The Little Book of Smiley Plots

A collection of ancient DNA patterns and their causes

The SPAAM Community

2025-01-05

Table of contents

Preface

A key part of any ancient DNA project is to show that the DNA is exactly that - that the DNA is ancient, rather than from modern contamination.

A key authentication method is to show the presence of elevated C to T deamination patterns (and the complementary G to A) at the end of DNA molecules - known as damage patterns - originally reported by (Briggs et al. 2007).

These patterns can be plotted in what have been colloquially known as 'Smiley Plots. However, there can be a wide range of smiley plots, some which show valid ancient DNA, and others that do not - either due to not actually having true ancient DNA but also from laboratory and/or bioinformatic artifacts.

This book aims to act as a reference guide to interpreting ancient DNA damage plots, providing a wide range of example 'smiley plots', with descriptions of what the describe and what can cause them. As an added bit of fun, each type of 'smiley plot' comes with a artistic interpretation of the line shape contributed by members of the ancient DNA community.

Contributors

The following people contributed [1] either data, caricatures, or suggestions to the little book.

- James A. Fellows Yates
- Tessa Zeibig
- Aida Andrades Valtueña
- Pete D. Heintzman
- Katerina Guschanski
- Jaelle C. Brealey

[1] More information can be found on this blog post on the NEB website

Introduction

What are damage patterns

Damage patterns on ancient DNA molecules occur due to increased miscoding lesions at the end of molecules. When DNA molecules start to decompose (i.e., repair mechanisms are lost once an organism dies), the very long DNA molecules start to fragment due to 'nicks' occurring on the sugar-phosphate of one of the strands, weakening the structure and causing the molecule to cleave into two. However, this cleavage is not necessarily 'clean', i.e., occurs on both strands at the same position. Rather, when the two uneven 'nicks' cause the DNA molecule to cleave into two, this results in a 'jagged' break - with the molecules having 'overhangs' of one strand being longer than the other of each of the new two now-'independent' molecules.

The resulting single-stranded overhangs leave the nitrogenous-bases 'exposed' on the overhang to the surrounding environment. In such cases, of the four nucleotides, it was found that cytosines undergo deamination at a higher rate than the others via hydrolysis. The loss of a cytosines amine group results in a nucleotide structure normally found more often in RNA molecules - uracils. The reason why palaeogenomicists report 'C to T' damage patterns is because some polymerases will misread uracils as an adenine, and will incorporate a thymine on the opposite strand during DNA amplification. During each subsequent amplification cycle, the mis-incorporated T will propagate across the subsequent copies of the original DNA molecule.

To summarise, the unequal ends of fragmented DNA molecules results in the increased chance of damage to the nucleotides to the overhangs. This structural damage occurs more frequently in cytosines over the other bases, however these 'damaged' cytosines are misread by polymerases during DNA amplification to result on thymines on the opposite strand (rather than the expected complementary guanines).

It is important to note that the library construction method will influence damage, e.g., is the library constructed from double-stranded DNA or single-stranded DNA, is the polymerase in the initial library amplification proof-reading or not, and so on. Throughout the rest of this book, each damage pattern will be described in the context of the library construction method of the data used to generate each damage pattern.

The increased *frequency* to C to T was detected at the end of molecules could only be detected with the invention of 'Next Generation Sequencing' or 'NGS' (Shendure and Ji 2008). NGS allowed palaeogenomicists to easily sequence thousands to millions of DNA molecules in one

go in an untargeted manner, which subsequently meant that molecules from across *entire* genomes could be compared against a reference genome. Bioinformaticially, the increase of C to T miscoding lesions were detected by measuring the frequency of mutations at each position across each read, where each read was derived from a different place on the reference genome. As many different places across the reference genome would have different base composition, one would expect to see an approximately random distribution of mutations across the genome. However it was observed in Neanderthals DNA libraries that the frequency of C to T mutation in the first ~10 base pairs of the 5p end of double-stranded library molecules had a higher frequency than the expected approximate equal distribution across each type of mutation, as seen in Figure 1.

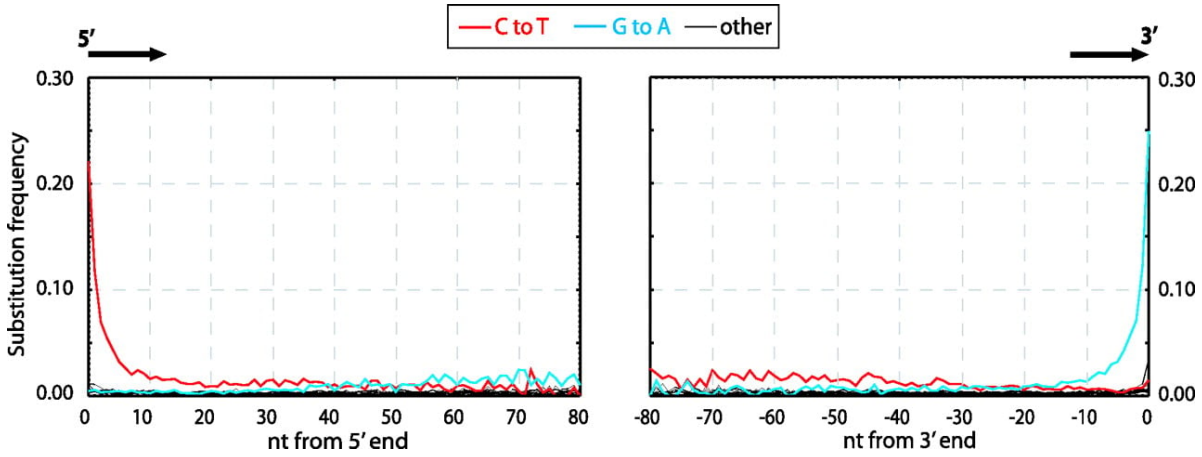

Figure 1: First reported misincorporation lesion 'smiley plot' from Neanderthal DNA (Briggs et al. 2007). Reproduced here under free access.

How are damage patterns analysed

There is a range of software that can generate damage pattern plots from ancient DNA NGS libraries. The vast majority of tools require to be of sequencing reads aligned to a reference genome or genomes. Here we make suggestions of some tools that you can use to generate such plots. The example damage patterns in this book will mostly be derived from genomics tools, as metagenomic damage plot generation may account for other factors than the 'classical' ancient DNA damage plot.

Genomics

These tools generally take BAM files as input (i.e., after mapping of FASTQ files to a reference genome using a short-read aligner):

- mapDamage
  - Source: https://github.com/ginolhac/mapDamagee
  - Documentation: https://ginolhac.github.io/mapDamage
  - Citation: (Jónsson et al. 2013)
- PMDtools
  - Source: https://github.com/pontussk/PMDtools
  - Documentation: https://github.com/pontussk/PMDtools
  - Citation: (Skoglund et al. 2014)
- DamageProfiler
  - Source: https://github.com/Integrative-Transcriptomics/DamageProfiler
  - Documentation: https://damageprofiler.readthedocs.io/en/latest/
  - Citation: (Neukamm, Peltzer, and Nieselt 2021)

Metagenomics

These tools may take different approaches to generating their alignments (or even alignment free methods).

- MaltExtract
  - Source: https://github.com/rhuebler/MaltExtract
  - Documentation: https://github.com/rhuebler/MaltExtract
  - Citation: (Hübler et al. 2019)
- pyDamage
  - Source: https://github.com/maxibor/pydamage
  - Documentation: https://pydamage.readthedocs.io/en/0.7/
  - Citation: (Borry et al. 2021)
- MetaDMG
  - Source: https://github.com/metaDMG-dev/metaDMG-core
  - Documentation: https://metadmg-dev.github.io/metaDMG-core/
  - Citation: (Michelsen et al. 2022)

Part I

Valid Smiley Plots

This section of the litte book of smiley plots shows damage patterns as they *should* be from a molecular biology point of view.

These are the ones that will immediately make *you* smile as these give you good indications of valid ancient DNA in your sequencing library!

1 Double stranded DNA libraries

Figure 1.1: Archibald by Tessa Zeibig

This is the 'classical' ancient DNA plot that you will see most often in palaeogenomics. You expect to see a smooth curve from the beginning of the read (position 1) to a flat line in the middle (e.g. positions 10-25 in mapDamage plots). At the 5' end this will be indicated by the original C to T deamination, whereas the 3' of the molecule will show the complementary G to A. You only see deamination the C to T (and complement G to A) at one end of the the molecule, as during typical double-stranded library construction protcols (Meyer and Kircher 2010) only *one* end of the single-ended overhangs of a DNA molecule is repaired by being

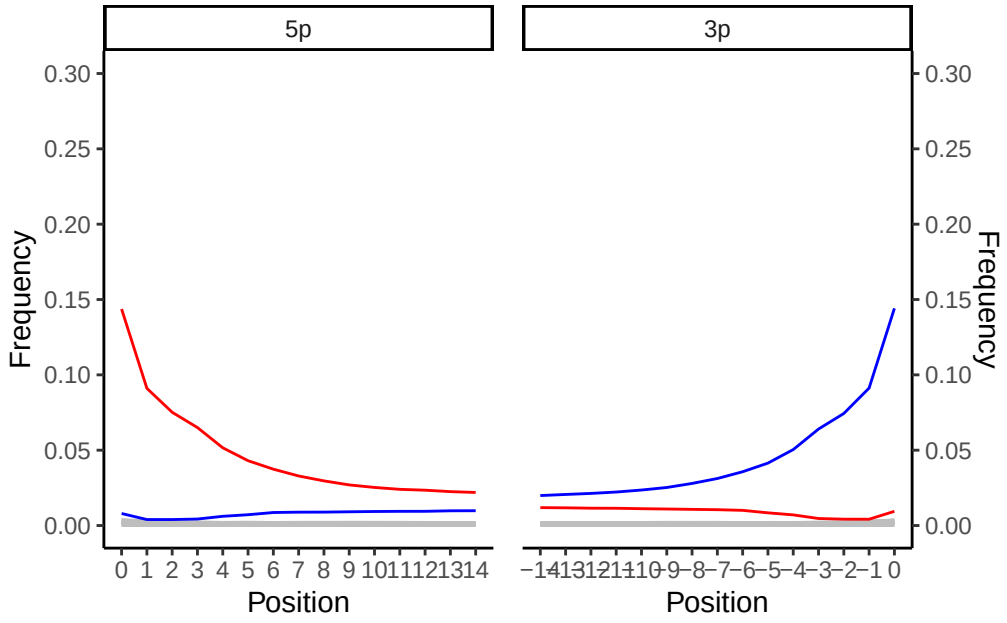

Figure 1.2: Example of a smiley plot of a double stranded DNA library. Data taken from library COD076E1bL1 (ERR1943600-ERR1943602) of (Star et al. 2017). Damage data generated using DamageProfiler and plotted using R and tidyverse packages (Wickham et al. 2019).

'filled in' (where the mis-reading of the deaminated C occurs). Overhangs at the other end of the molecule (which may also hold cytosine demination) are 'blunt-ended' by being trimmed off. Both fill-in and blunt-ending reactions are performed to allow ligation of next-generation-sequencing adapters and/or internal barcodes to both ends of the molecules. The highest frequency point of the curve can vary from 1% to 50% depending on the age and preservation of the sample.

If you get such a plot with smooth lines from ancient DNA double-stranded libraries, this is a good indication you have authentic ancient DNA!

2 Single stranded DNA Libraries

This is an increasingly common ancient DNA plot that is you will see more often as single-stranded library construction protocols become more popular. You expect to see a smooth curve from the beginning of the read (position 1) to a flat line in the middle (e.g. positions 10-25 in mapDamage plots). As with double-stranded ancient DNA libraries, the 5' end will have the expect original C to T deamination curve. In contrast to the double stranded protocol, the 3' of the molecule will *also* show the complementary C to T curve. You see the C to T deamination at both ends of the the molecule in this case, as during typical single-stranded library construction protcols (e.g. Gansauge et al. 2017) the entire DNA molecule is denatured, immobilised and then the complementary strand reconstructed - i.e., without any filling in or blunt ending. The highest frequency point of the curve can vary from 1% to ???% depending on the age and preservation of the sample.

If you get such a plot with smooth lines from ancient DNA single-stranded libraries, this is a good indication you have authentic ancient DNA!

Figure 2.1: Homage to the best moustache ever (Salvador Dalí) by Nihan D. Dagtas. Based upon 'Salvador Dalí' by Philippe Halsman (1953) here under Fair Use as per Wikiart definition and is used for educational purposes

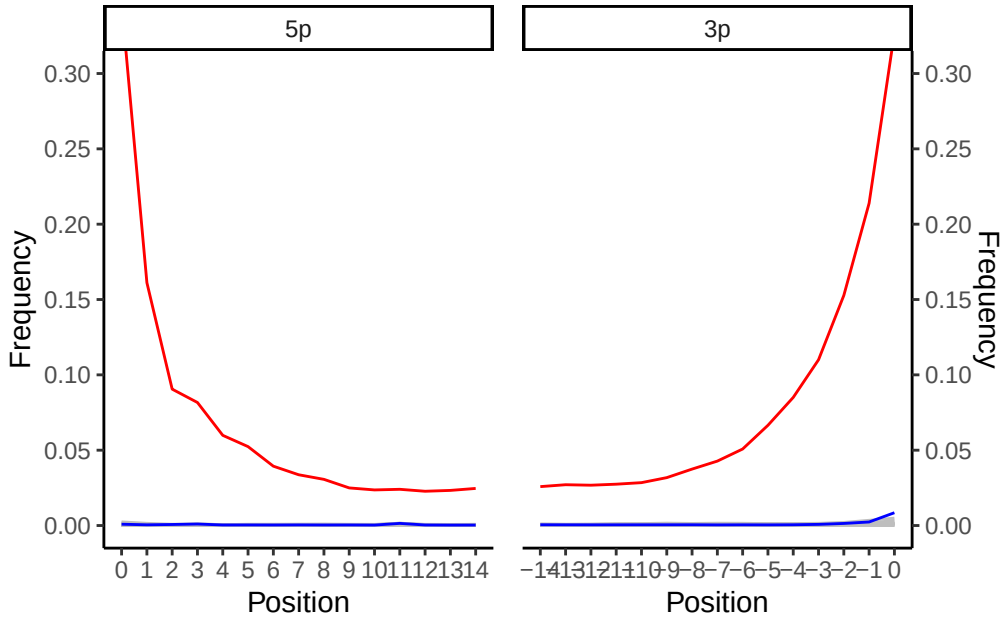

Figure 2.2: Example of a smiley plot of a double stranded DNA library. Data taken from library VEL003.B0101 of (Andrades Valtueña et al. 2022). Damage data generated using DamageProfiler and plotted using R and tidyverse packages (Wickham et al. 2019).

3 Partial UDG (double-stranded)

CARICATURE PLOT GOES HERE

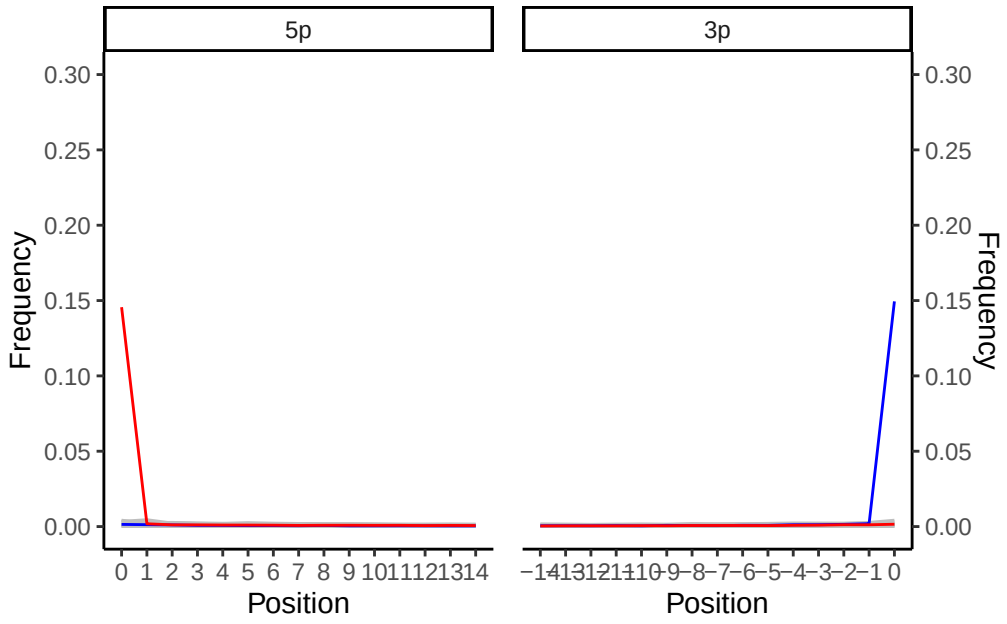

Figure 3.1: Example of a smiley plot of a double-stranded 'partial UDG' or 'UDG half' library. Data taken from library VLI092.A0101 (ERR8958796) of (Andrades Valtueña et al. 2022). Damage data generated using DamageProfiler and plotted using R and tidyverse packages (Wickham et al. 2019).

The smiley plot presented here has been generated from a double-stranded 'partial UDG' or 'UDG-half' library. UDG or USER Treatment consists of enzymatically cleaving off ends of molecules when a uracil is present. Whereas complete-UDG UDG full runs the protocol long enough for all DNA molecules to be enzymatically modified, partial UDG stops early, meaning that a small number of reads retain uracils on the last base. This allows for simultaneous authentication of damage, but makes it easier to then *in silico* remove damage by 'trimming' off one base from either end of each read, so you can ensure you do not incorporate damaged bases into downstream analyses.

Part II

Half Valid Smiley Plots

This section of the little book of smiley plots shows all the slightly odd damage profiles that don't look as expected, but still retain enough of a pattern that you can take as a positive indicator of ancient DNA.

The descriptions hopefully will provide guidance on how to interpret and when to maybe consider re-doing libraries or mapping.

4 Proofreading enzymes

CARICATURE PLOT GOES HERE

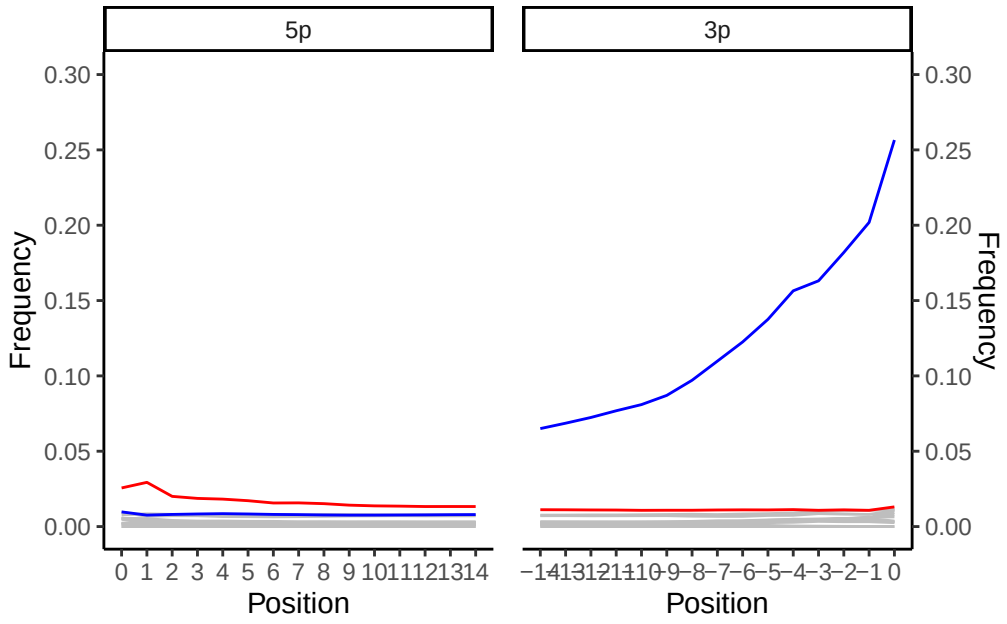

Figure 4.1: Example of a smiley plot of a double stranded library amplified after adapter ligation with a proofreading NEBNext Ultra II Q5 Master polymerase. Data taken from library HH06D of (Cai et al. 2022). Damage data generated using DamageProfiler and plotted using R and tidyverse packages (Wickham et al. 2019).

Sometimes the type of polymerase you use during library construction will influence the type of damage pattern you will receive.

In the example above, Cai et al. (2022) found a funny smiley plot, where while the 3' G to A patterns look like a classic non-UDG (i.e. non-USER treated, thus retaining damage) plot, the frequency of C to T on the 5' end was extremely reduced.

They identified the problem as the choice of polymerase. The Q5 polymerase is a 'high fidelity' enzyme, which corresponds to being a 'proofreading' enzyme. This means that when the

enzyme hits an incorrect nucleotide (such as a deaminated cytosine), it will instead remove the nucleotide on it's 3' to 5' exonuclease activity [1].

While having a low error rate is great for modern genomics, this can be less optimal for preserving ancient DNA damage for profiling later on.

In the case of this particular enzyme, it is maybe not so much of a problem as you retain the damage signal on the 3' for proving authenticity. However this may be problematic for downstream aDNA validation tools that may have an expected 'model' of ancient DNA damage.

> **Tip**
>
> The choice of enzyme only matters during the first amplification after adapter ligation. At subsequent amplifications, the (misincorporated) thymines have already been integrated into the template molecules, so it doesn't matter which enzyme you use.

[1] More information can be found on this blog post on the NEB website

5 Internal Barcode Ligation Bias

CARICATURE PLOT GOES HERE

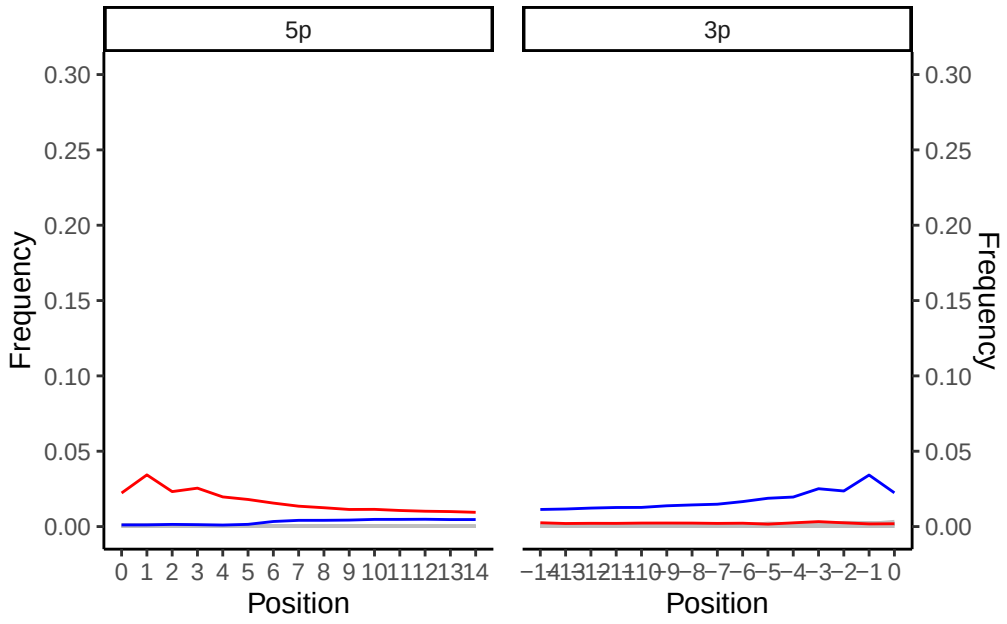

Figure 5.1: Example of a smiley plot of an internal barcoded double-stranded library with a ligation bias of certain barcodes. Data taken from sample Ua9 (ERS4545914) of (Brealey et al. 2020). Damage data generated using DamageProfiler and plotted using R and tidyverse packages (Wickham et al. 2019).

In this smiley plot, you see slightly spiky ends of the damage curves; mainly that some of the last couple of bases are often lower than expected from the rest of the otherwise classical damage curve of a constant decrease. This was observed in Brealey et al. (2020), who associated this to the explanation of reduced ligation efficiency to DNA molecules of internal barcode with a terminal G or C (i.e., short synthetic oligos with known sequences added directly prior addition of library adapters and indices) as suggested by Rohland et al. (2015).

In other words, certain internal barcode sequences with a terminal G or C do not ligate as as well to an ancient DNA molecule with a deaminated C on the terminus, and thus those

read will be lost during the second round of demultiplexing as it will not contain the barcode associated with that library.

Ultimately you should not be too worried about this plot if you get it - you probably still have aligned true aDNA reads, however you may have lost a small fraction of true ancient DNA reads in that particular library. If you wish to ensure you have retained as many aDNA reads as possible, you should re-build the library from an extract but with different internal barcodes without a terminal G or C.

Part III

Invalid Smiley Plots

This section of the little book of smiley plots shows all the weird and wonderful *strange* smiley plots that represent problems or artefacts in the sequencing libraries which will make you bemused…

The descriptions hopefully will provide guidance on how to interpret and remedy such problems.

6 Insufficient reads

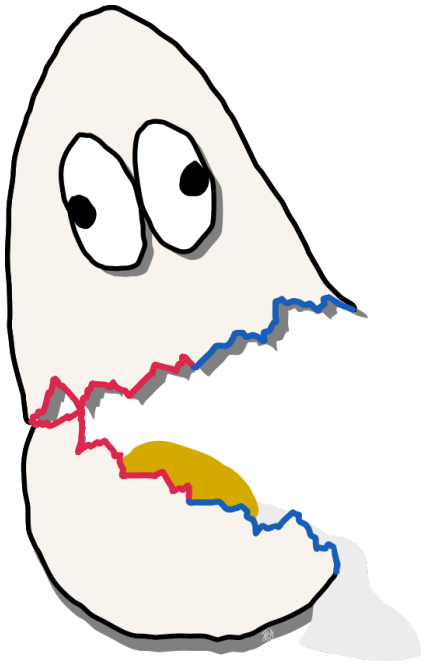

Figure 6.1: Egbert by James A. Fellows Yates

When you get random spikey lines in both 5p and 3p ends of the smiley plot, this more often than not indicates that insufficient reads are present to generate the damage profile. Given the plots are based on frequency, sufficient numbers of reads are needed to visualise the 'fraction' of C to T misincorporations versus the reference, if there are too few reads, this produces 'noise' in the line.

In this case of the example above, the aligned DNA *does* have a true damage signal (as indicated by the high frequency of the C-T misincorporations on the 0 and 1 positions of the 5p plot) so *may* give you a teeny-weeny hint of the presence of true ancient DNA. However the rest of line and also the 3p show random spikes making it very difficult to make any firm conclusion.

When you receive a plot like this, you normally need to increase the number of reads in your alignment against the reference genome (deeper sequencing, relaxing alignment parameters), or possibly you have the wrong reference genome (meaning it is not similar enough to align the reads in your library against it).

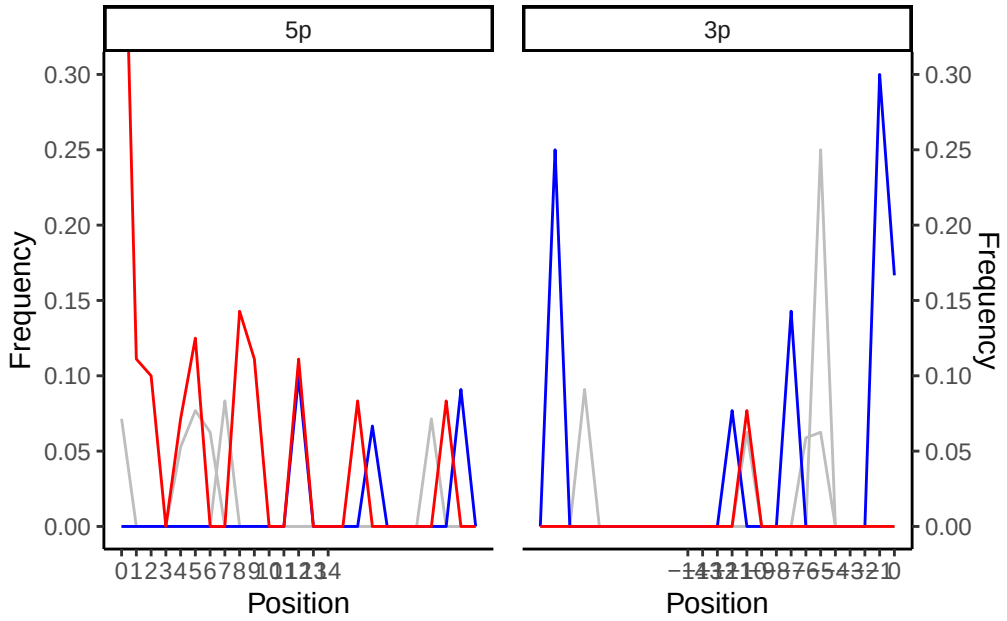

Figure 6.2: Example of a smiley plot of an alignment with insufficient reads to generate a confident smiley plot. Data taken from a non-UDG library of a captured Woolly Mammoth mitochondrial genome (JK2782) from (Fellows Yates et al. 2017), and sampled aligned reads down to 50 reads. Damage data generated using DamageProfiler and plotted using R and tidyverse packages (Wickham et al. 2019).

7 Internal barcodes not removed

CARICATURE PLOT GOES HERE

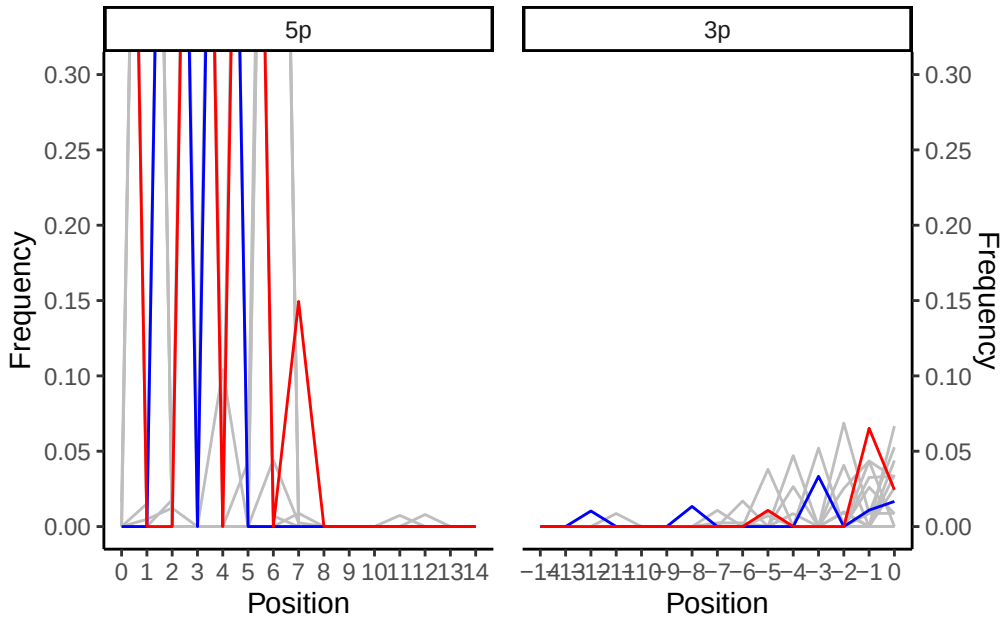

Figure 7.1: Example of a smiley plot of a double stranded library with barcodes that were not removed prior mapping. Data taken from library GEN72 (ERR2112579) of (Andrades Valtueña et al. 2017). Damage data generated using DamageProfiler and plotted using R and tidyverse packages (Wickham et al. 2019).

In this case, this smiley plot is not very smiley as the mapped reads have not had 'internal barcode' removed prior mapping. As a side effect, you see a very spiky initial 'curve', and then the rest of the read being flat.

In this case, this library only has a single barcode on the 5p end, but some labs may add internal barcodes to both ends of molecules.

References

Andrades Valtueña, Aida, Alissa Mittnik, Felix M Key, Wolfgang Haak, Raili Allmäe, Andrej Belinskij, Mantas Daubaras, et al. 2017. "The Stone Age Plague and Its Persistence in Eurasia." *Current Biology: CB* 27 (23): 3683–3691.e8. https://doi.org/10.1016/j.cub.2017.10.025.

Andrades Valtueña, Aida, Gunnar U Neumann, Maria A Spyrou, Lyazzat Musralina, Franziska Aron, Arman Beisenov, Andrey B Belinskiy, et al. 2022. "Stone Age Yersinia Pestis Genomes Shed Light on the Early Evolution, Diversity, and Ecology of Plague." *Proceedings of the National Academy of Sciences of the United States of America* 119 (17): e2116722119. https://doi.org/10.1073/pnas.2116722119.

Borry, Maxime, Alexander Hübner, Adam B Rohrlach, and Christina Warinner. 2021. "PyDamage: Automated Ancient Damage Identification and Estimation for Contigs in Ancient DNA de Novo Assembly." *PeerJ* 9 (July): e11845. https://doi.org/10.7717/peerj.11845.

Brealey, Jaelle C, Henrique G Leitão, Tom van der Valk, Wenbo Xu, Katia Bougiouri, Love Dalén, and Katerina Guschanski. 2020. "Dental Calculus as a Tool to Study the Evolution of the Mammalian Oral Microbiome." *Molecular Biology and Evolution* 37 (10): 3003–22. https://doi.org/10.1093/molbev/msaa135.

Briggs, Adrian W, Udo Stenzel, Philip L F Johnson, Richard E Green, Janet Kelso, Kay Prüfer, Matthias Meyer, et al. 2007. "Patterns of Damage in Genomic DNA Sequences from a Neandertal." *Proceedings of the National Academy of Sciences of the United States of America* 104 (37): 14616–21. https://doi.org/10.1073/pnas.0704665104.

Cai, Dawei, Siqi Zhu, Mian Gong, Naifan Zhang, Jia Wen, Qiyao Liang, Weilu Sun, et al. 2022. "Radiocarbon and Genomic Evidence for the Survival of Equus Sussemionus Until the Late Holocene." *eLife* 11 (May). https://doi.org/10.7554/eLife.73346.

Fellows Yates, James A, Dorothée G Drucker, Ella Reiter, Simon Heumos, Frido Welker, Susanne C Münzel, Piotr Wojtal, et al. 2017. "Central European Woolly Mammoth Population Dynamics: Insights from Late Pleistocene Mitochondrial Genomes." *Scientific Reports* 7 (1): 17714. https://doi.org/10.1038/s41598-017-17723-1.

Gansauge, Marie-Theres, Tobias Gerber, Isabelle Glocke, Petra Korlevic, Laurin Lippik, Sarah Nagel, Lara Maria Riehl, Anna Schmidt, and Matthias Meyer. 2017. "Single-Stranded DNA Library Preparation from Highly Degraded DNA Using T4 DNA Ligase." *Nucleic Acids Research* 45 (10): e79. https://doi.org/10.1093/nar/gkx033.

Hübler, Ron, Felix M Key, Christina Warinner, Kirsten I Bos, Johannes Krause, and Alexander Herbig. 2019. "HOPS: Automated Detection and Authentication of Pathogen DNA in Archaeological Remains." *Genome Biology* 20 (1): 280. https://doi.org/10.1186/s13059-019-1903-0.

Jónsson, Hákon, Aurélien Ginolhac, Mikkel Schubert, Philip L F Johnson, and Ludovic Orlando. 2013. "mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters." *Bioinformatics* 29 (13): 1682–84. https://doi.org/10.1093/bioinformatics/btt193.

Meyer, Matthias, and Martin Kircher. 2010. "Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing." *Cold Spring Harbor Protocols* 2010 (6): db.prot5448. https://doi.org/10.1101/pdb.prot5448.

Michelsen, Christian, Mikkel Winther Pedersen, Antonio Fernandez-Guerra, Lei Zhao, Troels C Petersen, and Thorfinn Sand Korneliussen. 2022. "MetaDMG – a Fast and Accurate Ancient DNA Damage Toolkit for Metagenomic Data." *bioRxiv*. https://doi.org/10.1101/2022.12.06.519264.

Neukamm, Judith, Alexander Peltzer, and Kay Nieselt. 2021. "DamageProfiler: Fast Damage Pattern Calculation for Ancient DNA." *Bioinformatics* 37 (20): 3652–53. https://doi.org/10.1093/bioinformatics/btab190.

Rohland, Nadin, Eadaoin Harney, Swapan Mallick, Susanne Nordenfelt, and David Reich. 2015. "Partial Uracil-DNA-Glycosylase Treatment for Screening of Ancient DNA." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1660): 20130624. https://doi.org/10.1098/rstb.2013.0624.

Shendure, Jay, and Hanlee Ji. 2008. "Next-Generation DNA Sequencing." *Nature Biotechnology* 26 (10): 1135–45. https://doi.org/10.1038/nbt1486.

Skoglund, Pontus, Bernd H Northoff, Michael V Shunkov, Anatoli P Derevianko, Svante Pääbo, Johannes Krause, and Mattias Jakobsson. 2014. "Separating Endogenous Ancient DNA from Modern Day Contamination in a Siberian Neandertal." *Proceedings of the National Academy of Sciences of the United States of America* 111 (6): 2229–34. https://doi.org/10.1073/pnas.1318934111.

Star, Bastiaan, Sanne Boessenkool, Agata T Gondek, Elena A Nikulina, Anne Karin Hufthammer, Christophe Pampoulie, Halvor Knutsen, et al. 2017. "Ancient DNA Reveals the Arctic Origin of Viking Age Cod from Haithabu, Germany." *Proceedings of the National Academy of Sciences of the United States of America* 114 (34): 9152–57. https://doi.org/10.1073/pnas.1710186114.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.