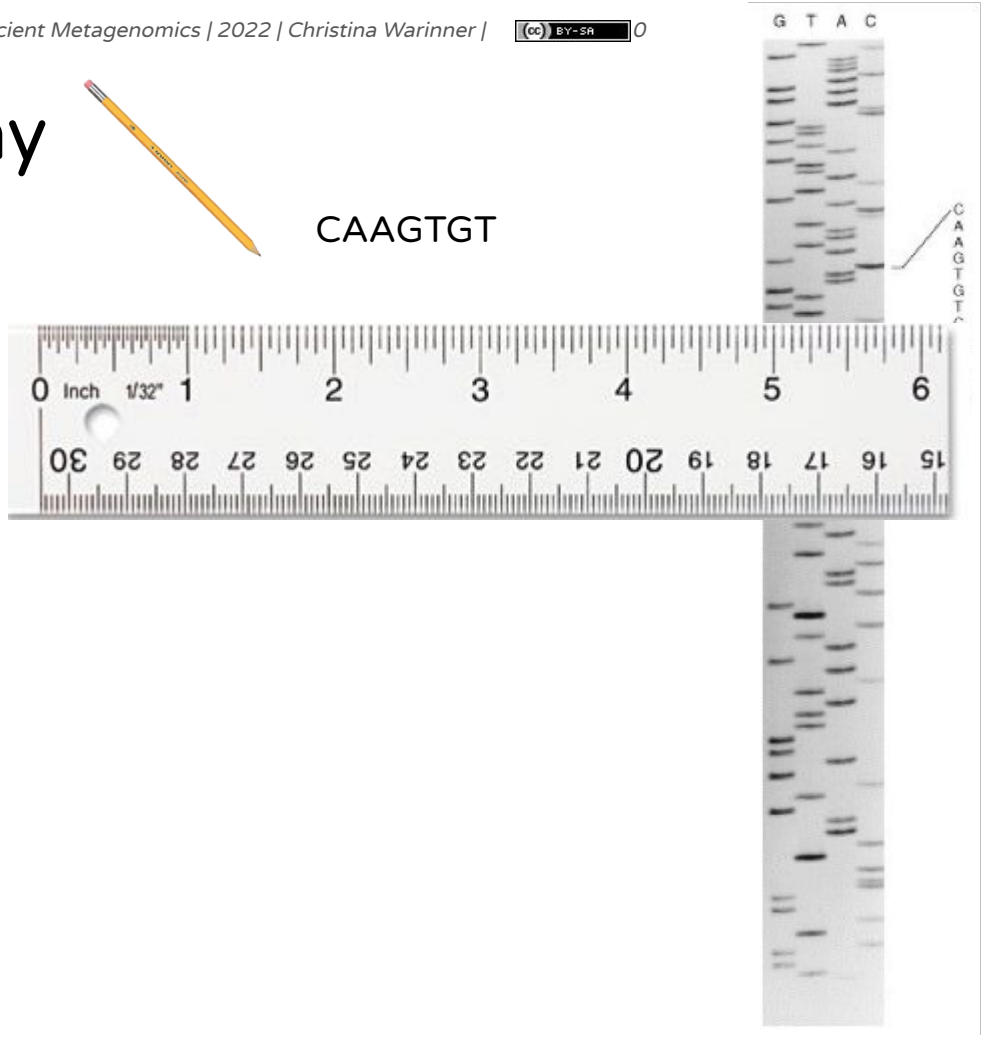


We've come a long way



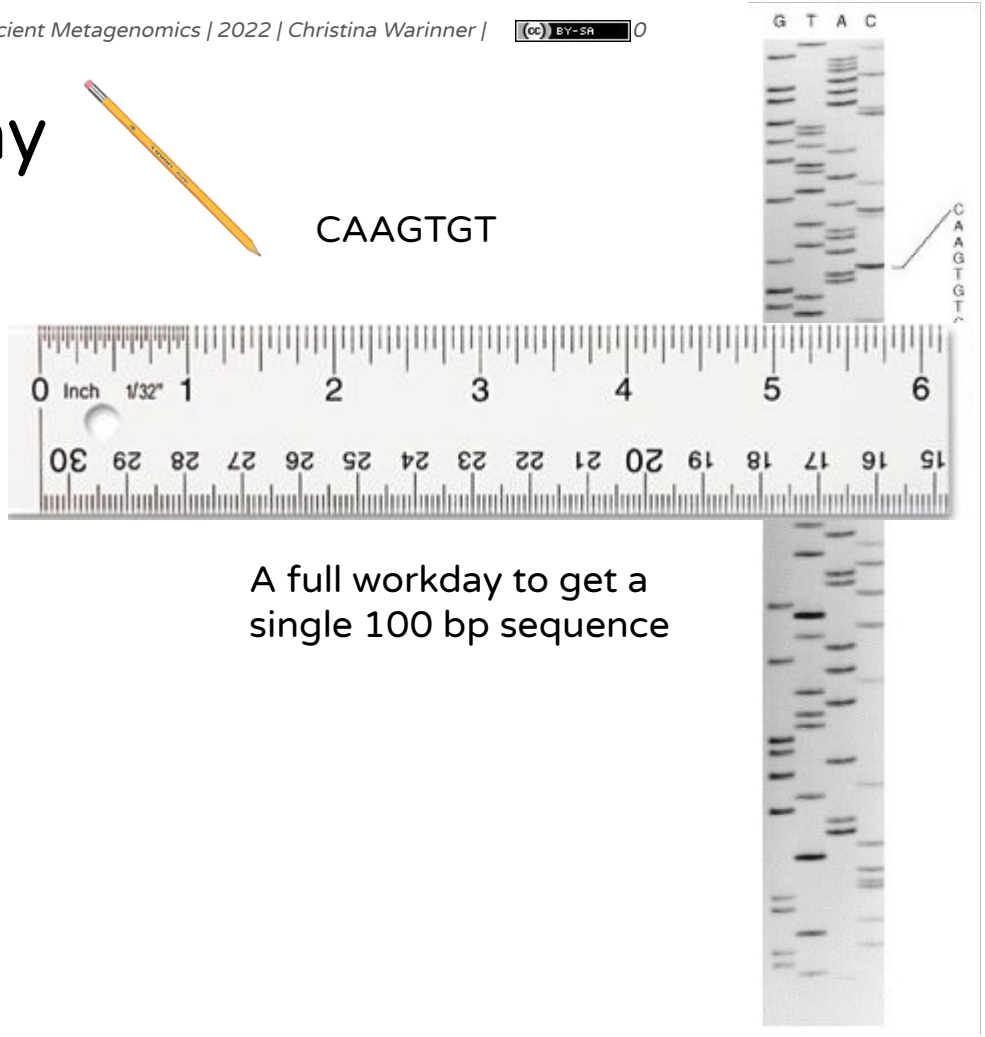
We've come a long way

CAAGTGT



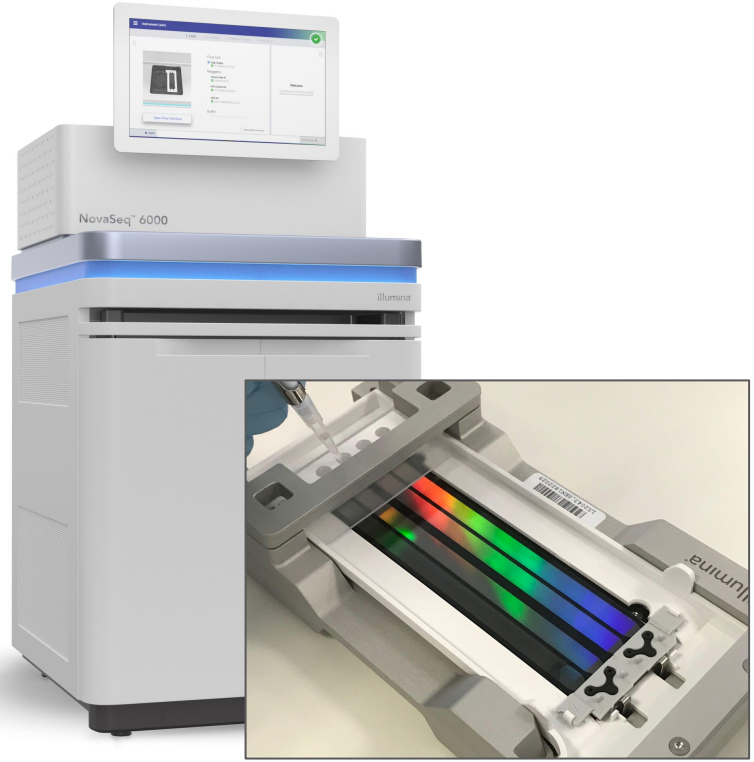
We've come a long way

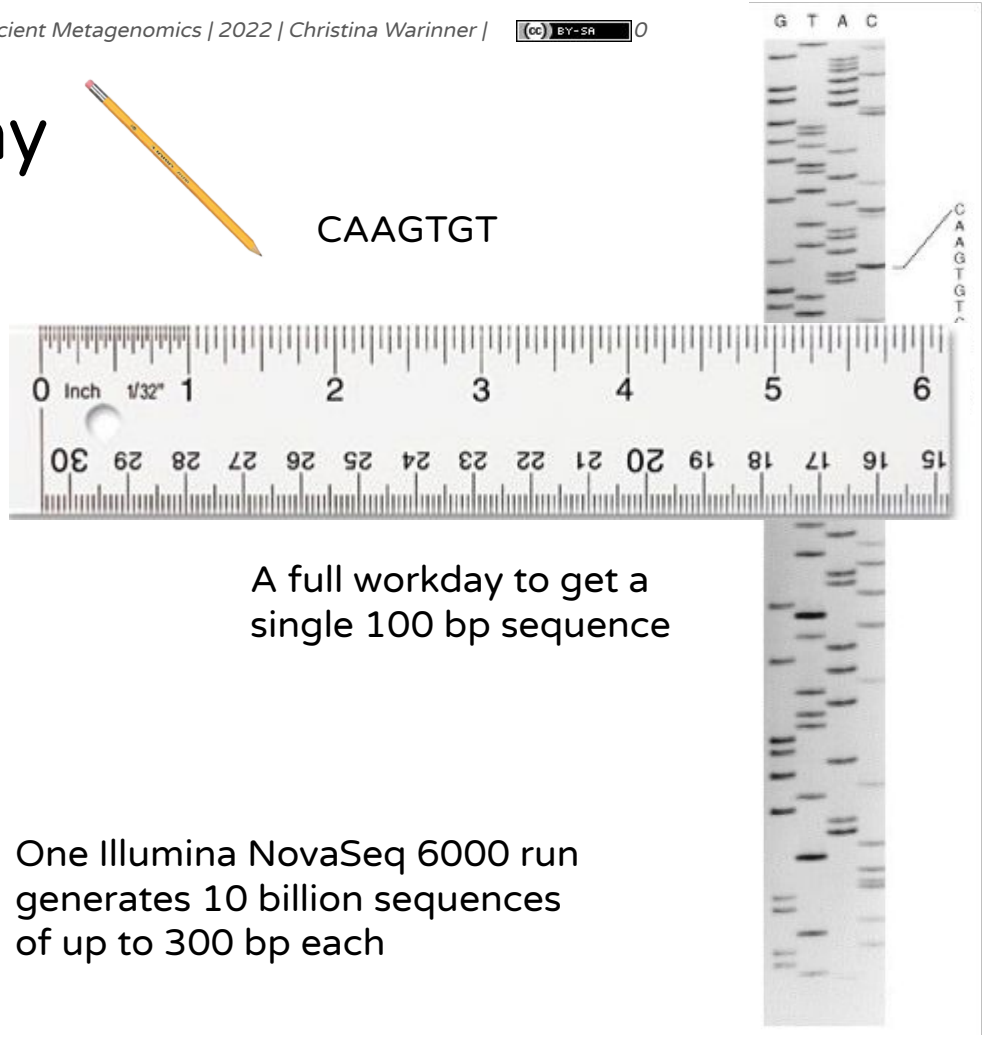
CAAGTGT



A full workday to get a single 100 bp sequence

We've come a long way





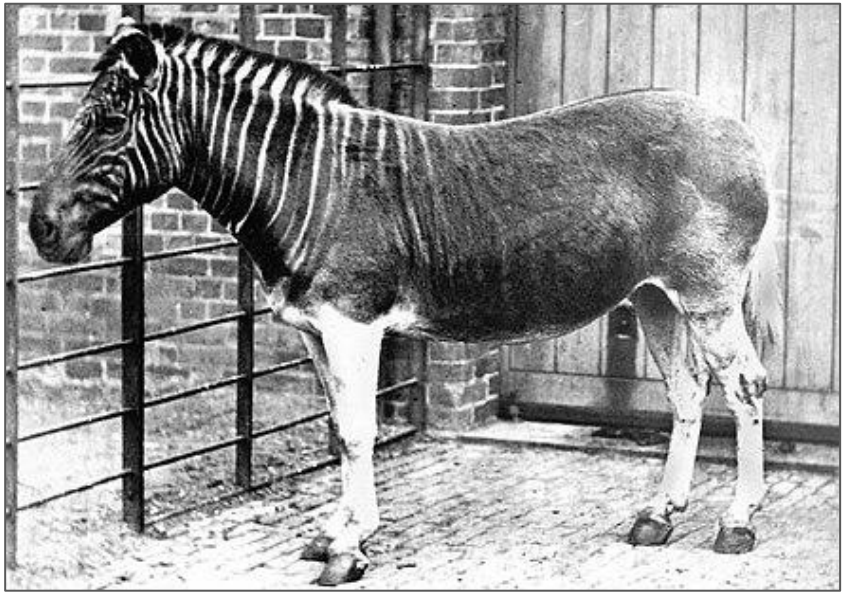
A full workday to get a single 100 bp sequence

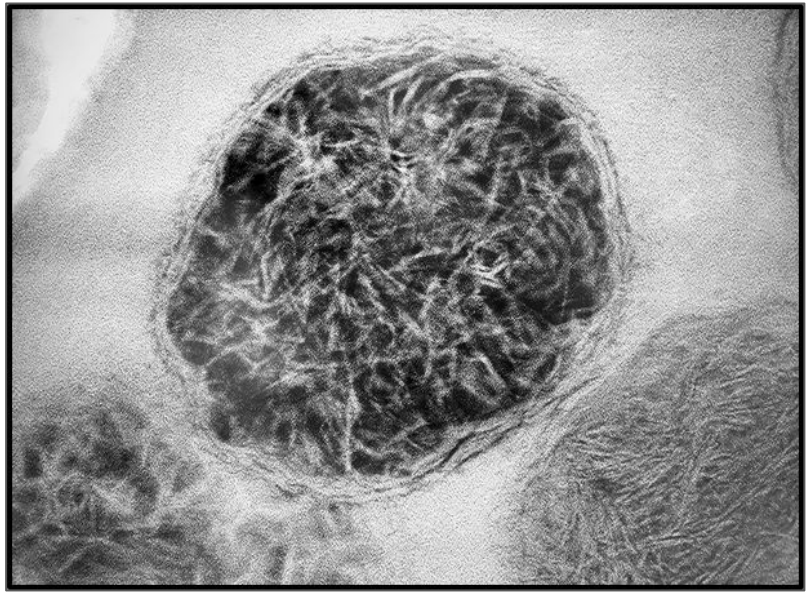
One Illumina NovaSeq 6000 run generates 10 billion sequences of up to 300 bp each

We've come a long way

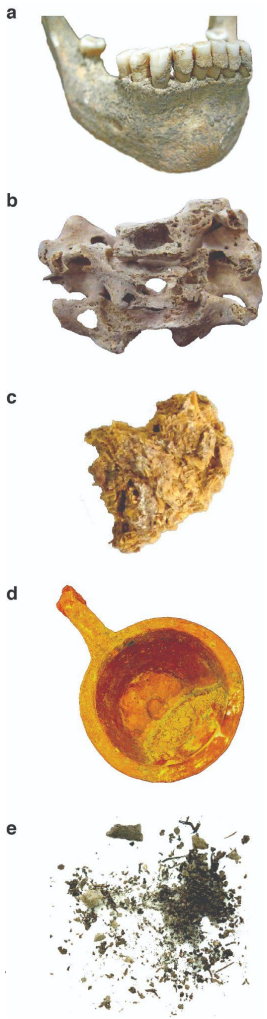
We've come a long way

From quagga to ancient microbes

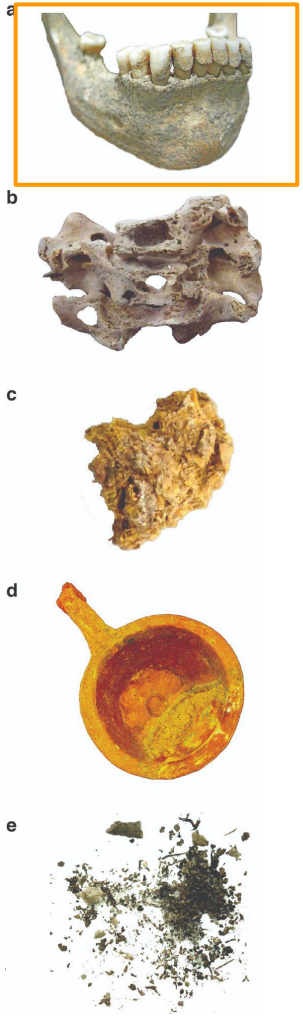


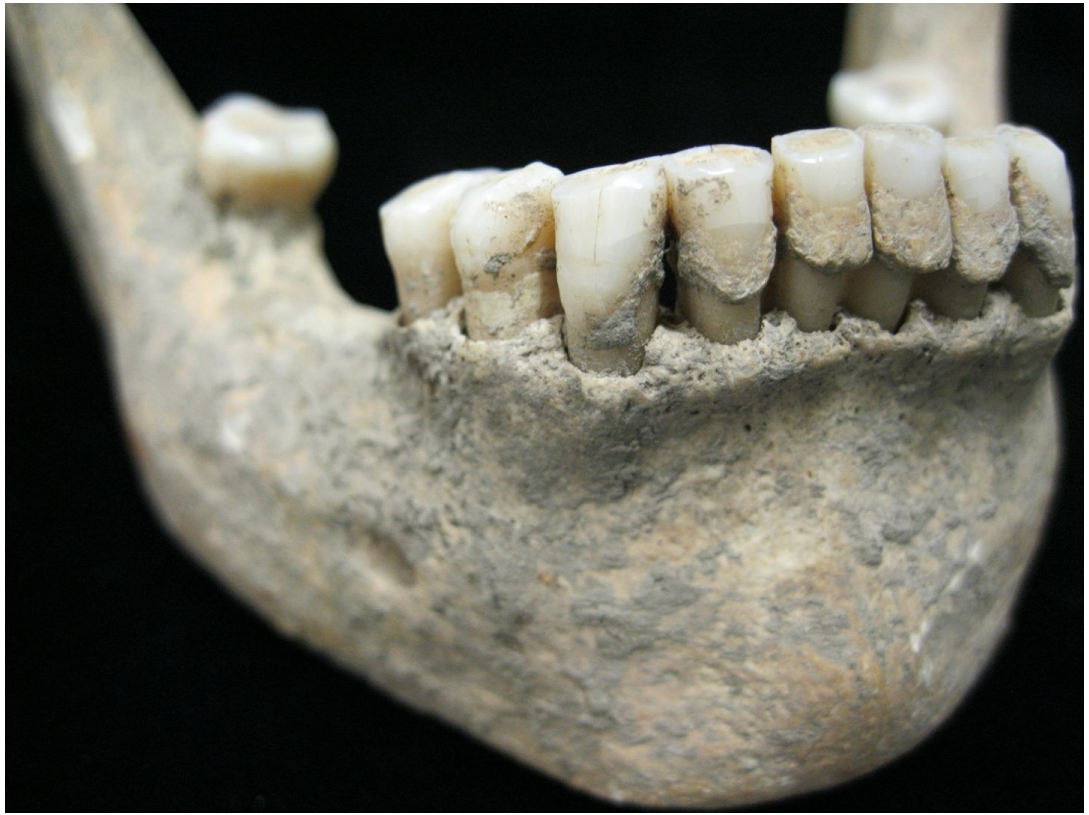


Where do we get ancient microbial DNA?

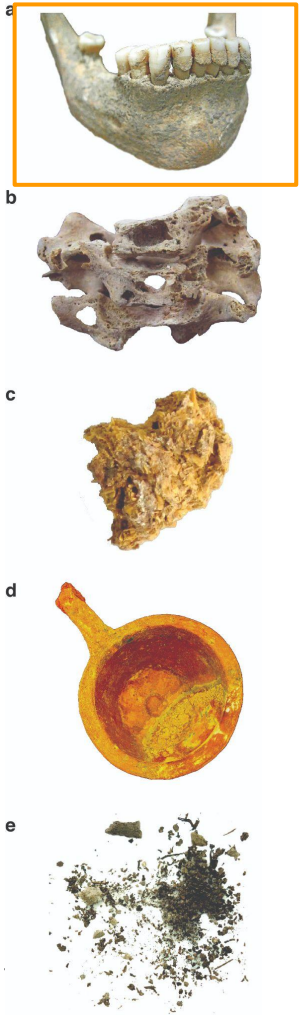


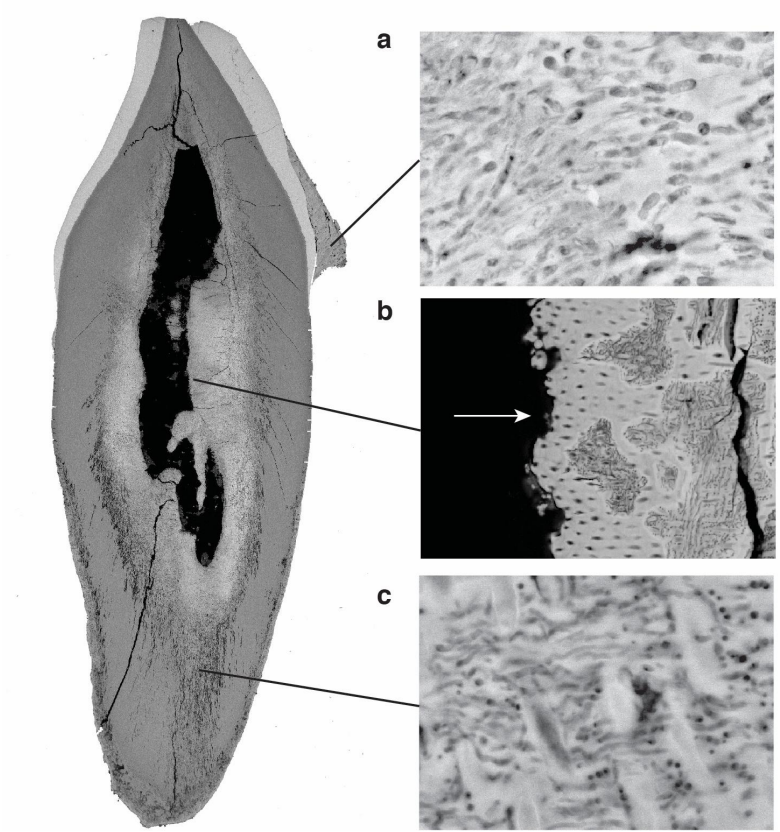
Where do we get ancient microbial DNA?





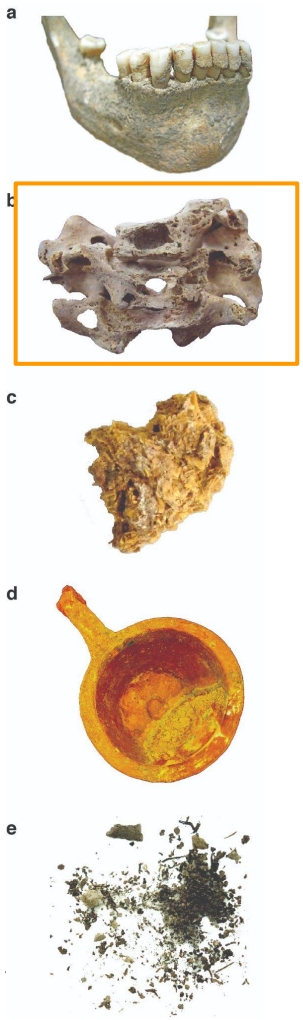
Where do we get ancient microbial DNA?





Germany, ca. 1100 CE
Warinner et al. 2014

Where do we get ancient microbial DNA?

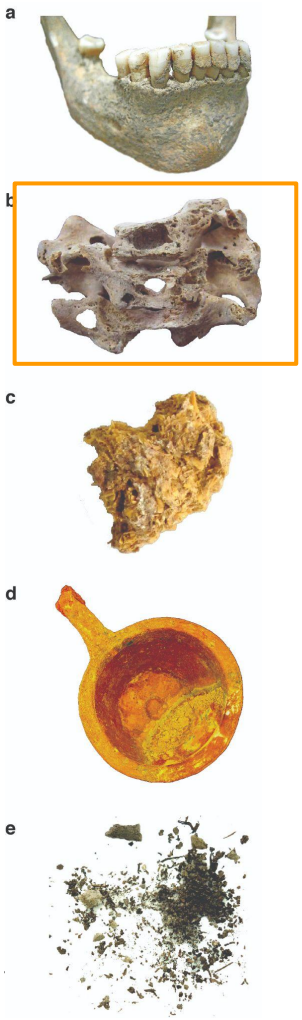


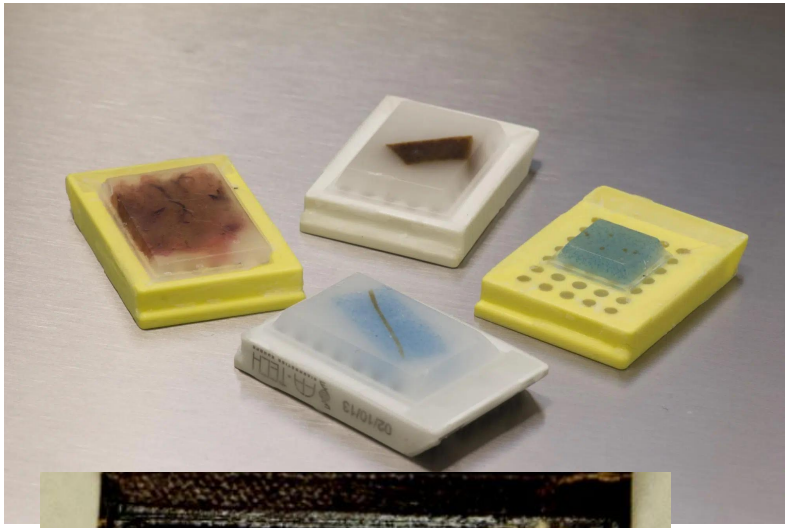
Tuberculosis, Peru
1000 CE, Bos et al. 2014

Tuberculosis, Egypt
250 BCE

Leprosy, England
ca. 1400 CE
Schünemann et al. 2018

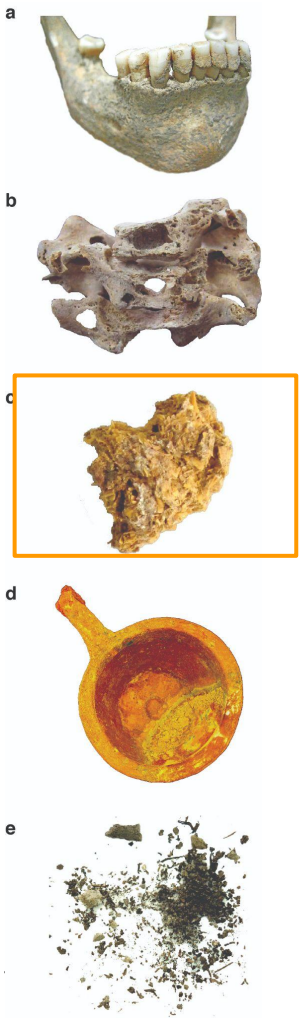
Where do we get ancient microbial DNA?



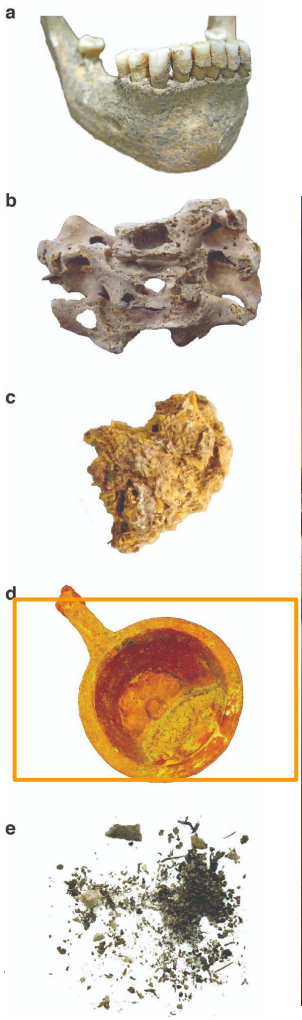


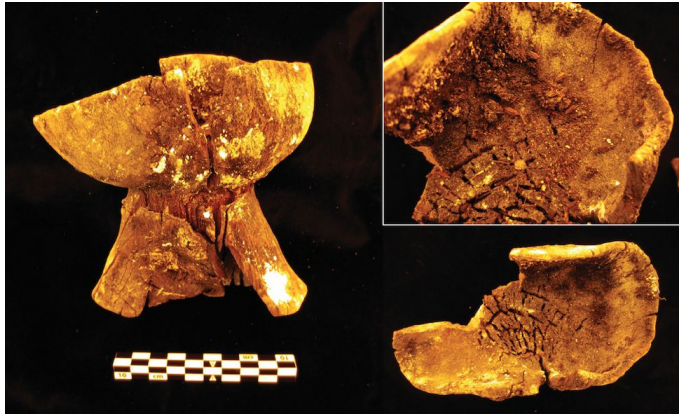
USA, 19th century, Duggan et al. 2020

Where do we get ancient microbial DNA?

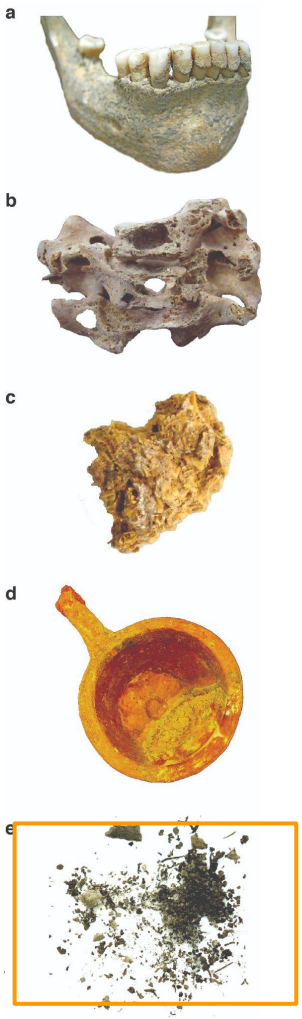


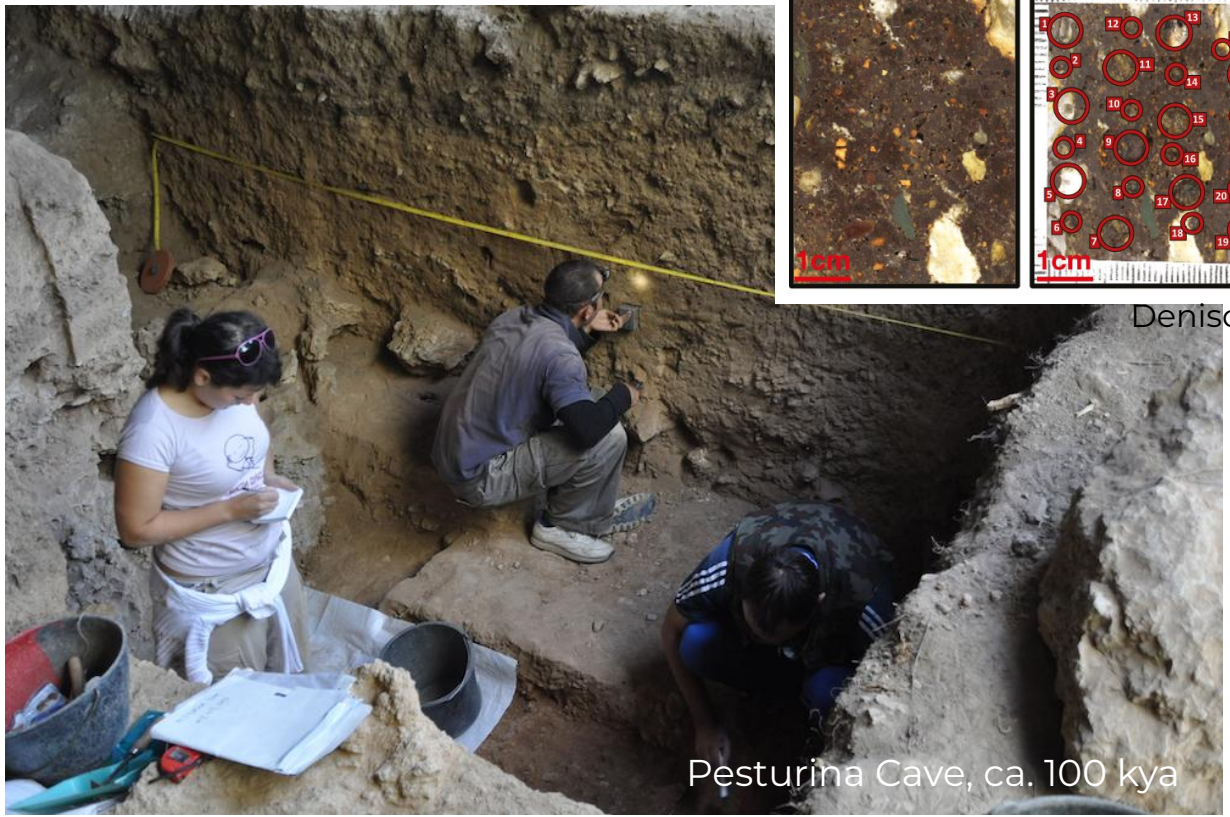
Where do we get ancient microbial DNA?

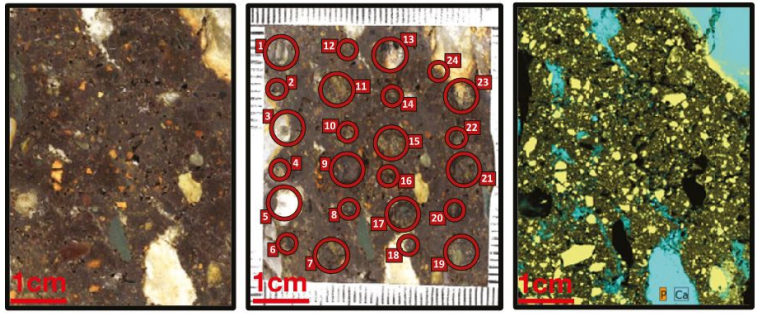




Where do we get ancient microbial DNA?

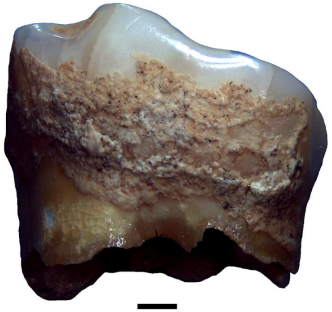






Denisova Cave, ca. 120 kya
Massilani et al. 2022

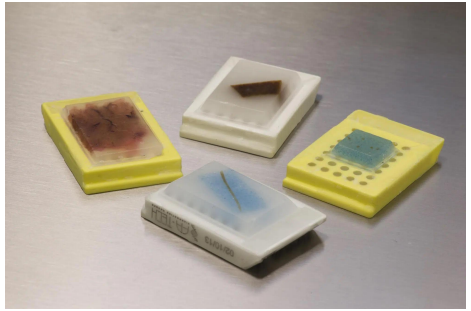
What is ancient DNA?



Any DNA from a non-living source that shows evidence of molecular degradation

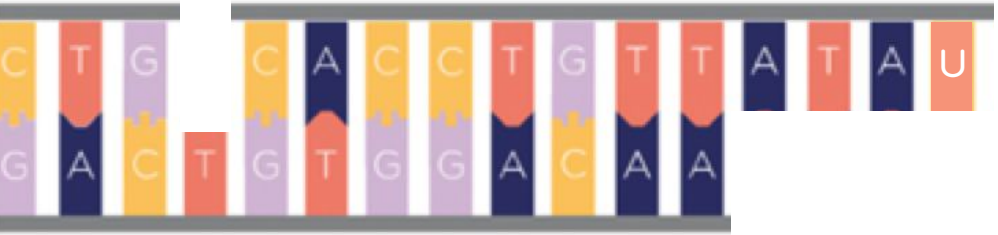
Not defined by a fixed age, but rather its condition

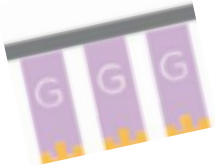
- 100,000-year-old Neanderthal oral microbiome DNA from dental calculus
- 5,000-year-old hepatitis B virus DNA from teeth
- 2,000-year-old gut microbiome DNA from paleofeces
- 600-year-old plague DNA from skeletons
- Oral bacterial DNA from 19th century gorillas in a museum
- Pathogen DNA from a 19th century medical specimen in alcohol
- Leprosy DNA from mid-20th century formalin-fixed paraffin embedded (FFPE) tissue blocks



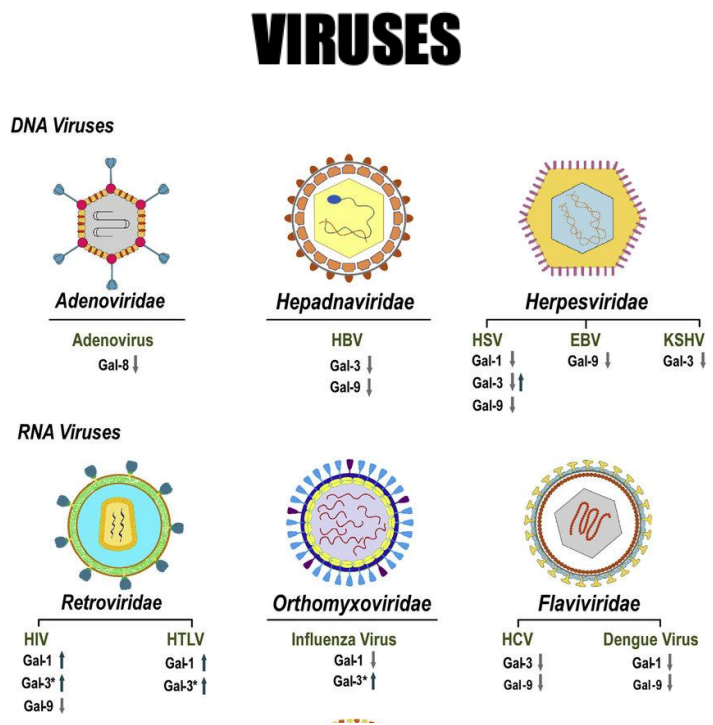
What is ancient DNA?

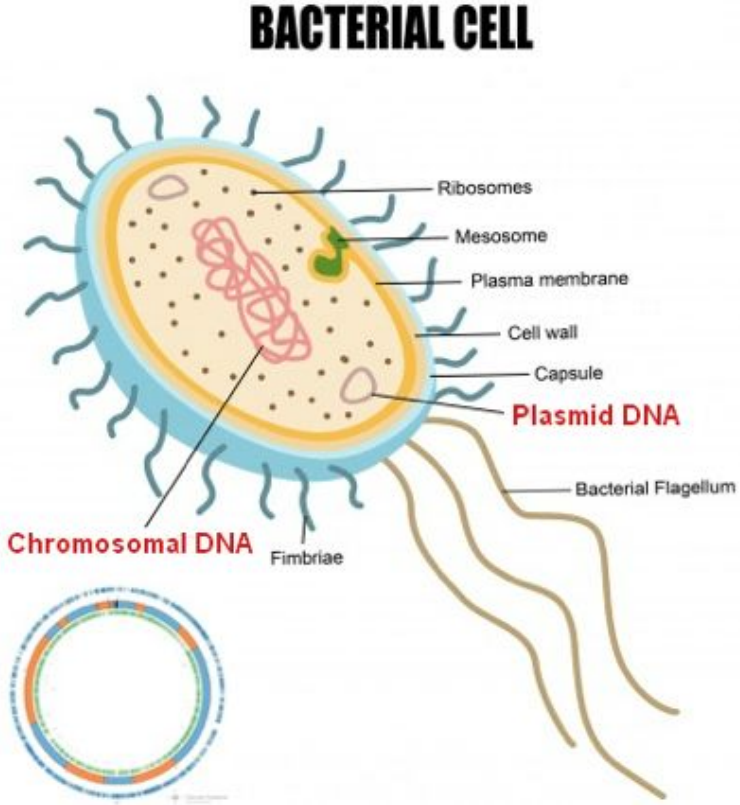
What is ancient DNA?



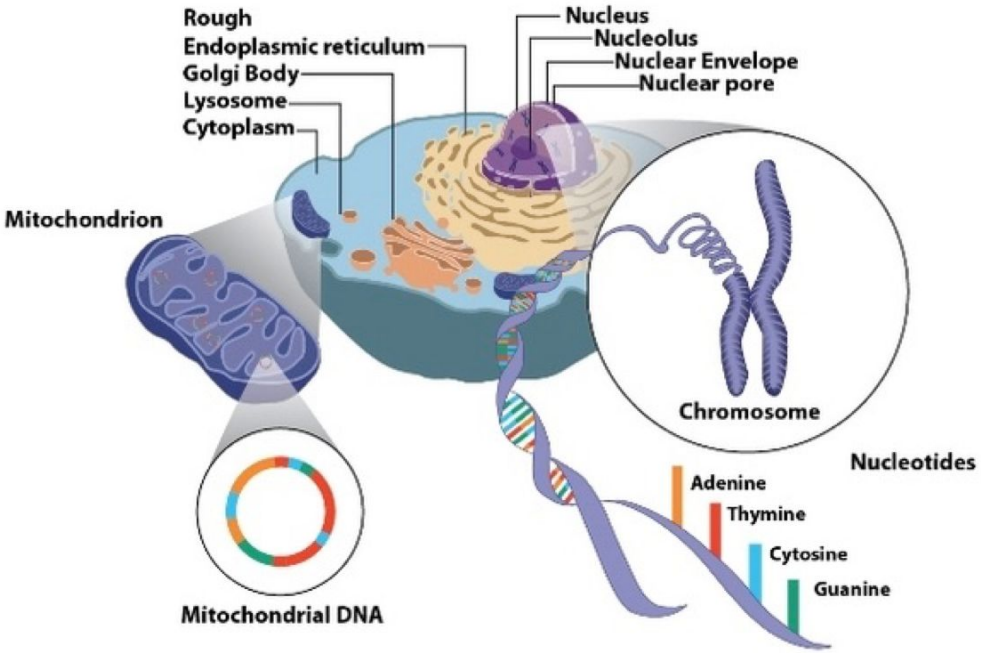


Genome basics

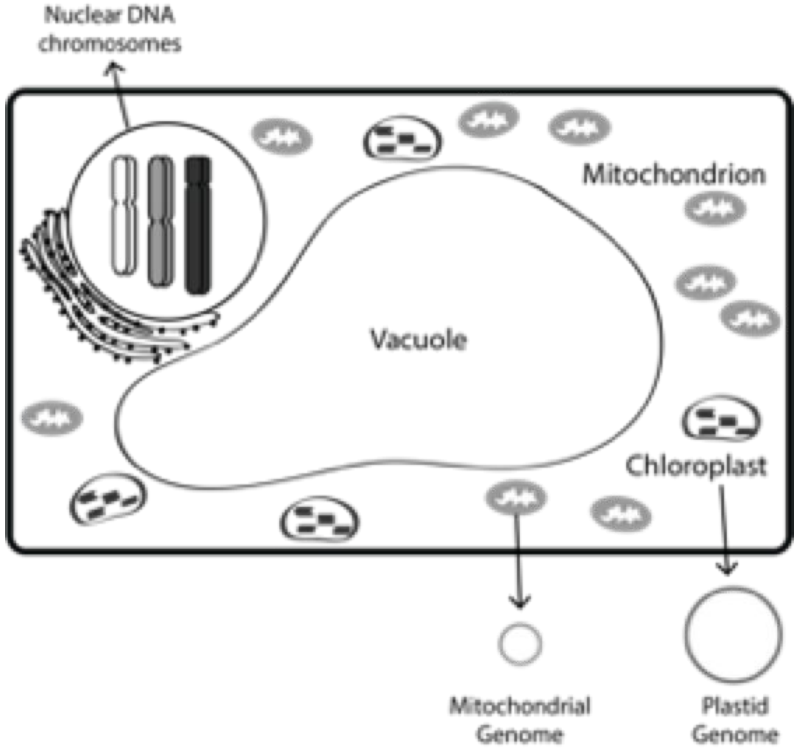




ANIMAL CELL



PLANT CELL



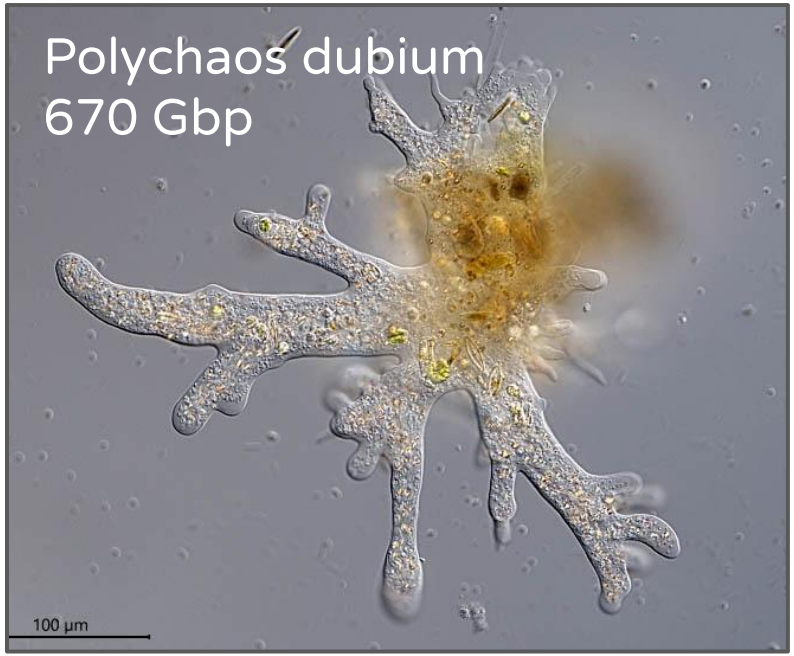
Relative genome sizes

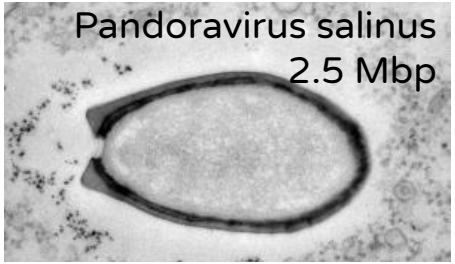
Viruses: 5-100 thousand bp (kbp)

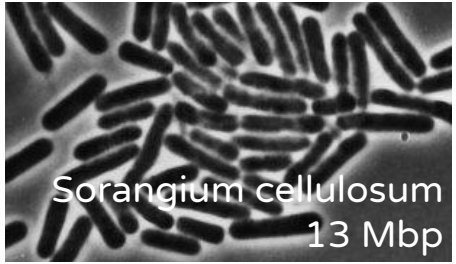
Bacteria: 1-5 million bp (Mbp)

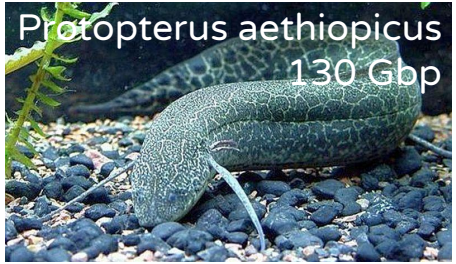
Animals: 3-6 billion bp (Gbp)

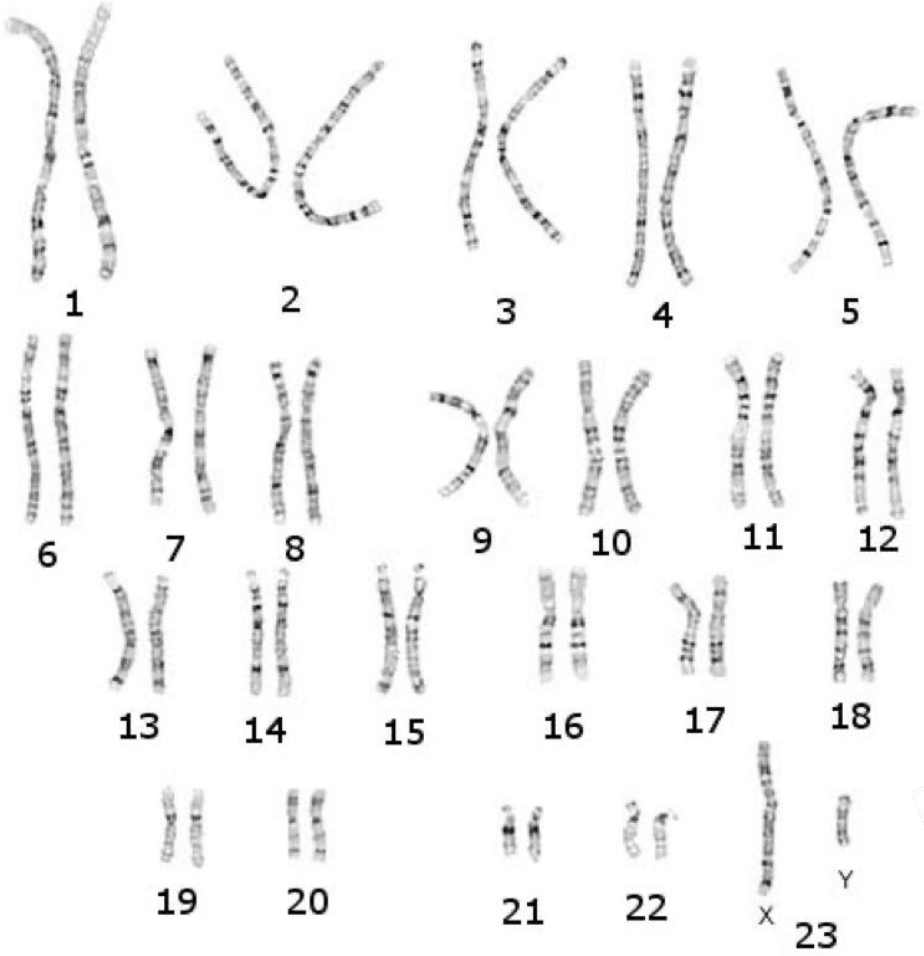
Plants: 6-18 billion bp (Gbp)











Human genome

3 Gbp

Copies: 2

Total: 6 Gbp

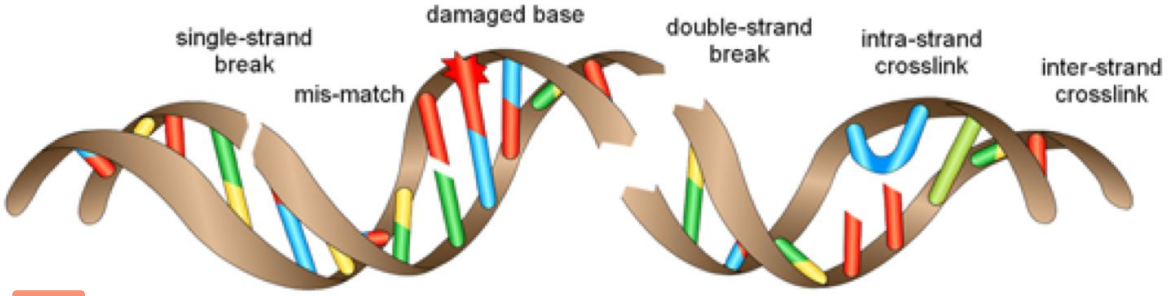
Chromosomes: 46 (23 pairs)

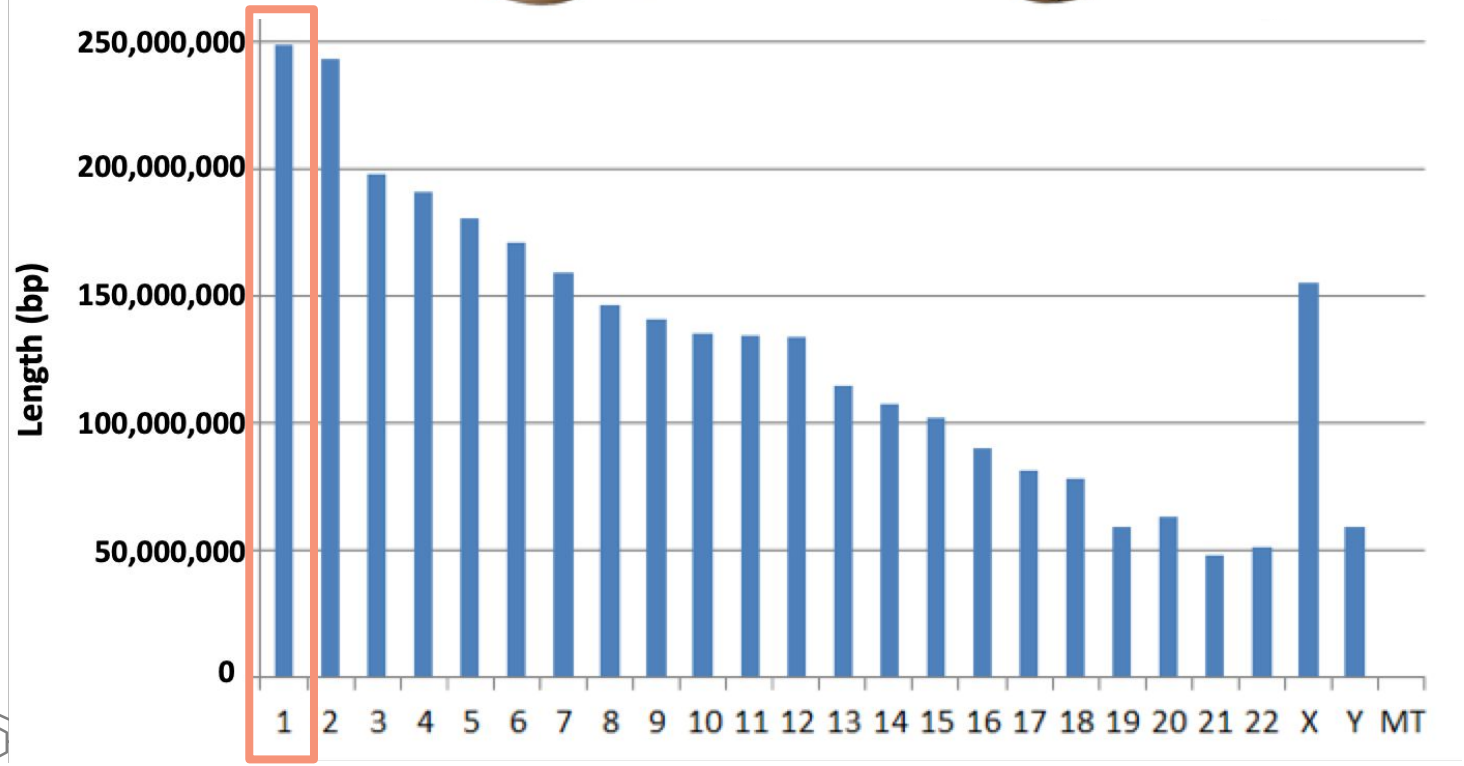
50-250 Mbp each

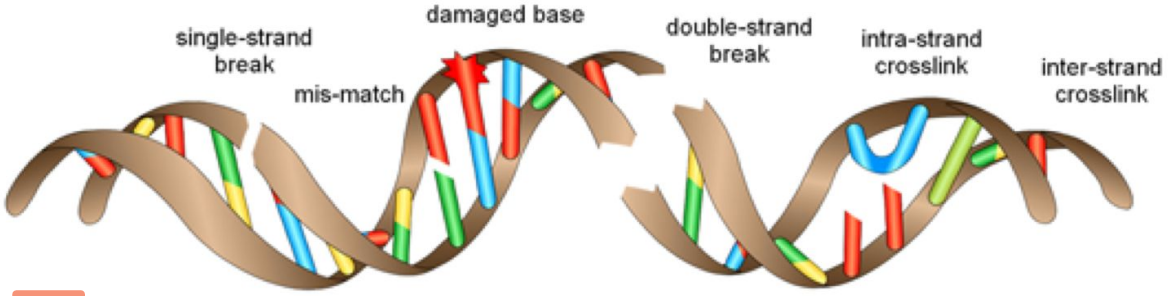
Mitogenome

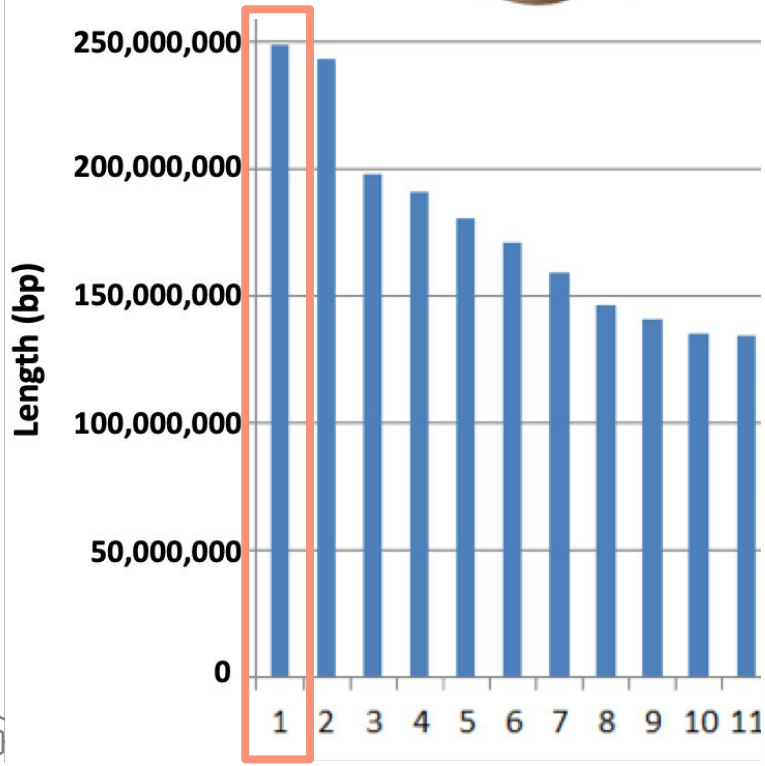
16.5 kbp

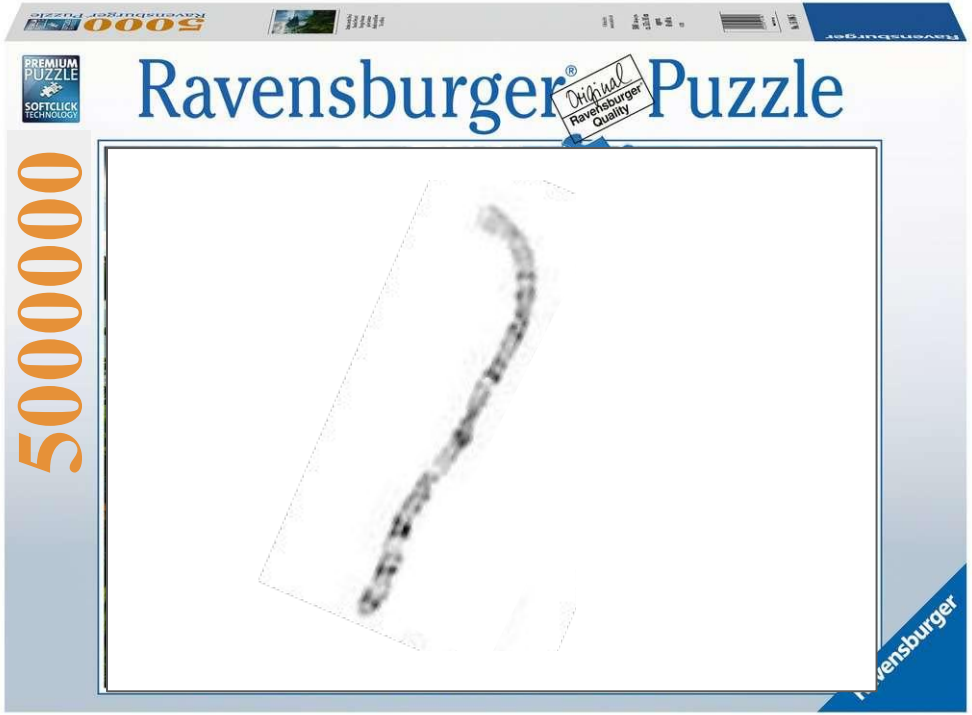
Copies: 1000+

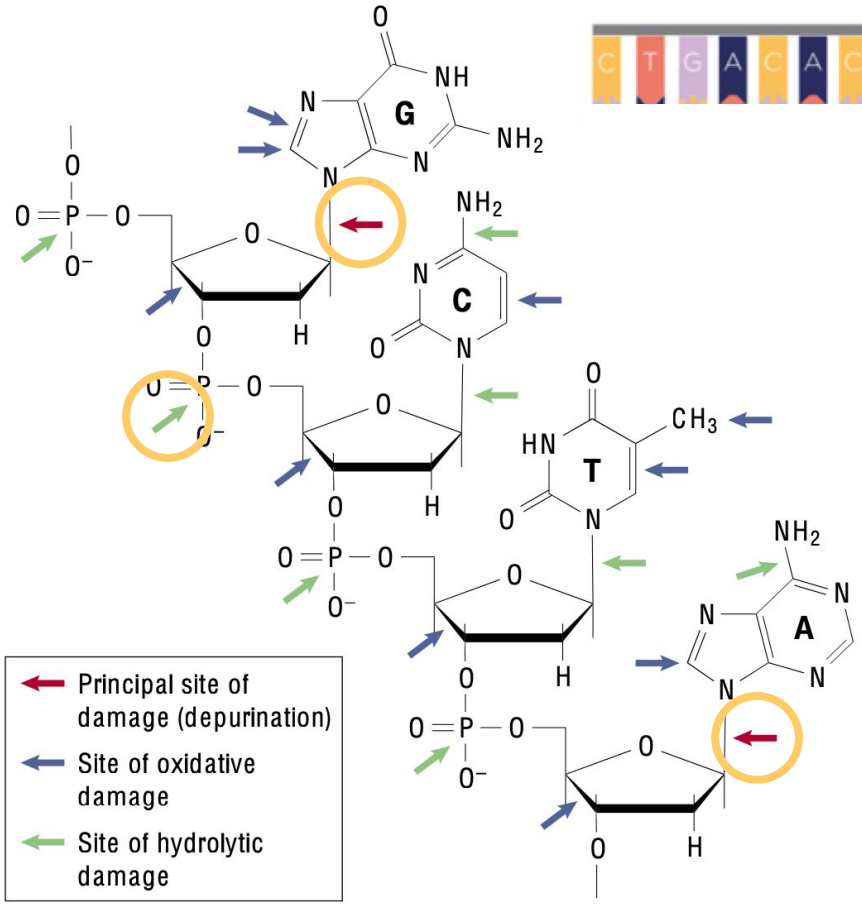

single-strand break
damaged base
mis-match
double-strand break
intra-strand crosslink
inter-strand crosslink

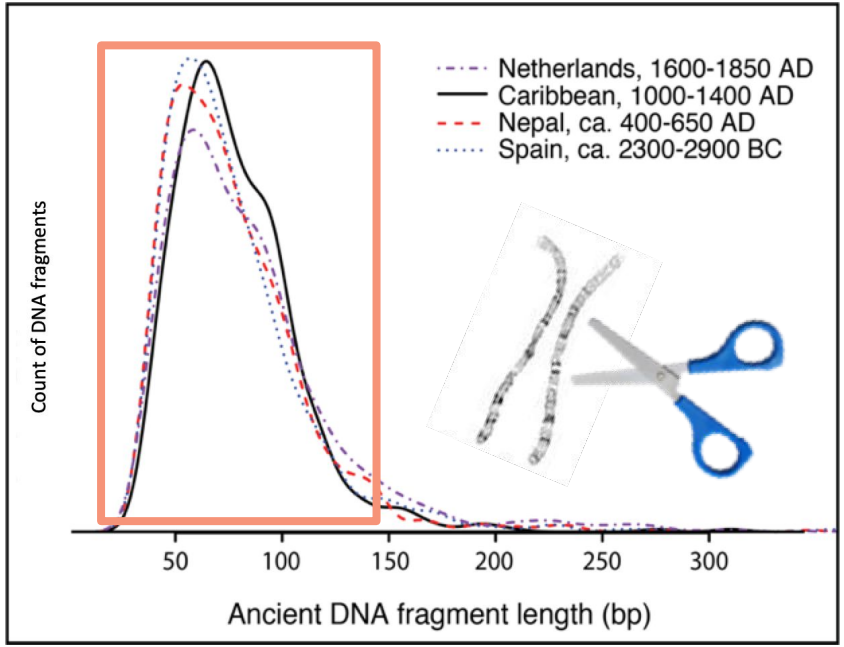

Length (bp)
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT


single-strand break
damaged base
mis-match
double-strand break
intra-strand crosslink
inter-strand crosslink

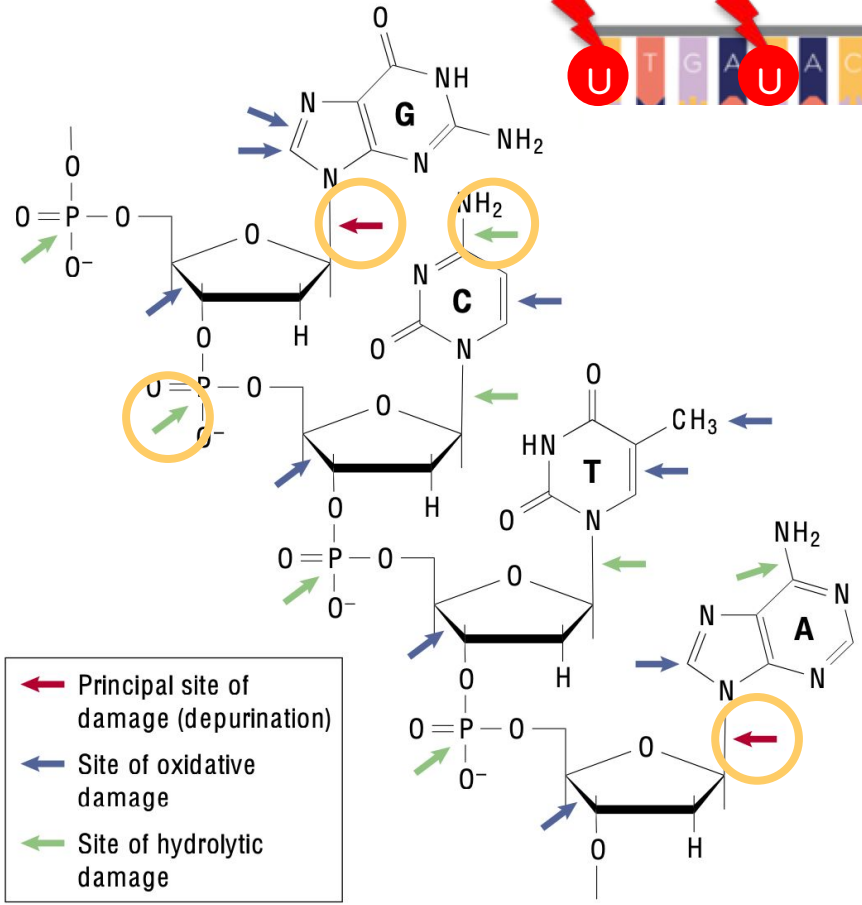

Length (bp)
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
0
1 2 3 4 5 6 7 8 9 10 11

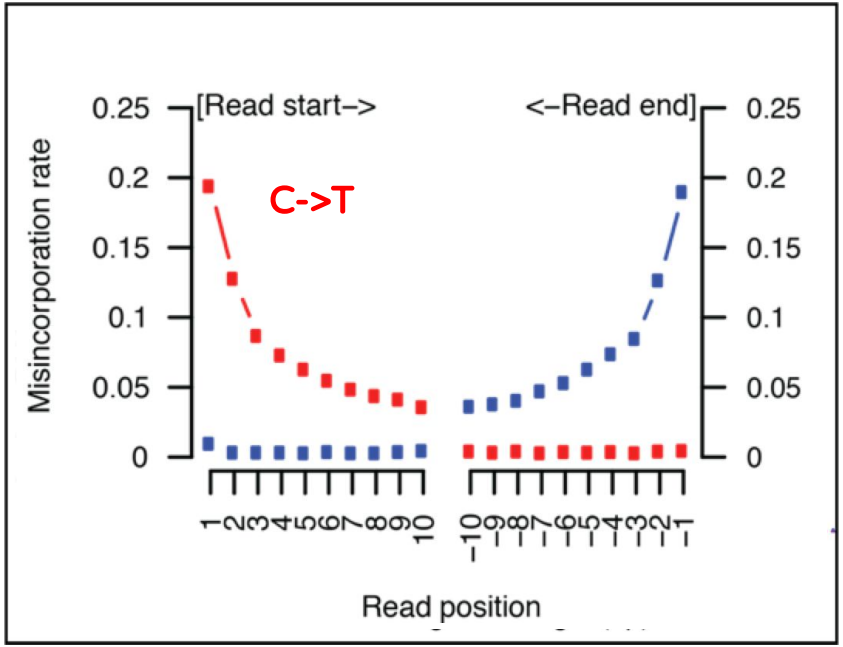

Ravensburger Puzzle
5000000
PREMIUM PUZZLE
SOFTCLICK TECHNOLOGY
Original Ravensburger Quality





Hofreiter et al. 2001





Hofreiter et al. 2001

DNA damage

1. **Depurination:**
   Random loss of A and G bases

2. **Nicking:**
   Hydrolytic attack of phosphate backbone at sites of depurination

3. **Fragmentation:**
   When two nicks on opposite strands are very close together, the hydrogen bonds between the bases aren't strong enough to hold the strands together and they separate, or "melt", causing fragmentation with single-stranded overhangs

4. **Deamination:**
   Cytosines on single-stranded overhangs undergo hydrolytic attack and lose their amine group, converting into uracil. DNA polymerases "read" the uracil as a thymine, introducing **C->T** errors in downstream sequences


PREMIUM PUZZLE
SOFTCLICK TECHNOLOGY
Ravensburger
Original Ravensburger Quality
Puzzle
60000
Ravensburger

How was this figured out?

pre-NGS era

Knew aDNA was fragmented but actual fragment length distribution was unknown (Pääbo et al. 2004)

Length of aDNA couldn't be precisely measured - short DNA easily lost during extraction, and DNA recovery was too low to see on a gel

Lots of guesses of "around 100 to 500 bp"

Early PCRs targeted DNA templates 300-500 bp long, but high PCR failure rate and vexing contamination problems (Hagelberg 1991; Champlot et al. 2010)

Known for some time that was an excess of C->T and G->A miscoding lesions in aDNA, but damage process was not well understood (Gilbert et al. 2003)

DNA damage was a "problem"

How was this figured out?

NGS era

Instead of requiring primer sites on the DNA template, NGS ligated primer binding sites onto the ends of molecules, making it possible for the first time to recover ALL of the DNA and measure the true size of aDNA

The order of damage processes could be determined and the process of DNA degradation could be defined (Briggs et al. 2007)

Improved extraction methods improved recovery of very short fragments, revealing that aDNA is very short, with an average of about 30-50 bp (Dabney et al. 2012)

The predictability of DNA damage became the “solution” to authenticating aDNA (Jónsson et al. 2013; Skoglund et al. 2014)

How was this figured out?

2007

Patterns of damage in genomic DNA sequences from a Neandertal

Adrian W. Briggs*†, Udo Stenzel*, Philip L. F. Johnson‡, Richard E. Green*, Janet Kelso*, Kay Prüfer*, Matthias Meyer*, Johannes Krause*, Michael T. Ronan§, Michael Lachmann*, and Svante Pääbo*†

*Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany; ‡Biophysics Graduate Group, University of California, Berkeley, CA 94720; and §454 Life Sciences, Branford, CT 06405

Contributed by Svante Pääbo, May 25, 2007 (sent for review April 25, 2007)

High-throughput direct sequencing techniques have recently opened the possibility to sequence genomes from Pleistocene organisms. Here we analyze DNA sequences determined from a Neandertal, a mammoth, and a cave bear. We show that purines are overrepresented at positions adjacent to the breaks in the ancient DNA, suggesting that depurination has contributed to its degradation. We furthermore show that substitutions resulting from miscoding cytosine residues are vastly overrepresented in the DNA sequences and drastically clustered in the ends of the molecules, whereas other substitutions are rare. We present a model where the observed substitution patterns are used to estimate the rate of deamination of cytosine residues in single- and double-stranded portions of the DNA, the length of single-stranded ends, and the frequency of nicks. The results suggest that reliable genome sequences can be obtained from Pleistocene organisms.

454 | deamination | depurination | paleogenomics

The retrieval of DNA sequences from long-dead organisms offers a unique perspective on genetic history by making information from extinct organisms and past populations available. However, three main technical challenges affect such studies. First, when DNA is preserved in ancient specimens, it is invariably degraded to a small average size (1). Second, chemical damage is present in ancient DNA (2) that may cause incorrect DNA sequences to be determined (3). Third, because ancient DNA is present in low amounts or absent in many specimens, traces of modern DNA from extraneous sources may cause modern DNA sequences to be mistaken for endogenous ancient DNA sequences (4–6). Recently, a DNA sequencing method based on highly parallel pyrosequencing of DNA templates generated by the PCR has been developed by 454 Life Sciences (454) (7). This method allows several hundred thousand DNA sequences of length 100 or 250 nt to be determined in a short time. It has been used to determine DNA sequences from the remains of three Pleistocene species: mammoths (8, 9), a cave bear (9), and a Neandertal (10). In all cases, the majority of DNA sequences retrieved are from microorganisms that have colonized the tissues after the death of the organisms. However, a fraction stem from the ancient organisms. In fact, the throughput of this technology, as well as other sequencing technologies currently becoming available (11), makes it possible to contemplate sequencing the complete genomes of extinct Pleistocene species (8, 10).

Here, we analyze DNA sequences determined on the 454 platform from an ≈38,000-year-old Neandertal specimen found at Vindija Cave, Croatia (10, 12), with respect to two features of particular significance for genomic studies of ancient DNA. First, we investigate the DNA sequence context around strand breaks in ancient DNA. This has not been previously possible, because when PCR is used to retrieve ancient DNA sequences, primers that target particular DNA sequences are generally used and thus the ends of the ancient DNA molecules are not revealed. Second, we investigate the patterns of nucleotide misincorporations in the ancient DNA sequences as a function of their position in ancient DNA fragments. Although there is strong evidence that the majority of such misincorporations are due to deamination of cytosine residues to uracil residues (3), which code as thymine residues, it is unclear whether other miscoding lesions are present in any appreciable frequency in ancient DNA or how miscoding lesions are distributed along ancient DNA molecules. When relevant, we use comparable data from an ≈43,000-year-old mammoth bone (9) from the Bol'shaya Kolopatkaya river, Russia, an ≈42,000-year-old cave bear bone from Ochsenhalt Cave, Austria (13), a contemporary human, and DNA sequences of the Vindija Neandertal cloned in a plasmid vector (14) to ask whether the patterns seen are general features of Pleistocene DNA sequences or are caused by the 454 sequencing process. Finally, we develop a model that allows us to estimate features of ancient DNA preservation and discuss the implications of our findings for the determination of complete genome sequences from Pleistocene organisms.

Results and Discussion

The 454 Process. Because aspects of the 454 sequencing process are of crucial importance for the analyses presented, we briefly review some of its essential features. In a first step, a double-stranded DNA extract is end-repaired and ligated to two different synthetic oligonucleotide adaptors termed A and B. From each successfully ligated molecule, one of the DNA strands is isolated and subjected to emulsion PCR, during which each template remains isolated from other templates on a Sepharose bead carrying oligonucleotides complementary to one of the adaptors, producing beads each coated with ≈10 million copies of one DNA molecule. Up to 800,000 such DNA-containing beads are then loaded onto a multiwell glass plate, and their sequences are determined by pyrosequencing (7).

The end repair of the template DNA and ligation of adapters, which are critical for the analyses in this paper, are described in more detail in Fig. 1. First, *T4* DNA polymerase is used to remove single-stranded 3′-overhanging ends and to fill in 5′-overhanging ends (Fig. 1*ii*). Simultaneously, 5′-ends are phos-

Author contributions: A.W.B., R.E.G., and S.P. designed research; J. Kelso, K.P., J. Krause, and M.T.R. contributed new reagents/analytic tools; A.W.B., U.S., P.L.F.J., R.E.G., M.M., M.L., and S.P. analyzed data; and A.W.B., P.L.F.J., R.E.G., and S.P. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: 454, 454 Life Sciences; mtDNA, mitochondrial DNA; C.I., confidence interval.

Data deposition: The sequences reported in this paper have been deposited as follows. Directly sequenced Neandertal and mammoth sequences have been deposited in the European Molecular Biology Laboratory database (Neandertal accession nos. CAAM02000001–CAAM02070991, mammoth accession nos. CAAM02000001–CAAM02064265) and in the National Center for Biotechnology Information trace archive under GenomeProject IDs 18313 (Neandertal) and 17621 (mammoth). Cave bear and contemporary human sequences have been deposited in the National Center for Biotechnology Information trace archive under GenomeProject IDs 19671 (cave bear) and 19675 (human).

†To whom correspondence should be addressed. E-mail: briggs@eva.mpg.de or paabo@eva.mpg.de.

This article contains supporting information online at www.pnas.org/cgi/content/full/0704665104/DC1.

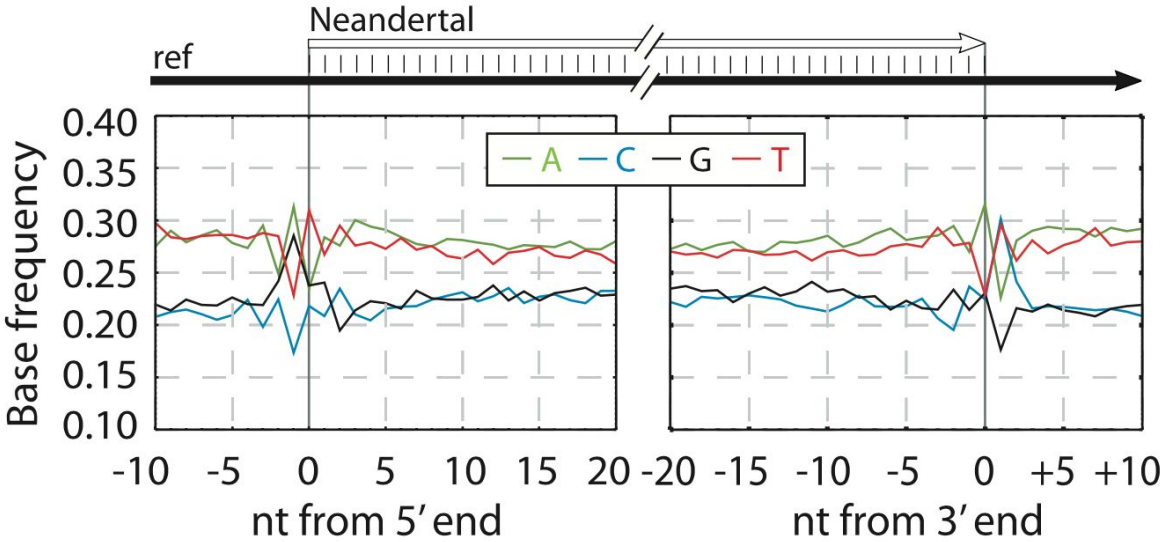


Fig. 2. Base composition at ends of Neandertal DNA sequences. The base composition of the human reference sequence is plotted as a function of distance from 5′- and 3′-ends of Neandertal sequences.

How was this figured out?

2007

Patterns of damage in genomic DNA sequences from a Neandertal

Adrian W. Briggs*†, Udo Stenzel*, Philip L. F. Johnson‡, Richard E. Green*, Janet Kelso*, Kay Prüfer*, Matthias Meyer*, Johannes Krause*, Michael T. Ronan§, Michael Lachmann*, and Svante Pääbo*†

*Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany; ‡Biophysics Graduate Group, University of California, Berkeley, CA 94720; and §454 Life Sciences, Branford, CT 06405

Contributed by Svante Pääbo, May 25, 2007 (sent for review April 25, 2007)

High-throughput direct sequencing techniques have recently opened the possibility to sequence genomes from Pleistocene organisms. Here we analyze DNA sequences determined from a Neandertal, a mammoth, and a cave bear. We show that purines are overrepresented at positions adjacent to the breaks in the ancient DNA, suggesting that depurination has contributed to its degradation. We furthermore show that substitutions resulting from miscoding cytosine residues are vastly overrepresented in the DNA sequences and drastically clustered in the ends of the molecules, whereas other substitutions are rare. We present a model where the observed substitution patterns are used to estimate the rate of deamination of cytosine residues in single- and double-stranded portions of the DNA, the length of single-stranded ends, and the frequency of nicks. The results suggest that reliable genome sequences can be obtained from Pleistocene organisms.

454 | deamination | depurination | paleogenomics

The retrieval of DNA sequences from long-dead organisms offers a unique perspective on genetic history by making information from extinct organisms and past populations available. However, three main technical challenges affect such studies. First, when DNA is preserved in ancient specimens, it is invariably degraded to a small average size (1). Second, chemical damage is present in ancient DNA (2) that may cause incorrect DNA sequences to be determined (3). Third, because ancient DNA is present in low amounts or absent in many specimens, traces of modern DNA from extraneous sources may cause modern DNA sequences to be mistaken for endogenous ancient DNA sequences (4–6). Recently, a DNA sequencing method based on highly parallel pyrosequencing of DNA templates generated by the PCR has been developed by 454 Life Sciences (454) (7). This method allows several hundred thousand DNA sequences of length 100 or 250 nt to be determined in a short time. It has been used to determine DNA sequences from the remains of three Pleistocene species: mammoths (8, 9), a cave bear (9), and a Neandertal (10). In all cases, the majority of DNA sequences retrieved are from microorganisms that have colonized the tissues after the death of the organisms. However, a fraction stem from the ancient organisms. In fact, the throughput of this technology, as well as other sequencing technologies currently becoming available (11), makes it possible to contemplate sequencing the complete genomes of extinct Pleistocene species (8, 10).

Here, we analyze DNA sequences determined on the 454 platform from an ≈38,000-year-old Neandertal specimen found at Vindija Cave, Croatia (10, 12), with respect to two features of particular significance for genomic studies of ancient DNA. First, we investigate the DNA sequence context around strand breaks in ancient DNA. This has not been previously possible, because when PCR is used to retrieve ancient DNA sequences, primers that target particular DNA sequences are generally used and thus the ends of the ancient DNA molecules are not revealed. Second, we investigate the patterns of nucleotide misincorporations in the ancient DNA sequences as a function of their position in ancient DNA fragments. Although there is strong evidence that the majority of such misincorporations are due to deamination of cytosine residues to uracil residues (3), which code as thymine residues, it is unclear whether other miscoding lesions are present in any appreciable frequency in ancient DNA or how miscoding lesions are distributed along ancient DNA molecules. When relevant, we use comparable data from an ≈43,000-year-old mammoth bone (9) from the Bol'shaya Kolopatkaya river, Russia, an ≈42,000-year-old cave bear bone from Ochsenhalt Cave, Austria (13), a contemporary human, and DNA sequences of the Vindija Neandertal cloned in a plasmid vector (14) to ask whether the patterns seen are general features of Pleistocene DNA sequences or are caused by the 454 sequencing process. Finally, we develop a model that allows us to estimate features of ancient DNA preservation and discuss the implications of our findings for the determination of complete genome sequences from Pleistocene organisms.

Results and Discussion

The 454 Process. Because aspects of the 454 sequencing process are of crucial importance for the analyses presented, we briefly review some of its essential features. In a first step, a double-stranded DNA extract is end-repaired and ligated to two different synthetic oligonucleotide adaptors termed A and B. From each successfully ligated molecule, one of the DNA strands is isolated and subjected to emulsion PCR, during which each template remains isolated from other templates on a Sepharose bead carrying oligonucleotides complementary to one of the adaptors, producing beads each coated with ≈10 million copies of one DNA molecule. Up to 800,000 such DNA-containing beads are then loaded onto a multiwell glass plate, and their sequences are determined by pyrosequencing (7).

The end repair of the template DNA and ligation of adapters, which are critical for the analyses in this paper, are described in more detail in Fig. 1. First, *T4* DNA polymerase is used to remove single-stranded 3′-overhanging ends and to fill in 5′-overhanging ends (Fig. 1*ii*). Simultaneously, 5′-ends are phos-

Author contributions: A.W.B., R.E.G., and S.P. designed research; J. Kelso, K.P., J. Krause, and M.T.R. contributed new reagents/analytic tools; A.W.B., U.S., P.L.F.J., R.E.G., M.M., M.L., and S.P. analyzed data; and A.W.B., P.L.F.J., R.E.G., and S.P. wrote the paper.

The authors declare no conflict of interest.

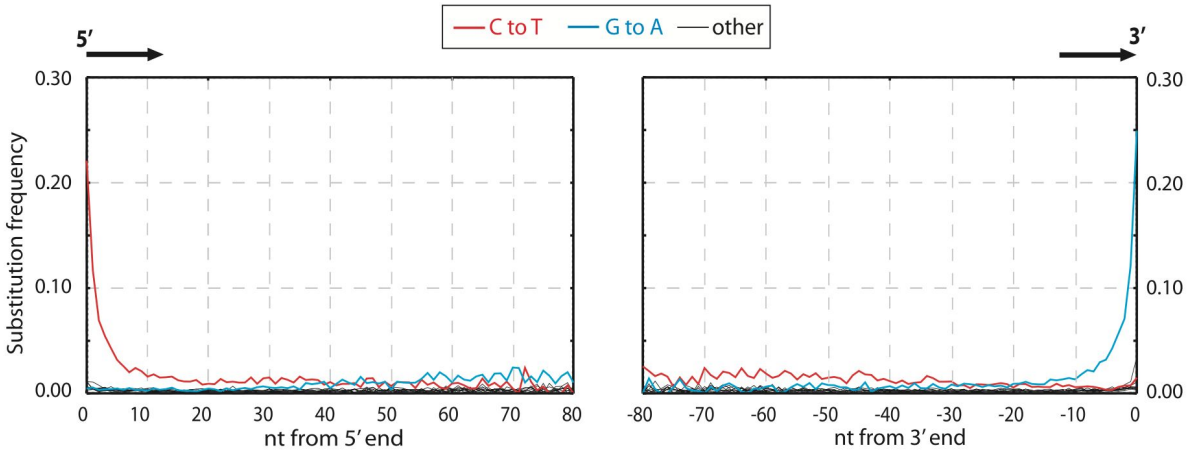
Abbreviations: 454, 454 Life Sciences; mtDNA, mitochondrial DNA; C.I., confidence interval.

Data deposition: The sequences reported in this paper have been deposited as follows. Directly sequenced Neandertal and mammoth sequences have been deposited in the European Molecular Biology Laboratory database (Neandertal accession nos. CAAM02000001–CAAM02470991, mammoth accession nos. CAAM02000001–CAAM02064265) and in the National Center for Biotechnology Information trace archive under GenomeProject IDs 18313 (Neandertal) and 17621 (mammoth). Cave bear and contemporary human sequences have been deposited in the National Center for Biotechnology Information trace archive under GenomeProject IDs 19671 (cave bear) and 19675 (human).

†To whom correspondence should be addressed. E-mail: briggs@eva.mpg.de or paabo@eva.mpg.de.

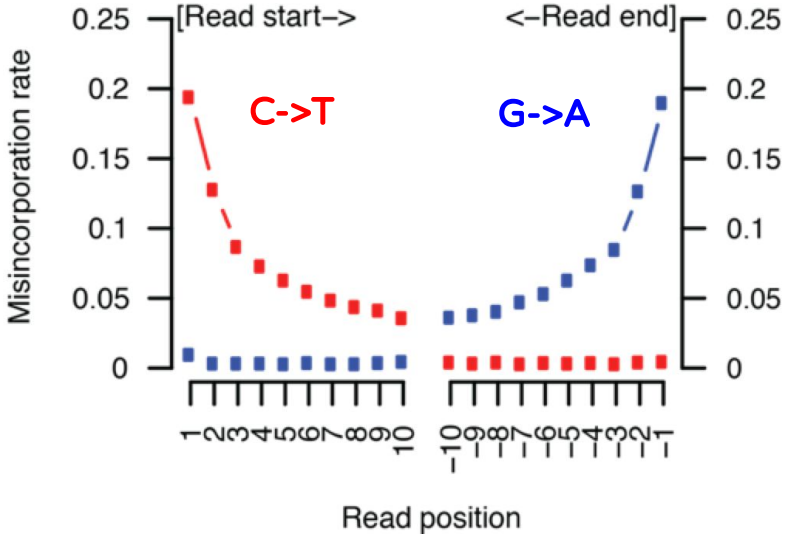
This article contains supporting information online at www.pnas.org/cgi/content/full/0704665104/DC1.

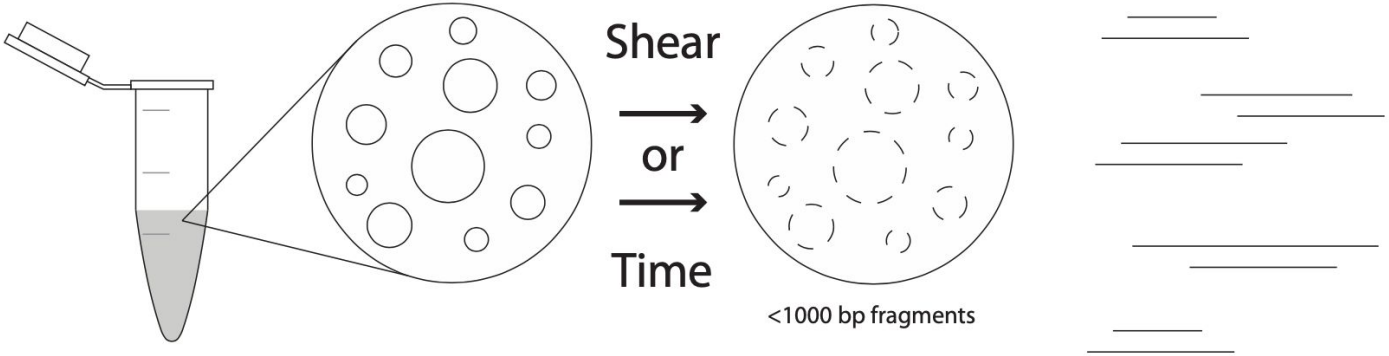
nicknamed "smile plot"



Why a “smile” plot?

Randomness of nicking (causes overhangs)

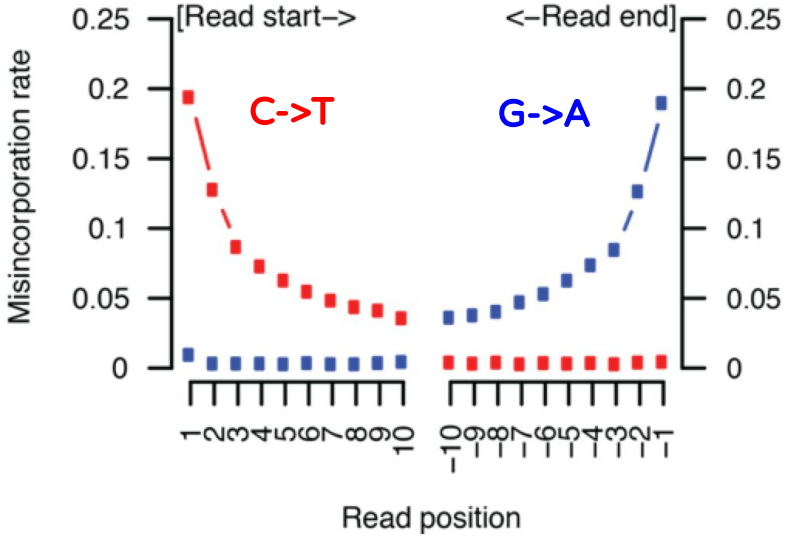


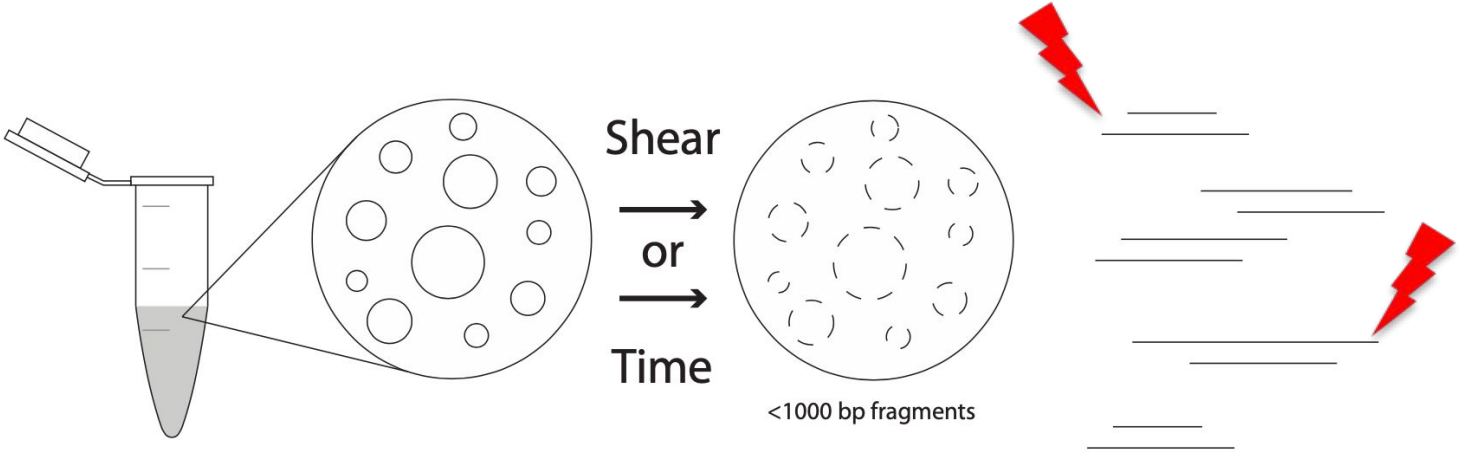


Why a “smile” plot?

Randomness of nicking (causes overhangs)

Cytosine deaminates 1000x faster when on overhang



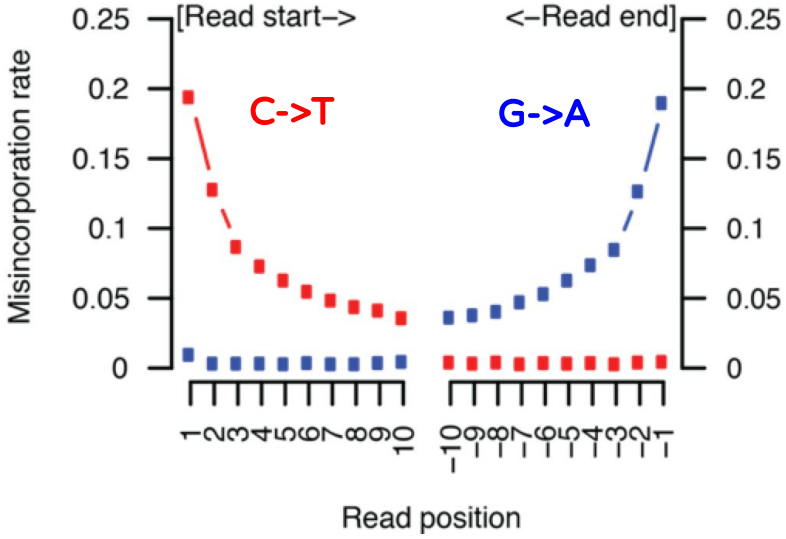


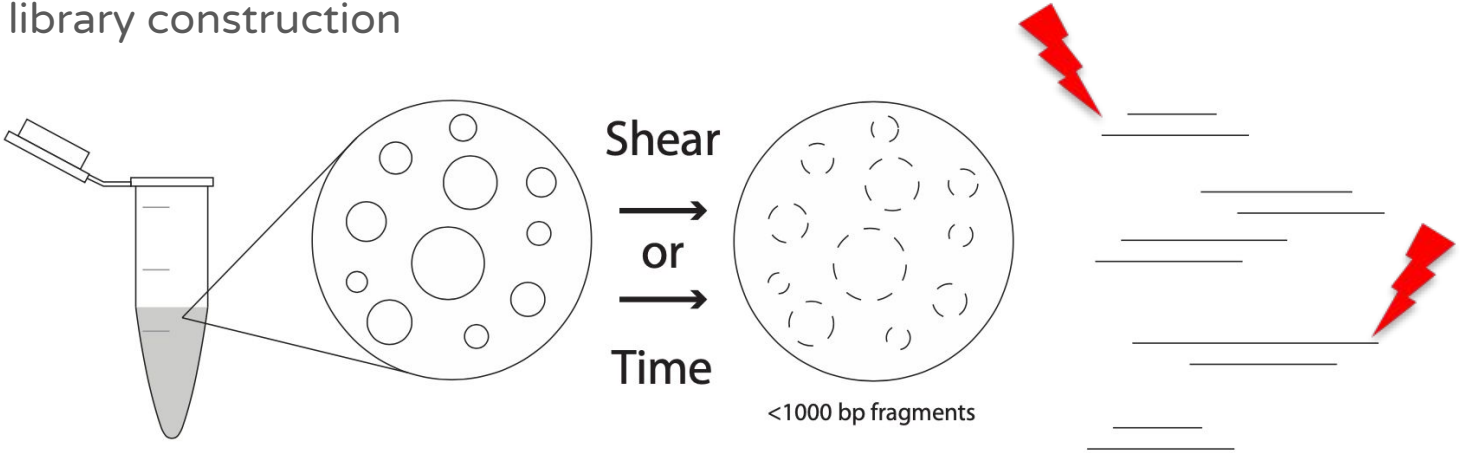
Why a “smile” plot?

Randomness of nicking (causes overhangs)

Cytosine deaminates 1000x faster when on overhang

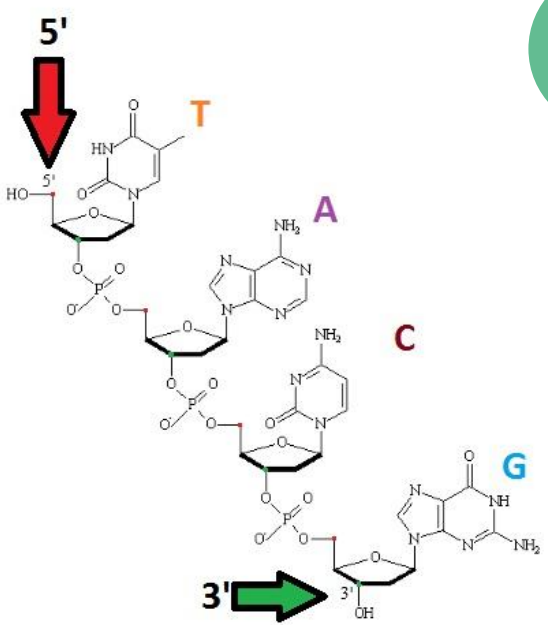
Asymmetric behavior of repair enzymes during blunt end library construction



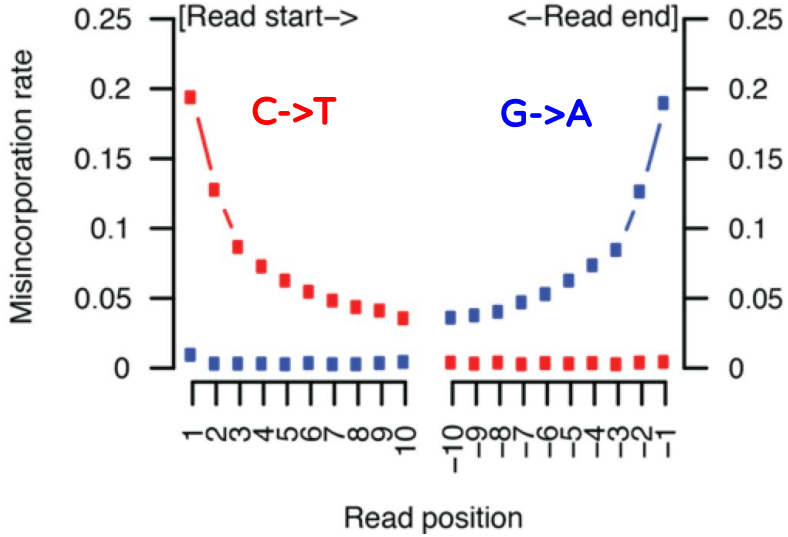


Why a “smile” plot?

DNA has a 5’ -> 3’ orientation:

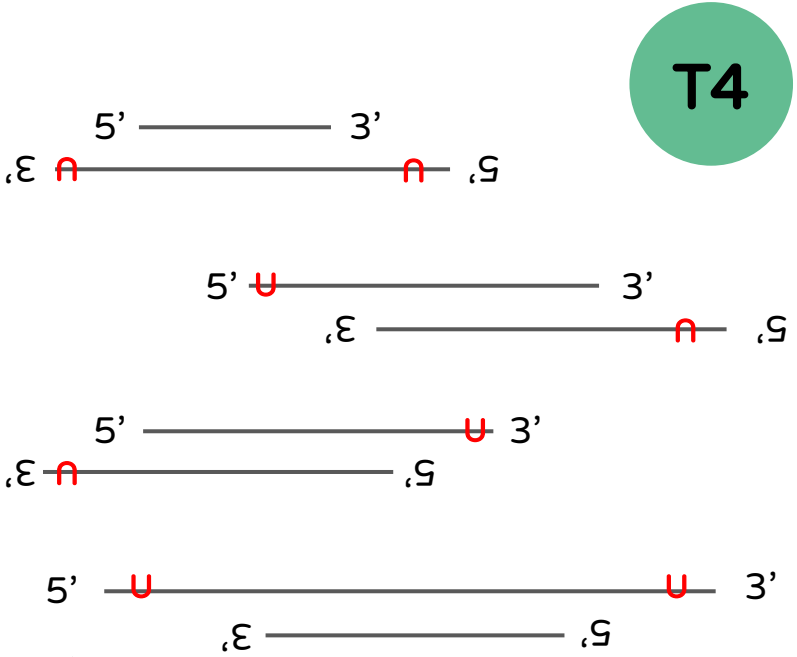


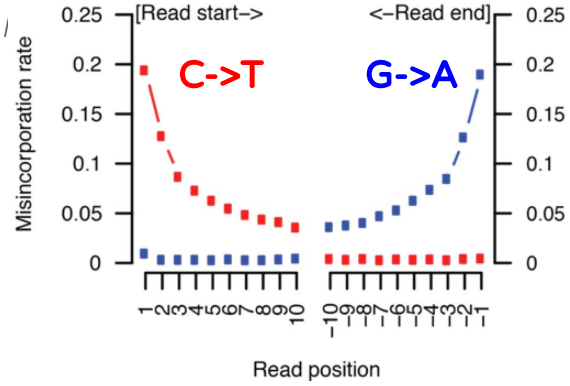
T4



Why a “smile” plot?

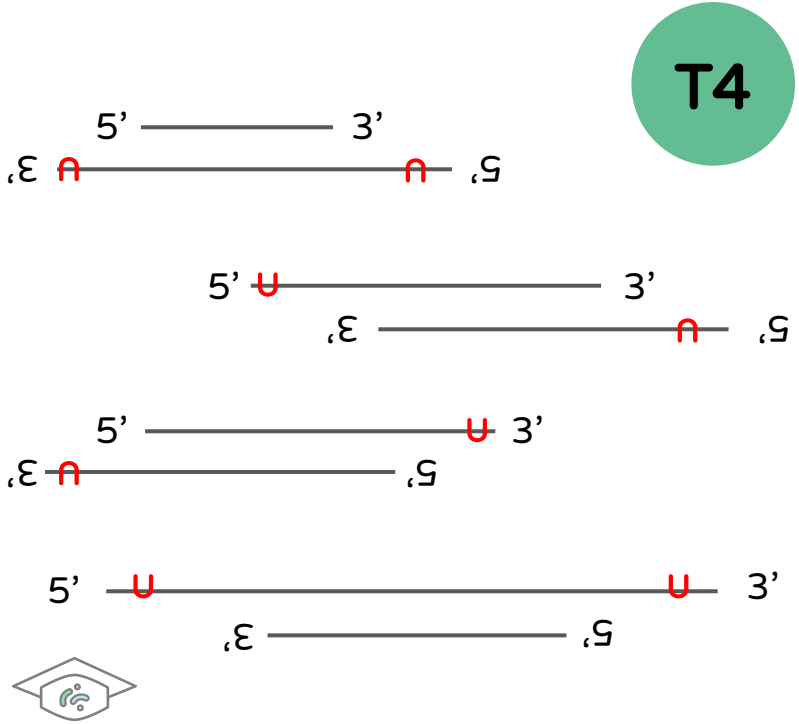
DNA has a 5’ -> 3’ orientation:





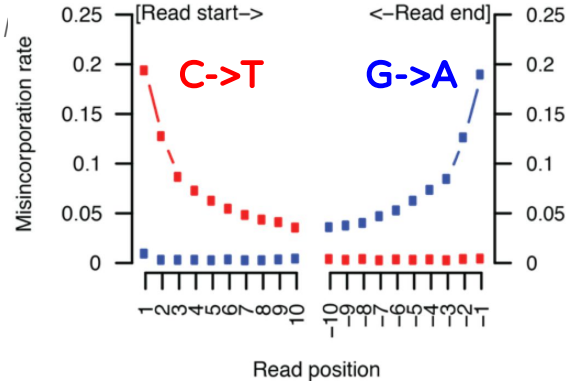
Why a “smile” plot?

DNA has a 5’ -> 3’ orientation:

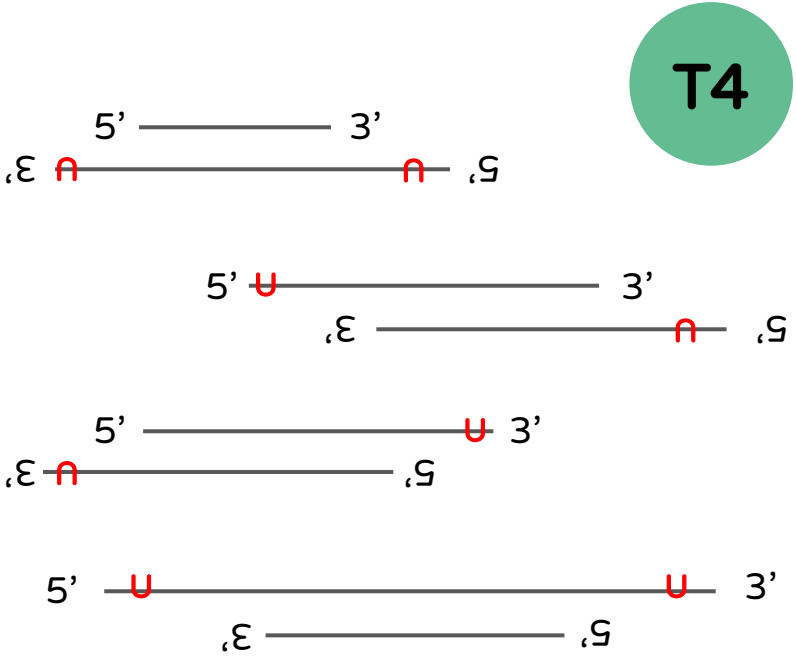


First step of NGS library construction is DNA repair to make strands fully double stranded with blunt ends

Why a “smile” plot?



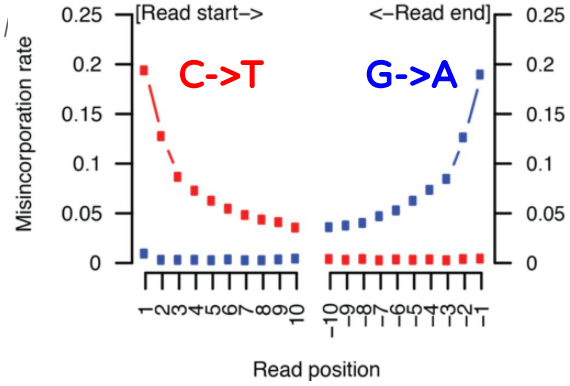
DNA has a 5’ -> 3’ orientation:



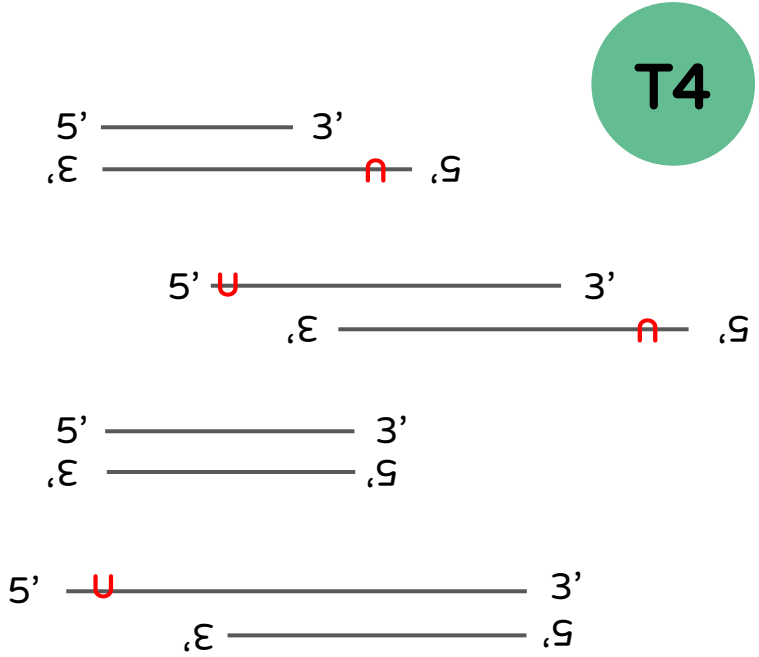
First step of NGS library construction is DNA repair to make strants fully double stranded with blunt ends

T4 polymerase cuts off **3’ overhangs** and fills in **5’ overhangs**

Why a “smile” plot?



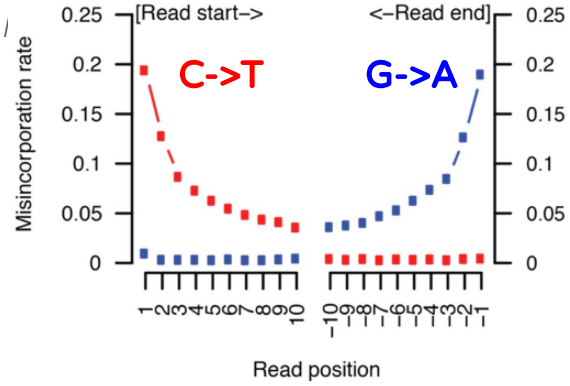
DNA has a 5’ -> 3’ orientation:



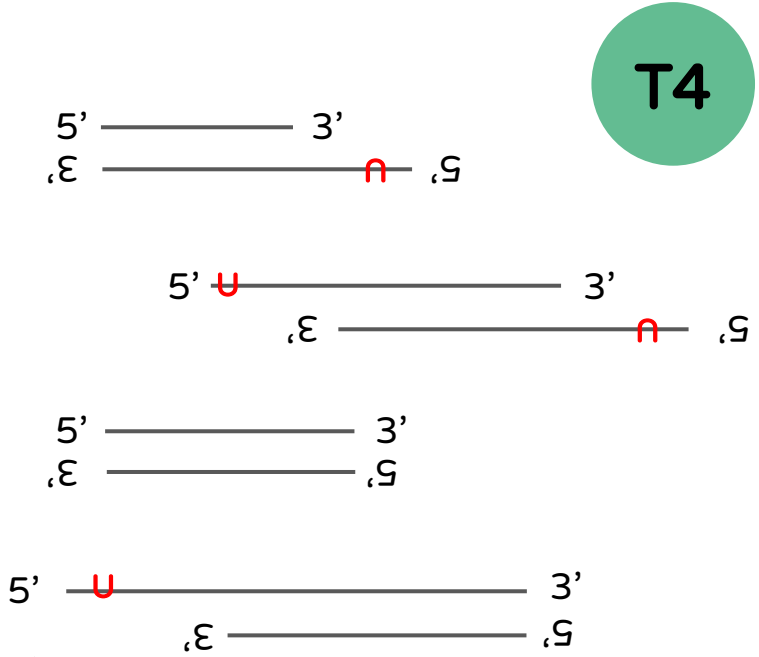
First step of NGS library construction is DNA repair to make strants fully double stranded with blunt ends

T4 polymerase cuts off **3’ overhangs** and fills in **5’ overhangs**

Why a “smile” plot?



DNA has a 5’ -> 3’ orientation:

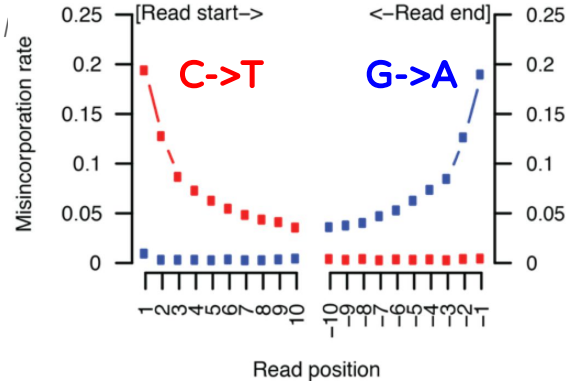


First step of NGS library construction is DNA repair to make strants fully double stranded with blunt ends

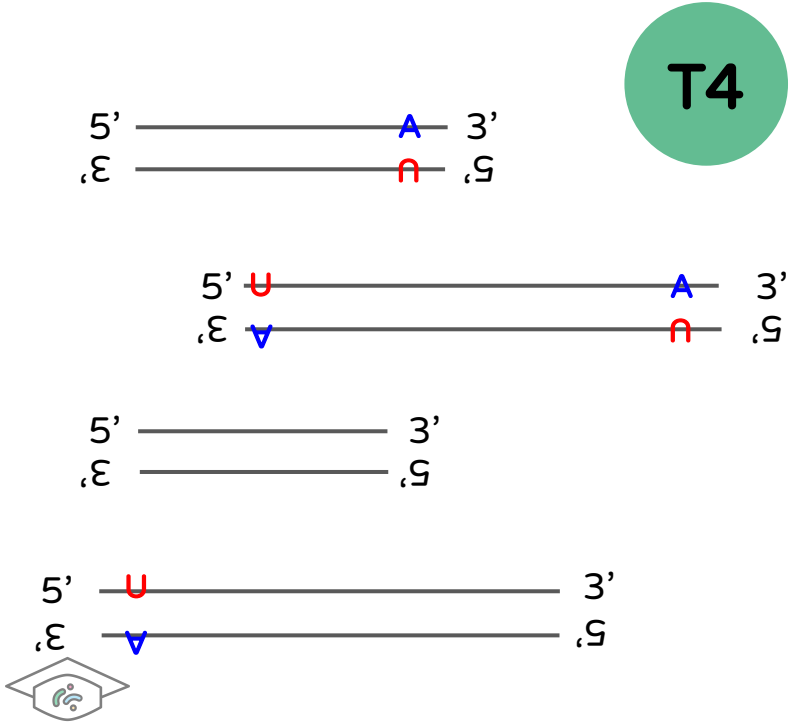
T4 polymerase cuts off **3’ overhangs** and fills in **5’ overhangs**

Then T4 polymerase fills in the **5’ overhangs**

Why a “smile” plot?



DNA has a 5’ -> 3’ orientation:



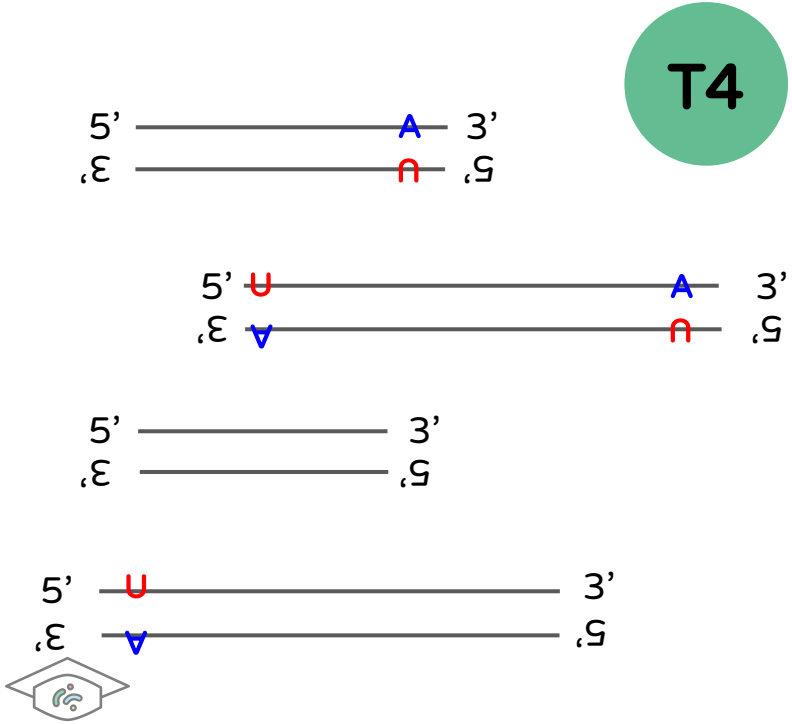
First step of NGS library construction is DNA repair to make strants fully double stranded with blunt ends

T4 polymerase cuts off **3’ overhangs** and fills in **5’ overhangs**

Then T4 polymerase fills in the **5’ overhangs**

Why a “smile” plot?

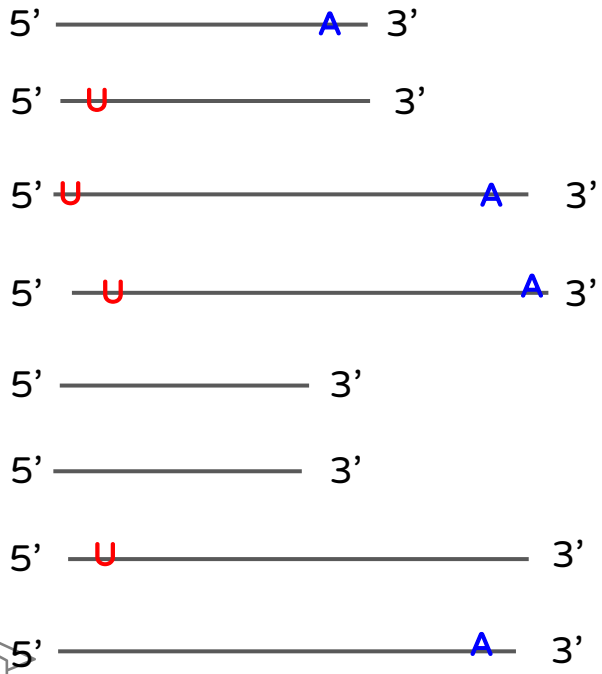
DNA has a 5’ -> 3’ orientation:



And later when the strands are melted and reoriented 5’ to 3’ for sequencing…

Why a “smile” plot?

DNA has a 5’ -> 3’ orientation:

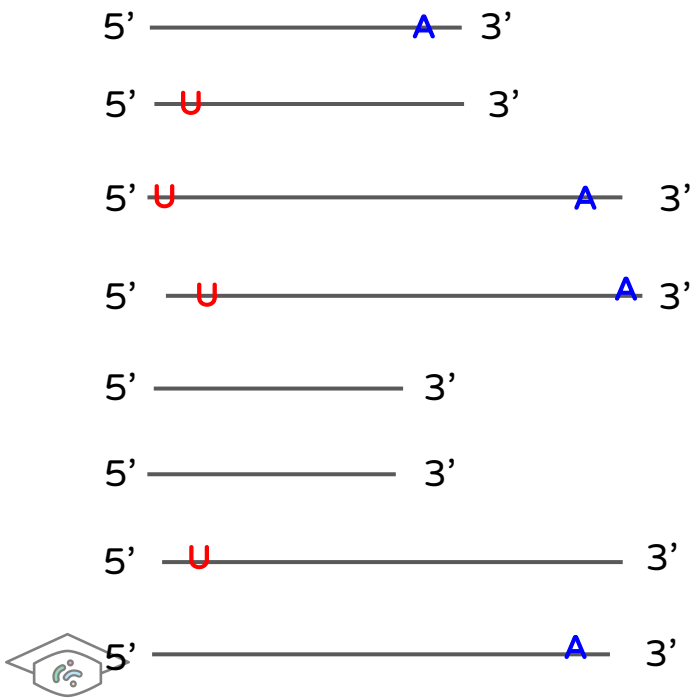


And later when the strands are melted and reoriented 5’ to 3’ for sequencing…

All the T miscoding lesions are on the 5’ end, and all the complementary As are on the 3’ end.

Why a “smile” plot?

DNA has a 5’ -> 3’ orientation:



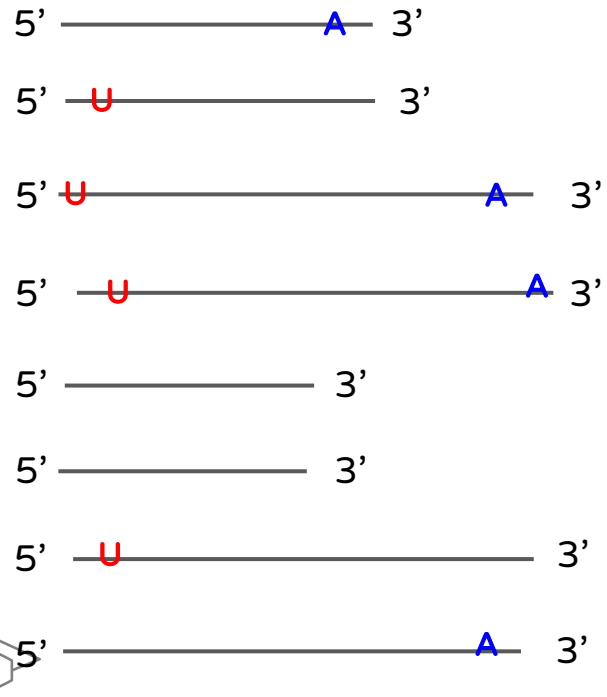
And later when the strands are melted and reoriented 5’ to 3’ for sequencing…

All the T miscoding lesions are on the 5’ end, and all the complementary As are on the 3’ end.

The only damage is C->T, but because of the T4 polymerase, you only “see” the 5’ Ts in the data, and the As are just the complement.

Why a “smile” plot?

DNA has a 5’ -> 3’ orientation:



Fun fact:

Because damage typically only occurs on single-stranded overhangs, the misincorporation rate can never reach 1, and the maximum rate under normal circumstances is 0.5.

DNA damage as authentication tool

mapDamage (2011) & mapDamage 2.0 (2013)

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 13 2013, pages 1682–1684
doi:10.1093/bioinformatics/btt193

Sequence analysis Advance Access publication April 23, 2013

mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters

Hákon Jónsson[1,*], Aurélien Ginolhac[1], Mikkel Schubert[1], Philip L. F. Johnson[2] and Ludovic Orlando[1]

[1]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350 København K, Denmark and [2]Department of Biology, Emory University, Atlanta, GA 30322, USA

Associate Editor: Michael Brudno

PMD tools (2014)

PNAS

Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal

Pontus Skoglund[a,1], Bernd H. Northoff[b,2], Michael V. Shunkov[c], Anatoli P. Derevianko[c], Svante Pääbo[b], Johannes Krause[b,d], and Mattias Jakobsson[a,e]

[a]Department of Evolutionary Biology and [e]Science for Life Laboratory, Uppsala University, 75236 Uppsala, Sweden; [b]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany; [c]Palaeolithic Department, Institute of Archaeology and Ethnography, Russian Academy of Sciences Siberian Branch, Novosibirsk 630090, Russia; and [d]Institute for Archaeological Sciences, University of Tuebingen, 72070 Tuebingen, Germany

Edited by Richard G. Klein, Stanford University, Stanford, CA, and approved December 27, 2013 (received for review October 9, 2013)

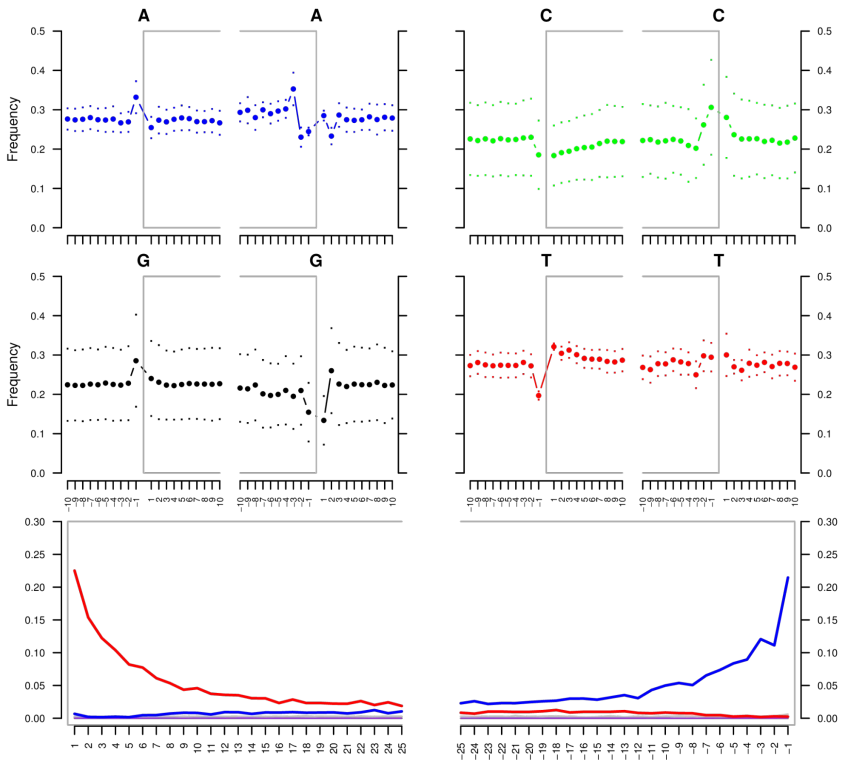
DamageProfiler (2021)

Sequence analysis

DamageProfiler: fast damage pattern calculation for ancient DNA

Judith Neukamm [1,2,3,*], Alexander Peltzer[2,4] and Kay Nieselt[2,*]

[1]Institute of Evolutionary Medicine, University of Zurich, 8057 Zurich, Switzerland, [2]Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany, [3]Institute for Archaeological Sciences, University of Tübingen, 72070 Tübingen, Germany and [4]Max Planck Institute for the Science of Human History, 07745 Jena, Germany



DNA damage as authentication tool

mapDamage (2011) & mapDamage 2.0 (2013)

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 13 2013, pages 1682–1684
doi:10.1093/bioinformatics/btt193

Sequence analysis Advance Access publication April 23, 2013

mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters

Hákon Jónsson[1,*], Aurélien Ginolhac[1], Mikkel Schubert[1], Philip L. F. Johnson[2] and Ludovic Orlando[1]

[1]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350 København K, Denmark and [2]Department of Biology, Emory University, Atlanta, GA 30322, USA

Associate Editor: Michael Brudno

PMD tools (2014)

PNAS

Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal

Pontus Skoglund[a,1], Bernd H. Northoff[b,2], Michael V. Shunkov[c], Anatoli P. Derevianko[c], Svante Pääbo[b], Johannes Krause[b,d], and Mattias Jakobsson[a,e]

[a]Department of Evolutionary Biology and [e]Science for Life Laboratory, Uppsala University, 75236 Uppsala, Sweden; [b]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany; [c]Palaeolithic Department, Institute of Archaeology and Ethnography, Russian Academy of Sciences Siberian Branch, Novosibirsk 630090, Russia; and [d]Institute for Archaeological Sciences, University of Tuebingen, 72070 Tuebingen, Germany

Edited by Richard G. Klein, Stanford University, Stanford, CA, and approved December 27, 2013 (received for review October 9, 2013)

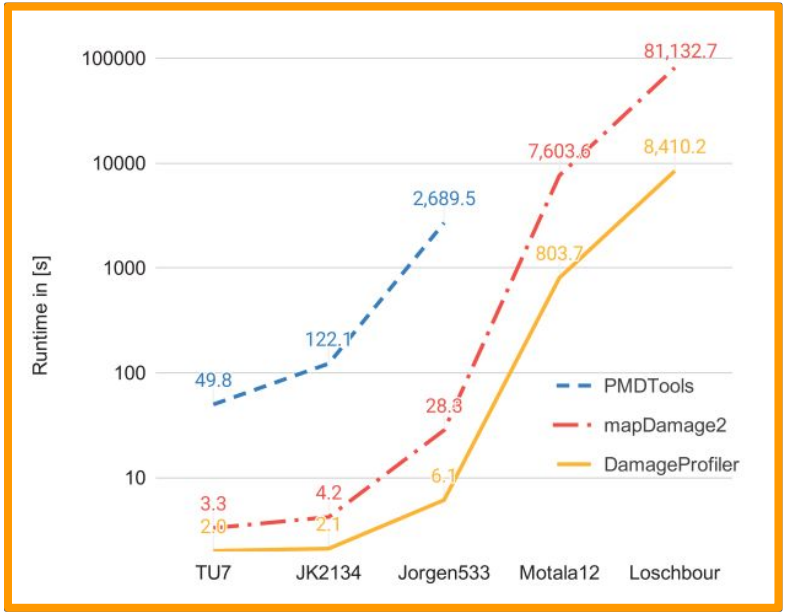
DamageProfiler (2021)

Sequence analysis

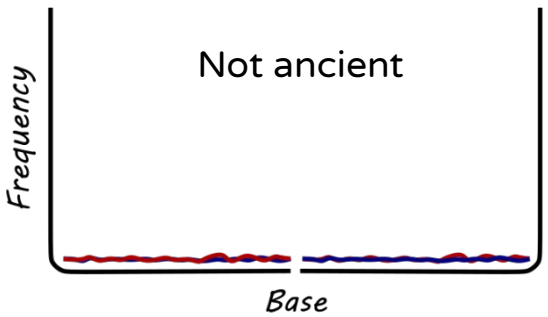
DamageProfiler: fast damage pattern calculation for ancient DNA

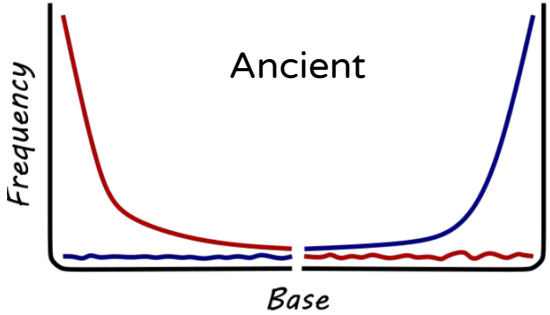
Judith Neukamm [1,2,3,*], Alexander Peltzer[2,4] and Kay Nieselt[2,*]

[1]Institute of Evolutionary Medicine, University of Zurich, 8057 Zurich, Switzerland, [2]Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany, [3]Institute for Archaeological Sciences, University of Tübingen, 72070 Tübingen, Germany and [4]Max Planck Institute for the Science of Human History, 07745 Jena, Germany

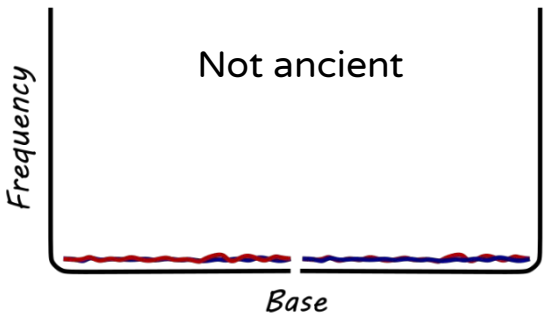


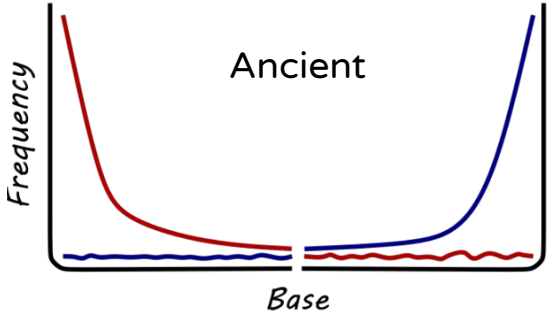
DNA damage as authentication tool

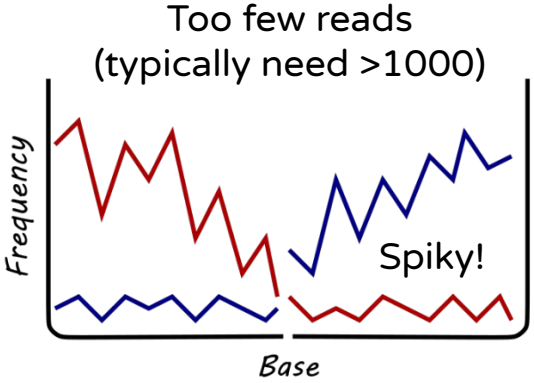




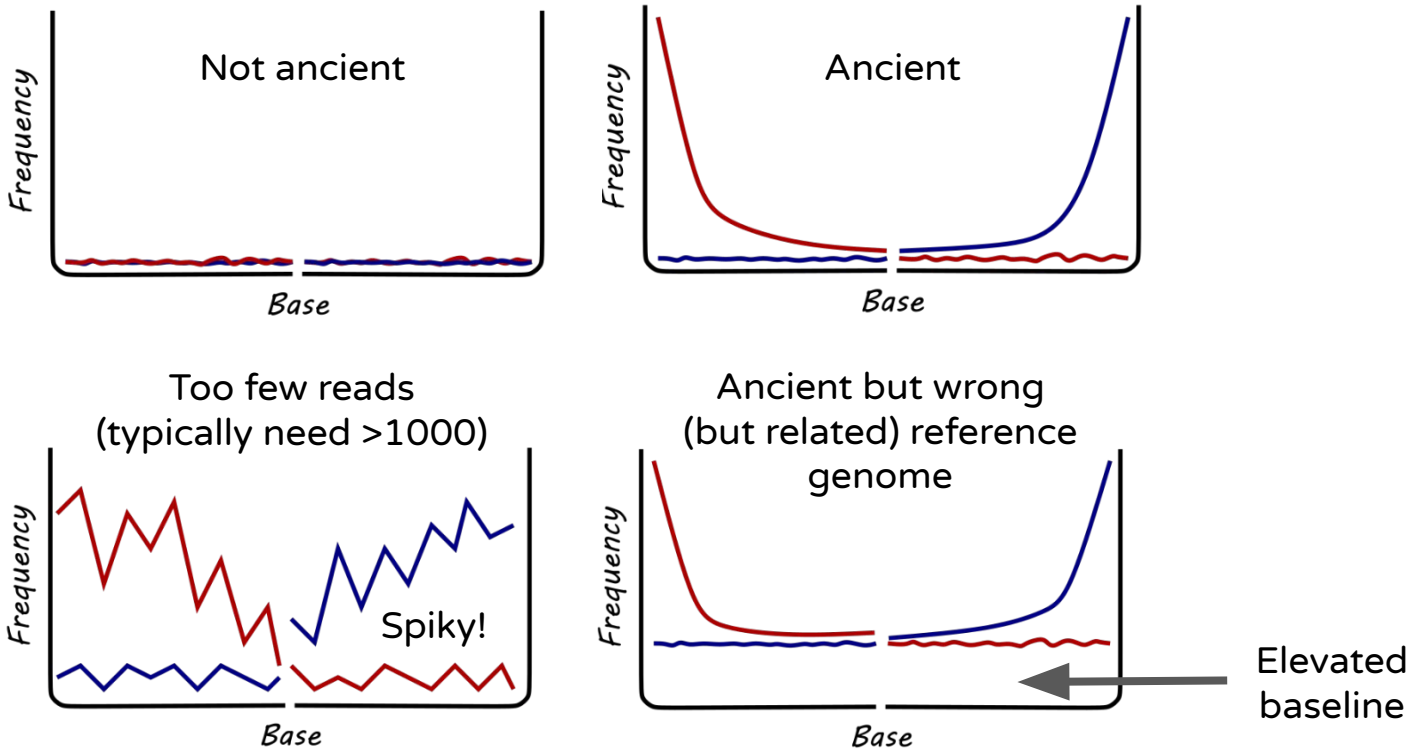
DNA damage as authentication tool







DNA damage as authentication tool



DNA damage as a clock?

DNA damage as a clock?

…sort of, but not really

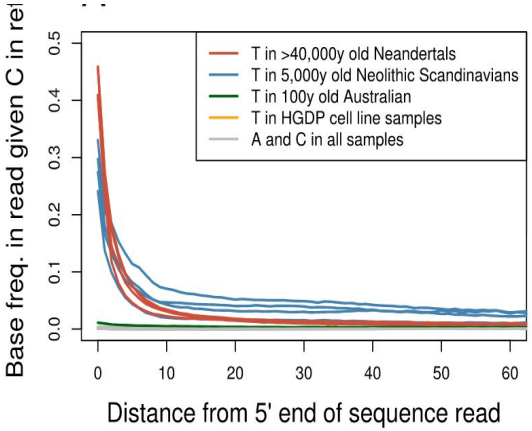
More like a clock that only says “today” or “a while ago”

DNA damage as a clock?

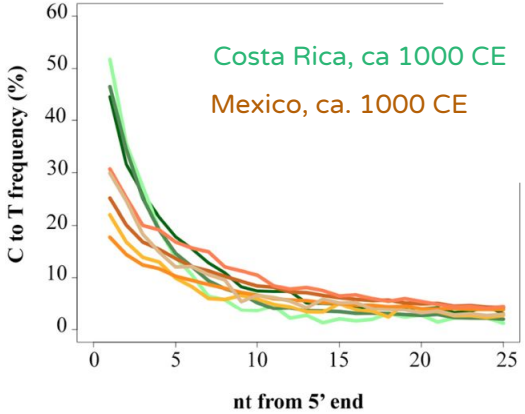
…sort of, but not really

More like a clock that only says “today” or “a while ago”

Skoglund et al. 2014



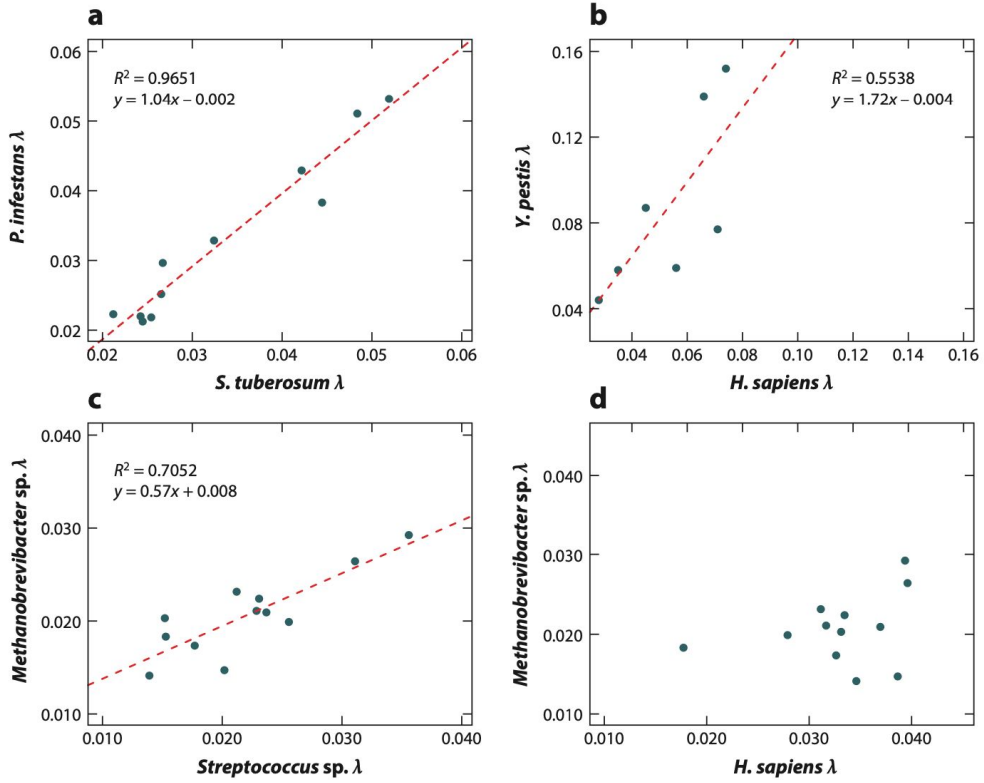
Morales-Arce et al. 2017



Relationship to time not linear

DNA damage highly dependent on local temperature and humidity

DNA damage as a clock?



Warinner et al. 2017

And varies by organism - even within the same sample

DNA damage is a relative indicator

Removing damage - UDG

Damage is useful for authentication, but sometimes you don't want it - especially for sensitive genotyping and tree building analyses when base calling accuracy is important.

You can remove damaged cytosines with the enzyme cocktail **USER**, which contains **uracil–DNA–glycosylase (UDG)** and **endonuclease VIII** (Briggs et al. 2009)

Enzyme	Effect
PNK	5' U U 3'
UDG	AP AP
endo VIII	AP AP
PNK/T4 pol	
T4 ligase Bst pol	A B

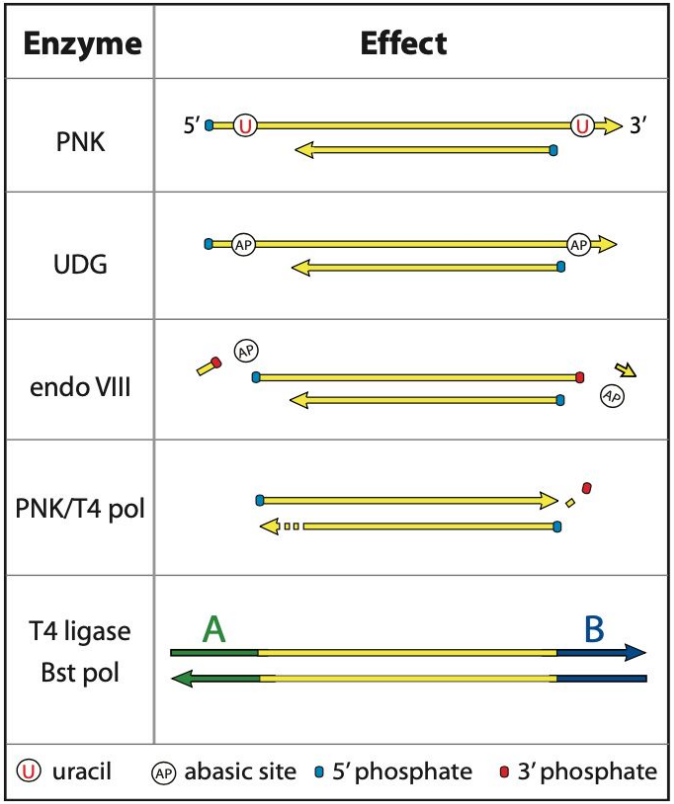
U uracil AP abasic site 5' phosphate 3' phosphate

Removing damage - UDG

Damage is useful for authentication, but sometimes you don't want it - especially for sensitive genotyping and tree building analyses when base calling accuracy is important.

You can remove damaged cytosines with the enzyme cocktail **USER**, which contains **uracil–DNA–glycosylase (UDG)** and **endonuclease VIII** (Briggs et al. 2009)

UDG clips out the uracil base, leaving an abasic site (X)

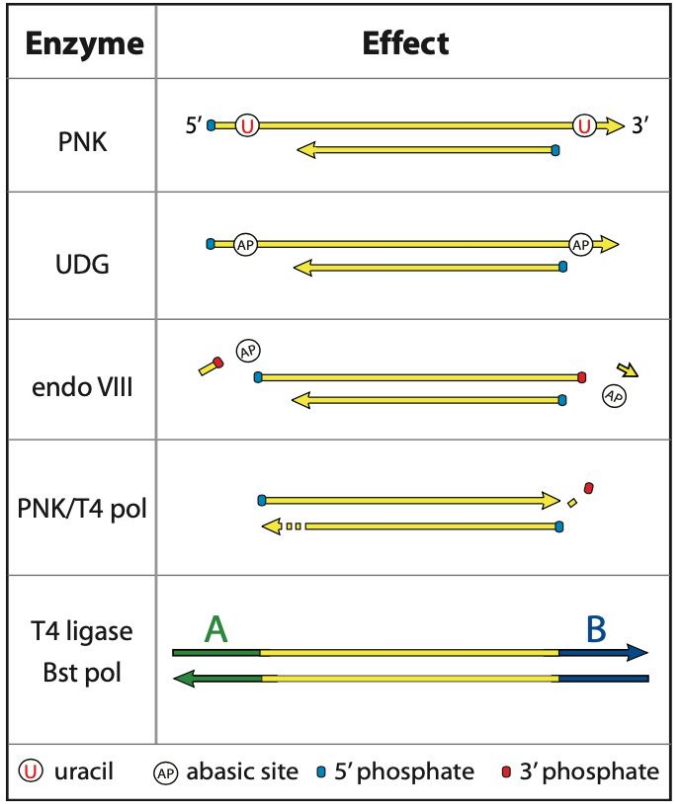


Removing damage - UDG

Damage is useful for authentication, but sometimes you don't want it - especially for sensitive genotyping and tree building analyses when base calling accuracy is important.

You can remove damaged cytosines with the enzyme cocktail **USER**, which contains **uracil–DNA–glycosylase (UDG)** and **endonuclease VIII** (Briggs et al. 2009)

Endo VIII clips the DNA backbone at the abasic site, shortening the DNA

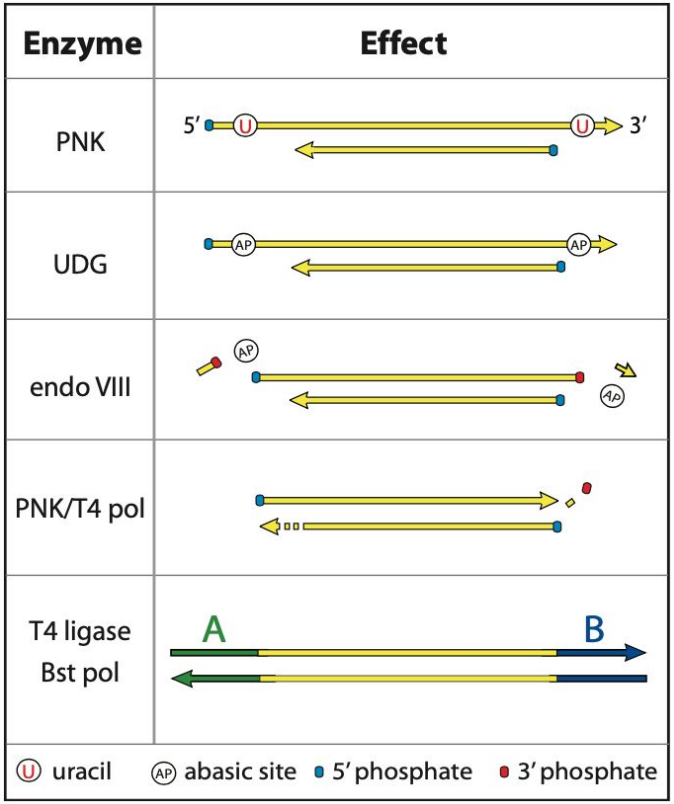


Removing damage - UDG

Damage is useful for authentication, but sometimes you don't want it - especially for sensitive genotyping and tree building analyses when base calling accuracy is important.

You can remove damaged cytosines with the enzyme cocktail **USER**, which contains **uracil–DNA–glycosylase (UDG)** and **endonuclease VIII** (Briggs et al. 2009)

T4 polymerase trims the 3' overhang

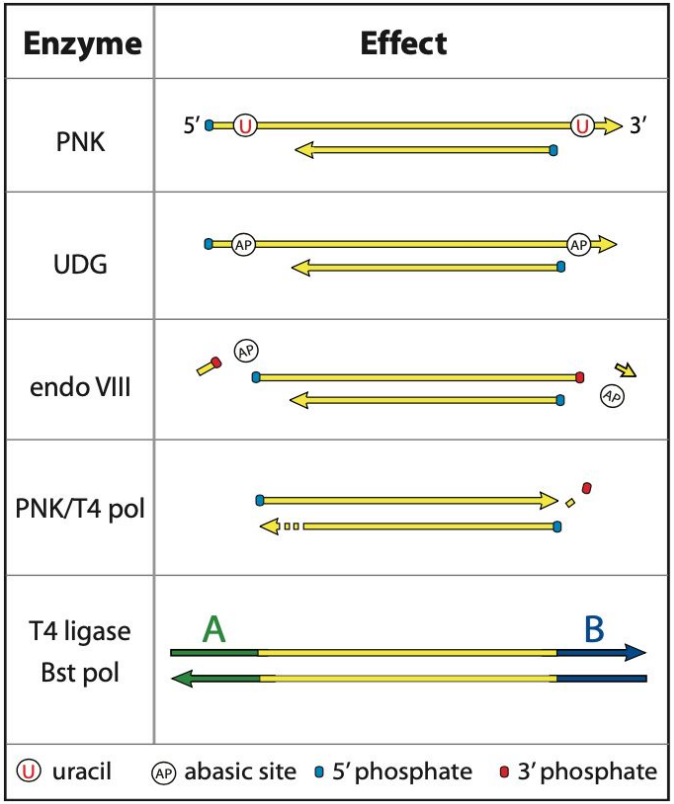


Removing damage - UDG

Damage is useful for authentication, but sometimes you don't want it - especially for sensitive genotyping and tree building analyses when base calling accuracy is important.

You can remove damaged cytosines with the enzyme cocktail **USER**, which contains **uracil–DNA–glycosylase (UDG)** and **endonuclease VIII** (Briggs et al. 2009)

T4 polymerase fills in the 5' overhang



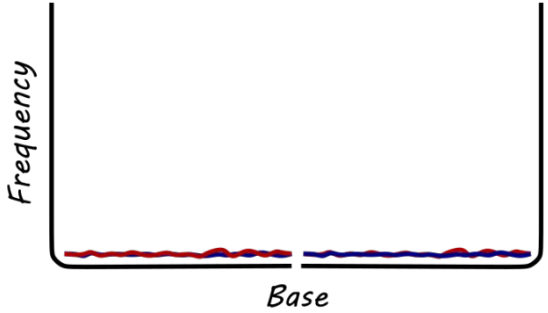
Removing damage - UDG

Damage is useful for authentication, but sometimes you don't want it - especially for sensitive genotyping and tree building analyses when base calling accuracy is important.

You can remove damaged cytosines with the enzyme cocktail **USER**, which contains **uracil–DNA–glycosylase (UDG)** and **endonuclease VIII** (Briggs et al. 2009)

Cytosine damage is now gone

DNA will have no damage and be a little bit shorter



Removing damage - UDG-half

Sometimes you don't want to remove all of the damage. Maybe you want to remove *almost all* of the damage (to improve sequence accuracy) but leave just one damaged base at the end (for authentication).

Can you have your cake and eat it too? Yes!

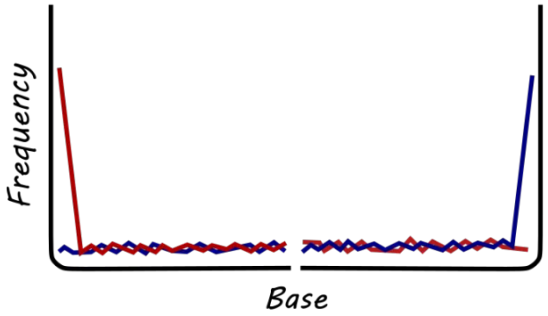
Removing damage - UDG-half

Sometimes you don't want to remove all of the damage. Maybe you want to remove *almost all* of the damage (to improve sequence accuracy) but leave just one damaged base at the end (for authentication).

Can you have your cake and eat it too? Yes!

You can remove all but the innermost damaged cytosines using a **partial UDG protocol**, also called UDG-half protocol (Rohland et al. 2015)

Damage will only be on the first base



Removing damage - UDG-half

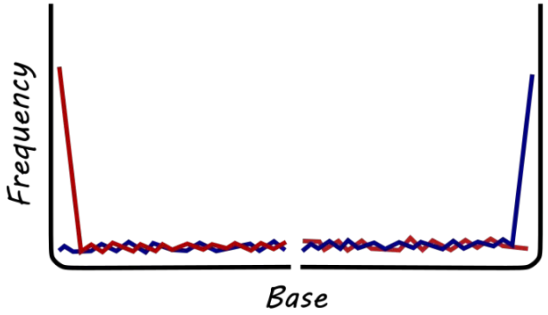
Sometimes you don't want to remove all of the damage. Maybe you want to remove *almost all* of the damage (to improve sequence accuracy) but leave just one damaged base at the end (for authentication).

Can you have your cake and eat it too? Yes!

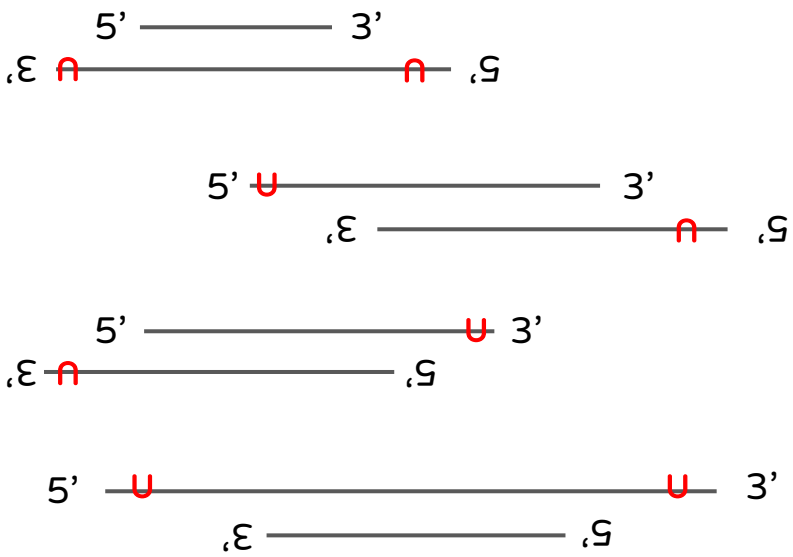
You can remove all but the innermost damaged cytosines using a **partial UDG protocol**, also called UDG-half protocol (Rohland et al. 2015)

Note: *the damage after partial UDG treatment is always lower than no treatment - can you think why?*

Damage will only be on the first base

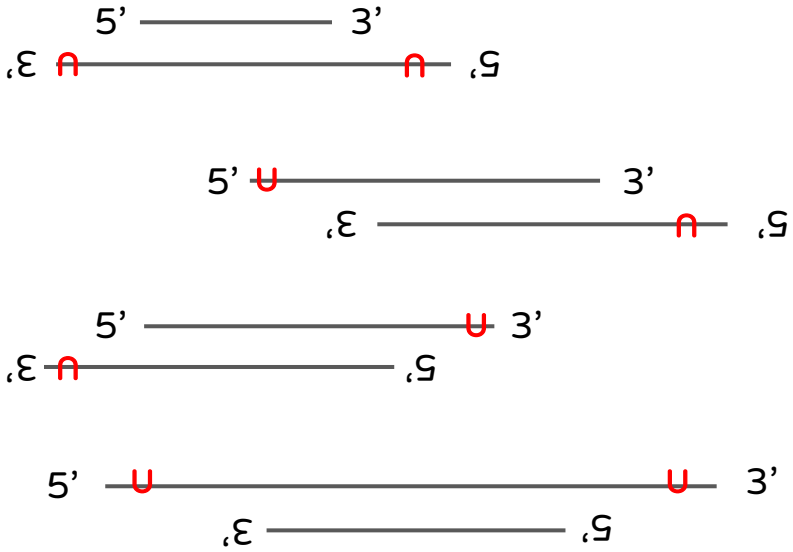


Single stranded libraries



Okay, everything we've talked about so far is valid for DNA sequence data generated from standard double stranded DNA libraries (Meyer and Kircher 2010)

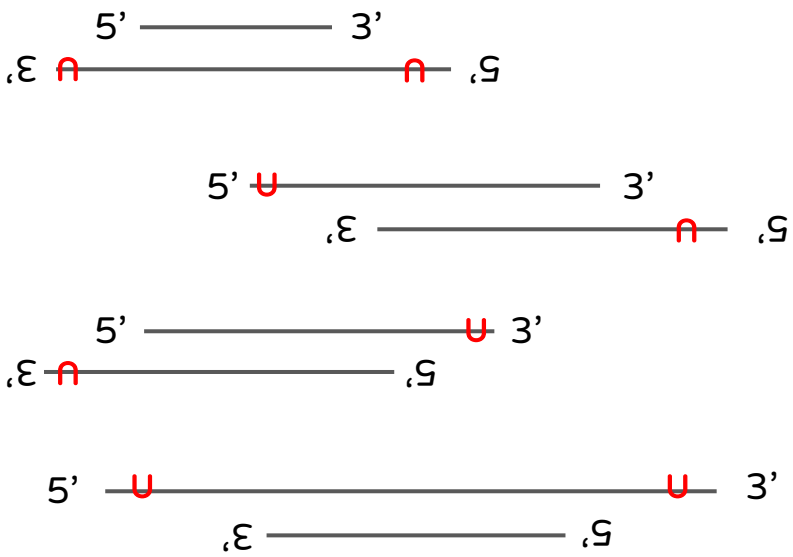
Single stranded libraries



Okay, everything we've talked about so far is valid for DNA sequence data generated from standard double stranded DNA libraries (Meyer and Kircher 2010)

But you can also make libraries using a single-stranded DNA library construction protocol (Gansauge and Meyer 2013, 2019)

Single stranded libraries

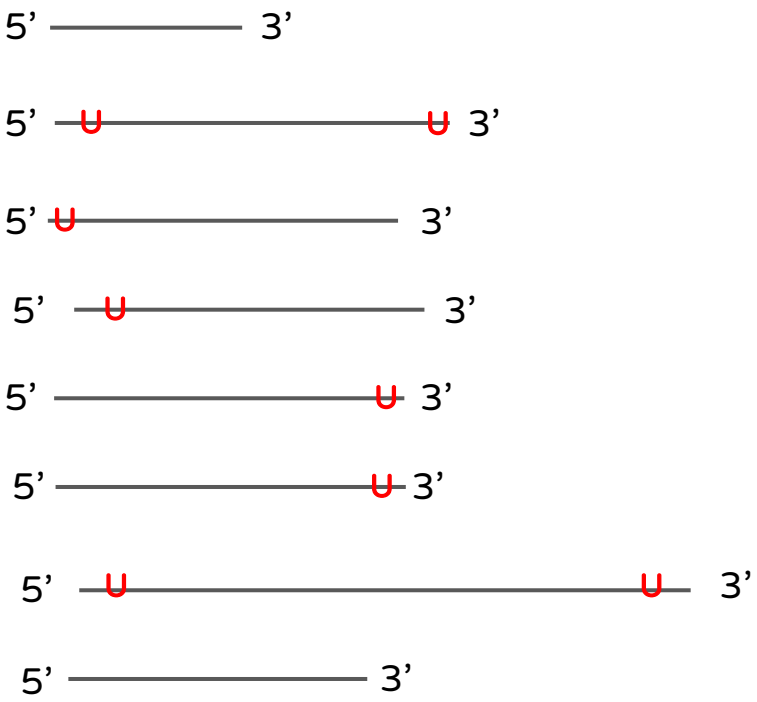


Okay, everything we've talked about so far is valid for DNA sequence data generated from standard double stranded DNA libraries (Meyer and Kircher 2010)

But you can also make libraries using a single-stranded DNA library construction protocol (Gansauge and Meyer 2013, 2019)

This protocol does not clip 3' overhangs so you keep all of your original damage

Single stranded libraries

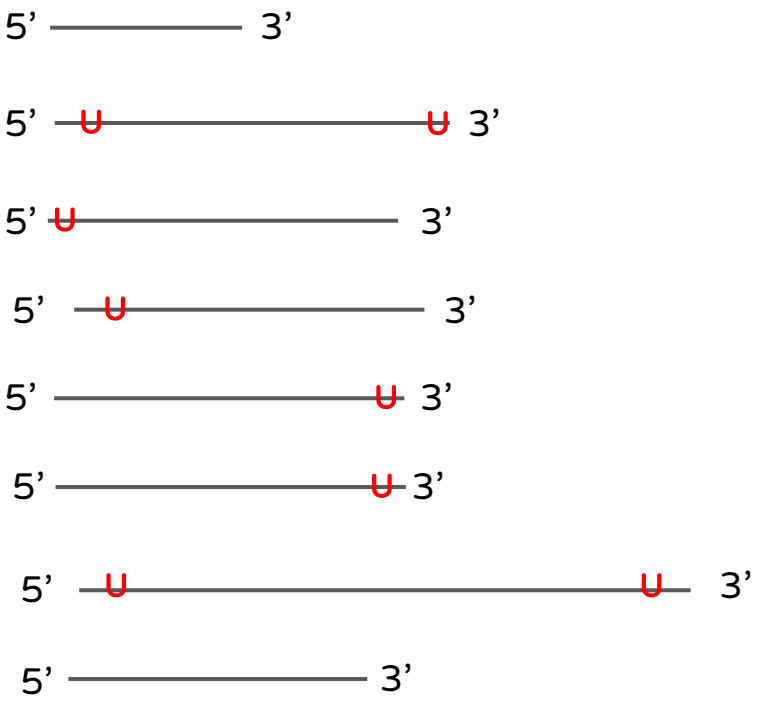


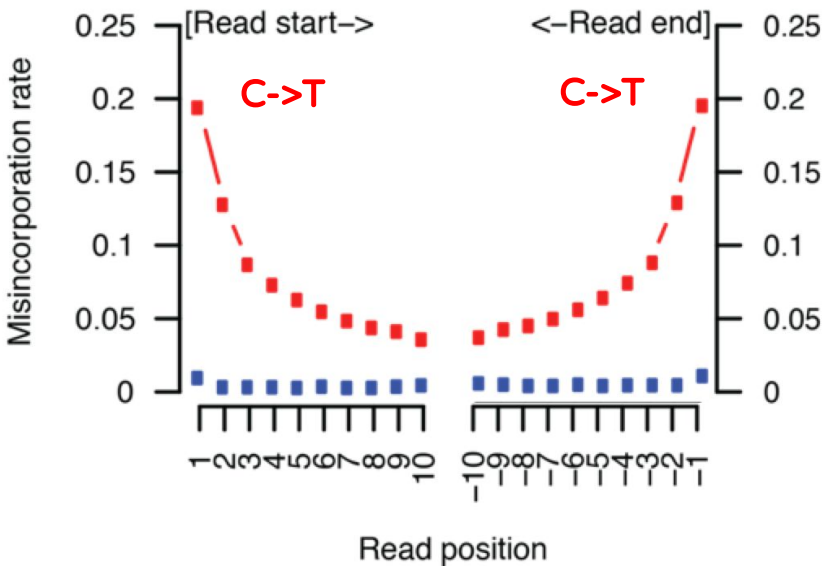
Okay, everything we’ve talked about so far is valid for DNA sequence data generated from standard double stranded DNA libraries (Meyer and Kircher 2010)

But you can also make libraries using a single-stranded DNA library construction protocol (Gansauge and Meyer 2013, 2019)

This protocol does not clip 3’ overhangs so you keep all of your original damage

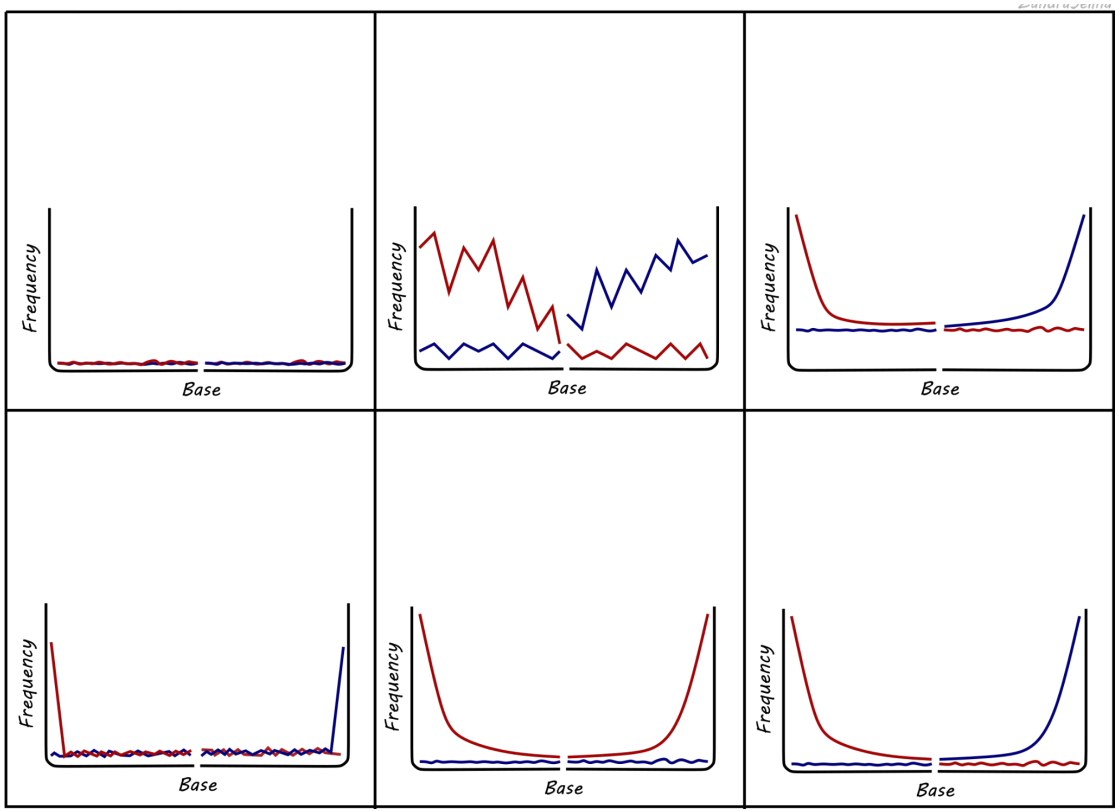
Single stranded libraries





As a result, smile plots are C->T on both sides

Damage wrap-up



Enzyme alert!

As you know, uracil (U) is not a normal component of DNA

So far, we've discussed how enzymes like T4 polymerase treats uracil (U) like a thymine (T), introducing C->T misincorporations

NOT ALL ENZYMES DO THIS

Enzyme alert!

As you know, uracil (U) is not a normal component of DNA

So far, we've discussed how enzymes like T4 polymerase treats uracil (U) like a thymine (T), introducing C->T misincorporations

NOT ALL ENZYMES DO THIS

Some enzymes just … 

… when they encounter a U.

Enzyme alert!

As you know, uracil (U) is not a normal component of DNA

So far, we've discussed how enzymes like T4 polymerase treats uracil (U) like a thymine (T), introducing C->T misincorporations

NOT ALL ENZYMES DO THIS

Some enzymes just … 

… when they encounter a U.

The damage present in ancient DNA (fragmentation and deamination) requires the use of specialized library protocols specifically for ancient DNA

Enzyme alert!

DNA polymerases come in two flavors:

- Non-proofreading - treat U like a T
- Proofreading - stop at U

For ancient DNA, it is **critical** to use a non-proofreading polymerase for library construction and the indexing PCR in order to lock in the damage by turning U into T

Later amplifications can use a proofreading polymerase

Note: *if you use a proofreading enzyme for library construction, your damaged aDNA molecules will not be sequenced, which may bias your dataset towards contamination. However, UDG-treated aDNA is compatible with proofreading enzymes because its DNA damage has already been removed.*

Enzyme alert!

Why use proofreading enzymes at all?

Proofreading enzymes are more accurate

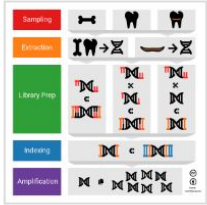
So we use proofreading enzymes for every step **except** the two key steps in which the polymerase encounters the original damaged cytosines (U):

- non-proofreading T4 polymerase for DNA repair
- non-proofreading polymerase (e.g., Pfu Turbo Cx) for library indexing amplification

Subsequent amplifications, reamplifications, and reconditioning steps are all performed using a proofreading enzyme (e.g., Herculase II)

Enzyme alert!

For more information about library protocols and enzymes, check out our online bench protocols:



Version 2

Jun 15, 2021

Bookmark

Copy / Fork

A-Z of ancient DNA protocols for shotgun Illumina Next Generation Sequencing V.2

James A Fellows Yates[1], Franziska Aron[2], Gunnar U Neumann[3], Irina Velsko[1], Eirini Skourtanioti[1], Eleftheria Orfanou[3], Zandra Fagernäs[3], Raphaela Stahl, Aida Andrades Valtuena[3], Christina Warinner[3], Wolfgang Haak[3], Guido Brandt[3]

[1]Max Planck Institute for Evolutionary Anthropology; [2]Friedrich-Schiller Universität Jena;
[3]Max Planck Institute for the Science of Human History

2 *Works for me* Share

dx.doi.org/10.17504/protocols.io.bvt9n6r6

WarinnerGroup MPI EVA Archaeogenetics

James Fellows Yates
Max Planck Institute for Evolutionary Anthropology

Big picture: Why does DNA damage matter?

Big picture: Why does DNA damage matter?

Allows DNA authentication of:

- Individual species (Jonsson et al. 2013)
- Metagenomic assemblies (Borry et al. 2021)
- Individual reads (Skoglund et al. 2014)

Big picture: Why does DNA damage matter?

Allows DNA authentication of:

- Individual species (Jonsson et al. 2013)
- Metagenomic assemblies (Borry et al. 2021)
- Individual reads (Skoglund et al. 2014)

Poses major challenges for:

- Taxonomic identification of sequences
- Accurate genome mapping
- Metagenomic assembly

Big picture: Why does DNA damage matter?

Allows DNA authentication of:

- Individual species (Jonsson et al. 2013)
- Metagenomic assemblies (Borry et al. 2021)
- Individual reads (Skoglund et al. 2014)

Poses major challenges for:

- Taxonomic identification of sequences
- Accurate genome mapping
- Metagenomic assembly

Turns out the biggest challenge is not C deamination, but fragment length

Big picture: Why does DNA damage matter?

Taxonomic identification of sequences

- DNA fragments <30 bp lack sufficient specificity for taxonomic assignment - they align to too many genomes with no phylogenetic coherence

Big picture: Why does DNA damage matter?

Taxonomic identification of sequences

- DNA fragments <30 bp lack sufficient specificity for taxonomic assignment - they align to too many genomes with no phylogenetic coherence
- 1-million-year limit of aDNA is not how long DNA survives, but how long DNA sequences >30 bp survive (van der Valk et al. 2022)

Article

Million-year-old DNA sheds light on the genomic history of mammoths

https://doi.org/10.1038/s41586-021-03224-9
Received: 3 July 2020
Accepted: 11 January 2021
Published online: 17 February 2021
Check for updates

Tom van der Valk[1,2,3,17], Patrici
Anders Bergström[5], Jonas Opp
Jessica A. Thomas[8], Marianne
Shanlin Liu[11], Mehmet Somel[9],
Pontus Skoglund[6], Michael Ho
Love Dalén[1,2,4,18]

Big picture: Why does DNA damage matter?

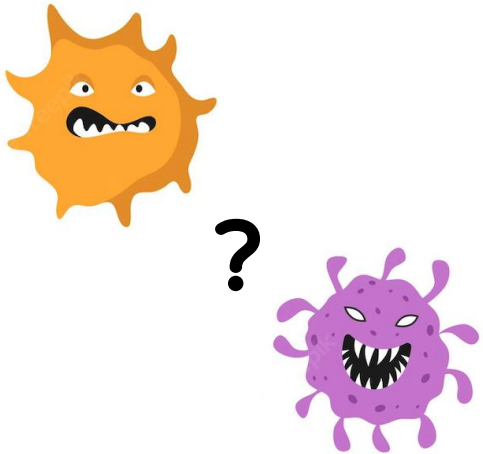
Accurate genome mapping

- DNA sequences <100 bp often lack taxonomic specificity within clades, leading to **cross-mapping** within groups of related microbial taxa

Big picture: Why does DNA damage matter?

Accurate genome mapping

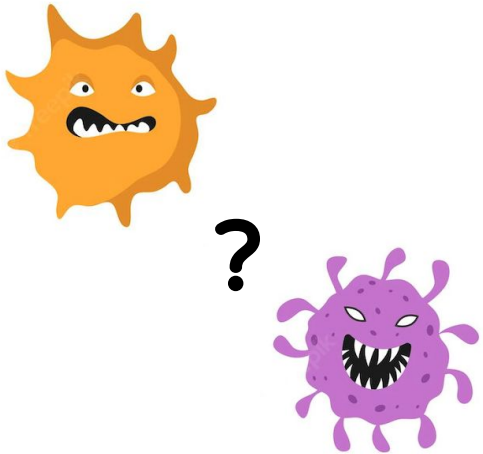
- DNA sequences <100 bp often lack taxonomic specificity within clades, leading to **cross-mapping** within groups of related microbial taxa
- When there are insufficient reference genomes for a given species or genus, these short sequences can easily be **misassigned** to the wrong strain or species (Warinner et al. 2017; Velsko et al. 2018)



Big picture: Why does DNA damage matter?

Accurate genome mapping

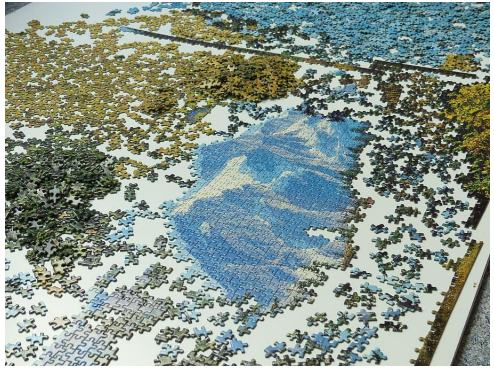
- DNA sequences <100 bp often lack taxonomic specificity within clades, leading to **cross-mapping** within groups of related microbial taxa
- When there are insufficient reference genomes for a given species or genus, these short sequences can easily be **misassigned** to the wrong strain or species (Warinner et al. 2017; Velsko et al. 2018)
- Causes big problems for genotyping, building phylogenies, and inferring evolutionary histories (Fellows-Yates et al. 2021)



Big picture: Why does DNA damage matter?

Metagenomic assembly

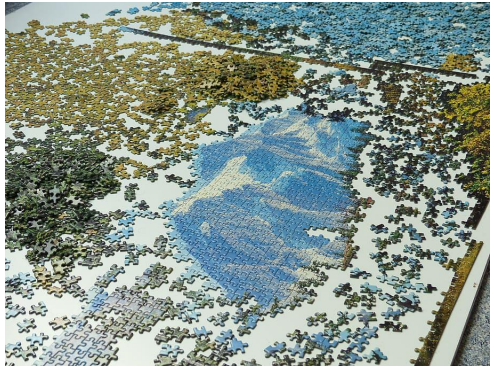
- DNA sequences <250 bp are challenging to *de novo* assemble



Big picture: Why does DNA damage matter?

Metagenomic assembly

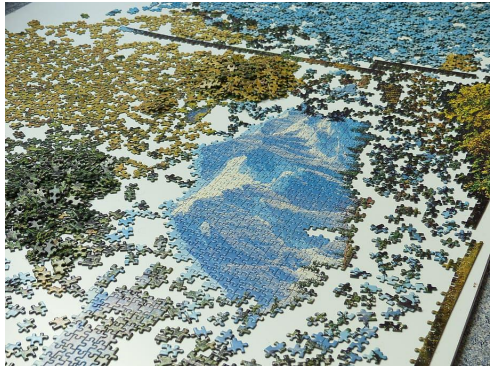
- DNA sequences <250 bp are challenging to *de novo* assemble
- Result in many short contigs because the reads aren't long enough to span repetitive elements



Big picture: Why does DNA damage matter?

Metagenomic assembly

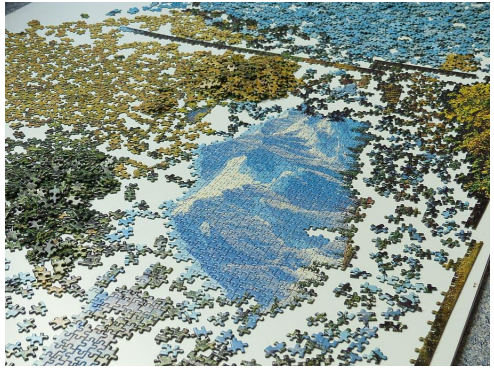
- DNA sequences <250 bp are challenging to *de novo* assemble
- Result in many short contigs because the reads aren't long enough to span repetitive elements
- Many assemblers automatically discard short sequences - so be sure to change default settings!

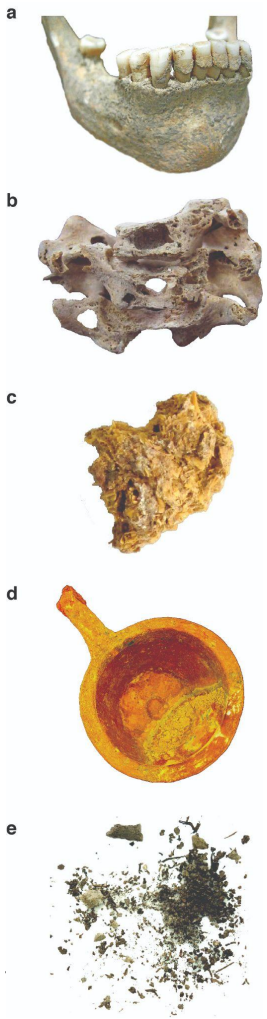


Big picture: Why does DNA damage matter?

Metagenomic assembly

- DNA sequences <250 bp are challenging to *de novo* assemble
- Result in many short contigs because the reads aren't long enough to span repetitive elements
- Many assemblers automatically discard short sequences - so be sure to change default settings!
- Metagenome-assembled genomes (MAGs) are possible, but require pipelines fine-tuned for aDNA

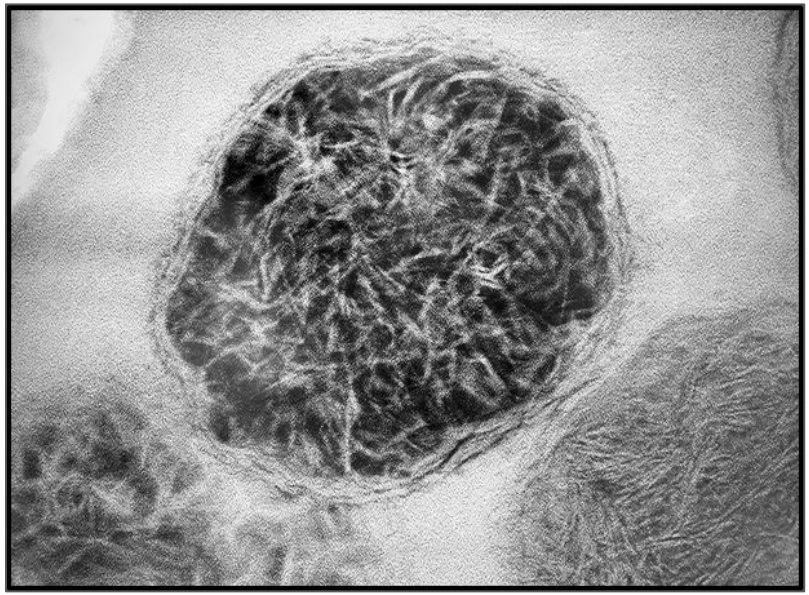




Ancient DNA review

1. Ancient DNA has changed enormously since its beginnings in the early 1980s!
2. Gone are the days of radiographic films and rulers for DNA sequencing; now we have machines capable of churning out 10 billion sequences at a time
3. This means archaeogeneticists today must learn coding and scripting
4. Genomes are big but they fragment into thousands or millions of pieces once the organism dies
5. The shortness of the DNA fragments - mode 30-50 bp, with max ~150 bp - makes taxonomic identification, genome mapping, and metagenomic assembly hard
6. Ancient DNA accumulates damage, and we can characterize fragmentation and cytosine deamination as indicators of authenticity, but not precise age
7. Ancient DNA requires specialized laboratory and library protocols in order to handle DNA damage
8. We now have options to remove damage with UDG or we can recover even more damage with ssDNA library protocols, depending on the application
9. DNA fragmentation is our biggest challenge in ancient metagenomics

Questions?



Want to read more?

Blevins, K.E., Crane, A.E., Lum, C., Furuta, K., Fox, K. and Stone, A.C., 2020. Evolutionary history of Mycobacterium leprae in the Pacific Islands. *Philosophical Transactions of the Royal Society B*, *375*(1812), p.20190582.

Borry, M., Hübner, A., Rohrlach, A. B., & Warinner, C. (2021). PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA de novo assembly. *PeerJ*, *9*, e11845.

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., ... & Krause, J. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, *514*(7523), 494-497.

Briggs, A.W., Stenzel, U., Meyer, M., Krause, J., Kircher, M. and Pääbo, S., 2010. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Research*, *38*(6), pp.e87-e87.

Campillo-Balderas, J. A., Lazcano, A., & Becerra, A. (2015). Viral genome size distribution does not correlate with the antiquity of the host lineages. *Frontiers in Ecology and Evolution*, *3*, 143.

Duggan, A. T., Klunk, J., Porter, A. F., Dhody, A. N., Hicks, R., Smith, G. L., ... & Poinar, H. N. (2020). The origins and genomic diversity of American Civil War Era smallpox vaccine strains. *Genome Biology*, *21*(1), 1-11.

Fellows Yates, J. A., Velsko, I. M., Aron, F., Posth, C., Hofman, C. A., Austin, R. M., ... & Warinner, C. (2021). The evolution and changing ecology of the African hominid oral microbiome. *Proceedings of the National Academy of Sciences*, *118*(20), e2021655118.

Gansauge, M.T. and Meyer, M., 2013. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols*, *8*(4), pp.737-748.

Gansauge, M.T. and Meyer, M., 2019. A method for single-stranded ancient DNA library preparation. In *Ancient DNA* (pp. 75-83). Humana Press, New York, NY.

Want to read more?

Gilbert, M. T. P., Willerslev, E., Hansen, A. J., Barnes, I., Rudbeck, L., Lynnerup, N., & Cooper, A. (2003). Distribution patterns of postmortem damage in human mitochondrial DNA. *The American Journal of Human Genetics*, *72*(1), 32-47.

Gregory, T. R., Nicol, J. A., Tamm, H., Kullman, B., Kullman, K., Leitch, I. J., ... & Bennett, M. D. (2007). Eukaryotic genome size databases. *Nucleic Acids Research*, *35*(suppl_1), D332-D338.

Hagelberg, E., & Clegg, J. B. (1991). Isolation and characterization of DNA from archaeological bone. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *244*(1309), 45-50.

Hidalgo, O., Pellicer, J., Christenhusz, M., Schneider, H., Leitch, A.R. and Leitch, I.J., 2017. Is there an upper limit to genome size? *Trends in Plant Science*, *22*(7), pp.567-573.

Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., & Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. *Nature*, *312*(5991), 282-284.

Hofreiter, M., Serre, D., Poinar, H.N., Kuch, M. and Pääbo, S., 2001. Ancient DNA. *Nature Reviews Genetics*, *2*(5), pp.353-359.

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., & Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, *29*(13), 1682-1684.

Land, M., Hauser, L., Jun, S. R., Nookaew, I., Leuze, M. R., Ahn, T. H., ... & Ussery, D. W. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, *15*(2), 141-161.

Massilani, D., Morley, M. W., Mentzer, S. M., Aldeias, V., Vernot, B., Miller, C., ... & Meyer, M. (2022). Microstratigraphic preservation of ancient faunal and hominin DNA in Pleistocene cave sediments. *Proceedings of the National Academy of Sciences*, *119*(1), e2113666118.

Want to read more?

Meyer, M. and Kircher, M., 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pp.pdb-prot5448.

Morales-Arce, A.Y., Hofman, C.A., Duggan, A.T., Benfer, A.K., Katzenberg, M.A., McCafferty, G. and Warinner, C., 2017. Successful reconstruction of whole mitochondrial genomes from ancient Central America and Mexico. *Scientific Reports*, *7*(1), pp.1-13.

Neukamm, J., Peltzer, A. and Nieselt, K., 2021. DamageProfiler: Fast damage pattern calculation for ancient DNA. *Bioinformatics*, *37*(20), pp.3652-3653.

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., ... & Warinner, C. (2021). Ancient DNA analysis. *Nature Reviews Methods Primers*, *1*(1), 1-26.

Pääbo, Svante, Hendrik Poinar, David Serre, Viviane Jaenicke-Després, Juliane Hebler, Nadin Rohland, Melanie Kuch, Johannes Krause, Linda Vigilant, and Michael Hofreiter. "Genetic analyses from ancient DNA." *Annual Review of Genetics* 38, no. 1 (2004): 645-679.

Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. and Reich, D., 2015. Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1660), p.20130624.

Sanger, F., Nicklen, S. and Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), pp.5463-5467.

Schuenemann, V. J., Avanzi, C., Krause-Kyora, B., Seitz, A., Herbig, A., Inskip, S., ... & Krause, J. (2018). Ancient genomes reveal a high diversity of Mycobacterium leprae in medieval Europe. *PLoS Pathogens*, *14*(5), e1006997.

Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S., Krause, J., & Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, *111*(6), 2229-2234.

Want to read more?

van der Valk, T., Pečnerová, P., Díez-del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., ... & Dalén, L. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, *591*(7849), 265-269.

Velsko, I. M., Frantz, L. A., Herbig, A., Larson, G., & Warinner, C. (2018). Selection of appropriate metagenome taxonomic classifiers for ancient microbiome research. mSystems, 3(4), e00080-18.

Warinner, C., Herbig, A., Mann, A., Yates, J. A. F., Weiß, C. L., Burbano, H. A., ... & Krause, J. (2017). A robust framework for microbial archaeology. *Annual Review of Genomics and Human Genetics*, *18*, 321.

Warinner, C., Rodrigues, J. F. M., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., ... & Cappellini, E. (2014). Pathogens and host immunity in the ancient human oral cavity. *Nature Genetics*, *46*(4), 336-344.