NIELS BOHR INSTITUTE

UNIVERSITY OF
COPENHAGEN

PH.D. THESIS

by

Christian Michelsen

Biological Data Science

Ancient genomics, anesthesiology, epidemiology,
and a bit in between

Submitted: November 16, 2022

*This thesis has been submitted to the
PhD School of The Faculty of Science,
University of Copenhagen*

Supervisor	Troels C. Petersen	Niels Bohr Institute
Cosupervisor	Thorfinn S. Korneliussen	Globe Institute

Christian Michelsen,
Biological Data Science:
ancient genomics, anesthesiology, epidemiology, and a bit in between,
November 16, 2022.

Til kvinderne i mit liv

Table of Contents

FRONT

Preface

This Ph.D. thesis summarizes my scientific research in collaboration with the Niels Bohr Institute (NBI) and the Globe Institute, University of Copenhagen, and was funded by the Lundbeck Foundation. The research was supervised by Associate Professor Troels C. Petersen (NBI) and Assistant Professor Thorfinn S. Korneliussen (Globe Institute).

Being a multi-disciplinary project, the research presented in this thesis is multi-faceted and covers a wide range of topics with the main scope being the development and integration of a novel statistical methods and machine learning models for the analysis of large-scale biological data. The thesis is organized as follows. First I present a brief introduction to the statistical methods and machine learning models used in the thesis. Then I present the research in the form of four papers, each of which reflects a different aspect of the research.

The first paper presents a novel method I developed for detecting and classifying ancient DNA damage in metagenomic samples taking the full taxanomic information into account. While the first paper focuses on the development of the statistical model in the field of ancient genomics, the second paper focuses on the use of modern machine learning models in medicine and how advanced boosted decision trees can not only improve the accuracy of identifying patients at risk of being readmitted after knee or hip surgery, but doing so in a way that is interpretable as well.

In the beginning of 2020 we all experienced how COVID-19 suddenly changed our lives and impacted our societies in dramatic ways. During this time, I ended up working for Statens Serum Institut, the Danish CDC, on a project to develop an agent based model capable of simulating the spread of COVID-19 in Denmark. This model is presented in the third paper and was used to inform the Danish government on how to best handle the pandemic in the early stages and the effect of contact tracing.

Finally, in the fourth paper I show how advanced Bayesian methods can be utilized to better estimate the diffusion coefficients of molecules in the cell nucleus in XXX experiments.

Acknowledgements

First of all I want to express my sincere gratitude to my long time supervisor, captain, and friend: Troels. You are truly an inspiration to work with. I want to thank you for opening so many doors for me, both academically, professionally, and nautically. I am looking forward to our future adventures together. I also want to thank my co-supervisor, Thorfinn. I want to thank you for introducing me to the field of bioinformatic and helping me to develop my skills in this area. I also want to thank you for your patience and guiding me through the endless amount of (near) identical biological concepts and helping me to understand the minute differences.

I have been fortunate to work with people from a wide range of backgrounds and disciplines during my Ph.D. The author lists on the papers in this thesis include a particle physicist, bioinformaticians, a clinical professor, epidemiologists, a medical doctor, a bio-physicist, a mathematician, a biologist, and the president of the Royal Danish Academy of sciences and letters. Before anything else, I want to thank all of my co-authors for their work and contributions to these papers and for allowing me to be a part of their projects. I have learned a lot from all of you, and I hope that I have been able to contribute something to your work as well.

I am thankful for all the people who have helped me with my work and listened to my complaints when I was stuck, when the code did not compile, or when the small bug was almost impossible to find (which was not a small amount of time). In particular I want to thank the people at Globe who I have spent the most time with; Rasa, Alba, and Rasmus. I also want to thank the Korneliussen Group and the people in my office for helpful advice, suggestions and discussions. This also includes Daniel Nielsen and Rasmus Ørsøe from NBI. Finally, I want to thank Mathias Heltberg for many years of fruitful collaboration and for including me in his projects.

This project would not have been possible had it not been for the Lundbeck Foundation which funded my Ph.D. In addition to the funding itself, I am grateful for the inter-disciplinary aspect of project which has allowed me to meet so many inspiring and talented people and for the freedom to pursue my own interests within the project.

I would also like to express my gratitude to Professor Guido Sanguinetti from the International School for Advanced Studies, SISSA, in Trieste, Italy, for hosting me in his group during the Winter of 2021/2022. My gratitude also goes out to Kosio, Sara, Max, Romina, Noor, Viplove, Anne-Marie, and all the other wonderful

people that I met during in Trieste. Thanks for making my stay in Italy so enjoyable and for welcoming me in a way that only non-Danes can do.

I want to thank my friends for always being there for me. A special thanks to my friends from NBI and Borchen who I know that I can always count on, whether or not that includes a trip in the party bus of the Sea, taking Artemis out for a sail, or board games and beer. Thank you for always being there. I also want to thank my family, especially my parents for their unconditional support and encouragement. I am grateful for the opportunities that they have given me and for the sacrifices that they have made for me.

Lastly, I want to thank my future wife and mother of our child, Anna. I would not have been able to do this without you. Thank you for your patience and support. I am looking forward to our future together. I love you to the moon and back.

Abstract

Basically a thesis (book?) class for Tufte lovers like myself. I am aware that `tufte-latex` already exists but I just wanted to create my own thing.

Dansk Abstract

Her et dansk abstract.

Publications

The work presented in this thesis is based on the following publications:

Paper 1: **Christian Michelsen**[†], Mikkel W. Pedersen[†], Antonio Fernandez-Guerra, Lei Zhao, Troels C. Petersen, Thorfinn S. Korneliussen (2022). *"metaDMG: An Ancient DNA Damage Toolkit"*. Submitted to Methods in Ecology and Evolution.

Paper 2: **Christian Michelsen**[†], Christoffer C. Jorgensen[†], Mathias Heltberg, Mogens H. Jensen, Alessandra Lucchetti, Pelle B. Petersen, Troels C. Petersen, Henrik Kehlet (2022). *"Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty – a machine learning based approach."*. Accepted and in review at BMJ Open.

Paper 3: Mathias S. Heltberg[†], **Christian Michelsen**[†], Emil S. Martiny, Lasse E. Christensen, Mogens H. Jensen, Tariq Halasa and Troels C. Petersen (2022). *"Spatial Heterogeneity Affects Predictions from Early-Curve Fitting of Pandemic Outbreaks: A Case Study Using Population Data from Denmark"*. Published in: Royal Society Open Science 9.9. issn: 2054-5703. doi: 10.1098/rsos.220018.

Paper 4: Susmita Sridar[†], Mathias S. Heltberg[†], **Christian Michelsen**[†], Judith M. Hattab, Angela Taddei (2022). *"Microscopic single molecule dynamics suggest underlying physical properties of the silencing foci"*. Paper draft.

In the cases of shared first authorship, this is indicated with a dagger (†) next to the name.

MAIN

1 *Introduction*

The primary content of my thesis is the four papers that are included in the thesis. This chapter is meant as a brief introduction to the background needed to understand the basics of the methods used throughout the papers. As such, this chapter is not meant to be a comprehensive guide to statistics and bioinformatics used in the papers. The original research motivation supporting the funding of this Ph.D. was very multi-disciplinary and the papers included in my thesis are also clearly influenced by this.

In Section 1.1, I will shortly introduce the field of ancient genomics and the statistical methods used to identify ancient DNA will be explained. Paper I, see Chapter 2, utilize modern Bayesian methods to classify which species are ancient and which ones are not. Bayesian methods are great when possible, however, they also rely on some statistical model being defined. In the case of Paper I, the model is a Beta-Binomial distribution combined with an exponential-decay damage model.

Sometimes, however, the model is not known and the data generating process has to be inferred by other means. This is the case in Paper II, see Chapter 3, where we utilize machine learning methods to extract this information. This paper deals with estimating the individual risk scores for each patient being re-hospitalized after a knee or hip operation. Section 1.2 introduces the reader to basic classification with machine learning models.

While the former two papers are based on real life data, Paper III, see Chapter 4, concerns the development of a new agent based model for COVID-19. The model is based on the SIR model, but with a more detailed description of the disease and the transmission process. The model is used to simulate the spread of the virus in Denmark and to estimate the effect of contact tracing. The model is also used to simulate and predict the spread of the "alpha" variant of COVID-19 in Denmark. Section 1.3 introduces the reader to the basics of agent based models.

Finally, the method of Bayesian model comparison of different diffusion models is introduced in Paper IV, see Chapter 5. In particular, this paper deals with different mixture-models of independent Rayleigh-distributions and how they can be used to extract important information about the underlying diffusion processes of a polymer bridging model in cell nuclei, see Section 1.4.

1.1 *Ancient DNA and Bayesian Statistics*

Previously, the only way to study ancient animals, plants, and other species was by studying their fossils. This changed in the middle of the 1980's when the first DNA was recovered from almost 5000 years old ancient mummies, showing that it was indeed possible to extract and sequence ancient DNA (Pääbo, 1985a; Pääbo, 1985b). This discovery, along with a dozen other pushing the boundary for what is scientifically possible with ancient DNA, led to Svante Pääbo being awarded with the Nobel Prize in Physiology or Medicine in 2022 for "his discoveries concerning the genomes of extinct hominins and human evolution" (Karolinska Institutet, 2022).

The field of ancient DNA (aDNA) was drastically changed with the invention of the Polymerase Chain Reaction, PCR, method (Mullis et al., 1986) along with the Next Generation Sequencing technology which revolutionized the speed and throughput of genomic sequencing while decimating the cost (Slatko, Gardner, and Ausubel, 2018). This technological advance has lead to better understanding of human migration and the genealogical tree of modern humans including the previously unknown human (sub)species; the Denisova hominin (Krause et al., 2010).

Leaving the homocentric world view, aDNA also allows for the study of archaic animals. In recent years, the boundary of how old DNA can be sequenced has been severely pushed; in 2013 with the early Middle Pleistocene 560–780 kyr BP horse (Orlando et al., 2013) and in 2021 with the million-year-old mammoths (van der Valk et al., 2021). High-throughput sequencing not only allows for the sequencing of single genomes – like single humans, animals, or plants – but also for sequencing of entire communities of organisms, so-called metagenomics. By analysing the DNA in environmental samples, environmental DNA, one can survey the rich plant and animal assemblages of a given area and at a specific time in the past. A new paper published in Nature shows that it is now possible to perform metagenomic sequencing on environmental DNA that is 2 million years old, see Appendix A. This is a direct application of the statistical method developed in Paper I, see Chapter 2, showing that `metaDMG`, the method, can help to push the boundary of what is possible with ancient DNA.

Ancient DNA is difficult to work with since it often contains only a limited amount of biological material due to bad preservation, leading to low endogenous content with high duplication rates, making high-depth sequencing difficult (Renaud et al., 2019). Genotype likelihoods are often used to alleviate the problem of low-coverage data (Nielsen et al., 2011). In addition to this, the DNA is often highly degraded. In particular, the two prominent issues with aDNA is fragmentation

and deamination (Dabney, Meyer, and Pääbo, 2013; Peyrégne and Prüfer, 2020). Fragmentation refers to the fact that the DNA is broken into very short fragments, often with a fragment size of less than 50 bp. This leads to low-quality mapping issues and reference biases, which can somewhat be mitigated by the use variant graphs (Martiniano et al., 2020). Deamination is a process in which cytosine (C) in the single-stranded overhangs in the end of the DNA molecules is often hydrolized to uracil (U) which is then read as thymine (T) by the DNA polymerase. This particular type of postmortem damage is known as cytosine deamination, or C-to-T transitions, and is a one of the main reasons behind nucleotide misincorporations in ancient DNA (Briggs et al., 2007). Due to the short fragment sizes in ancient DNA, they will often contain overhangs with over-expressed C-to-T frequency. In the case of single-genome analysis, previous solutions have been to either remove all transitions and only keep transversions, or apply trimming at the read ends (Schubert et al., 2012). For an illustration of both fragmentation and deamination of ancient DNA, see Figure 1.

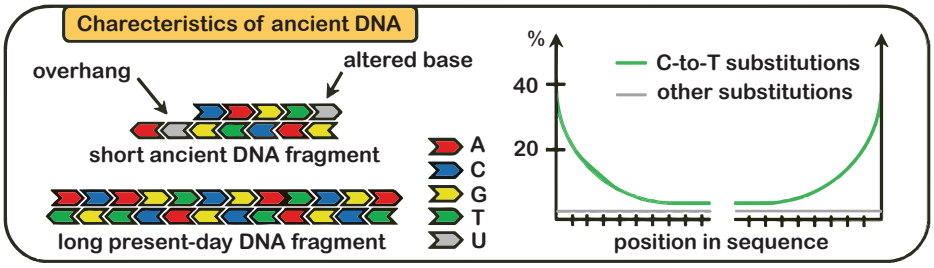


Figure 1. Illustration of DNA damage. Ancient DNA is often highly fragmented with short reads compared to modern, present-day DNA, and can contain uracils (U). These uracils will then be misread as thymines (T) while sequencing leading to C-T nucleotide misincorporations. This is primaryly happening at the end of the reads. Modified from (Peyrégne and Prüfer, 2020).

Currently, a handfull of different methods for quantifying ancient DNA damage exists. In particular, the mapDamage 2.0 software has been the gold standard for how to measure ancient DNA damage in the field (Jónsson et al., 2013), however, it uses slow algorithms leading to unfeasible runtimes for large datasets. Newer, faster methods are being developed all of the time, such as PyDamage (Borry et al., 2021) which tackle some of mapDamage's limitations, although even faster methods suited at metagenomic analysis for large-scale datasets are still lacking.

In Paper I, see Chapter 2, we introduce the `metaDMG` software which utilizes the C-to-T deamination pattern[1] to identify ancient DNA damage. One of the key features of this method is the beta-binomial model which allows the uncertainty to fitted independently of the mean leading to improved accuracy of the damage estimation. Since the data is based on misincorporation counts, in particular the number of C-to-T transitions, k, out of N total C's, the classical likelihood to use

1 for the forward strand and the G-to-A deamination pattern for the reverse strand

for this type of data is a binomial distribution. The mean and variance of the binomial distribution is given by:

$$\begin{aligned} \mathbb{E}[k] &= Np \\ \mathbb{V}[k] &= Np(1-p), \end{aligned} \tag{1}$$

where p is the probability of success (a C-to-T substitution). One of the issues, however, is that the variance of the binomial distribution is proportional to the mean. The binomial distribution is thus not flexible enough to accommodate large amounts of variance in the data, so-called overdispersion (McElreath, 2020). One way to accommodate overdispersion is to instead use a beta-binomial model. The beta-binomial model is a generalization of the binomial distribution where the variance is allowed to be flexible. Technically, the beta-binomial model assumes that p is a random variable which follows a beta distribution $p \sim \text{Beta}(\mu, \varphi)$ where the beta distribution is parameterized[2] in terms of its mean, μ, and dispersion parameter, φ, (Cepeda-Cuervo and Cifuentes-Amado, 2017). The mean and variance of this beta-binomial model is then given by:

2 This can be reparameterization in term of the classical α, β parameterization by: $\mu = \alpha/(\alpha + \beta)$ and $\varphi = \alpha + \beta$.

$$\begin{aligned} \mathbb{E}[k] &= N\mu \\ \mathbb{V}[k] &= N\mu(1-\mu)\frac{\varphi + N}{\varphi + 1}. \end{aligned} \tag{2}$$

Comparing Equation 1 and Equation 2, we see that the variance of the beta-binomial model is no longer (strictly) proportional to the mean, but instead is a function of the dispersion parameter, φ, allowing for higher variance than the binomial-only model. When $\varphi = 0$, the variance of the beta-binomial model is N times larger, and when $\varphi \to \infty$ the variance reduces to the variance of the binomial model, showing that the beta-binomial model is a generalization of the binomial model.

Equation 2 shows how we model the C-to-T damage at a specific base position in the read. We model the position-dependent damage frequency, $f(x) = k(x)\ N(x)$, see Figure 1, as a function of the distance from the end of the read, x, with an exponential decay:

$$f(x; A, q, c) = A(1-q)^{x-1} + c. \tag{3}$$

Here A is the scale factor, or amplitude, q is the decay rate, and c is a constant offset, the baseline damage. Since x is discrete, this is similar to a (modified) geometric sequence starting from $x = 1$. The combination of Equation 2 and Equation 3 is illustrated in Figure 2, which shows the position-dependent, decreasing damage

frequency. The figure also shows the increase in uncertainty in the beta-binomial model compared to the binomial-only model.

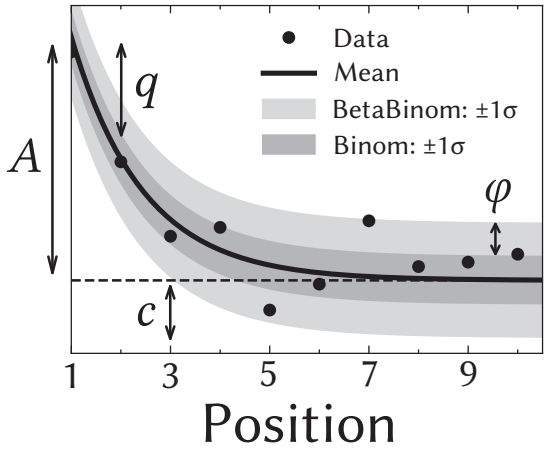


Figure 2.
Illustration of the damage model. The figure shows data points as circles and the damage frequency, $f(x)$, as a solid line. The amplitude of the damage is A, the offset is c, and the relative decrease in damage pr. position is given by q. The damage uncertainty for a binomial model is shown in dark grey and the uncertainty for a beta-binomial model in light grey.

The damage framework described above is based on the nucleotide misincorporations, i.e. the C-to-T transitions. The background for this data can be from either sequence files (fasta or fastq files) mapped to a single genome or from metagenomic data consisting of multiple mapped reads. As such, the damage framework is a general tool for estimating damage. In the metagenomic case where single DNA reads are mapped to multiple species, `metaDMG` performs a simple lowest common ancestor based on the ngsLCA algorithm (Wang et al., n.d.). This means that for each read that map to multiple reference, i.e. has multiple alignments, the taxonomic tree is traversed for each alignment until a common ancestor is found. This is the so-called lowest common ancestor (LCA). Figure 3 illustrates the LCA for a read that maps to different (sub)species. In this example, the LCA of alignment 1 and 2 is the Subspecies I while the LCA for all four alignments is the Genus X. `metaDMG` works by default with the NCBI taxanomic database but can also be used with custom databases.

Given the nucleotide misincorporations, either coming from a single-reference alignment file or after LCA in the metagenomic case, we fit eq. (2) and (3) with a Bayesian model. This is done to ensure the optimal inference of the parameters, A, q, and c, and to account for the uncertainty in the data. Bayesian inference also allows for the inclusion of domain knowledge in the form of the prior distribution by Bayes theorem. Bayes theorem is based on the law of conditional probability (Barlow, 1993) stating that the probability of two events, A and B, both happening,

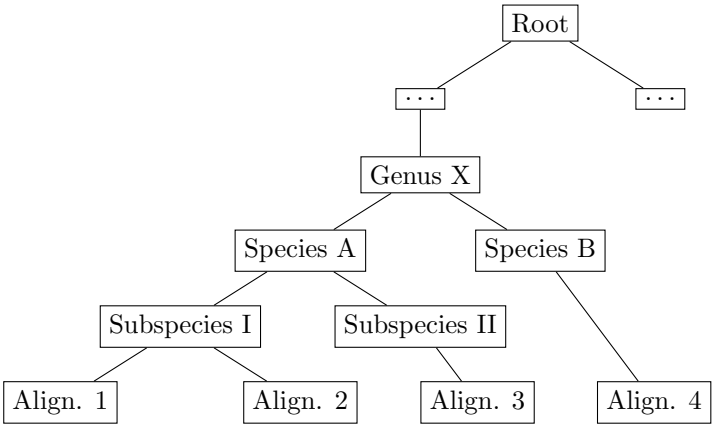


Figure 3.
Illustration of the lowest common ancestor (LCA) for taxonomic trees. Here the LCA of alignment 1 and 2 is Subspecies I, while the LCA of all four reads is Genus X. The dots (…) refers to other taxonomic levels, e.g. family and order.

$P(A \cap B)$, is given by the probability of B, $P(B)$ times the probability of A given B, $P(A|B)$:

$$P(A \cap B) = P(B)P(A|B). \tag{4}$$

Similarly, $P(A \cap B)$ can also be expressed in terms of the probability of A:

$$P(A \cap B) = P(A)P(B|A). \tag{5}$$

Combining Equation 4 and Equation 5 and rearranging terms gives the Bayes theorem:

$$P(\theta|x) = \frac{P(\theta)P(x|\theta)}{P(x)}, \tag{6}$$

with a change of variables where x now refers to the observed data and θ the parameter(s) of the model. The first term in the numerator, $P(\theta)$, is the prior distribution and describes the probability distribution assigned to θ before observing any data. The second term is the likelihood function, $P(x|\theta)$, which is the probability of observing the data, x, given the parameter(s), θ. Together these two terms combine to a compromise between data and prior information.

The numerator, $P(x)$, also known as the evidence, can be treated as a data-related normalization factor. In the case of continuous θ, this can calculated as the marginalization of the likelihood function over θ:

$$P(x) = \int_\theta P(x|\theta)P(\theta)\,\mathrm{d}\theta. \tag{7}$$

This equation, however, is often intractable to compute in the higher-dimensional case. Luckily, it can be shown that Markov Chain Monte Carlo (MCMC) sampling can approximate the posterior distribution, $P(\theta|x)$, and asymptotically converge to the correct distribution (Gelman et al., 2015).

Traditionally MCMC methods such as Metropolis Hastings (MH) or Gibbs sampling have been used for Bayesian inference, however, these methods are often slow and require a lot of tuning. In the last decades, a new class of MCMC methods have been developed, namely Hamiltonian Monte Carlo (HMC) methods. While traditional MH uses a Gaussian random walk, HMC is a gradient-based MCMC method that uses Hamiltonian dynamics to guide the sampling. This makes HMC more efficient than traditional MCMC methods and allows for sampling from high-dimensional distributions (Betancourt, 2018; Neal, 2011). A particularly efficient variant of HMC is the No-U-Turn Sampler (NUTS). NUTS is a variant of HMC that automatically tunes the step size and number of steps to take in the Hamiltonian dynamics (Homan and Gelman, 2014).

Most statistical domain-specific languages (DSL) such as Stan (Carpenter et al., 2017), Pyro (Bingham et al., 2019), NumPyro (Phan, Pradhan, and Jankowiak, 2019) or Turing.jl (Ge, Xu, and Ghahramani, 2018), implement HMC and in particular the NUTS algorithm. Since `metaDMG` is implemented in Python, we use NumPyro for the Bayesian inference of the damage model as it is easy to implement and computationally efficient since it which uses JAX (Bradbury et al., 2018) under the hood for automatic differentiation and just-in-time (JIT) compilation.

Even though NumPyro is fast and `metaDMG` is efficiently implemented, the Bayesian inference of the damage model is still computationally expensive. Thus, we have decided to also include a faster, approximate method of Bayesian inference: the maximum a posteriori (MAP) estimate. The MAP estimate is the point estimate of the posterior distribution that maximizes the posterior probability density function, i.e. the posterior mode:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} P(\theta|x) = \arg\max_{\theta} P(\theta)P(x|\theta), \tag{8}$$

where the second equality is due to the evidence being independent of θ. Since this is a point estimate, $\hat{\theta}_{\text{MAP}}$ does not fully explain the full posterior, however, it is often a good approximation*. Comparing $\hat{\theta}_{\text{MAP}}$ to the maximum likelihood estimate (MLE):

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} P(x|\theta), \tag{9}$$

* Especially when the posterior is unimodal, which it generally is in the case of `metaDMG`.

the MAP estimate can be seen as a regularized version of the MLE estimate (Murphy, 2012). To further optimize the computational efficacy of the MAP estimation in `metaDMG`, we JIT compile the MAP estimation function using Numba (Lam, Pitrou, and Seibert, 2015) and mathematically optimize the function with iMinuit (Dembinski et al., 2021).

One of the limitations of the `metaDMG` software is XXX.

1.2 *Anestesiology – a Machine Learning Approach*

1.3 *COVID-19 and Agent Based Models*

asdasdasdads

1.4 *Diffusion Models and Bayesian Model Comparison*

asdasdas

Bibliography

Barlow, R. J. (1993). *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*. Chichester, England ; New York: Wiley. 222 pp. ISBN: 978-0-471-92295-7.

Betancourt, Michael (2018). "A Conceptual Introduction to Hamiltonian Monte Carlo". arXiv: 1701.02434 [stat].

Bingham, Eli et al. (2019). "Pyro: Deep Universal Probabilistic Programming". In: *Journal of Machine Learning Research* 20, 28:1–28:6. URL: http://jmlr.org/papers/v20/18-403.html.

Borry, Maxime et al. (2021). "PyDamage: Automated Ancient Damage Identification and Estimation for Contigs in Ancient DNA de Novo Assembly". In: *PeerJ* 9, e11845. ISSN: 2167-8359. DOI: 10.7717/peerj.11845.

Bradbury, James et al. (2018). *JAX: Composable Transformations of Python NumPy Programs*. Version 0.2.5. URL: http://github.com/google/jax.

Briggs, Adrian W. et al. (2007). "Patterns of Damage in Genomic DNA Sequences from a Neandertal". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.37, pp. 14616–14621. ISSN: 0027-8424. DOI: 10.1073/pnas.0704665104. pmid: 17715061.

Carpenter, Bob et al. (2017). "Stan: A Probabilistic Programming Language". In: *Journal of statistical software* 76.1.

Cepeda-Cuervo, Edilberto and MARíA VICTORIA Cifuentes-Amado (2017). "Double Generalized Beta-Binomial and Negative Binomial Regression Models". In: *Revista Colombiana de Estadística* 40.1, pp. 141–163. ISSN: 0120-1751. DOI: 10.15446/rce.v40n1.61779.

Dabney, Jesse, Matthias Meyer, and Svante Pääbo (2013). "Ancient DNA Damage". In: *Cold Spring Harbor Perspectives in Biology* 5.7, a012567. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a012567. pmid: 23729639.

Dembinski, Hans et al. (2021). *Scikit-Hep/Iminuit: V2.8.2*. Version v2.8.2. Zenodo. DOI: 10.5281/ZENODO.3949207.

Ge, Hong, Kai Xu, and Zoubin Ghahramani (2018). "Turing: A Language for Flexible Probabilistic Inference". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1682–1690. URL: https://proceedings.mlr.press/v84/ge18b.html (visited on 2022).

Gelman, Andrew et al. (2015). *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC. 675 pp. ISBN: 978-0-429-11307-9. DOI: 10.1201/b16018.

Homan, Matthew D. and Andrew Gelman (2014). "The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *The Journal of Machine Learning Research* 15.1, pp. 1593–1623. ISSN: 1532-4435.

Jónsson, Hákon et al. (2013). "mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters". In: *Bioinformatics* 29.13, pp. 1682–1684. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btt193.

Karolinska Institutet, The Nobel Assembly at (2022). *The Nobel Prize in Physiology or Medicine 2022*. NobelPrize.org. URL: https://www.nobelprize.org/prizes/medicine/2022/press-release/ (visited on 2022).

Krause, Johannes et al. (2010). "The Complete Mitochondrial DNA Genome of an Unknown Hominin from Southern Siberia". In: *Nature* 464.7290 (7290), pp. 894–897. ISSN: 1476-4687. DOI: 10.1038/nature08976.

Lam, Siu Kwan, Antoine Pitrou, and Stanley Seibert (2015). "Numba: A LLVM-based Python JIT Compiler". In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*. LLVM '15. New York, NY, USA: Association for Computing Machinery, pp. 1–6. ISBN: 978-1-4503-4005-2. DOI: 10.1145/2833157.2833162. URL: https://github.com/numba/numba.

Martiniano, Rui et al. (2020). "Removing Reference Bias and Improving Indel Calling in Ancient DNA Data Analysis by Mapping to a Sequence Variation Graph". In: *Genome Biology* 21.1, p. 250. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02160-7.

McElreath, Richard (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd ed. CRC Texts in Statistical Science. Boca Raton: Taylor and Francis, CRC Press. ISBN: 978-0-367-13991-9.

Mullis, K. et al. (1986). "Specific Enzymatic Amplification of DNA in Vitro: The Polymerase Chain Reaction". In: *Cold Spring Harbor Symposia on Quantitative Biology* 51 Pt 1, pp. 263–273. ISSN: 0091-7451. DOI: 10.1101/sqb.1986.051.01.032. pmid: 3472723.

Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press. ISBN: 0-262-01802-0.

Neal, Radford M. (2011). *MCMC Using Hamiltonian Dynamics*. Routledge Handbooks Online. ISBN: 978-1-4200-7941-8 978-1-4200-7942-5. DOI: 10.1201/b10905-7. URL: https://www.routledgehandbooks.com/doi/10.1201/b10905-7 (visited on 2022).

Nielsen, Rasmus et al. (2011). "Genotype and SNP Calling from Next-Generation Sequencing Data". In: *Nature reviews. Genetics* 12.6, pp. 443–451. ISSN: 1471-0056. DOI: 10.1038/nrg2986. pmid: 21587300.

Orlando, Ludovic et al. (2013). "Recalibrating Equus Evolution Using the Genome Sequence of an Early Middle Pleistocene Horse". In: *Nature* 499.7456 (7456), pp. 74–78. ISSN: 1476-4687. DOI: 10.1038/nature12323.

Pääbo, Svante (1985a). "Molecular Cloning of Ancient Egyptian Mummy DNA". In: *Nature* 314.6012 (6012), pp. 644–645. ISSN: 1476-4687. DOI: 10.1038/314644a0.

— (1985b). "Preservation of DNA in Ancient Egyptian Mummies". In: *Journal of Archaeological Science* 12.6, pp. 411–417. ISSN: 0305-4403. DOI: 10.1016/0305-4403(85)90002-0.

Peyrégne, Stéphane and Kay Prüfer (2020). "Present-Day DNA Contamination in Ancient DNA Datasets". In: *BioEssays* 42.9, p. 2000081. ISSN: 1521-1878. DOI: 10.1002/bies.202000081.

Phan, Du, Neeraj Pradhan, and Martin Jankowiak (2019). "Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro". arXiv: 1912.11554 [cs, stat].

Renaud, Gabriel et al. (2019). "Authentication and Assessment of Contamination in Ancient DNA". In: *Ancient DNA: Methods and Protocols*. Ed. by Beth Shapiro et al. Methods in Molecular Biology. New York, NY: Springer, pp. 163–194. ISBN: 978-1-4939-9176-1. DOI: 10.1007/978-1-4939-9176-1_17.

Schubert, Mikkel et al. (2012). "Improving Ancient DNA Read Mapping against Modern Reference Genomes". In: *BMC Genomics* 13, p. 178. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-178. pmid: 22574660.

Slatko, Barton E., Andrew F. Gardner, and Frederick M. Ausubel (2018). "Overview of Next Generation Sequencing Technologies". In: *Current protocols in molecular biology* 122.1, e59. ISSN: 1934-3639. DOI: 10.1002/cpmb.59. pmid: 29851291.

Van der Valk, Tom et al. (2021). "Million-Year-Old DNA Sheds Light on the Genomic History of Mammoths". In: *Nature* 591.7849 (7849), pp. 265–269. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03224-9.

Wang, Yucheng et al. (n.d.). "ngsLCA—A Toolkit for Fast and Flexible Lowest Common Ancestor Inference and Taxonomic Profiling of Metagenomic Data". In: *Methods in Ecology and Evolution* n/a.n/a (). ISSN: 2041-210X. DOI: 10.1111/2041-210X.14006.

2 *Paper I*

The following pages contain the article:

Christian S. Michelsen, Mikkel W. Pedersen, Antonio Fernandez-Guerra, Lei Zhao, Troels C. Petersen, Thorfinn S. Korneliussen (2022). “metaDMG: An Ancient DNA Damage Toolkit”.

3 *Paper II*

The following pages contain the article:

Christian Michelsen, Christoffer C. Jorgensen, Mathias Heltberg, Mogens H. Jensen, Alessandra Lucchetti, Pelle B. Petersen, Troels C. Petersen, Henrik Kehlet (2022). “Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty - a machine learning based approach.”.

4 *Paper III*

The following pages contain the article:

Mathias S. Heltberg, Christian Michelsen, Emil S. Martiny, Lasse E. Christensen, Mogens H. Jensen, Tariq Halasa and Troels C. Petersen (2022). "Spatial Heterogeneity Affects Predictions from Early-Curve Fitting of Pandemic Outbreaks: A Case Study Using Population Data from Denmark". In: Royal Society Open Science 9.9. issn: 2054-5703. doi: 10.1098/rsos.220018.

5 *Paper IV*

The following pages contain the article:

Susmita Sridar, Mathias S. Heltberg, Christian S. 6 Michelsen Judith M. Hattab, Angela Taddei (2022). “Microscopic single molecule dynamics suggest underlying physical properties of the silencing foci”.

APPENDIX

A *Kap København*

The following pages contain the paper published in Nature 2022:

> Kurt H. Kjær, Mikkel W. Pedersen, Bianca De Sanctis, Binia De Cahsan, Thorfinn S. Korneliussen, **Christian Michelsen**, Karina K. Sand, Stanislav Jelavić, Anthony H. Ruter, Astrid M. Z. Bonde, Kristian K. Kjeldsen, Alexey S. Tesakov, Ian Snowball, John C. Gosse, Inger G. Alsos, Yucheng Wang, Christoph Dockter, Magnus Rasmussen, Morten E. Jørgensen, Birgitte Skadhauge, Ana Prohaska, Jeppe Å. Kristensen, Morten Bjerager, Morten E. Allentoft, Eric Coissac, PhyloNorway Consortium, Alexandra Rouillard, Alexandra Simakova, Antonio Fernandez-Guerra, Chris Bowler, Marc Macias-Fauria, Lasse Vinner, John J. Welch, Alan J. Hidy, Martin Sikora, Matthew J. Collins, Richard Durbin, Nicolaj K. Larsen & Eske Willerslev, *"A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA"* (Published in Nature, 2022, doi: 10.1038/s41586-022-05453-y).

The paper use the `metaDMG` tool to identify ancient species and classify the amount of ancient damage in these species. This shows, that modern modern statistical methods combined with excellent work in the ancient DNA labs can provide new insights into the past – even on data that are more than two millions years old.

Article

A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA

https://doi.org/10.1038/s41586-022-05453-y

Received: 30 September 2021

Accepted: 18 October 2022

Open access

Check for updates

Kurt H. Kjær[1,23]✉, Mikkel W. Pedersen[1,23], Bianca De Sanctis[2,3], Binia De Cahsan[4], Thorfinn S. Korneliussen[1], Christian S. Michelsen[1,5], Karina K. Sand[1], Stanislav Jelavić[1,6], Anthony H. Ruter[1], Astrid M. Z. Bonde[7], Kristian K. Kjeldsen[8], Alexey S. Tesakov[9], Ian Snowball[10], John C. Gosse[11], Inger G. Alsos[12], Yucheng Wang[1,2], Christoph Dockter[13], Magnus Rasmussen[13], Morten E. Jørgensen[13], Birgitte Skadhauge[13], Ana Prohaska[1], Jeppe Å. Kristensen[9,14], Morten Bjerager[9], Morten E. Allentoft[1,15], Eric Coissac[12,16], PhyloNorway Consortium*,**, Alexandra Rouillard[1,17], Alexandra Simakova[9], Antonio Fernandez-Guerra[1], Chris Bowler[18], Marc Macias-Fauria[19], Lasse Vinner[1], John J. Welch[3], Alan J. Hidy[20], Martin Sikora[1], Matthew J. Collins[21], Richard Durbin[3], Nicolaj K. Larsen[1] & Eske Willerslev[1,2,22]✉

Late Pliocene and Early Pleistocene epochs 3.6 to 0.8 million years ago[1] had climates resembling those forecasted under future warming[2]. Palaeoclimatic records show strong polar amplification with mean annual temperatures of 11–19 °C above contemporary values[3,4]. The biological communities inhabiting the Arctic during this time remain poorly known because fossils are rare[5]. Here we report an ancient environmental DNA[6] (eDNA) record describing the rich plant and animal assemblages of the Kap København Formation in North Greenland, dated to around two million years ago. The record shows an open boreal forest ecosystem with mixed vegetation of poplar, birch and thuja trees, as well as a variety of Arctic and boreal shrubs and herbs, many of which had not previously been detected at the site from macrofossil and pollen records. The DNA record confirms the presence of hare and mitochondrial DNA from animals including mastodons, reindeer, rodents and geese, all ancestral to their present-day and late Pleistocene relatives. The presence of marine species including horseshoe crab and green algae support a warmer climate than today. The reconstructed ecosystem has no modern analogue. The survival of such ancient eDNA probably relates to its binding to mineral surfaces. Our findings open new areas of genetic research, demonstrating that it is possible to track the ecology and evolution of biological communities from two million years ago using ancient eDNA.

The Kap København Formation is located in Peary Land, North Greenland (82° 24′ N 22° 12′ W) in what is now a polar desert. The upper depositional sequence contains well-preserved terrestrial animal and plant remains washed into an estuary during a warmer Early Pleistocene interglacial cycle[7] (Fig. 1). Nearly 40 years of palaeoenvironmental and climate research at the site provide a unique perspective into a period when the site was situated at the boreal Arctic ecotone with reconstructed summer and winter average minimum temperatures of 10 °C and −17 °C respectively—more than 10 °C warmer than the present[7–11]. These conditions must have driven substantial ablation of the Greenland Ice Sheet, possibly producing one of the last ice-free intervals[7] in the last 2.4 million years (Myr). Although the Kap København Formation is known to yield well-preserved macrofossils from a coniferous boreal forest and a rich insect fauna, few traces of vertebrates have been found. To date, these comprise remains from lagomorph genera, their coprolites and *Aphodius* beetles, which live in and on mammalian dung[10,11]. However, the approximately 3.4 Myr old Fyles Leaf bed and Beaver Pond on Ellesmere Island in Arctic Canada preserve fossils of

[1]Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark. [2]Department of Zoology, University of Cambridge, Cambridge, UK. [3]Department of Genetics, University of Cambridge, Cambridge, UK. [4]Section for Evolutionary Genomics, Faculty of Health and Medical Sciences, The Globe Institute, Copenhagen K, Denmark. [5]Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark. [6]Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, IRD, Université Gustave Eiffel, ISTerre, Grenoble, France. [7]Halsnaes Kommune, Frederiksværk, Denmark. [8]GEUS, Geological Survey of Denmark and Greenland, Copenhagen K, Denmark. [9]Geological Institute, Russian Academy of Sciences, Moscow, Russia. [10]Department of Earth Sciences, Uppsala University, Uppsala, Sweden. [11]Department of Earth and Environmental Sciences, Dalhousie University, Halifax, Canada. [12]The Arctic University Museum of Norway, UiT—The Arctic University of Norway, Tromsø, Norway. [13]Carlsberg Research Laboratory, Copenhagen V, Denmark. [14]Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, UK. [15]Trace and Environmental DNA (TrEnD) Laboratory, School of Molecular and Life Sciences, Curtin University, Perth, Western Australia, Australia. [16]University of Grenoble-Alpes, Université Savoie Mont Blanc, CNRS, LECA, Grenoble, France. [17]Department of Geosciences, UiT—The Arctic University of Norway, Tromsø, Norway. [18]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM Université PSL, Paris, France. [19]School of Geography and the Environment, University of Oxford, Oxford, UK. [20]Center for Accelerator Mass Spectrometry, Lawrence Livermore National Laboratory, Livermore, CA, USA. [21]Department of Archaeology, University of Cambridge, Cambridge, UK. [22]MARUM, University of Bremen, Bremen, Germany. [23]These authors contributed equally: Kurt H. Kjær, Mikkel W. Pedersen. *A list of authors and their affiliations appears at the end of the paper. A full list of members and their affiliations appears in the Supplementary Information. ✉e-mail: kurtk@sund.ku.dk; ew482@cam.ac.uk

Article

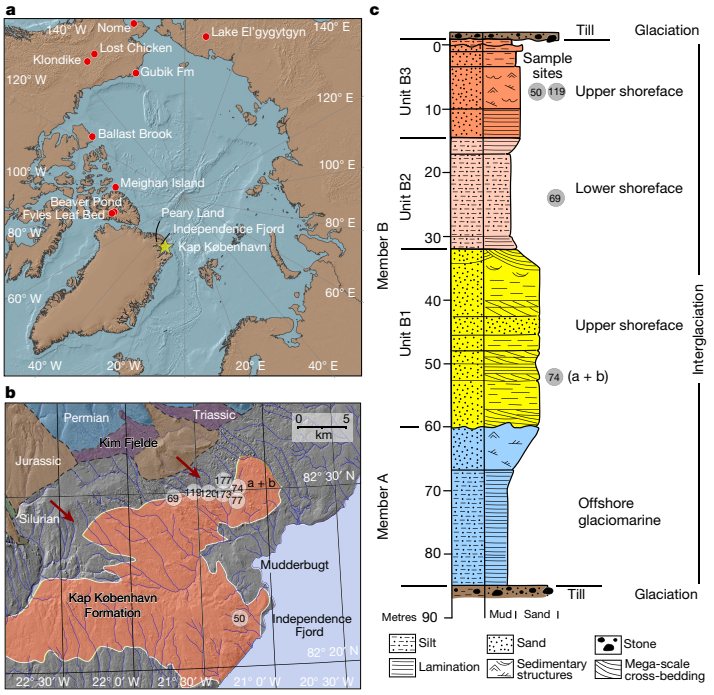


Fig. 1 | Geographic location and depositional sequence. a. Location of Kap København Formation in North Greenland at the entrance to the Independence Fjord (82° 24′ N 22° 12′ W) and locations of other Arctic Plio-Pleistocene fossil-bearing sites (red dots). **b**, Spatial distribution of the erosional remnants of the 100-m thick succession of shallow marine near-shore sediments between Mudderbugt and the low mountains towards the north. **c**, Glacial–interglacial division of the depositional succession of clay Member A and units B1, B2 and B3 constituting sandy Member B. Sampling intervals for all sites are projected onto the sedimentary succession of locality 50. Sedimentological log modified after ref. [7]. Circled numbers on the map mark sample sites for environmental DNA analyses, absolute burial dating and palaeomagnetism. Numbered sites refer to previous publications[7,10,11,14,98].

mammals that potentially could have colonized Greenland, such as the extinct bear (*Protarctos abstrusus*), giant beavers (*Dipoides* sp.), the small canine *Eucyon* and Arctic giant camelines[4,12,13] (similar to *Paracamelus*). Whether the Nares Strait was a sufficient barrier to isolate northern Greenland from colonization by this fauna remains an open question.

The Kap København Formation is formally subdivided into two members[7] (Fig. 1). The lower Member A consists of up to 50 m of laminated mud with an Arctic ostracod, foraminifera and mollusc fauna deposited in an offshore glaciomarine environment[14]. The overlying Member B consists of 40–50 m of sandy (units B1 and B3) and silty (unit B2) deposits, including thin organic-rich beds with an interglacial macrofossil fauna that were deposited closer to the shore in a shallow marine or estuarine environment represented by upper and lower shoreface sedimentary facies[7].

The specific depositional environments are also reflected in the mineralogy of the units, where the proximal B3 locality has the lowest clay and highest quartz contents (Sample compositions in Supplementary Tables 4.2.1 and 4.2.2 and unit averages in Supplementary Tables 4.2.3 and 4.2.4). The architecture of the basin infill suggests that Member B units thicken towards the present coast—that is, distal to the sediment source in the low mountains in the north (Fig. 1). Abundant organic detritus horizons are recorded in units B1 and B3, which also contain beds rich in arctic and boreal plant and invertebrate macrofossils, as well as terrestrial mosses[10,15]. Therefore, the taphonomy of the DNA most probably reflects the biological communities eroded from a range of habitats, fluvially transported to the foreshore and concentrated as organic detritus mixed into sandy near-shore sediments within units B1 and B3. Conversely, the deeper water facies from Member A and unit B2 have a stronger marine signal. This scenario is supported by the similarities in the mineralogic composition between Kap København Formation sediments and Kim Fjelde sediments (Supplementary Tables 4.2.1 and 4.2.5).

Geological age

A series of complementary studies has successively narrowed the depositional age bracket of the Kap København Formation from 4.0–0.7 million years ago (Ma) to a 20,000-year-long age bracket around 2.4 Ma (see Supplementary Information, sections 1–3). This was achieved by a combination of palaeomagnetism, biostratigraphy and allostratigraphy[7,14,16–18]. Notably, the last appearance data of the mammals, foraminifera and molluscs in the stratigraphic record show an age close to 2.4 Myr (see Supplementary Information, section 2). Within this overall framework, we add new palaeomagnetic data showing that Member A has reversed magnetic polarity and the main part of the overlying unit B2 has normal magnetic polarity. In the context of previous work, this is consistent with three magnetostratigraphic intervals in the Early Pleistocene where there is a reversal: 1.93 Myr (scenario 1), 2.14 Myr (scenario 2) or 2.58 Myr (scenario 3) (Supplementary

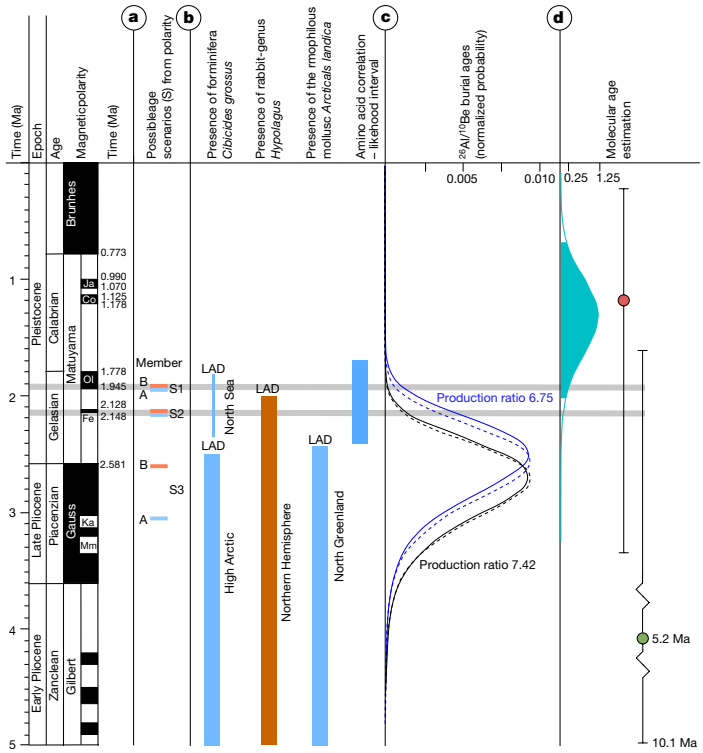


Fig. 2 | Age proxies for the Kap København Formation. a, Revised palaeomagnetic analysis shows unit B2 to have normal polarity and unlocks three possible age scenarios (S1–S3) including Members A (blue) and B (brown). Normal polarity is coloured black and reverse polarity is shown in white. Ja, Jaramillo; Co, Cobb Mountain; Ol, Olduvai; Fe, Feni; Ka, Kaena; Mm, Mammoth. **b**, Presence and last appearance datum (LAD) for marine foraminifera *Cibicides grossus*, rabbit-genus *Hypolagus* and the mollusc *Arctica islandica* in the High Arctic, Northern Hemisphere and North Greenland, respectively. The blue band on the far right indicates the age range for Member A estimated from amino acid ratios on shells[7]. **c**, Convolved probability distribution functions for cosmogenic burial ages calculated for two different production ratios (7.42 (black) and 6.75 (blue)). The dashed line and the solid line show the distributions for steady erosion and zero erosion, respectively. These distributions are all maximum ages. **d**, Molecular dating of *Betula* sp. yielding a median age of the DNA in the sediment of 1.323 Myr, with whiskers confining the 95% height posterior density (HPD) of 0.68 to 2.02 Myr (blue density plot), running Markov chain Monte Carlo estimation over for 100 million iterations. The red dot is the median molecular age estimate found using the Mastodon mitochondrial genome restricting to radiocarbon-dated specimens, whereas the green area includes molecular clock estimated specimens in BEAST, running Markov chain Monte Carlo estimation for 400 million iterations. Whiskers confine the 95% HPD.

Information, section 1). Furthermore, we constrain the age using cosmogenic ^{26}Al:^{10}Be burial dating of Member B at four sites in this study (Supplementary Information, section 3). The recommended maximum burial age for the Kap København Formation is 2.70 ± 0.46 Myr (Fig. 2; Methods). However, we discard the older scenario 3 as it contradicts the evidence for a continuous sedimentation across Members A and B during a single glacial–interglacial depositional cycle[7,14,16,18,19]. This leaves two possible scenarios (scenarios 1 and 2), in which scenario 1 supports an age of 1.9 Myr and scenario 2 supports an age of 2.1 Myr.

DNA preservation

DNA degrades with time owing to microbial enzymatic activity, mechanical shearing and spontaneous chemical reactions such as hydrolysis and oxidation[20]. The oldest known DNA obtained to date has been recovered from a permafrost-preserved mammoth molar dated to 1.2–1.1 Ma using geological methods and 1.7 Ma (95% highest posterior density, 2.1–1.3 Ma) using molecular clock dating[21]. To explore the likelihood of recovering DNA from sediments at the Kap København formation, we calculated the thermal age of the DNA and its expected degree of depurination at the Kap København Formation. Using the mean average temperature[22] (MAT) of −17 °C, we found a thermal age of 2.7 $kyr_{DNA@10\,°C}$—that is, 741 times less than the age of 2.0 Myr (Supplementary Information, section 4 and Supplementary Table 4.4.1). Using the rate of depurination from Moa bird fossils[23], we found it plausible that DNA with an average size of 50 base pairs (bp) could survive at the Kap København Formation, assuming that the site remained frozen (Supplementary Information, section 4 and Supplementary Table 4.4.2). Mechanisms that preserve DNA in sediments are likely to be different from that of bone. Adsorption at mineral surfaces modifies the DNA conformation, probably impeding molecular recognition by enzymes, which effectively hinders enzymatic degradation[24–27]. To investigate whether the minerals found in Kap København Formation could have retained DNA during the deposition and preserved it, we determined the mineralogic composition of the sediments using X-ray diffraction and measured their adsorption capacities. Our findings

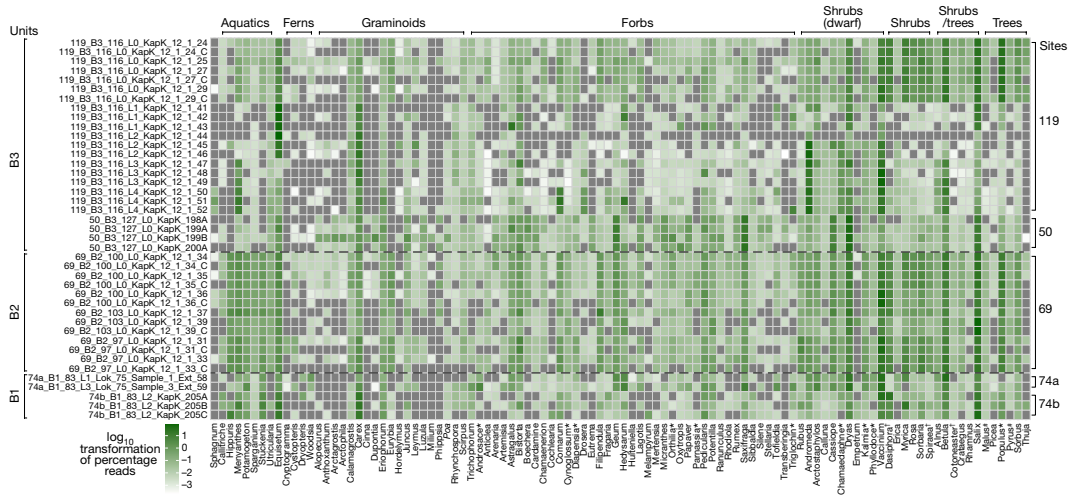


Fig. 3 | Early Pleistocene plants of Northern Greenland. Metagenomic taxonomic profiles of the plant assemblage. Taxa in bold are genera only found as DNA and not as macrofossil or pollen. Asterisks indicate those that are found at other Pliocene arctic sites. Extinct species as identified by either macrofossils or phylogenetic placements are marked with a dagger. Reads classified as *Pyrus* and *Malus* are marked with a pound symbol, and are probably over-classified DNA sequences belonging to another species within Rosaceae that are not present as a reference genome.

highlight that the marine depositional environment favours adsorption of extracellular DNA on the mineral surfaces (Supplementary Information, section 4 and Supplementary Table 4.3.1.1). Specifically, the clay minerals (9.6–5.5 wt%) and particularly smectite (1.2–3.7 wt%), have higher adsorption capacity compared to the non-clay minerals (59–75 wt%). At a DNA concentration representative of the natural environments[28] (4.9 ng ml^{-1} DNA), the DNA adsorption capacity of smectite is 200 times greater than for quartz. We applied a sedimentary eDNA extraction protocol[29] on our mineral-adsorbed DNA samples, and retrieved only 5% of the adsorbed DNA from smectite and around 10% from the other clay minerals (Methods and Supplementary Information, section 4). By contrast, we retrieved around 40% of the DNA adsorbed to quartz. The difference in adsorption capacity and extraction yield from the different minerals demonstrates that mineral composition may have an important role in ancient eDNA preservation and retrieval.

Kap København metagenomes

We extracted DNA[29] from 41 organic-rich sediment samples at five different sites within the Kap København Formation (Supplementary Information, section 6 and Source Data 1), which were converted into 65 dual-indexed Illumina sequencing libraries[30]. First, we tested 34 of the 65 libraries for plant plastid DNA by screening for the conserved photosystem II D2 (*psbD*) gene using droplet digital PCR (ddPCR) with a gene-targeting primer and probe spanning a 39-bp region and a P7 index primer. Further, we screened for the *psbA* gene using a similar assay targeting the Poaceae (Methods and Supplementary Fig. 6.12.1). A clear signal in 31 out of 34 samples tested confirmed the presence of plant plastid DNA in these libraries (Source Data 1, sheets 5 and 6). Additionally, we subjected 34 of the 65 libraries to mammalian mtDNA capture enrichment using the Arctic PaleoChip 1.0[31] and shotgun sequenced all libraries (initial and captured) using the Illumina HiSeq 4000 and NovaSeq 6000. A total of 16,882,114,068 reads were sequenced, which after adaptor trimming, filtering for ≥30 bp and a minimum phred quality of 30 and duplicate removal resulted in 2,873,998,429 reads. These were analysed for *k*mer comparisons using simka[32] (Supplementary Information, section 6) and then parsed for taxonomic classification using competitive mapping with HOLI (https://github.com/miwipe/KapCopenhagen.git), which includes a recently published dataset of more than 1,500 genome skims of Arctic and boreal plant taxa[33,34] (Methods and Supplementary Information, section 6). Considering the age of the samples and thus the potential genetic distance to recent reference genomes, we allowed each read to have a similarity between 95–100% for it to be taxonomically classified using ngsLCA[35]. The metaDMG (v.0.14.0) program (https://metadmg-dev.github.io/metaDMG-core/index.html) was subsequently used to quantify and filter each taxonomic node for postmortem DNA damage for all the metagenomic samples (Methods). This method estimates the average damage at the termini position (D-max) and a likelihood ratio (λ-LR) that quantifies how much better the damage model (that is, more damage at the beginning of the read) fits the data compared with a null model (that is, a constant amount of damage; see Supplementary Information, section 6). We found the DNA damage to be highly increased, especially for eukaryotes (mean D-max = 40.7%). From this we set D-max ≥25% as a filtering threshold for a taxonomic node to be parsed for further downstream analysis as well as a λ-LR higher or equal to 1.5. We furthermore set a threshold requiring that the minimum number of reads per taxon exceeded the median of reads assigned across all taxa divided by two to filter for taxa in low abundance. Similarly, for a sample to be considered, the total number of reads for a sample had to exceed the median number of reads per sample divided by two, to filter for samples with fewest reads. Lastly, we filtered out taxa with fewer than three replicates and subsequently reads were normalized by conversion to proportions (Figs. 3 and 4a).

DNA, pollen and macrofossils comparison

Greenland's coasts extend from around 60° to 83° N and include bioclimatic zones from the subarctic to the northern polar desert[36,37]. There are 175 vascular plant genera native to Greenland, excluding historically introduced species[38–40]. Of these, 70 (40%) were detected

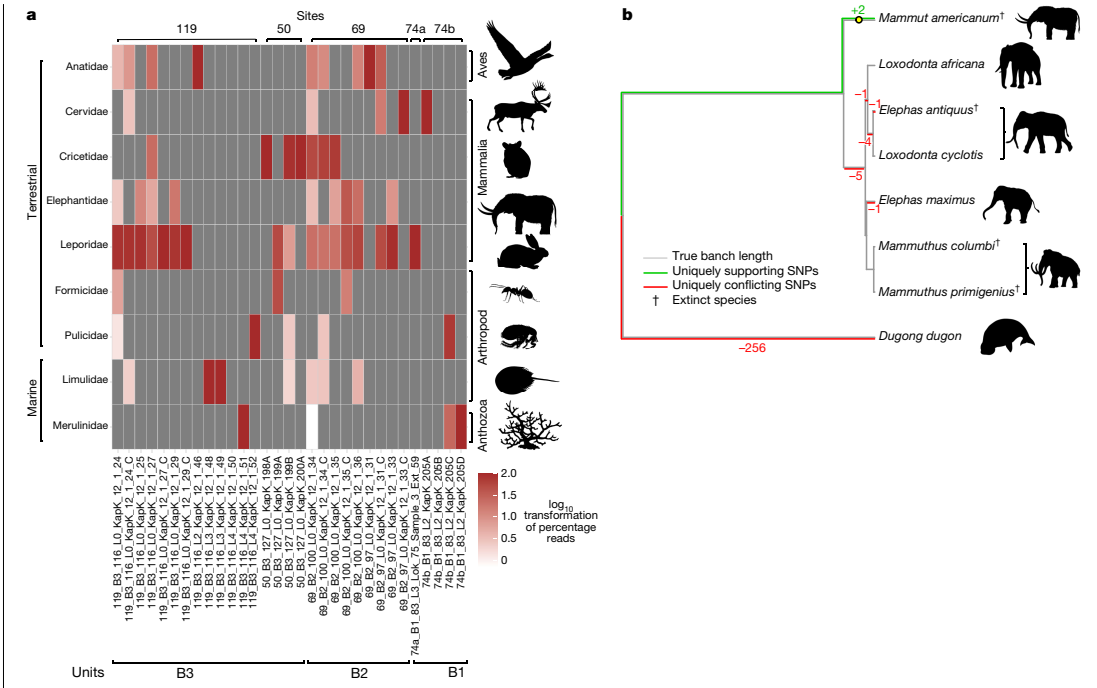


Fig. 4 | Early Pleistocene animals of Northern Greenland. a, Metagenomic taxonomic profiles of the animal assemblage from units B1, B2 and B3. Taxa in bold are genera only found as DNA. **b**, phylogenetic placement and pathPhynder[88] results of mitochondrial reads uniquely classified to Elephantidae or lower (Source Data 1).

by the metagenomic analysis (Fig. 3); the majority of these genera are today confined to bioclimatic zones well to the south of Kap København's polar desert (see ref. [41] and references therein), for example, all aquatic macrophytes. Reads assigned to *Salix*, *Dryas*, *Vaccinium*, *Betula*, *Carex* and *Equisetum* dominate the assemblage, and of these genera, *Equisetum*, *Dryas*, *Salix arctica* and two species of *Carex* (*Carex nardina* and *Carex stans*) grow there currently, whereas only a few records of *Vaccinium uliginosum* are found above 80° N, and *Betula nana* are found above 74° N (ref. [42]). Out of the 102 genera detected in the Kap København ancient eDNA assemblage, 39% no longer grow in Greenland but do occur in the North American boreal (for example, *Picea* and *Populus*) and northern deciduous and maritime forests (for example, *Crataegus*, *Taxus*, *Thuja* and *Filipendula*). Many of the plant genera in this diverse assemblage do not occur on permafrost substrates and require higher temperatures than those at any latitude on Greenland today.

In addition to the DNA, we counted pollen in six samples from locality 119, unit B3 (Methods and Supplementary Fig. 4.1.1). Percentages were calculated for 4 of the samples with pollen sums ranging from 71–225 terrestrial grains (mean = 170.25). Upland herbs, including taxa in the Cyperaceae, Ericales and Rosaceae comprised around 40% of sample 4. Samples 5 and 6 were dominated by arboreal taxa, particularly *Betula*. The Polypodiopsida (for example, *Equisetum*, *Asplenium* and *Athyrium filix-mas*) and Lycopodiopsida (*Lycopodium annotinum* and *Selaginella rupestris*) were also well represented and comprised over 30% of the assemblage in samples 1, 4 and 6.

A total of 39 plant genera out of the 102 identified by DNA also occurred as macrofossils or pollen at the genus level. A further 39 taxa were potentially identified as macrofossil or pollen but not to the same taxonomic level[10,15] (Source Data 1, sheets 1 and 2). For example, 12 genera of Poaceae were identified by DNA (*Alopecurus*, *Anthoxanthum*, *Arctagrostis*, *Arctophila*, *Calamagrostis*, *Cinna*, *Dupontia*, *Hordelymus*, *Leymus*, *Milium*, *Phippsia* and *Poa*), of these only *Hordelymus* is not found in the Arctic today (http://panarcticflora.org/), but these were only distinguished to family level in the pollen analysis and only one Poaceae macrofossil was found. There were 24 taxa that were recorded only as DNA. These included the boreal tree *Populus* and a few shrubs and dwarf shrubs, but mainly herbaceous plants. Of the 73 plant genera recovered as macrofossils[10,15], only 24 were not detected in the DNA analysis. Because macrofossils and DNA have similar taphonomies—as both are deposited locally—more overlap is expected between them than between DNA and pollen, which is typically dispersed regionally[43]. Nine of the taxa absent in DNA were bryophytes, probably owing to poor representation of this group within the genomic reference databases. Furthermore, the extinct taxon Araceae is not present in the reference databases. The remaining undetected genera were vascular plants, and all except two (*Oxyria* and *Cornus*) were rare in the macrofossil record. Because the detection of rare taxa is challenging in both macrofossil and DNA records[44], we argue that this overlap between the DNA and macrofossil records is as high as can be expected on the basis of the limitations of both methods.

An additional 19 taxa were recorded in the pollen record presented here and in that of Bennike[45] including four trees or shrubs, five ferns, three club mosses, and one each of algae, fungi and liverwort. We also find pollen from anemophilous trees, particularly gymnosperms, which can be distributed far north of the region where the plants actually grow[10]. Bennike[45] also notes a high proportion of club mosses and ferns and suggests they may be overrepresented owing to their spore wall

being resistant to degradation. Furthermore, if these taxa were preferentially distributed along streams flowing into the estuary, their spores could be relatively more concentrated in the alluvium than the pollen of more generally distributed taxa. Thus, both decay resistance and alluvial deposition could contribute to the relative frequencies we observe. This same alluvial dynamic might also have contributed to the very large read counts for *Salix*, *Betula*, *Populus*, *Carex* and *Equisetum* in the metagenomic record, implying that neither the proportion of these taxa in the pollen records nor read counts necessarily correlate with their actual abundance in the regional vegetation in terms of biomass or coverage.

Finally, we sought to date the age of the plant DNA by phylogenetic placement of the chloroplast DNA. We examined data for the genera *Betula*, *Populus* and *Salix*, because these had both sufficiently high chloroplast genome coverage (with mean depth 24.16×, 57.06× and 27.04×, respectively) and sufficient present-day whole chloroplast reference sequences (Methods). Owing to their age and hence potential genetic distance from the modern reference genomes, we lowered the similarity threshold of uniquely classified reads to 90% and merged these by unit to increase coverage. Both *Betula* and *Salix* placed basally to most of the represented species in the respective genera, and the *Populus* placement results showed support for a mixture of different species related to *P. trichocarpa* and *P. balsamifera* (Extended Data Figs. 7–9).

We used the *Betula* chloroplast reads for a molecular dating analysis, because they were placed confidently on a single edge of the phylogenetic tree (that is, not a mixture as in *Populus*), had a large number of reference sequences, and had high coverage in the ancient sample. We used BEAST[46] v1.10.4 to obtain a molecular clock date estimate for our ancient *Betula* chloroplast sample (see Methods, 'Molecular dating methods' for details). We included 31 modern *Betula* and one *Alnus* chloroplast reference sequences, used only sites that had a depth of at least 20 in the ancient sample, and included a previously estimated *Betula*–*Alnus* chloroplast divergence time[47] of 61.1 Myr for calibration of the root node. Our BEAST analysis was robust to both different priors on the age of the ancient sample, and to different nucleotide substitution models (Supplementary Fig. 10). This yielded a median age estimate of 1.323 Myr, with a 95% HPD of (0.6786, 2.0172) Myr (Fig. 2).

Animal DNA results

The metazoan mitochondrial and nuclear DNA record was much less diverse than that of the plants but contained one extinct family, one that is absent from Greenland today, and four vertebrate genera native to Greenland as well as representatives of four invertebrate families (Fig. 4a). Assignments were based on incomplete and variable representation of reference genomes, so we identified reads to family level, and only where sufficient mitochondrial reads were present, we refined the assignment to genus level by matching these into mitochondrial phylogenies based on more complete present-day mitochondrial sequences (Supplementary Information, section 6). As for the plant reads, uniquely classified animal reads with more than 90% similarity were parsed and merged by unit to increase coverage for phylogenetic placement.

Q19 Most notably, we found reads in unit B2 and B3 assigned to the family Elephantidae, which includes elephants and mammoths, but taxonomically not mastodon (*Mammut* sp.)—which are, however, in the NCBI taxonomy, and therefore our analysis reads classified to Elephantidae or below therefore include *Mammut* sp. A consensus genome of our Elephantidae mitochondrial reads falls on the *Mammut* sp. branch (Fig. 4b) and is placed basal to all clades of mastodons. However, we note that this placement within the mastodons depends on only two transition single nucleotide polymorphisms (SNPs), with the first one supported by a read depth of three and the second by only one (Extended Data Fig. 4, Methods and Supplementary Information, section 6). Furthermore, we attempted dating the recovered mastodon mitochondrial genome using BEAST[48]. We implemented two dating approaches, one was based on using radiocarbon-dated specimens alone, while the other used radiocarbon- and molecular-dated mastodons. The first analysis yielded a median age estimate for our mastodon mitogenome of 1.2 Myr (95% HPD: 191,000 yr–3.27 Myr), the second approach resulted in a median age estimate of 5.2 Myr (95% HPD: 1.64–10.1 Myr) (Supplementary Fig. 6.8.5 and Supplementary Information, section 6).

Similarly, reads assigned to the Cervidae support a basal placement on the *Rangifer* (reindeer and caribou) branch (Extended Data Fig. 3). Mitochondrial reads mapping to Leporidae (hares and rabbits) place near the base to the Eurasian hare clade (Extended Data Fig. 2), which is the only mammal found in the fossil record[7]. *Lepus*, specifically *Lepus arcticus*, is also the only genus in the Leporidae living in Greenland today. Mitochondrial reads assigned to Cricetidae cover only one informative transversion SNP, which places them as deriving from the subfamily Arvicolinae (voles, lemmings and muskrats) (Extended Data Fig. 6). For the only avian taxon represented in our dataset—Anatidae, the family of geese and swans—we found a robust basal placement to the genus *Branta* of black geese, supported by three transversion SNPs with read depths ranging between two and four (Extended Data Fig. 5). The refined vertebrate assignments based on mitochondrial references are more biogeographically conserved than for plants. *Dicrostonyx*—specifically *Dicrostonyx groenlandicus* (the Nearctic collared lemming)—is the only genus of the Cricetidae native to Greenland today, just as *Rangifer*—specifically *Rangifer tarandus groenlandicus* (the barren-ground caribou)—is the only member of the Cervidae. The mastodon is the exception, as no member of the Elephantidae lives in present-day Greenland.

Ancient DNA from marine organisms

The other metazoan taxa identified in the DNA record were a single reef-building coral (Merulinidae) and several arthropods, with matches to two insects—Formicidae (ants) and Pulicidae (fleas)—and one marine family—Limulidae (horseshoe crabs). This is somewhat unexpected, given the rich insect macrofossil record from the Kap København Formation, which comprises more than 200 species, including *Formica* sp. The marine taxa are less abundant than the terrestrial taxa, and no mitochondrial DNA was identified from marine metazoans. The read lengths, DNA damage and the fact that the reads assigned distribute evenly across the reference genomes suggests that these are not artefacts but may be over-matched DNA sequences of closely related, potentially extinct species within the families that are currently absent from our reference databases owing to poor taxonomic representation. By contrast, Limulidae, in the subphylum *Chelicerata*, is unlikely to be misidentified as this distinct genus is the only surviving member within its order and thus deeply diverged from other extant organisms.

The probable source of these reads is a population of *Limulus polyphemus*, the only Atlantic member of the genus, which would have spawned directly onto the sediment as it accumulated. Today this genus does not spawn north of the Bay of Fundy (about 45° N), suggesting warmer surface water conditions in the Early Pleistocene at Kap København consistent with the +8 °C annual sea surface temperature anomaly reconstructed for the Pleistocene of the coast of northeast Greenland[49]. By aligning our reads against the *Tara* Oceans eukaryotic metagenomic assembled genomes (SMAGs) data (Methods), we further reveal the presence of 24 marine planktonic taxa in 14 samples, covering both zooplankton and phytoplankton (Fig. 5). These detected SMAGs belong to the supergroups Opisthokonta (6), Stramenopila (15) and Archaeplastida (3). The majority of these signals are from SMAGs associated with cold regions in the modern ocean (that is, the Arctic Ocean and Southern Ocean), such as diatoms (Bacillariophyta), Chrysophyceae and the MAST-4 group (Supplementary

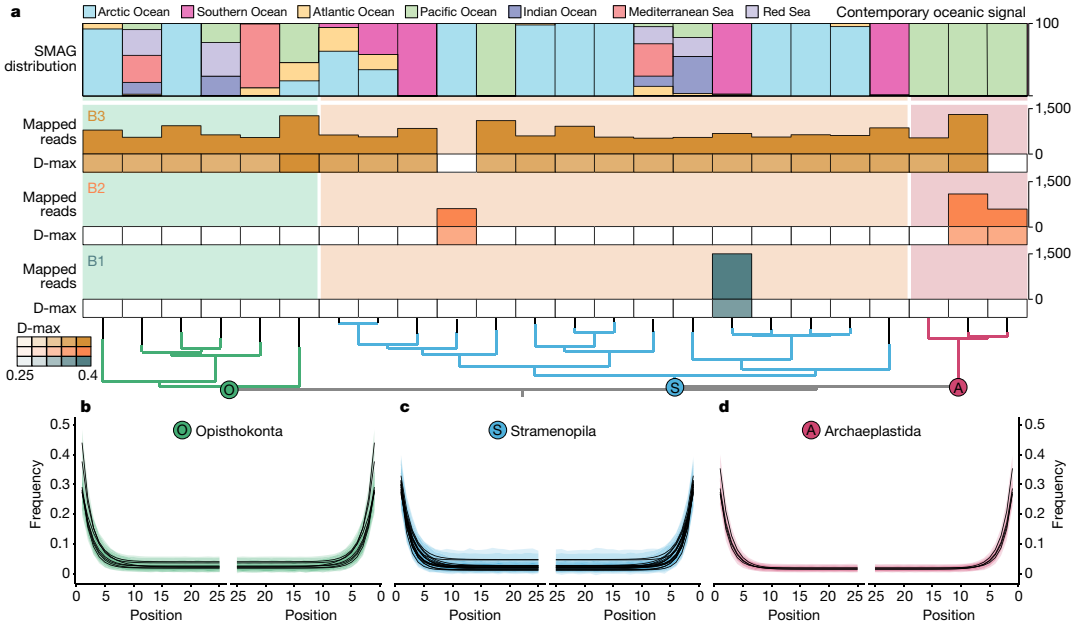


Fig. 5 | Marine planktonic eukaryotes identified at the Kap København Formation. a, Detection of SMAGs and average damage (D-max) of a SMAG within a member unit. Top, the SMAG distribution in contemporary oceans based on the data of Delmont et al.[73]. The SMAGs are ordered on the basis of phylogenomic inference from Delmont et al.[73]. **b**–**d**, Distribution of DNA damage among the taxonomic supergroup Opisthokonta (**b**), Stramenopila (**c**) and Archaeplastida (**d**) (Source Data 1).

Table 6.11.1), as we expected. However, a few are cosmopolitan, whereas others, such as Archaeplastida (green microalgae), have an oceanic signal that is today confined to more temperate waters in the Pacific Ocean (Fig. 5). Although we do not know whether modern day ecologies can be extrapolated to ancient ecosystems, the abundance of green microalgae is believed to be increasing in Arctic regions, which tends to be associated with warming surface waters.

Discussion

The Kap København ancient eDNA record is extraordinary for several reasons; the upper limit of the 95% highest posterior density of the estimated molecular age is 2.0 Myr and independently supports a geological age of approximately 2 Myr (Fig. 2). This implies that the DNA is considerably older than any previously sequenced DNA[21]. Our DNA results detected five times as many plant genera as previous studies using shotgun sequencing of ancient sediments[29,34,50,51], which is well within the range of the richest northern boreal metabarcoding records[52]. The accuracy of the assignments is strengthened by the observation that 76% of the taxa identified to the level of genus or family also occurred in macrofossil and/or pollen assemblages from the same units. Our results demonstrate the potential of ancient environmental metagenomics to reconstruct ancient environments, phylogenetically place and date ancient lineages from diverse taxa from around 2 Ma (Supplementary Information, section 6). Finally, the DNA identified a set of additional plant genera, which occur as macrofossils at other Arctic Late Pliocene and Early Pleistocene sites (Figs. 1 and 3a and Supplementary Information, section 5) but not as fossils at Kap København, thereby expanding the spatiotemporal distribution of these ancient floras.

Of note, the detection of both *Rangifer* (reindeer and caribou) and *Mammut* (mastodon) forces a revision of earlier palaeoenvironmental reconstructions based on the site's relatively impoverished faunal record, entailing both higher productivity and habitat diversity for much of the deposition period. Because all the vertebrate taxa identified by DNA are herbivores, their representation may be a function of relative biomass (see discussion on taphonomy in Supplementary Information, section 6). Caribou, geese, hares and rodents can all be abundant, at least seasonally, in boreal environments. Additionally, the excrement of large herbivores (such as caribou and particularly mastodons) can be a significant component of sediments[34]. By contrast, carnivores are not represented, consistent with their smaller total biomass. This dynamic also explains the dominance of plant reads over metazoans and to some extent differences in representation of various plant genera (Supplementary Information, section 6). In the general absence of fossils, DNA may prove the most effective tool for reconstructing the biogeography of vertebrates through the Early Pleistocene. DNA from mastodon must imply a viable population of this large browsing megaherbivore, which would require a more productive boreal habitat than that inferred in earlier reconstructions based primarily on plant macrofossils[7]. Mastodon dung from a site in central Nova Scotia from around 75,000 years ago contained macrofossils from sedges, cattail, bulrush, bryophytes and even charophytes, but was dominated by spruce needles and birch samaras[53]. The Kap København units with mastodon DNA yielded macrofossils and DNA from *Betula* as well as more thermophilic arboreal taxa including *Thuja*, *Taxus*, *Cornus* and *Viburnum*, none of which range into Greenland's hydric Arctic tundra or polar deserts today. The co-occurrence of these taxa in multiple units compels a revision of previous temperature estimates as well as the presence of permafrost.

Article

No single modern plant community or habitat includes the range of taxa represented in many of the macrofossil and DNA samples from Kap København. The community assemblage represents a mixture of modern boreal and Arctic taxa, which has no analogue in modern vegetation[10,15]. To some degree, this is expected, as the ecological amplitudes of modern members of these genera have been modified by evolution[54]. Furthermore, the combination of the High Arctic photoperiod with warmer conditions and lower atmospheric CO_2 concentrations[55] made the Early Pleistocene climate of North Greenland very different from today. The mixed character of the terrestrial assemblage is also reflected in the marine record, where Arctic and more cosmopolitan SMAGs of Ophistokonta and Stramenophila are found together with horseshoe crabs, corals and green microalgae (Archaeplastida), which today inhabit warmer waters at more southern latitudes.

Megaherbivores, particularly mastodons, could have had a significant impact on an interglacial taiga environment, even providing a top-down trophic control on vegetation structure and composition at this high latitude. The presence of mastodons[56,57] coupled with the absence of anthropogenic fire, which has had a role in some Holocene boreal habitats[58], are important differences. Another important factor is the proximity and biotic richness of the refugia from which pioneer species were able to disperse into North Greenland when conditions became favourable at the beginning of interglacials. The shorter duration of Early Pleistocene glaciations produced less extensive ice sheets allowing colonization from relatively species-rich coniferous-deciduous woodlands in northeastern Canada[12,59]. More extensive glaciation later in the Pleistocene increasingly isolated North Greenland and later re-colonizations were from increasingly distant and/or less diverse refugia.

In summary, we show the power of ancient eDNA to add substantial detail to our knowledge of this unique, ancient open boreal forest community intermixed with Arctic species, a community composition that has no modern analogues and included mastodons and reindeer, among others. Similar detailed flora and vertebrate DNA records may survive at other localities. If recovered, these would advance our understanding of the variability of climate and biotic interactions during the warmer Early Pleistocene epochs across the High Arctic.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05453-y.

1. Salzmann, N. et al. Glacier changes and climate trends derived from multiple sources in the data scarce Cordillera Vilcanota region, southern Peruvian Andes. *Cryosphere* **7**, 103–118 (2013).
2. IPCC *Climate Change 2013: The Physical Science Basis* (eds Stocker, T. F. et al.) (Cambridge Univ. Press, 2013).
3. Brigham-Grette, J. et al. Pliocene warmth, polar amplification, and stepped Pleistocene cooling recorded in NE Arctic Russia. *Science* **340**, 1421–1427 (2013).
4. Gosse, J. C. et al. PoLAR-FIT: Pliocene Landscapes and Arctic Remains—Frozen in Time. *Geosci. Can.* **44**, 47–54 (2017).
5. Matthews, J. V., Telka, A. Jr & Kuzmina, S. A. Late Neogene insect and other invertebrate fossils from Alaska and Arctic/Subarctic Canada. *Zool. Bespozvon.* **16**, 126–153 (2019).
6. Willerslev, E. et al. Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* **300**, 791–795 (2003).
7. Funder, S. et al. Late Pliocene Greenland—the Kap København formation in North Greenland. *Bull. Geol. Soc. Den.* **48**, 117–134 (2001).
8. Funder, S. & Hjort, C. A reconnaissance of the Quaternary geology of eastern North Greenland. *Rapp. Grønlands Geol. Unders.* **99**, 99–105 (1980).
9. Fredskild, B. & Røen, U. Macrofossils in an interglacial peat deposit at Kap København, North Greenland. *Boreas* **11**, 181–185 (2008).
10. Bennike, O. & Böcher, J. Forest-tundra neighbouring the North Pole: plant and insect remains from the Plio-Pleistocene Kap København Formation, North Greenland. *Arctic* **43**, 331–338 (1990).
11. Böcher, J. *Palaeoentomology of the Kap Kobenhavn Formation, a Pilo-Pleistocene sequence in Peary Land, North Greenland* (Museum Tusculanum Press, 1995).
12. Rybczynski, N. et al. Mid-Pliocene warm-period deposits in the High Arctic yield insight into camel evolution. *Nat. Commun.* **4**, 1550 (2013).
13. Wang, X., Rybczynski, N., Harington, C. R., White, S. C. & Tedford, R. H. A basal ursine bear (*Protarctos abstrusus*) from the Pliocene High Arctic reveals Eurasian affinities and a diet rich in fermentable sugars. *Sci Rep.* **7**, 17722 (2017).
14. Simonarson, L. A., Petersen, K. S. & Funder, S. *Molluscan palaeontology of the Pliocene-Pleistocene Kap København Formation, North Greenland*. Arct. Antarct. Alp. Res. 32, (1998).
15. Mogensen, G. S. Pliocene or Early Pleistocene mosses from Kap København, North Greenland. *Lindbergia* **10**, 19–26 (1984).
16. Funder, S., Abrahamsen, N., Bennike, O. & Feyling-Hanssen, R. W. Forested Arctic: evidence from North Greenland. *Geology* **13**, 542–546 (1985).
17. Abrahamsen, N. & Marcussen, C. Magnetostratigraphy of the Plio-Pleistocene Kap København Formation, eastern North Greenland. *Phys. Earth Planet. Inter.* **44**, 53–61 (1986).
18. Bennike, O. *The Kap København Formation: Stratigraphy and Palaeobotany of a Plio-Pleistocene Sequence in Peary Land, North Greenland*. Meddelelser om Grønland, Geoscience Vol. 23 (Kommissionen for Videnskabelige Umdersøgelser i Grønland, 1990).
19. Feyling-Hanssen, R. W. *Foraminiferal Stratigraphy in the Plio-Pleistocene Kap Kobenhavn Formation, North Greenland* (Museum Tusculanum Press, 1990).
20. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
21. van der Valk, T. et al. Million-year-old DNA sheds light on the genomic history of mammoths. *Nature* **591**, 265–269 (2021).
22. Klimaet i Grønland. https://www.dmi.dk/klima/temaforside-klimaet-frem-til-i-dag/klimaet-i-groenland/ (DMI, 2021).
23. Allentoft, M. E. et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. Biol. Sci.* **279**, 4724–4733 (2012).
24. Nguyen, T. H. & Elimelech, M. Plasmid DNA adsorption on silica: kinetics and conformational changes in monovalent and divalent salts. *Biomacromolecules* **8**, 24–32 (2007).
25. Melzak, K. A., Sherwood, C. S., Turner, R. F. B. & Haynes, C. A. Driving forces for DNA adsorption to silica in perchlorate solutions. *J. Colloid Interface Sci.* **181**, 635–644 (1996).
26. Cai, P., Huang, Q.-Y. & Zhang, X.-W. Interactions of DNA with clay minerals and soil colloidal particles and protection against degradation by DNase. *Environ. Sci. Technol.* **40**, 2971–2976 (2006).
27. Fang, Y. & Hoh, J. H. Early intermediates in spermidine-induced DNA condensation on the surface of mica. *J. Am. Chem. Soc.* **120**, 8903–8909 (1998).
28. Karl, D. M. & Bailiff, M. D. The measurement and distribution of dissolved nucleic acids in aquatic environments. *Limnol. Oceanogr.* **34**, 543–558 (1989).
29. Pedersen, M. W. et al. Postglacial viability and colonization in North America's ice-free corridor. *Nature* **537**, 45–49 (2016).
30. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, db.prot5448 (2010).
31. Murchie, T. J. et al. Optimizing extraction and targeted capture of ancient environmental DNA for reconstructing past environments using the PalaeoChip Arctic-1.0 bait-set. *Quat. Res.* **99**, 305–328 (2021).
32. Benoit, G. et al. Multiple comparative metagenomics using multiset *k*-mer counting. Preprint at https://arxiv.org/abs/1604.02412 (2016).
33. Pedersen, M. W. et al. Environmental genomics of Late Pleistocene black bears and giant s hort-faced bears. *Curr. Biol.* **31**, 2728–2736.e8 (2021).
34. Wang, Y. et al. Late Quaternary dynamics of Arctic Biota from ancient environmental genomics. *Nature* **600**, 86–92 (2021).
35. Wang, Y. et al. ngsLCA—A toolkit for fast and flexible lowest common ancestor inference and taxonomic profiling of metagenomic data. *Methods Ecol. Evol.* https://doi.org/10.1111/2041-210X.14006 (2022).
36. Raynolds, M. K. et al. A raster version of the Circumpolar Arctic Vegetation Map (CAVM). *Remote Sens. Environ.* **232**, 111297 (2019).
37. Bay, C. Floristical and ecological characterization of the polar desert zone of Greenland. *J. Veg. Sci.* **8**, 685–696 (1997).
38. Boertmann, D. & Bay, C. *Grønlands Rødliste 2018: Fortegnelse over Grønlandske Dyr og Planters Trusselstatus* (Grønlands Naturinstitut, Aarhus Universitet, 2018).
39. Böcher, T. W., Holman, K. & Jakobson, K. *Grønlands Flora*, 3rd Edn (Forlaget Haase & Søn, 1978).
40. Elven, R., Murray, D. F., Razzhivin, V. Y. & Yurtsev, B. A. *Annotated Checklist of the Panarctic Flora* (PAF) (2011).
41. Bay, C. Four decades of new vascular plant records for Greenland. *PhytoKeys* **145**, 63–92 (2020).
42. Bay, C. *A Phytogeographical Study of the Vascular Plants of Northern Greenland—North of 74 Northern Latitude*, Vol. 36 (Kommissionen for Videnskabelige Undersøgelser i Grønland, 1992).
43. Parducci, L. et al. Ancient plant DNA in lake sediments. *New Phytol.* **214**, 924–942 (2017).
44. Alsos, I. G. et al. Plant DNA metabarcoding of lake sediments: How does it represent the contemporary vegetation. *PLoS ONE* **13**, e0195403 (2018).
45. Bennike, O. *The Kap København Formation: Stratigraphy and Palaeobotany of a Plio-Pleistocene Sequence in Peary Land, North Greenland* (Kommissionen for Videnskabelige Undersøgelser i Grønland, 1990).
46. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
47. Yang, X.-Y. et al. Plastomes of Betulaceae and phylogenetic implications. *J. Syst. Evol.* **57**, 508–518 (2019).
48. Bouckaert, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
49. Dowsett, H. J., Chandler, M. A., Cronin, T. M. & Dwyer, G. S. Middle Pliocene sea surface temperature variability. *Paleoceanography* **20**, https://doi.org/10.1029/2005PA001133 (2005).
50. Graham, R. W. et al. Timing and causes of mid-Holocene mammoth extinction on St Paul Island, Alaska. *Proc. Natl Acad. Sci. USA* **113**, 9310–9314 (2016).
51. Parducci, L. et al. Shotgun environmental DNA, pollen, and macrofossil analysis of lateglacial lake sediments from southern Sweden. *Front. Ecol. Evol.* **7**, 189 (2019).

52. Rijal, D. P. et al. Sedimentary ancient DNA shows terrestrial plant richness continuously increased over the Holocene in northern Fennoscandia. *Sci. Adv.* **7**, eabf9557 (2021).
53. Cocker, S. L. et al. Dung analysis of the East Milford mastodons: dietary and environmental reconstructions from central Nova Scotia at ~75 ka yr BP. Can. J. Earth Sci. https://doi.org/10.1139/cjes-2020-0164 (2021).
54. Fletcher, T. L., Telka, A., Rybczynski, N. & Matthews, J. V. Jr. Neogene and early Pleistocene flora from Alaska, USA and Arctic/Subarctic Canada: new data, intercontinental comparisons and correlations. *Palaeontol. Electronica* 24, https://doi.org/10.26879/1121 (2021).
55. Feng, R. et al. Amplified Late Pliocene terrestrial warmth in northern high latitudes from greater radiative forcing and closed Arctic Ocean gateways. *Earth Planet. Sci. Lett.* **466**, 129–138 (2017).
56. Galetti, M. et al. Ecological and evolutionary legacy of megafauna extinctions. *Biol. Rev. Camb. Philos. Soc.* **93**, 845–862 (2018).
57. Malhi, Y. et al. Megafauna and ecosystem function from the Pleistocene to the Anthropocene. *Proc. Natl Acad. Sci. USA* **113**, 838–846 (2016).
58. Rolstad, J., Blanck, Y.-L. & Storaunet, K. O. Fire history in a western Fennoscandian boreal forest as influenced by human land use and climate. *Ecol. Monogr.* **87**, 219–245 (2017).
59. Elias, S. A. & Matthews, J. V. Jr Arctic North American seasonal temperatures from the latest Miocene to the Early Pleistocene, based on mutual climatic range analysis of fossil beetle assemblages. *Can. J. Earth Sci.* **39**, 911–920 (2002).

PhyloNorway Consortium

Inger Greve Alsos[12] & Eric Coissac[12,18]

Article

Methods

Sampling

Sediment samples were obtained from the Kap København Formation in North Greenland (82° 24′ 00″ N 22° 12′ 00″ W) in the summers of 2006, 2012 and 2016 (see Supplementary Table 3.1.1). Sampled material consisted of organic-rich permafrost and dry permafrost. Prior to sampling, profiles were cleaned to expose fresh material. Samples were hereafter collected vertically from the slope of the hills either using a 10 cm diameter diamond headed drill bit or cutting out ~40 × 40 × 40 cm blocks. Sediments were kept frozen in the field and during transportation to the lab facility in Copenhagen. Disposable gloves and scalpels were used and changed between each sample to avoid cross-contamination. In a controlled laboratory environment, the cores and blocks were further sub-sampled for material taking only the inner part of sediment cores, leaving 1.5–2 cm between the inner core and the surface that provided a subsample of approximately 6–10 g. Subsequently, all samples were stored at temperatures below −22 °C.

We sampled organic-rich sediment by taking samples and biological replicates across the three stratigraphic units B1, B2 and B3, spanning 5 different sites, site: 50 (B3), 69 (B2), 74a (B1), 74b (B1) and 119 (B3). Each biological replicate from each unit at each site was further sampled in different sublayers (numbered L0–L4, Source Data 1, sheet 1).

Absolute age dating

In 2014, Be and Al oxide targets from 8×1 kg quartz-rich sand samples collected at modern depths ranging from 3 to 21 m below stream cut terraces were analysed by accelerator mass spectrometry and the cosmogenic isotope concentrations interpreted as maximum ages using a simple burial dating approach[1] (^{26}Al:^{10}Be versus normalized ^{10}Be). The ^{26}Al and ^{10}Be isotopes were produced by cosmic ray interactions with exposed quartz in regolith and bedrock surfaces in the mountains above Kap København prior to deposition. We assume that the ^{26}Al:^{10}Be was uniform and steady for long time periods in the upper few metres of these gradually eroding palaeo-surfaces. Once eroded by streams and hillslope processes, the quartz sand was deposited in sandy braided stream sediment, deltaic distributary systems, or the near-shore environment and remained effectively shielded from cosmic ray nucleons buried (many tens of metres) under sediment, intermittent ice shelf or ice sheet cover, and—at least during interglacials—the marine water column until final emergence. The simple burial dating approach assumes that the sand grains experienced only one burial event. If multiple burial events separated by periods of re-exposure occurred, then the starting ^{26}Al:^{10}Be before the last burial event would be less than the initial production ratio (6.75 to 7.42, see discussion below) owing to the relatively faster decay of ^{26}Al during burial, and therefore the calculated burial age would be a maximum limiting age. Multiple burial events can be caused by shielding by thick glacier ice in the source area, or by sediment storage in the catchment prior to final deposition. These shielding events mean that the ^{26}Al:^{10}Be is lower, and therefore a calculated burial age assuming the initial production ratio would overestimate the final burial duration. We also consider that once buried, the sand grains may have been exposed to secondary cosmogenic muons (their depth would be too great for submarine nucleonic production). As sedimentation rates in these glaciated near-shore environments are relatively rapid, we show that even the muonic production would be negligible (see Supplemental Information). However, once the marine sediments emerged above sea level, in-situ production by both nucleogenic and muogenic production could alter the ^{26}Al:^{10}Be. The ^{26}Al versus ^{10}Be isochron plot reveals this complex burial history (Supplementary Information, section 3) and the concentration versus depth composite profiles for both ^{26}Al and ^{10}Be reveal that the shallowest samples may have been exposed during a period of time (~15,000 years ago) that is consistent with deglaciation in the area (Supplemental Information). While we interpret the individual simple burial age of all samples as a maximum limiting age of deposition of the Kap København Formation Member B, we recommend using the three most deeply shielded samples in a single depth profile to minimize the effect of post-depositional production. We then calculate a convolved probability distribution age for these three samples (KK06A, B and C). However, this calculation depends on the ^{26}Al:^{10}Be production ratio we use (that is, between 6.75 and 7.42) and on whether we adjust for erosion in the catchment. So, we repeat the convolved probability distribution function age for the lowest and highest production ratio and zero to maximum possible erosion rate, to obtain the minimum and maximum limiting age range at 1σ confidence (Supplementary Information, section 3). Taking the midpoint between the negative and positive 3σ confidence limits, we obtain a maximum burial age of 2.70 ± 0.46 Myr. This age is also supported by the position of those three samples on the isochron plot, which suggests the true age may not be significantly different that this maximum limiting age.

Thermal age

The extent of thermal degradation of the Kap København DNA was compared to the DNA from the Krestovka Mammoth molar. Published kinetic parameters for DNA degradation[60] were used to calculate the relative rate difference over a given interval of the long-term temperature record and to quantify the offset from the reference temperature of 10 °C, thus estimating the thermal age in years at 10 °C for each sample (Supplementary Information, section 4). The mean annual air temperature (MAT) for the the Kap København sediment was taken from Funder et al. (2001)[6] and for the Krestovka Mammoth the MAT was calculated using temperature data from the Cerskij Weather Station (WMO no. 251230) 68.80° N 161.28° E, 32 m from the IRI Data Library (https://iri.columbia.edu/) (Supplementary Table 4.4.1).

We did not correct for seasonal fluctuation for the thermal age calculation of the Kap København sediments or from the Krestovka Mammoth. We do provide theoretical average fragment length for four different thermal scenarios for the DNA in the Kap København sediments (Supplementary Table 4.4.2). A correction in the thermal age calculation was applied for altitude using the environmental lapse rate (6.49 °C km^{-1}). We scaled the long-term temperature model of Hansen et al. (2013)[61] to local estimates of current MATs by a scaling factor sufficient to account for the estimates of the local temperature decline at the last glacial maximum and then estimated the integrated rate using an Ea of 127 kJ mol^{-1} (ref. [60]).

Mineralogic composition

The minerals in each of the Kap København sediment samples were identified using X-ray diffraction and their proportions were quantified using Rietveld refinement. The samples were homogenized by grinding ~1 g of sediment with ethanol for 10 min in a McCrone Mill. The samples were dried at 60 °C and added corundum (CR-1, Baikowski) as the internal standard to a final concentration of 20.0 wt%. Diffractograms were collected using a Bruker D8 Advance (Θ–Θ geometry) and the LynxEye detector (opening 2.71°), with Cu $K_{\alpha1,2}$ radiation (1.54 Å; 40 kV, 40 mA) using a Ni-filter with thickness of 0.2 mm on the diffracted beam and a beam knife set at 3 mm. We scanned from 5–90° 2θ with a step size of 0.1° and a step time of 4 s while the sample was spun at 20 rpm. The opening of the divergence slit was 0.3° and of the antiscatter slit 3°. Primary and secondary Soller slits had an opening of 2.5° and the opening of the detector window was 2.71°. For the Rietveld analysis, we used the Profex interface for the BGMN software[62,63]. The instrumental parameters and peak broadening were determined by the fundamental parameters ray-tracing procedure[64]. A detailed description of identification of clay minerals can be found in the supporting information.

Adsorption

We used pure or purified minerals for adsorption studies. The minerals used and treatments for purifying them are listed in Supplementary

Table 4.2.6. The purity of minerals was checked using X-ray diffraction with the same instrumental parameters and procedures as listed in the above section i.e., mineralogical composition. Notes on the origin, purification and impurities can be found in the supplementary information section 4. We used artificial seawater[65] and salmon sperm DNA (low molecular weight, lyophilized powder, Sigma Aldrich) as a model for eDNA adsorption. A known amount of mineral powder was mixed with seawater and sonicated in an ultrasonic bath for 15 min. The DNA stock was then added to the suspension to reach a final concentration between 20–800 μg ml^{-1}. The suspensions were equilibrated on a rotary shaker for 4 h. The samples were then centrifuged and the DNA concentration in the supernatant determined with UV spectrometry (Biophotometer, Eppendorf), with both positive and negative controls. All measurements were done in triplicates, and we made five to eight DNA concentrations per mineral. We used Langmuir and Freundlich equations to fit the model to the experimental isotherm and to obtain adsorption capacity of a mineral at a given equilibrium concentration.

Pollen

The pollen samples were extracted using the modified Grischuk protocol adopted in the Geological Institute of the Russian Academy of Science which utilizes sodium pyrophosphate and hydrofluoric acid[66]. Slides prepared from 6 samples were scanned at 400× magnification with a Motic BA 400 compound microscope and photographed using a Moticam 2300 camera. Pollen percentages were calculated as a proportion of the total palynomorphs including the unidentified grains. Only 4 of the 6 samples yielded terrestrial pollen counts ≥50. In these, the total palynomorphs identified ranged from 225 to 71 (mean = 170.25; median = 192.5). Identifications were made using several published keys[67,68]. The pollen diagram was initially compiled using Tilia version 1.5.12[69] but replotted for this study using Psimpoll 4.10[70].

DNA recovery

For recovery calculation, we saturated mineral surfaces with DNA. For this, we used the same protocol as for the determination of adsorption isotherms with an added step to remove DNA not adsorbed but only trapped in the interstitial pores of wet paste. This step was important because interstitial DNA would increase the amount of apparently adsorbed DNA and overestimate the recovery. To remove trapped DNA after adsorption, we redispersed the minerals in seawater. The process of redispersing the wet paste in seawater, ultracentrifugation and removal of supernatant lasted less than 2.5 min. After the second centrifugation, the wet pastes were kept frozen until extraction. We used the same extraction protocol as for the Kap København sediments. After the extraction, the DNA concentration was again determined using UV spectrometry.

Metagenomes

A total of 41 samples were extracted for DNA[71] and converted to 65 dual-indexed Illumina sequencing libraries (including 13 negative extraction- and library controls)[30]. 34 libraries were thereafter subjected to ddPCR using a QX200 AutoDG Droplet Digital PCR System (Bio-Rad) following manufacturer's protocol. Assays for ddPCR include a P7 index primer (5′-AGCAGAAGACGGCATAC-3′) (900nM), gene-targeting primer (900 nM), and a gene-targeting probe (250nM). We screened for Viridiplantae psbD (primer: 5′-TCATAATTGGACGTTGAACC-3′, probe: 5′-(FAM)ACTCCCATCATATGAAA(BHQ1)-3′) and Poaceae psbA (primer: 5′-CTCACAACTTCCCTCTAGAC-3′, probe 5′-(HEX) AGCTGCTGTTGAAGTTC(BHQ1)-3′). Additionally, 34 of the 65 libraries were enriched using targeted capture enrichment, for mammalian mitochondrial DNA using the PaleoChip Arctic1.0 bait-set[31] and all libraries were hereafter sequenced on an Illumina HiSeq 4000 80 bp PE or a NovaSeq 6000 100 bp PE. We sequenced a total of 16,882,114,068 reads which, after low complexity filtering (Dust = 1), quality trimming ($q \geq 25$), duplicate removal and filtering for reads longer than 29 bp (only paired read mates for NovaSeq data) resulted in 2,873,998,429 reads that were parsed for further downstream analysis. We next estimated *k*mer similarity between all samples using simka[32] (setting heuristic count for max number of reads (-max-reads 0) and a *k*mer size of 31 (-kmer-size 31)), and performed a principal component analysis (PCA) on the obtained distance matrix (see Supplementary Information, 'DNA'). We hereafter parsed all QC reads through HOLI[33] for taxonomic assignment. To increase resolution and sensitivity of our taxonomic assignment, we supplemented the RefSeq (92 excluding bacteria) and the nucleotide database (NCBI) with a recently published Arctic-boreal plant database (PhyloNorway) and Arctic animal database[34] as well as searched the NCBI SRA for 139 genomes of boreal animal taxa (March 2020) of which 16 partial-full genomes were found and added (Source Data 1, sheet 4) and used the GTDB microbial database version 95 as decoy. All alignments were hereafter merged using samtools and sorted using gz-sort (v. 1). Cytosine deamination frequencies were then estimated using the newly developed metaDMG, by first finding the lowest common ancestor across all possible alignments for each read and then calculating damage patterns for each taxonomic level (https://metadmg-dev.github.io/metaDMG-core/index.html) (Supplementary Information, section 6). In parallel, we computed the mean read length as well as number of reads per taxonomic node (Supplementary Information, section 6). Our analysis of the DNA damage across all taxonomic levels pointed to a minimum filter for all samples at all taxonomic levels with a D-max ≥ 25% and a likelihood ratio (λ-LR) ≥ 1.5. This ensured that only taxa showing ancient DNA characteristics were parsed for downstream profiling and analysis and resulted in no taxa within any controls being found (Supplementary Information, section 6).

Marine eukaryotic metagenome

We sought to identify marine eukaryotes by first taxonomically labelling all quality-controlled reads as Eukaryota, Archaea, Bacteria or Virus using Kraken 2[72] with the parameters '--confidence 0.5 --minimum-hit-groups 3' combined with an extra filtering step that only kept those reads with root-to-leaf score >0.25. For the initial Kraken 2 search, we used a coarse database created by the taxdb-integration workflow (https://github.com/aMG-tk/taxdb-integration) covering all domains of life and including a genomic database of marine planktonic eukaryotes[73] that contain 683 metagenome-assembled genomes (MAGs) and 30 single-cell genomes (SAGs) from *Tara* Oceans[74], following the naming convention in Delmont et al.[73], we will refer to them as SMAGs. Reads labelled as root, unclassified, archaea, bacteria and virus were refined through a second Kraken 2 labelling step using a high-resolution database containing archaea, bacteria and virus created by the taxdb-integration workflow. We used the same Kraken 2 parameters and filtering thresholds as the initial search. Both Kraken 2 databases were built with parameters optimized for the study read length (--kmer-len 25 --minimizer-len 23 --minimizer-spaces 4).

Reads labelled as eukaryota, root and unclassified were hereafter mapped with Bowtie2[75] against the SMAGs. We used MarkDuplicates from Picard (https://github.com/broadinstitute/picard) to remove Q23
duplicates and then we calculated the mapping statistics for each SMAG in the BAM files with the filterBAM program (https://github.com/aMG-tk/bam-filter). We furthermore estimated the postmortem damage of the filtered BAM files with the Bayesian methods in metaDMG and selected those SMAGs with a D-max ≥ 0.25 and a fit quality (λ-LR) higher than 1.5. The SMAGs with fewer than 500 reads mapped, a mean read average nucleotide identity (ANI) of less than than 93% and a breadth of coverage ratio and coverage evenness of less than 0.75 were removed. We followed a data-driven approach to select the mean read ANI threshold, where we explored the variation of mapped reads as a function of the mean read ANI values from 90% to 100% and identified the elbow point in the curve (Supplementary Fig. 6.11.1). We used anvi'o[76] in manual mode to plot the mapping and damage results using the SMAGs phylogenomic tree inferred by Delmont et al.

as reference. We used the oceanic signal of Delmont et al. as a proxy to the contemporary distribution of the SMAGs in each ocean and sea (Fig. 5 and Supplementary Information, section 6).

Comparison of DNA, macrofossil and pollen

To allow comparison between records in DNA, macrofossil and pollen, the taxonomy was harmonized following the Pan Arctic Flora checklist[42] and NCBI. For example, since Bennike (1990)[18], *Potamogeton* has been split into *Potamogeton* and *Stuckenia*, *Polygonym* has been split to *Polygonum* and *Bistorta*, and *Saxifraga* was split to *Saxifraga* and *Micranthes*, whereas others have been merged, such as *Melandrium* with *Silene*[39]. Plant families have changed names—for instance, Gramineae is now called Poaceae and Scrophulariaceae has been re-circumscribed to exclude Plantaginaceae and Orobancheae[77]. We then classified the taxa into the following: category 1 all identical genus recorded by DNA and macrofossils or pollen, category 2 genera recorded by DNA also found by macrofossils or pollen including genus contained within family level classifications, category 3 taxa only recorded by DNA, category 4 taxa only recorded by macrofossils or pollen (Source Data 1).

Phylogenetic placement

We sought to phylogenetically place the set of ancient taxa with the most abundant number of reads assigned, and with a sufficient number of reference sequences to build a phylogeny. These taxa include reads mapped to the chloroplast genomes of the flora genera *Salix*, *Populus* and *Betula*, and to the mitochondrial genomes of the fauna families Elephantidae, Cricetidae, Leporidae, as well as the subfamilies Capreolinae and Anserinae. Although the evolution of the chloroplast genome is somewhat less stable than that of the plant mitochondrial genome, it has a faster rate of evolution, and is non-recombining, and hence is more likely to contain more informative sites for our analysis than the plant mitochondria[78]. Like the mitochondrial genome, the chloroplast genome also has a high copy number, so that we would expect a high number of sedimentary reads mapping to it.

For each of these taxa, we downloaded a representative set of either whole chloroplast or whole mitochondrial genome fasta sequences from NCBI Genbank[79], including a single representative sequence from a recently diverged outgroup. For the *Betula* genus, we also included three chloroplast genomes from the PhyloNorway database[34,80]. We changed all ambiguous bases in the fasta files to N. We used MAFFT[81] to align each of these sets of reference sequences, and inspected multiple sequence alignments in NCBI MSAViewer to confirm quality[82]. We trimmed mitochondrial alignments with insufficient quality due to highly variable control regions for Leporidae, Cricetidae and Anserinae by removing the d-loop in MegaX[83].

The BEAST suite[48] was used with default parameters to create ultrametric phylogenetic trees for each of the five sets of taxa from the multiple sequence alignments (MSAs) of reference sequences, which were converted from Nexus to Newick format in Figtree (https://github.com/rambaut/figtree). We then passed the multiple sequence alignments to the Python module AlignIO from BioPython[84] to create a reference consensus fasta sequence for each set of taxa. Furthermore, we used SNPSites[85] to create a vcf file from each of the MSAs. Since SNPSites outputs a slightly different format for missing data than needed for downstream analysis, we used a custom R script to modify the vcf format appropriately. We also filtered out non-biallelic SNPs.

From the damage filtered ngsLCA output, we extracted all readIDs uniquely classified to reference sequences within these respective taxa or assigned to any common ancestor inside the taxonomic group and converted these back to fastq files using seqtk (https://github.com/lh3/seqtk). We merged reads from all sites and layers to create a single read set for each respective taxon. Next, since these extracted reads were mapped against a reference database including multiple sequences from each taxon, the output files were not on the same coordinate system. To circumvent this issue and avoid mapping bias, we re-mapped each read set to the consensus sequence generated above for that taxon using bwa[86] with ancient DNA parameters (bwa aln -n 0.001). We converted these reads to bam files, removed unmapped reads, and filtered for mapping quality > 25 using samtools[87]. This produced 103.042, 39.306, 91.272, 182 and 129 reads for *Salix*, *Populus*, *Betula*, Elephantidae and Capreolinae, respectively. Q24

We next used pathPhynder[88], a phylogenetic placement algorithm that identifies informative markers on a phylogeny from a reference panel, evaluates SNPs in the ancient sample overlapping these markers, and traverses the tree to place the ancient sample according to its derived and ancestral SNPs on each branch. We used the transversions-only filter to avoid errors due to deamination, except for *Betula*, *Salix* and *Populus* in which we used no filter due to sufficiently high coverage. Last, we investigated the pathPhynder output in each taxon set to determine the phylogenetic placement of our ancient samples (see supplementary information for discussion on phylogenetic placement).

Based on the analysis described above we further investigated the phylogenetic placement within the genus *Mammut*, or mastodons. To avoid mapping reference biases in the downstream results, we first built a consensus sequence from all comparative mitochondrial genomes used in said analysis and mapped the reads identified in ngsLCA as Elephantidae to the consensus sequence. Consensus sequences were constructed by first aligning all sequences of interest using MAFFT[81] and taking a majority rule consensus base in Geneious v2020.0.5 (https://www.geneious.com). We performed three analyses for phylogenetic placement of our sequence: (1) Comparison against a single representative from each Elephantidae species including the sea cow (*Dugong dugon*) as outgroup, (2) Comparison against a single representative from each Elephantidae species, and (3) Comparison against all published mastodon mitochondrial genomes including the Asian elephant *as* outgroup.

For each of these analyses we first built a new reference tree using BEAST v1.10.4 (ref. [46]) and repeated the previously described pathPhynder steps, with the exception that the pathPhynder tree path analysis for the *Mammut* SNPs was based on transitions and transversions, not restricting to only transversions due to low coverage.

***Mammut americanum*.** We confirmed the phylogenetic placement of our sequence using a selection of Elephantidae mitochondrial reference sequences, GTR+G, strict clock, a birth-death substitution model, and ran the MCMC chain for 20,000,000 runs, sampling every 20,000 steps. Convergence was assessed using Tracer[89] v1.7.2 and an effective sample size (ESS) > 200. To determine the approximate age of our recovered mastodon mitogenome we performed a molecular dating analysis with BEAST[46] v1.10.4. We used two separate approaches when dating our mastodon mitogenome, as demonstrated in a recent publication[90]. First, we determined the age of our sequence by comparing against a dataset of radiocarbon-dated specimens ($n = 13$) only. Secondly, we estimated the age of our sequence including both molecularly ($n = 22$) and radiocarbon-dated ($n = 13$) specimens using the molecular dates previously determined[90]. We utilized the same BEAST parameters as Karpinski et al.[90] and set the age of our sample with a gamma distribution (5% quantile: 8.72×10^4, Median: 1.178×10^6; 95% quantile: 5.093×10^6; initial value: 74,900; shape: 1; scale: 1,700,000). In short, we specified a substitution model of GTR+G4, a strict clock, constant population size, and ran the Markov Chain Monte Carlo chain for 50,000,000 runs, sampling every 50,000 steps. Convergence of the run was again determined using Tracer.

Molecular dating methods

In this section, we describe molecular dating of the ancient birch (*Betula*) chloroplast genome using BEAST v1.10.4 (ref. [46]). In principle, the genera *Betula*, *Populus* and *Salix* had both sufficiently high chloroplast genome coverage (with mean depth 24.16×, 57.06× and

27.04×, respectively, although this coverage is highly uneven across the chloroplast genome) and enough reference sequences to attempt molecular dating on these samples. Notably, this is one of the reasons we included a recently diverged outgroup with a divergence time estimate in each of these phylogenetic trees. However, our *Populus* sample clearly contained a mixture of different species, as seen from its inconsistent placement in the pathPhynder output. In particular, there were multiple supporting SNPs to both *Populus balsamifera* and *Populus trichocarpa*, and both supporting and conflicting SNPs on branches above. Furthermore, upon inspection, our *Salix* sample contained a surprisingly high number of private SNPs which is inconsistent with any ancient or even modern age, especially considering the number of SNPs assigned to the edges of the phylogenetic tree leading to other *Salix* sequences. We are unsure what causes this inconsistency but hypothesize that our *Salix* sample is also a mixed sample, containing multiple *Salix* species that diverged from the same placement branch on the phylogenetic tree at different time periods. This is supported by looking at all the reads that cover these private SNP sites, which generally appear to be from a mixed sample, with reads containing both alternate and reference alleles present at a high proportion in many cases. Alternatively, or potentially jointly in parallel, this could be a consequence of the high number of nuclear plastid DNA sequences (NUPTs) in *Salix*[91]. Because of this, we continued with only *Betula*.

First, we downloaded 27 complete reference *Betula* chloroplast genome sequences and a single Alnus chloroplast genome sequence to use as an outgroup from the NCBI Genbank repository, and supplemented this with three *Betula* chloroplast sequences from the PhyloNorway database generated in a recent study[29], for a total of 31 reference sequences. Since chloroplast sequences are circular, downloaded sequences may not always be in the same orientation or at the same starting point as is necessary for alignment, so we used custom code (https://github.com/miwipe/KapCopenhagen) that uses an anchor string to rotate the reference sequences to the same orientation and start them all from the same point. We created a MSA of these transformed reference sequences with Mafft[81] and checked the quality of our alignment by eye in Seqotron[92] and NCBI MsaViewer. Next, we called a consensus sequence from this MSA using the BioAlign consensus function[84] in Python, which is a majority rule consensus caller. We will use this consensus sequence to map the ancient *Betula* reads to, both to avoid reference bias and to get the ancient *Betula* sample on the same coordinates as the reference MSA.

From the last common ancestor output in metaDMG[93], we extracted read sets for all units, sites and levels that were uniquely classified to the taxonomic level of *Betula* or lower, with at a minimum sequence similarity of 90% or higher to any *Betula* sequence, using Seqtk[94]. We mapped these read sets against the consensus *Betula* chloroplast genome using BWA[86] with ancient DNA parameters (-o 2 -n 0.001 -t 20), then removed unmapped reads, quality filtered for read quality ≥25, and sorted the resulting bam files using samtools[86]. For the purpose of molecular dating, it is appropriate to consider these read sets as a single sample, and so we merged the resulting bam files into one sample using samtools. We used bcftools[86] to make an mpileup and call a vcf file, using options for haploidy and disabling the default calling algorithm, which can slightly biases the calls towards the reference sequence, in favour of a majority call on bases that passed the default base quality cut-off of 13. We included the default option using base alignment qualities[95], which we found greatly reduced the read depths of some bases and removed spurious SNPs around indel regions. Lastly, we filtered the vcf file to include only single nucleotide variants, because we do not believe other variants such as insertions or deletions in an ancient environmental sample of this type to be of sufficiently high confidence to include in molecular dating.

We downloaded the gff3 annotation file for the longest *Betula* reference sequence, MG386368.1, from NCBI. Using custom R code[96], we parsed this file and the associated fasta to label individual sites as protein-coding regions (in which we labelled the base with its position in the codon according to the phase and strand noted in the gff3 file), RNA, or neither coding nor RNA. We extracted the coding regions and checked in Seqotron[92] and R that they translated to a protein alignment well (for example, no premature stop codons), both in the reference sequence and the associated positions in the ancient sequence. Though the modern reference sequence's coding regions translated to a high-quality protein alignment, translating the associated positions in the ancient sequence with no depth cut-off leads to premature stop codons and an overall poor quality protein alignment. On the other hand, when using a depth cut-off of 20 and replacing sites in the ancient sequence which did not meet this filter with N, we see a high-quality protein alignment (except for the N sites). We also interrogated any positions in the ancient sequence which differed from the consensus, and found that any suspicious regions (for example, with multiple SNPs clustered closely together spatially in the genome) were removed with a depth cut-off of 20. Because of this, we moved forward only with sites in both the ancient and modern samples which met a depth cut-off of at least 20 in the ancient sample, which consisted of about 30% of the total sites.

Next, we parsed this annotation through the multiple sequence alignment to create partitions for BEAST[46]. After checking how many polymorphic and total sites were in each, we decided to use four partitions: (1) sites belonging to protein-coding positions 1 and 2, (2) coding position 3, (3) RNA, or (4) non-coding and non-RNA. To ensure that these were high confidence sites, each partition also only included those positions which had at least depth 20 in the ancient sequence and had less than 3 total gaps in the multiple sequence alignment. This gave partitions which had 11,668, 5,828, 2,690 and 29,538 sites, respectively. We used these four partitions to run BEAST[46] v1.10.4, with unlinked substitution models for each partition and a strict clock, with a different relative rate for each partition. (There was insufficient information in these data to infer between-lineage rate variation from a single calibration). We assigned an age of 0 to all of the reference sequences, and used a normal distribution prior with mean 61.1 Myr and standard deviation 1.633 Myr for the root height[47]; standard deviation was obtained by conservatively converting the 95% HPD to *z*-scores. For the overall tree prior, we selected the coalescent model. The age of the ancient sequence was estimated following the overall procedures of Shapiro et al. (2011)[97]. To assess sensitivity to prior choice for this unknown date, we used two different priors, namely a gamma distribution metric towards a younger age (shape = 1, scale = 1.7); and a uniform prior on the range (0, 10 Myr). We also compared two different models of rate variation among sites and substitution types within each partition, namely a GTR+G with four rate categories, and base frequencies estimated from the data, and the much simpler Jukes Cantor model, which assumed no variation between substitution types nor sites within each partition. All other priors were set at their defaults. Neither rate model nor prior choice had a qualitative effect on results (Extended Data Fig. 10). We also ran the coding regions alone, since they translated correctly and are therefore highly reliable sites and found that they gave the same median and a much larger confidence interval, as expected when using fewer sites (Extended Data Fig. 10). We ran each Markov chain Monte Carlo for a total of 100 million iterations. After removing a burn-in of the first 10%, we verified convergence in Tracer[89] v1.7.2 (apparent stationarity of traces, and all parameters having an Effective Sample Size > 100). We also verified that the resulting MCC tree from TreeAnnotator[46] had placed the ancient sequence phylogenetically identically to pathPhynder[88] placement, which is shown in Extended Data Fig. 9. For our major results, we report the uniform ancient age prior, and the GTR+G_4 model applied to each of the four partitions. The associated XML is given in Source Data 3. The 95% HPD was (2.0172,0.6786) for the age of the ancient *Betula* chloroplast sequence, with a median estimate of 1.323 Myr, as shown in Fig. 2.

Article

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw sequence data is available through the ENA project accession PRJEB55522. Pollen counts are available through https://github.com/miwipe/KapCopenhagen.git. Source data are provided with this paper.

Code availability

All code used is available at https://github.com/miwipe/KapCopenhagen.git.

60. Lindahl, T. & Nyberg, B. Rate of depurination of native deoxyribonucleic acid. *Biochemistry* **11**, 3610–3618 (1972).
61. Hansen, J., Sato, M., Russell, G. & Kharecha, P. Climate sensitivity, sea level and atmospheric carbon dioxide. *Philos. Trans. A* **371**, 20120294 (2013).
62. Taut, T., Kleeberg, R. & Bergmann, J. The new Seifert Rietveld program BGMN and its application to quantitative phase analysis. *Mater. Struct.* **5**, 57–66 (1998).
63. Doebelin, N. & Kleeberg, R. Profex: a graphical user interface for the Rietveld refinement program BGMN. *J. Appl. Crystallogr.* **48**, 1573–1580 (2015).
64. Cheary, R. W. & Coelho, A. A fundamental parameters approach to X-ray line-profile fitting. *J. Appl. Crystallogr.* **25**, 109–121 (1992).
65. Kester, D. R., Duedall, I. W., Connors, D. N. & Pytkowicz, R. M. Preparation of artificial seawater1. *Limnol. Oceanogr.* **12**, 176–179 (1967).
66. Grichuk, E. D. & Zaklinskaya, V. P. *The Analysis of Fossil Pollen and Spore and Using these Data in Paleogeography* (GeographGIZ Press, 1948).
67. Kupriyanova, L. A. & Aleshina, L. A. *Pollen and Spores of the European USSR Flora* (Nauka, 1972).
68. Moore, P. D., Webb, J. A. & Collinson, M. E. *Pollen Analysis*. (Blackwell Scientific, 1991).
69. Grimm, E. C. *Tilia and Tiliagraph* (Illinois State Museum, 1991).
70. Bennett, K. D. Manual for psimpoll and pscomb. http://www.chrono.qub.ac.uk/psimpoll/psimpoll.html (2002).
71. Ardelean, C. F. et al. Evidence of human occupation in Mexico around the Last Glacial Maximum. *Nature* **584**, 87–92 (2020).
72. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
73. Delmont, T. O. et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. Preprint at *bioRxiv* https://doi.org/10.1101/2020.10.15.341214 (2021).
74. Karsenti, E. et al. A holistic approach to marine eco-systems biology. *PLoS Biol.* **9**, e1001177 (2011).
75. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
76. Eren, A. M. et al. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2021).
77. The Angiosperm Phylogeny Group.An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot. J. Linn. Soc.* **141**, 399–436 (2003).
78. Chevigny, N., Schatz-Daas, D., Lotfi, F. & Gualberto, J. M. DNA repair and the stability of the plant mitochondrial genome. *Int. J. Mol. Sci.* **21**, 328 (2020).
79. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44**, D67–D72 (2016).
80. Alsos, I. G. et al. Last Glacial Maximum environmental conditions at Andøya, northern Norway; evidence for a northern ice-edge ecological 'hotspot'. *Quat. Sci. Rev.* **239**, 106364 (2020).
81. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
82. Yachdav, G. et al. MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics* **32**, 3501–3503 (2016).
83. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
84. Cock, P. J. A. et al. BioPython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
85. Page, A. J. et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* **2**, e000056 (2016).
86. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
87. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
88. Martiniano, R., De Sanctis, B., Hallast, P. & Durbin, R. Placing ancient DNA sequences into reference phylogenies. *Mol. Biol. Evol.* **39**, msac017 (2022).
89. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
90. Karpinski, E. et al. American mastodon mitochondrial genomes suggest multiple dispersal events in response to Pleistocene climate oscillations. *Nat. Commun.* **11**, 4048 (2020).
91. Huang, Y., Wang, J., Yang, Y., Fan, C. & Chen, J. Phylogenomic analysis and dynamic evolution of chloroplast genomes in Salicaceae. *Front. Plant Sci.* **8**, 1050 (2017).
92. Fourment, M. & Holmes, E. C. Seqotron: a user-friendly sequence editor for Mac OS X. *BMC Res. Notes* **9**, 106 (2016).
93. Michelsen, C. et al. metaDMG: a fast and accurate ancient DNA damage toolkit for metagenomic data. *Nat. Methods*.
94. Li, H. et al. Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences (2013).
95. Li, H. Improving SNP discovery by base alignment quality. *Bioinformatics* **27**, 1157–1158 (2011).
96. R Core Team. R: A language and environment for statistical computing (R Foundation for Statistical Computing, 2022).
97. Shapiro, B. et al. A Bayesian phylogenetic method to estimate unknown sequence ages. *Mol. Biol. Evol.* **28**, 879–887 (2011).
98. Feyling-Hanssen, R. W. A remarkable foraminiferal assemblage from the Quaternary of northeast Greenland. *Bull. Geol. Soc. Denmark* **38**, 101–107 (1989).
99. Huang, D. I., Hefer, C. A., Kolosova, N., Douglas, C. J. & Cronk, Q. C. B. Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, *Populus balsamifera* and *P. trichocarpa* (Salicaceae). *New Phytol.* **204**, 693–703 (2014).
100. Levsen, N. D., Tiffin, P. & Olson, M. S. Pleistocene speciation in the genus *Populus* (salicaceae). *Syst. Biol.* **61**, 401–412 (2012).
101. Zhang, L., Xi, Z., Wang, M., Guo, X. & Ma, T. Plastome phylogeny and lineage diversification of Salicaceae with focus on poplars and willows. *Ecol. Evol.* **8**, 7817–7823 (2018).

Acknowledgements We acknowledge support from the Carlsberg Foundation for logistics to carry-out two expeditions to Kap København in 2006 and 2012 (S. Funder, principal investigator for Carlsberg foundation grant to LongTerm and Kap København—the age). The fieldwork in 2016 was supported by a grant to N.K.L. from the Villum Foundation. E.W. and K.H.K. thank the Danish National Research Foundation (DNRF) and the Lundbeck Foundation for providing long-term funds to develop the necessary DNA technology that eventually made it possible to retrieve environmental DNA from these ancient deposits in the Kap København Formation. M.W.P. acknowledges support from the Carlsberg Foundation (CF17-0275). K.K.S. and S.J. acknowledge support from VILLUM FONDEN (00025352). I.G.A. and E.C. have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 819192). B.D.S. acknowledges support from the Wellcome Trust programme in Mathematical Genomics and Medicine (WT220023). J.Å.K. was supported by the Carlsberg Foundation (CF20-0238). C.B. acknowledges ERC Advanced Award Diatomic (grant agreement no. 835067). J.C.G. was supported by Natural Science and Engineering Research Council of Canada–Discovery Grant 06785 and Canada Foundation for Innovation grant 21305. M.J.C. acknowledges support from the Danish National Research Foundation DNRF128. We thank G. Yang for cosmogenic isotope AMS target chemistry; S. Funder for introducing us to the Kap København Formation and generating much of the platform that enabled us to conduct our research; T. O. Delmont for providing data and guidance on the SMAGs analysis; Minik Rosing for providing talc minerals; T. B. Zunic for providing tremolite, orthoclase and chlorite; Z. Vardanyan for help with the DNA extractions and library build; and L. B. Levy and D. Skov for their help collecting samples in 2016. This work was prepared in part by LLNL under contract DE-AC52-07NA27344; LLNL-JRNL-830653. E.W. thanks St Johns College, Cambridge for providing him with a stimulating environment for scientific thoughts and discussion.

Author contributions K.H.K. and E.W. conceived the idea. K.H.K., M.W.P. and E.W. designed the study. K.H.K., A.M.Z.B., A.S.T., N.K.L. and E.W. provided samples, context and carried out fieldwork. M.W.P. undertook the DNA laboratory analysis and taxonomic profiling. M.W.P., B.D.S. and B.D.C. performed the phylogenetic placement with the supervision of M.S. and R.D. B.D.S., M.W.P. and B.D.C. performed the genetic dating with the supervision of R.D. and J.J.W. M.W.P., T.S.K. and C.S.M. conceived, designed and performed the DNA damage estimates. K.K.S. and S.J., conceived and designed the DNA–mineral aspects of the study, interpreted, and wrote about the DNA–mineral data, and participated in the thermal age calculations. K.H.K., M.W.P., A.H.R., A.R., I.G.A. and E.W. undertook the floristic analysis and interpretations. K.K.K. performed cartography and GIS analysis. I.S. designed and carried out palaeomagnetic analysis and interpreted the results. J.C.G. prepared and analysed eight samples for cosmogenic ^{26}Al and ^{10}Be and interpreted their burial age. I.G.A., E.C. and Y.W. provided access to the PhyloNorway reference database, and gave input to the phylogenetic placement of the chloroplasts. A.S.T. counted pollen from the six additional samples. J.Å.K. supported sediment provenance evaluation. M.B. provided mineralogical data from North Greenland. C.D., M.R., M.E.J. and B.S. designed and carried out ddPCR based assays to detect and identify ancient plant DNA in samples. A.F.-G. contributed to the bioinformatic analysis of SMAGs and C.B. contributed to interpretation of marine metagenomic signals. M.J.C. contributed to the thermal age and DNA modelling. M.E.A. contributed to the DNA decay rate estimates. K.H.K., M.W.P., A.H.R. and E.W. interpreted the results and wrote the manuscript with contributions from K.K.S., S.J., A.R., B.D.S., B.D.C., I.G.A., J.C.G., .I.S. and N.K.L., with inputs from the other authors.

Competing interests The authors declare no competing interests.

Additional information

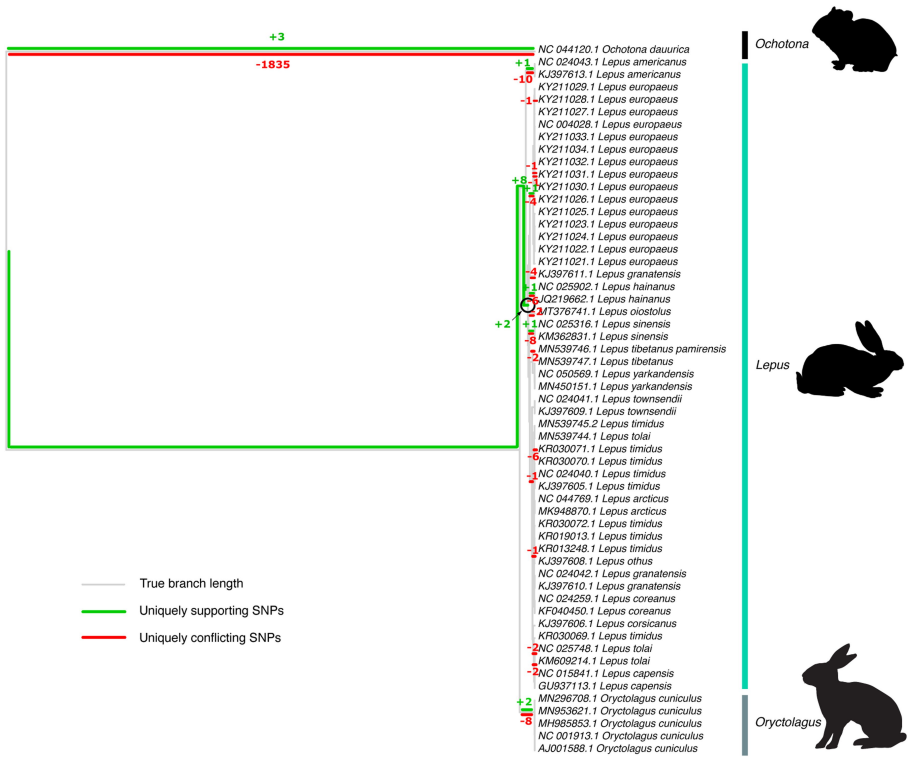
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05453-y.

Correspondence and requests for materials should be addressed to Kurt H. Kjær or Eske Willerslev.

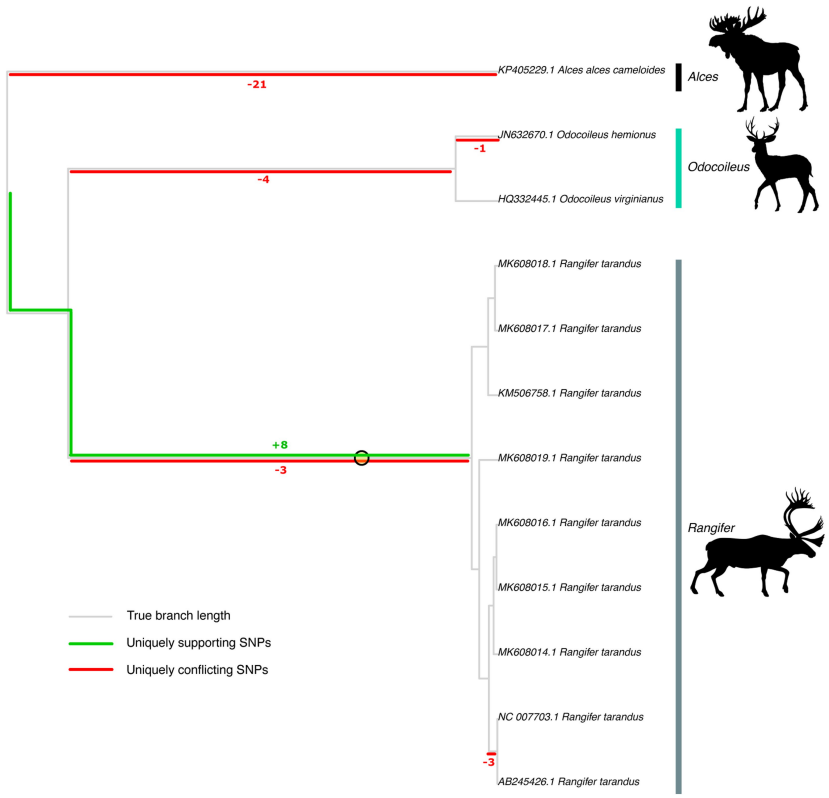
Peer review information *Nature* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

Extended Data Fig. 1 | Setting A. Type locality 50 indicating units in formation **b**. Overview locality 74a+b with a complete sediment sequence. **C**. Overview of locality 69. **D**. Detail of organic rich sediment in unit B3 before excavation and cleaning for ancient eDNA samples. **E**. Sampling in the permafrost within unit B3 at locality 50. **F**. Organic rich sediment at the base of mega-scale cross-bedding within unit B2 at locality 74a+b. White circles mark persons for scale.

Article

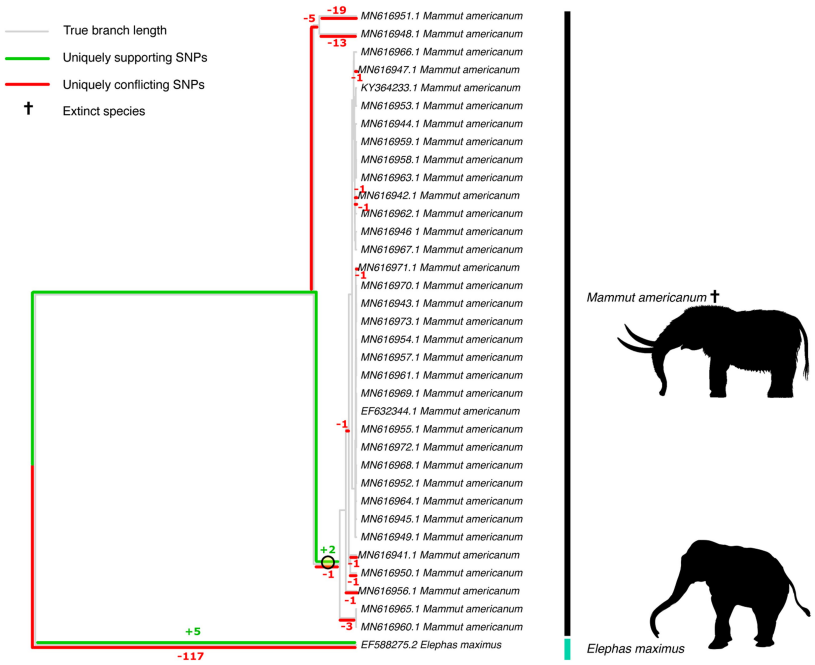


Extended Data Fig. 2 | Phylogenetic placement results of Leporidae mitochondrial reads, using transversion SNPs only. Reads have been merged from all layers and sites. The green numbers on each edge represent the number of supporting (+) SNPs, whereas the red numbers indicate conflicting (-) SNPs in the ancient Leporidae environmental mitochondrial genome overlapping the reference SNPs assigned to the respective edge. There is a clear placement for the ancient Leporidae environmental mitochondrial genome on the edge marked +2, basal to the extant *Lepus* lineage.

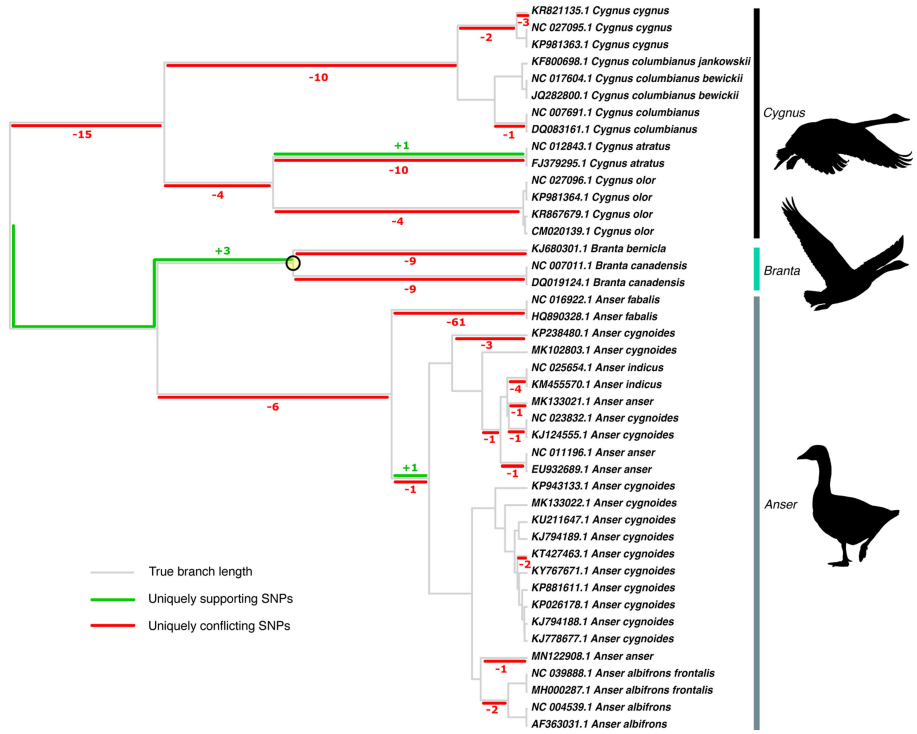


Extended Data Fig. 3 | Phylogenetic placement results for representatives of the Capreolinae mitochondrial reads, using transversion SNPs only. Reads have been merged from all layers and sites. The green numbers on each edge represent the number of supporting (+) SNPs, whereas the red numbers indicate conflicting (-) SNPs in the ancient Capreolinae environmental mitochondrial genome overlapping the reference SNPs assigned to the respective edge. There is a clear placement for the ancient Capreolinae environmental mitochondrial genome on the edge marked +8/−3, basal to the *Rangifer* genus.

Article

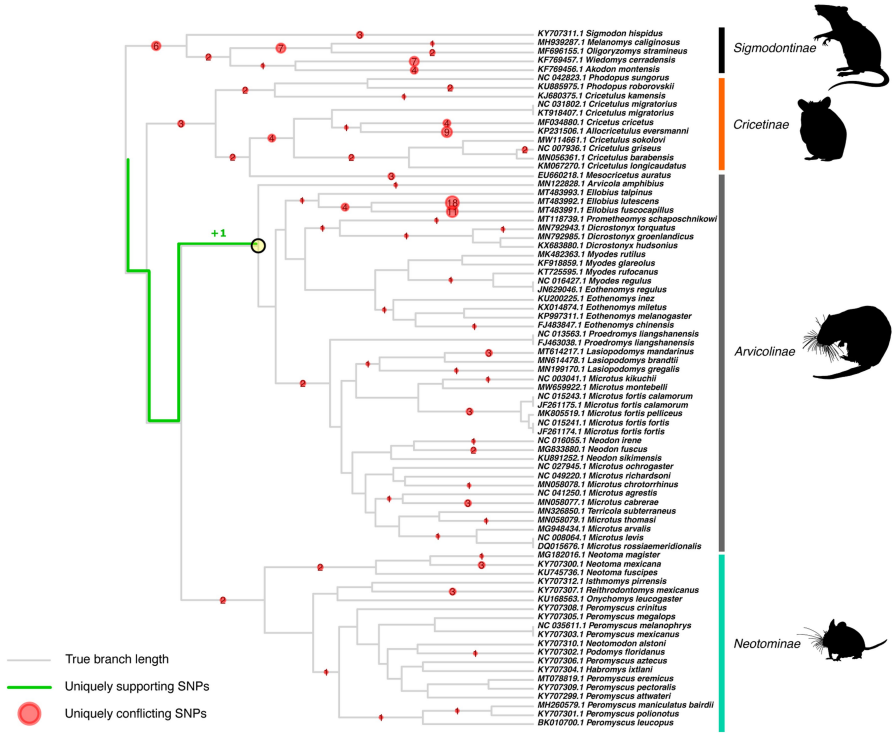


Extended Data Fig. 4 | Phylogenetic placement of Elephantidae mitochondrial reads within mastodons (*Mammut americanum*), using *Elephas maximus* as outgroup, including transitions and transversion SNPs. (Please note that the NCBI taxonomy includes the Mammut genus within Elephantidae). The reference dataset consisted of mitochondria from mastodons (*Mammut americanum*) only and one *Elephas maximus* as an outgroup. Reads have been merged from all layers and sites. The green numbers on each edge represent the number of supporting (+) SNPs, whereas the red numbers indicate conflicting (-) SNPs in the ancient Elephantidae environmental mitochondrial genome overlapping the reference SNPs assigned to the respective edge. There is a placement for the ancient Elephantidae environmental mitochondrial genome on the edge marked +2/−1, identifying it as basal to the mastodon (*Mammut americanum*) clade, which contains most of all mastodon reference mitochondrial genomes. Please note that this placement is based on two transition SNPs with a read depth of three reads per SNP.

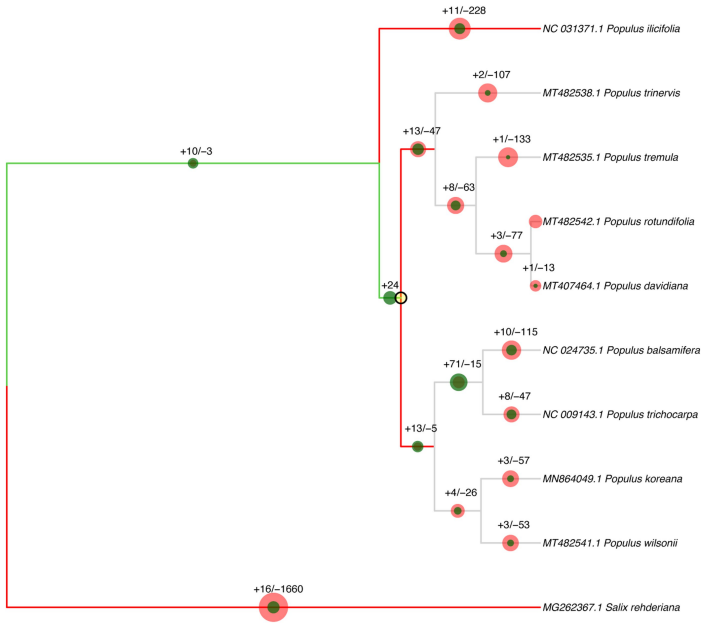


Extended Data Fig. 5 | Phylogenetic placement of mitochondrial reads assigned within Anatidae and placed with representatives of the Anatidae, using transversion SNPs only. Reads have been merged from all layers and sites. The green numbers on each edge represent the number of supporting (+) SNPs, whereas the red numbers indicate conflicting (-) SNPs in the ancient Anatidae environmental mitochondrial genome overlapping the reference SNPs assigned to the respective edge. There is a clear placement for the ancient Anatidae environmental mitochondrial genome on the edge marked +3, basal to the *Branta* genus.

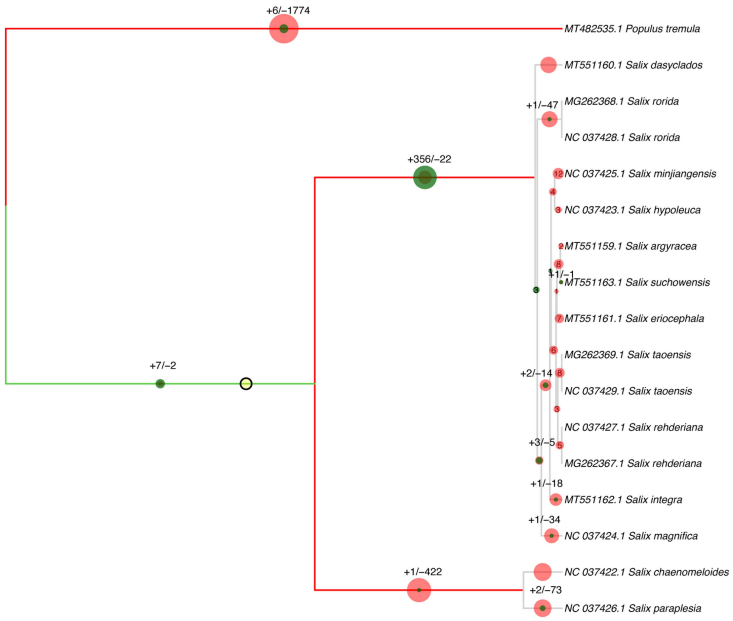
Article



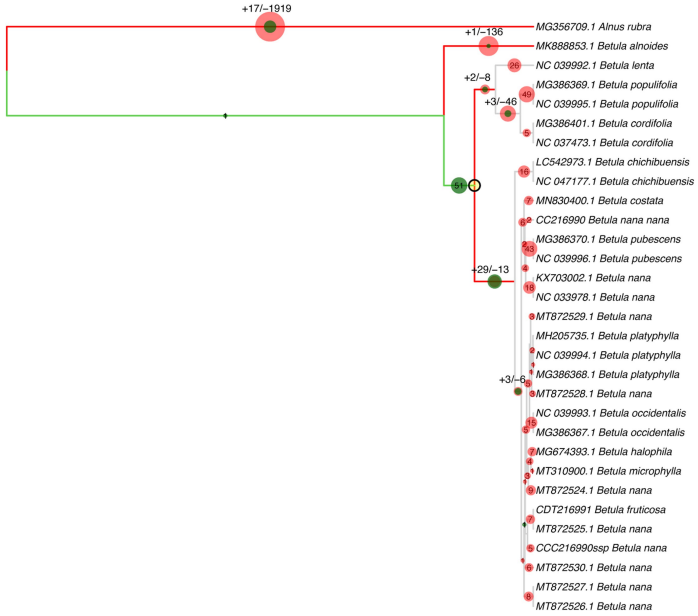
Extended Data Fig. 6 | Phylogenetic placement results of Cricetidae mitochondrial reads, using transversion SNPs only. Reads have been merged from all layers and sites. The green numbers on each edge represent the number of supporting (+) SNPs, whereas the red numbers in the red circles indicate conflicting (-) SNPs in the ancient Cricetidae environmental mitochondrial genome overlapping the reference SNPs assigned to the respective edge. There is a placement for the ancient Cricetidae environmental mitochondrial genome on the edge marked +1, basal to the Arvicolinae subfamily.



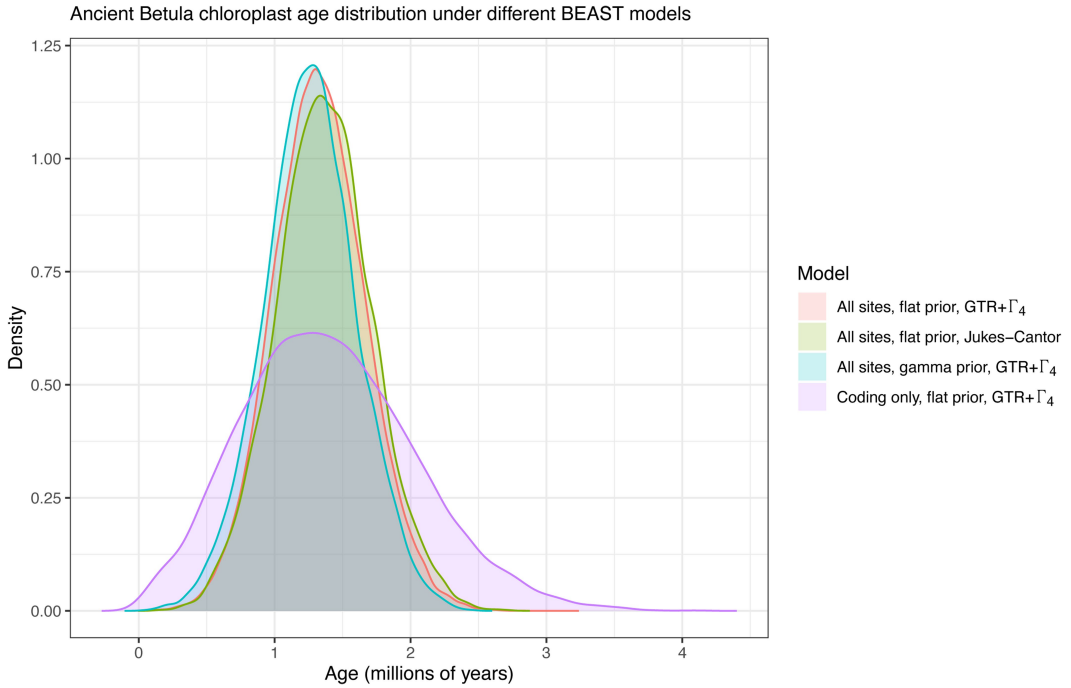
Extended Data Fig. 7 | Phylogenetic placement results for our *Populus* chloroplast reads, using both transition and transversion SNPs, and using reads merged from all layers and sites. The numbers on each edge represent the number of supporting (+) and conflicting (-) SNPs in the ancient *Populus* environmental genome overlapping the reference SNPs assigned to that edge. The ancient *Populus* environmental genome clearly contains a mixture of different species. The most likely placement is on the edge above *Populus trichocarpa* (NC 009143.1) and *Populus balsamifera* (NC 024735.1), with +71/−15 supporting and conflicting SNPs. However, we find some support for both branches directly leading to these species as well. *Populus balsamifera* and *P. trichocarpa* are considered sister species. They are both distributed in North America, as far North as Alaska, are known to hybridise both among themselves and other *Populus* species and are morphologically very similar[91,99,100]. Previous analyses found a very recent nuclear genome divergence time of only 75000 years ago for *Populus trichocarpa* and *P. balsamifera*[100], but a deep chloroplast genome divergence time of at least 6-7 Ma[99], which is an uncommon pattern in plants. Our ancient *Populus* sample could contain individuals either ancestral to, or hybridized from, the modern *Populus trichocarpa* and *P. balsamifera* species.



Extended Data Fig. 8 | Phylogenetic placement results for our *Salix* chloroplast reads, using both transition and transversion SNPs, and using reads merged from all layers and sites. The numbers on each edge represent the number of supporting (+) and conflicting (-) SNPs in the ancient *Salix* environmental genome overlapping the reference SNPs assigned to that edge. The ancient *Salix* environmental genome falls basal to a main *Salix* clade. Our ancient *Salix* sample is phylogenetically placed, with 356 supporting SNPs and 22 conflicting SNPs, on a basal branch leading to the main clade. Although the *Salix* chloroplast phylogeny is not considered fully resolved[91], the difficulties in resolution lie underneath our placement branch, and this along with the high number of SNPs on the placement branch allow us to be confident in the placement position. Our chloroplast phylogeny agrees roughly with[101], which estimated a divergence date between these two main *Salix* clades at 16.9 Ma, and a root age of the first clade, to which our ancient sample is basal to, of 8.1 Ma. It is reasonable, then, to conclude that our ancient *Salix* sample is at least 8.1 Ma diverged from modern *Salix* species, and probably represents an extinct species, or extant species without a reference genome sequenced, or a pool thereof.



Extended Data Fig. 9 | Phylogenetic placement results for our *Betula* chloroplast reads, using both transition and transversion SNPs, and using reads merged from all layers and sites. Our ancient *Betula* sample was placed basal to a main *Betula* clade, based on 29 supporting (green) and 13 conflicting (red) SNPs on its placement branch, and with very low numbers of supporting SNPs elsewhere in the tree other than those leading to this branch. This placement agrees with the BEAST molecular dating analysis (see Molecular Dating Methods).



Extended Data Fig 10 | Molecular age distribution. Results of running the ancient *Betula* chloroplast molecular dating analysis BEAST v1.10.4 (ref. [46]) with different priors and nucleotide substitution models. Using only coding regions, and therefore fewer total sites, gives a larger confidence interval as expected. Results reported in the text are for the red curve, with a flat prior and a GTR+Γ_4 substitution model.

Author Queries

Journal: **Nature**
Paper: **s41586-022-05453-y**
Title: **A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA**

AUTHOR:
The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query Reference	Reference
Q1	Please confirm or correct the city/country name inserted in affiliations 10,12,17.
Q2	Author: A single sentence summarizing your paper has been provided (editor's summary in the eproof), which will appear online on the table of contents and in e-alerts. Please check this sentence for accuracy and appropriate emphasis.
Q3	AUTHOR: Please check that the display items are as follows (ms no: 2021-09-15701C): Figs none (black & white); 5 (colour); Tables: None; Boxes: None; Extended Data display items: 10; SI: yes. The eproof contains the main-text figures edited by us and (if present) the Extended Data items (unedited except for minor formatting) and the Supplementary Information (unedited). Please check the edits to all main-text figures (and tables, if any) very carefully, and ensure that any error bars in the figures are defined in the figure legends. Extended Data items may be revised only if there are errors in the original submissions. If you need to revise any Extended Data items please upload these files when you submit your corrections to this eproof, and include a list of what has been changed.
Q4	Author: Your paper has been copy edited. (1) Please review every sentence to ensure that it conveys your intended meaning; if changes are required, please provide further clarification rather than reverting to the original text. Please note that formatting (including hyphenation, Latin words, and any reference citations that might be mistaken for exponents) and usage have been made consistent with our house style.(2) Check the title and the first paragraph with care, as they may have been re-written to aid accessibility for non-specialist readers.(3) Check the symbols for affiliations with care, and check all author names and Acknowledgements carefully to ensure that they are correct; check the email address of the corresponding author and the Competing Interests statement.(4) Check that there has been no corruption of mathematical symbols. Single-letter variables are set in italics (but not their subscripts unless these are also variables). Vectors are set as bold; matrices are set as italic only. We do not use italics for emphasis. Genetic material is set in italic and gene products are set upright. Please check that italicization and bolding are correct throughout.(5) Ensure that, where practicable, all figures, tables and other discrete elements of Supplementary Information are referred to at least once in the

Author Queries

Journal: **Nature**
Paper: **s41586-022-05453-y**
Title:**A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA**

AUTHOR:
The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query Reference	Reference
	paper at an appropriate place in the text or figure legends.(6) Please note, we reserve 'significant' and its derivatives for statistical significance. Please reword where this is not the intended meaning (for example, to important, notable, substantial).
Q5	Please note that we use as standard Myr for 'million years' and Ma for 'million years ago'. Please check that these have been changed correctly throughout.
Q6	Please check the edits to the sentence 'However, the approximately 3.4 Myr old Fyles Leaf bed...'.
Q7	(1) Please ensure that the following information is included in the figure legends where relevant. Sample size (exact n number); a statement of replicability (how many times was experiment replicated in the lab); description of sample collection (clarify whether technical or biological replicates and include how many animals, litters, cultures, etc.); state the statistical test used and give P values; define centre values (median or average) and error bars. (2) For figures/images that are reproduced or adapted from a third party, it is important that you confirm that permission has been obtained and that appropriate acknowledgement of the copyright holder is given. (3) Please note that we edit the main figures (but not the Extended Data figures) in house. There is no need to resupply any of the main figures to make minor changes to text labels to match the changes made in the text, as figures will have been edited accordingly. If you wish to make changes to any Extended Data figures, however, please resupply these, and please let us know what has changed.
Q8	Please clarify the meaning of the circled numbers and 'a + b' in Fig. 1b caption.
Q9	Please check your article carefully, coordinate with any co-authors and enter all final edits clearly in the eproof, remembering to save frequently. Once corrections are submitted, we cannot routinely make further changes to the article.

Author Queries

Journal: **Nature**
Paper: **s41586-022-05453-y**
Title: **A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA**

AUTHOR:
The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query Reference	Reference
Q10	Note that the eproof should be amended in only one browser window at any one time; otherwise changes will be overwritten.
Q11	Author surnames have been highlighted. Please check these carefully and adjust if the first name or surname is marked up incorrectly. Note that changes here will affect indexing of your article in public repositories such as PubMed. Also, carefully check the spelling and numbering of all author names and affiliations, and the corresponding email address(es).
Q12	You cannot alter accepted Supplementary Information files except for critical changes to scientific content. If you do resupply any files, please also provide a brief (but complete) list of changes. If these are not considered scientific changes, any altered Supplementary files will not be used, only the originally accepted version will be published.
Q13	If applicable, please ensure that any accession codes and datasets whose DOIs or other identifiers are mentioned in the paper are scheduled for public release as soon as possible, we recommend within a few days of submitting your proof, and update the database record with publication details from this article once available.
Q14	In Fig. 1a, the abbreviation for Mammoth was changed to Mm, as Ma is already in use within the figure. OK?
Q15	HPD was defined as '95% HPD'. Correct?
Q16	The sentence 'Using the mean average temperature22 (MAT) of…' ends with '741 times less than the age of 2.0 Ma.' Please check, as this seems to suggest the 2.0 Ma has a particular significance, but it isn't very clear what this might be.

Author Queries

Journal: **Nature**
Paper: **s41586-022-05453-y**
Title:**A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA**

AUTHOR:
The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query Reference	Reference
Q17	Please explain or write out the units 'kyr_DNA@10 °C'.
Q18	The original ref. 43 has been cited in the text and removed from the reference list, per style.
Q19	Please check the edits to the sentence 'Most notably, we found reads in unit B2...'.
Q20	(1) Please ensure that the following information is provided in the Methods section where relevant. Animal experiments require: statement about randomization; statement about blinding; statement of sex, age, species and strain of animals; statement of IRB approval for live vertebrate experimentation. For experiments involving humans: statement of IRB approval; statement of informed consent; statement of consent to publish any photos included in figures. Randomized clinical trials require trial registration. (2) We recommend that detailed protocols are deposited in Protocol Exchange, or a similar repository. (3) If custom computer code has been used and is central to the conclusions of this paper, please insert a section into the Methods titled 'Code availability' and indicate within this section whether and how the code can be accessed, including any restrictions to access. (4) If unpublished data are used, please obtain permission. (5) Please state whether statistical methods were used to predetermine sample size. (6) Please state whether blinding and randomization were used. (7) To address the issue of cell line misidentification and cross-contamination, for any cell lines mentioned in the paper please provide source of the cell lines and indicate whether the cell lines have been correctly identified/authenticated (if so, by what methods). Also, please state whether cell lines have been tested for mycoplasma contamination.
Q21	Please check the sentence 'The mean annual air temperature...', in particular '32 m from the IRI Data Library'.
Q22	Please define Ea in the sentence 'We scaled the long-term temperature model...'.

Author Queries

Journal: **Nature**
Paper: **s41586-022-05453-y**
Title: **A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA**

AUTHOR:
The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query Reference	Reference
Q23	Original ref. 77 has been cited in the text and removed from the reference list, per style.
Q24	Please check the numbers in the sentence 'This produced 103.042, 39.306, 91.272…' - are the periods supposed to indicate a decimal point, thus fractional read numbers? If not, please change to commas.
Q25	Please check the edits to the sentence 'This work was prepared in part by LLNL under…'. Also, please clarify who or what LLNL is - they are not listed as an author.
Q26	Author initials in Author contributions have been edited for consistency with the Author list. Please check and confirm they are correct.
Q27	Please confirm whether added details of ref. 4 are correct.
Q28	Please provide full details of ref. 14.
Q29	Please provide full details for ref. 40.
Q30	Please provide full details for ref. 93.
Q31	Please provide full details for ref. 94.
Q32	Please provide the version number for ref. 96.

nature portfolio

Corresponding author(s): Kurt H. Kjær and Eske Willerslev

Last updated by author(s): 01/09/2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software used for data collection.
Data analysis	Software: BWA (0.7.15), bowtie2 (1.2.3), samtools (1.10), fastq-tools (0.8), sga (0.10.15), gz-sort (1.0), ngsLCA (1.0.0), simka (1.5.3), mafft (7.427), FigTree (1.4.4), biopython (1.79), SNPSites (35), Seqtk-1.3 (r106), PathPhynder, BEAST (1.10.4), metaDMG (0.5.2). BEAST2 (2.6.4), Geneious Prime (2020.0.5), angsd (0.931). R packages: vegan (2.5-7), ggplot2 (3.3.5), ComplexHeatmap (2.4.3), taxize (0.9.99), IntClust (0.1.0), tidyverse (1.3.1), readxl (1.3.1), reshape2 (1.4.4), lattice (0.20-40), gplots (3.1.1), readr (2.0.1), limma (3.46.0), gghighlight (0.3.2), GGally (2.1.2), Hmisc (4.5-0). The custom scripts and code are available at https://github.com/miwipe/KapCopenhagen.git

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The datasets generated and analysed in this study are included in the paper, in the Extended Data Figures, its Supplementary Information, and the SourceData files 1-5. Raw sequence data is available through the ENA project accession PRJEB55522. Pollen counts are available through https://github.com/miwipe/KapCopenhagen.git.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data where this information has been collected, and consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We shotgun sequenced ancient environmental DNA from sediments found at the geological formation Kap København in Greenland, for paleo-environmental recontruction.
Research sample	41 samples obtained from bulk samples or directly in the profiles were used for DNA analysis. 8 1kg bulk samples were obtained for Cosmogenic nuclide burial dating. In addition, different types of minerals were used to test DNA adsorption and release. Sixty-nine samples were collected for determination of the polarity. All samples were taken during three field trips, and spanning 5 different localities within the same formation.
Sampling strategy	Samples were taken across the three units and from 5 different sites, within each site biological replicates were taken in the units both horizontally and vertically see DNA metadata.
Data collection	DNA processing was performed at Centre for GeoGenetics and sequenced at the Danish National Sequencing Centre on Illumina platforms (HiSeq 4000, NovaSeq6000).
Timing and spatial scale	DNA Data were collected from sediment samples from Kap København formation, the northern most Greenland which has been date to 2.0 Mya.
Data exclusions	The DNA results only includes samples that yielded sequenceable DNA. Some samples did not.

Reproducibility	The strongest evidence for reproducibility is that this study includes replicates of geological layers from the same unit but at different locations within the formation (sites) and the fact that they yield highly identical taxonomic profiles. Further, we had biological replicates within the same site and unit, as well as technical replicates of individual samples. All yielding near to identical results.
Randomization	Randomization is not relevant.
Blinding	Blinding is not relevant, as there is no presupposed hypothesis.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	Field works were performed by three different expedition groups. Details are supplied in Methods and SI.
Location	Kap København Formation in North Greenland (82° 24' 00" N 22° 12' 00" W)
Access & import/export	Sediment samples were collected and exported by different research groups from different countries, in agreement with the rules of the specific countries. All sediment samples were imported to Denmark as geological sediment samples for research, for which there is no specific permit required by the authorities.
Disturbance	The samples concerns small sediment samples, and didn't cause disturbance to the surrounding environment as a whole.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

B *SSI Ekspertrapport*

The following pages contain the report from Statens Serum Institut, the Danish CDC:

> Ekspertgruppen for matematisk modellering, *"Ekspertrapport af den 10. december 2020 – Effekten af kontaktopsporing"* (Statens Serum Institut, 2020).

The report is from December 10 2020 and is a summary on the effect of contact tracing related to COVID-19 in Denmark. The report is in Danish and is based on two agent based models, one from DTU and our model from NBI.

STATENS SERUM INSTITUT

Ekspertrapport af den 10. december 2020

Effekten af kontaktopsporing

Indhold

1. Sammenfatning og konklusion

I indeværende rapport har modelgruppen for matematisk modellering af COVID-19 estimeret hvilke delelementer af kontaktopsporing, som er afgørende for at opnå størst mulig effekt af kontaktopsporing af nære kontakter til COVID-19 smittede personer.

Rapporten præsenterer resultater fra to forskellige agentbaserede modeller, som er udviklet af eksperter fra Danmarks Tekniske Universitet (DTU) og Københavns Universitet, Niels Bohr Instituttet (NBI).

En agentbaseret model gør det muligt at modellere enkelte tiltag og deres effekt på smittespredningen af COVID-19. Forudsætningen for en præcis simulation er, at der er tilgængelige data, som kan informere modellen. Der er flere parametre, hvor der i nærværende arbejder er lavet antagelser på basis af de tilgængelige oplysninger. Det forventes, at nogle af disse kan belyses efterhånden som yderligere data frembringes. Hvor der ikke er specifikke eller komplette data, vil en agentbaseret model have unøjagtigheder eller risikere at være baseret på antagelser, som ikke nødvendigvis er retvisende. I modellerne anvendes der endvidere ens ventetidsfordelinger for alle agenter, selvom der i realiteten kan være lokale udsving i ventetider.

Sundhedsstyrelsen udkom d. 23. november 2020 med opdaterede retningslinjer for smitteopsporing af nære kontakter, herunder en udvidet definition af nære kontakter. Indeværende rapport er udviklet i henhold til de tidligere retningslinjer, og tager ikke højde for disse ændringer.

Der er i rapporten heller ikke taget højde for den stigende brug af private antigen test. Coronaopsporingen under STPS foretager også opsporing af nære kontakter, for primærtilfælde som er testet positiv for COVID-19 på sådanne antigen test.

Konklusion

Modellerne peger på, at den største reduktion i kontakttallet kan nås ved effektiv opsporing for flest mulige primærtilfælde. Gevinsten i form af en reduktion i kontakttallet er således større, såfremt der sikres effektiv opsporing for samtlige primærtilfælde, relativt til reduktionen i kontakttallet, som kan opnås ved at nedbringe ventetiden til test og testsvar for primærtilfældet.

Ventetiden til test og testsvar for et primærtilfælde med COVID-19, har stor betydning for den reduktion af kontakttallet, som kan opnås gennem kontaktopsporing. De to uafhængigt udviklede modeller fra hhv. DTU og NBI finder begge, at for hver dag ventetiden til test og testsvar forsinkes for primære tilfælde, stiger kontakttallet med 4%. DTU-modellen finder endvidere, at ventetiden til et primærtilfælde booker en test og samtidig går i isolation har stor betydning for reduktionen i kontakttallet.

Modellerne viser endvidere, at med de anvendte ventetidsfordelinger, vil størstedelen af de nære kontakter som opspores, bliver testet så sent, at det er en mindre del af smitten, som forhindres. Det er derfor vigtigt at opspore nære kontakter hurtigst muligt efter eksponering, så de kan isoleres og blive testet på dag 4 og 6. Dette vil igen afhænge af den samlede ventetid til test og testsvar for primærtilfældet, som er forudsætningen for at opsporingen af nære kontakter kan initieres.

Den agentbaserede model fra NBI finder, at der er yderligere gevinst at hente ved at opspore nære kontakter i de netværk en person indgår i uden for husstand, job og skole. Det skyldes, at relativt få kontakter uden for husstand, job og skole opspores, og at disse kontakter ofte starter nye smittekæder i ikke ellers relaterede netværk. En bredere smitteopsporing har den fordel, at den potentielt finder de nye smittede, som ikke udviser symptomer.

2. Formål og baggrund

2.1 Formål og baggrund for modelgruppen

Statens Serum Institut indgår i det operationelle beredskab for smitsomme sygdomme og yder rådgivning og bistand til regeringen i forbindelse med den aktuelle pandemi. Som en del af denne opgave har Statens Serum Institut nedsat og leder en ekspertgruppe, der har til formål at udvikle matematiske modeller til at belyse udviklingen i COVID-19 i Danmark. Medlemmerne af ekspertgruppen fremgår af bilag 5.

Ekspertgruppens modellering var i foråret 2020 baseret på en populationsmodel, der har fokus på den gennemsnitlige adfærd i befolkningen. Populationsmodellen er bedst egnet, når udviklingen beskrives godt ved gennemsnittet. Derimod er populationsmodellen ikke det bedste værktøj til at beskrive de stokastiske hændelser i lokale udbrud, som aktuelt driver smittespredningen af COVID-19 i Danmark.

Siden sommeren 2020 har modelgruppen derfor udviklet to agentbaserede modeller, som er platformen for de analyser, modelgruppen forventes at levere i den kommende periode. De agentbaserede modeller kan, modsat en populationsmodel, estimere effekten ved enkelte tiltag, såsom effekten ved at nedbringe forsamlingsforbuddet, eller effekten af kontaktopsporing.

2.2 Formål med rapporten

Opsporingen af nære kontakter, foretaget af Styrelsen for Patientsikkerhed (STPS), er løbende udbygget i Danmark siden foråret 2020. Opgaven er vokset betydeligt i takt med, at det daglige antal nye COVID-19 tilfælde stiger, som følge af både en opblussen af epidemien, men også af, at testkapaciteten i Danmark er væsentligt udbygget hen over sommeren. Der testes aktuelt omkring 70.000 personer dagligt.

Formålet med denne rapport er at belyse, hvilke faktorer der er afgørende for at sikre en effektiv kontaktopsporing. Dette belyses ved at estimere effekten af centrale elementer i kontaktopsporingen, såsom ventetid til test og testresultat hos primærtilfældet, samt ventetid til at nære kontakter bliver opsporet og testet.

3. Opsporing og håndtering af nære kontakter i Danmark

3.1 Forudsætninger for en effektiv kontaktopsporing

Den vigtigste forudsætning for, at kontaktopsporing er et effektivt redskab til at nedbringe smitten med COVID-19 er, at der til hver en tid identificeres flest mulige smittede personer, som der derved kan udføres smitteopsporing for. Jo lavere mørketallet er, desto flere vil kunne smitteopspores. Det er derfor afgørende, at der sikres nem og hurtig adgang til test, først og fremmest for personer med COVID-19 lignende symptomer, men også for øvrige personer, der kunne have mistanke om at være smittet med COVID-19. Den Nationale Prævalensundersøgelse i Danmark har vist, at op mod 40-50% af dem, som havde antistoffer mod SARS-CoV-2 i blodet, ingen erindring havde om at have haft COVID-19 lignende sygdom[1]. Ved at udbyde adgang til test for flest mulige personer, vil man også finde flere asymptomatiske smittebærere.

3.2 Definition af en nær kontakt

Sundhedsstyrelsen udkom d. 23. november 2020 med opdaterede retningslinjer for smitteopsporing af nære kontakter. Indeværende rapport er udviklet i henhold til de tidligere retningslinjer.

Der er således ikke taget højde for den udvidede definition af nære kontakter, eller indførslen af screeningsprøver for personer, som ikke umiddelbart opfylder kriteriet for nære kontakter, men som har været eksponeret i et omfang hvor der tilrådes en screeningstest.

Kontaktopsporingen af nære kontakter baserer sig på, at personer der testes positiv for COVID-19 isolerer sig, og dernæst at nære kontakter til den smittede opspores, isoleres og testes, for derved at afbryde smittekæder hurtigst muligt.

Definitionen af en nær kontakt er beskrevet i Sundhedsstyrelsens rapport om smitteopsporing af nære kontakter[2].

En nær kontakt er defineret som en af følgende personer:

- En person der bor sammen med en, der har fået påvist COVID-19
- En person der har haft direkte fysisk kontakt (fx kram) med en, der har fået påvist COVID-19
- En person med ubeskyttet og direkte kontakt til smittefarlige sekreter fra en person der har fået påvist COVID-19
- En person der har haft tæt "ansigt-til-ansigt" kontakt inden for en 1 meter i mere end 15 minutter (fx i samtale med personen) med en, der har fået påvist COVID-19
- Sundhedspersonale og andre som har deltaget i plejen af en patient med COVID-19, og som ikke har anvendt værnemidler på de forskrevne måder

[1] https://www.ssi.dk/-/media/arkiv/dk/aktuelt/nyheder/2020/notat---covid-19-prvalensundersgelsen.pdf?la=da
[2] https://www.sst.dk/da/Udgivelser/2020/COVID-19-Smitteopsporing-af-naere-kontakter

3.3 Periode for smitteopsporing

Der foretages smitteopsporing for perioden, hvor primærtilfældet vurderes at være smitsom. Smitteperioden er således afgrænset til 48 timer før symptomdebut til 48 timer efter symptomophør. For primære tilfælde der ikke har symptomer på COVID-19, er den smitsomme periode afgrænset til 48 timer før positiv test til 7 dage efter.

3.4 Opsporing af nære kontakter

Nære kontakter til en person der er smittet med COVID-19 kan opspores på følgende måder:

- De bliver kontaktet af STPS's Coronaopsporingen
- De bliver kontaktet ifm. kendte udbrud, eksempelvis på skoler
- De bliver kontaktet direkte af primærtilfældet
- De bliver notificeret om, at de har været tæt på en smittet person via appen Smitte|Stop

Nære kontakter opsporet af Coronaopsporingen

Coronaopsporingen under STPS kontakter smittede mhp. at hjælpe med at identificere og opspore nære kontakter til den smittede. Smittede kan også vælge selv at iværksætte opsporing af nære kontakter, og henvise dem til Coronaopsporingen, hvor de nære kontakter vil modtage rådgivning om, hvornår de bør testes, samt får adgang til at booke test på de pågældende dage.

Aktivitetsrapporter fra STPS viser, at der i hele opsporingsperioden i gennemsnit opspores ca. 5 nære kontakter for hvert primærtilfælde, der foretages kontaktopsporing for. Dette er et samlet gennemsnit for opsporede nære kontakter som STPS opsporer, og som primærtilfældet selv opsporer.

Til sammenligning er det estimeret i HOPE-projektet, at danskere henover sommeren i gennemsnit havde ca. 11 kontakter dagligt. Dette antal er nu faldet til ca. 7 kontakter dagligt, som opfylder kriterierne for en nær kontakt, se bilag 4.

Det skal dog pointeres, at Coronaopsporingen ikke er involveret i opsporing af nære kontakter i relation til udbrud i dagtilbud, skoler, plejehjem og hospitaler. Der vil der være opsporede kontakter fra sådanne udbrud, som kontakter Coronaopsporingen for at få rådgivning om hvilke dage de bør testes, samt for at få rekvisitioner til booking af test på de pågældende dage.

Nære kontakter anbefales at blive testet på dag 4 og dag 6 efter vurderet eksponering. Dette relaterer sig til latenstiden, som er perioden fra, at man bliver smittet, til at man er smitsom, og virus kan påvises. En person som er opsporet som nær kontakt til en smittet skal ifølge Sundhedsstyrelsens retningslinjer gå i selv-isolation, indtil der foreligger testsvar. Såfremt der foreligger et negativt testresultat på dag 4, kan den nære kontakt bryde isolationen, men skal fortsat testes på dag 6. Hvis testresultatet på dag 4 derimod er positivt, skal den nære kontakt ikke testes igen på dag 6, men forblive i isolation indtil 48 timer efter symptomophør.

Nære kontakter der ikke opspores af Coronaopsporingen

Der vil være nære kontakter, der ikke opspores og rådgives via Coronaopsporingen. Dette kunne fx være nære kontakter, der bliver opsporet af primærtilfældet selv, og som vælger at booke test på coronaprover.dk uden først at have rådført sig med Coronaopsporingen. Det kunne også være personer, som er opsporet via appen Smitte|Stop, eller personer der mener, at de på anden vis

kan være nære kontakter til en smittet – uden nødvendigvis at opfylde kriteriet for at være en nær kontakt.

I oktober måned blev der i alt testet 1.091.966 personer. Heraf havde 62% (n = 675.623) bestilt tid på coronaprover.dk. Blandt disse svarede 58% (n = 391.146) på spørgeskemaet på coronaprover.dk, hvoraf 25% (n = 99.389) anførte, at de blev testet fordi, de var nær kontakt til en smittet (herunder personer som er adviseret af Smitte|Stop app). Kun 13% (n = 12.706) af dem som svarede, at de blev testet fordi de var nær kontakt til en smittet, var testet på én af de rekvisitionskoder, som der anvendes i Coronaopsporingen (Tabel 1). Samlet set blev 45.616 personer testet på én af de rekvisitionskoder som anvendes i Coronaopsporingen i oktober måned, hvor test-positivprocenten var ca. 4%. Til sammenligning var positivprocenten for de personer, der svarede, at de var nær kontakt til en smittet på Coronaprover.dk omkring 2,5 %. Dette indikerer at Coronaopsporingen har større succes med at opspore de korrekte nære kontakter, sammenlignet med hvis befolkningen selv booker test som nær kontakt, uden forudgående rådgivning fra Coronaopsporingen.

Tabel 1. Oversigt over antal testede i oktober måned 2020.

	Oktober		
		Testpositive (1. test)	
	N	n	%
Testede personer	1.091.966	14.723	1,35
Total antal tests rekvireret via Coronaopsporingen	45.616	1.941	4,26
Bestilt på coronaprøver.dk	675.623	10.335	1,53
Svaret på spørgeskema	391.146	5.387	1,38
Ja, nær kontakt til smittet (herunder adviseret på Smitte|Stop app)	99.389	2.544	2,56
Rekvireret test via Coronaopsporingen	12.706	524	4,12

4. Agentbaserede modeller

4.1 Om agentbaserede modeller

I indeværende rapport er resultaterne for effekten af kontaktopsporing frembragt fra to forskellige agentbaserede modeller, som er udviklet på henholdsvis Danmarks Tekniske Universitet (DTU) og Niels Bohr Instituttet, Københavns Universitet (NBI).

En agentbaseret model simulerer et antal agenter (individer i en population) og deres interaktioner med andre agenter, svarende til de interaktioner som en befolkning normal vis har. Hver agent er således en person, som er knyttet til en lokation i Danmark, svarende til deres bopæl., Agenterne indgår i flere forskellige netværk, f.eks. husstand, job og skole hvor de har kontakt til andre personer. Desuden har de andre kontakter til tilfældige personer i samfundet i den tid, hvor personen ikke er hjemme, på job eller i skole.

Hvis en agent bliver smittet med SARS-CoV-2, er forløbet for den enkelte agent beskrevet således, at agenten først er eksponeret (E) og derefter infektiøs (I), hvorefter agenten ikke længere er smitsom og betragtes som rask (R). De gennemsnitlige tider i hvert sygdomsstadie kan findes i bilag 1 og 2.

Hver kontakt som en agent eksponeres for tildeles en sandsynlighed for at blive smittet af en anden agent, hvis denne er smitsom. Sandsynligheden er sat til et niveau, som afspejler den nuværende situation med et kontakttal omkring 1.

Ud fra de ovenstående generelle antagelser simuleres en epidemi. For en mere detaljeret beskrivelse af de agentbaserede modeller, herunder de inkluderede parametre, henvises til bilag 1 (NBI) og 2 (DTU).

4.2 Forbehold

Mens en agentbaseret model kan medtage mere detaljerede dynamikker i en epidemi, så kræver en præcis simulation input fra data, som ofte ikke er tilgængelige eller forefindes, fx hvem en person mødes med i løbet af en dag. Derfor kan en sådan model have unøjagtigheder eller bygge på antagelser, som ikke er retvisende. Det er ikke muligt at kvantificere den nøjagtige størrelse eller effekt af disse potentielle fejlkilder.

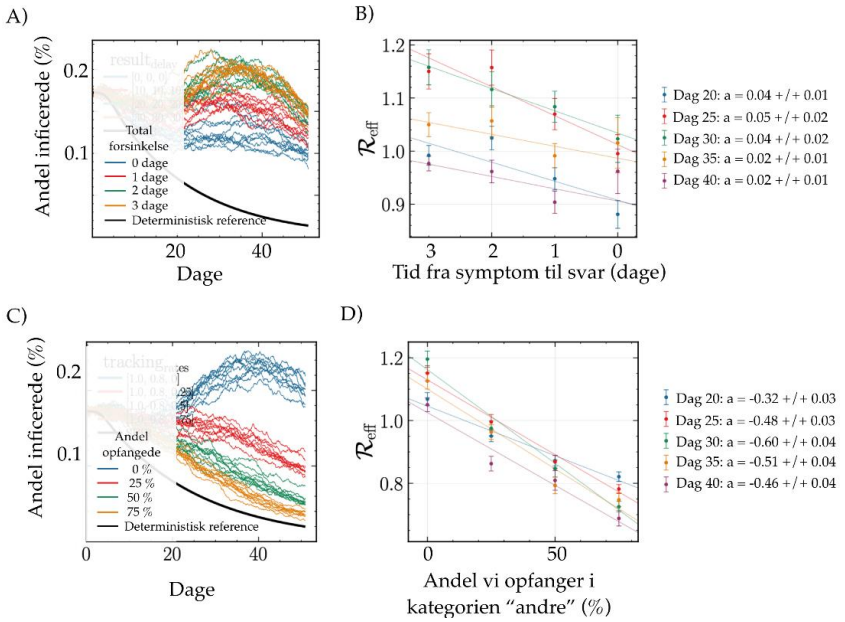
5. Resultater

5.1 Resultater fra den agentbaserede model udviklet af Niels Bohr Instituttet, Københavns Universitet.

Modelkørslerne viser, at når 80% af de sekundære tilfælde i netværkenes husstande, arbejde og skole opspores, vurderes det, at ville nedsætte kontakttallet med omkring 30% sammenlignet med et hypotetisk scenarie uden opsporing af nære kontakter. Dette fremgår af figur 1. Hvis det af logistiske eller kapacitetsmæssige årsager ikke lykkedes at kontakte alle nye COVID-19 tilfælde, vil det betyde en forøgelse af kontakttallet i proportion til dette tal. Dvs. hvis opsporingen ikke kommer i kontakt med 20% af nye COVID-19 tilfælde, vil man potentielt miste 6 procentpoint (0.2 x 0.3 = 0.06) af reduktionen i kontakttallet, som ellers kunne opnås ved kontaktopsporing.

Ventetiden fra at et primærtilfælde ønsker en COVID-19 test (fx hvis man har symptomer), til at vedkommende har modtaget resultatet fra en test har indflydelse på effekten af både selvisolation og kontaktopsporing. Ved en række simulationer med forskellige antagelser finder modellen, at for hver dag man forkorter tiden mellem bestilling af test og testresultat mindskes kontakttallet med omkring 4%. Effekten er lidt større ved højere kontakttal end 1.

Effekten af kontaktopsporing kan øges ved at opspore flere i netværket af øvrige kontakter (ud over husstand og job og skole). Den agentbaserede model viser, at hvis man opsporer 25% af øvrige kontakter, vil kontakttallet falde med omkring 10%. En mere komplet kontaktopsporing (evt. yderligere hjulpet af apps på mobiltelefoner) vil således nedsætte kontakttallet væsentligt. Tilsvarende resultater er fundet i lignende modeller (Plank et al. (september 2020) og Kretzschmar et al. (august 2020)).

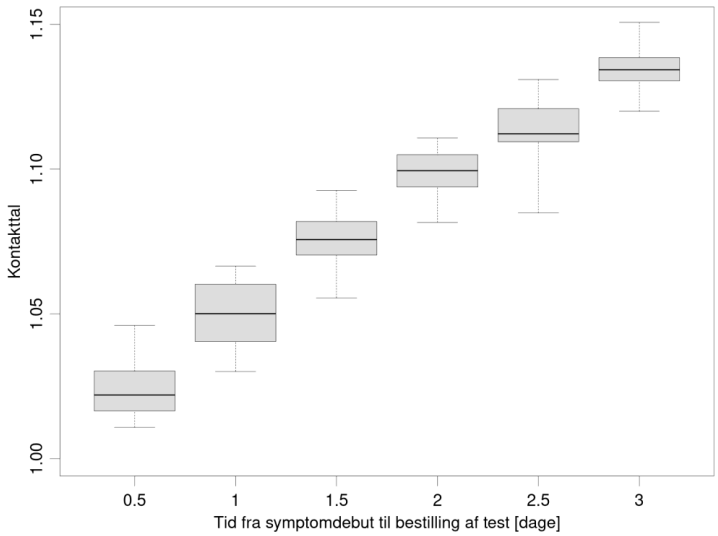


Figur 1: A) Simuleret model, hvor hver kørsel (markeret med samme farve) gentages 10 gange for forskellige værdier af tiden fra symptom til svar. B) Værdien af kontakttallet estimeret på forskellige tidspunkter i simulationen vist i A). Den lineære sammenhæng giver en værdi for hvor mange procent kontakttallet sænkes for hver dag, man gør opsporingen hurtigere. C) Samme som A, men her for forskellige værdier af hvor mange man opsporer blandt øvrige kontakter D) Samme som B) men som funktion af hvor mange man opsporer blandt øvrige kontakter.

5.2 Resultater fra den agentbaserede model udviklet af DTU Compute, Danmarks Tekniske Universitet

Denne agentbaserede model er baseret på tilhørsforhold til grupper (hjem, arbejdsplads, m.fl.). Modellen indeholder en række ventetider fra et primærtilfælde får symptomer til sekundære tilfælde er opsporet. Modellen er nærmere beskrevet i bilag 2. Modellen er kørt med en række forskellige kombinationer af parametre. For hver kombination er der lavet 40 gentagelser for at illustrere variabiliteten. For hver gentagelse simuleres 30 dage som en transient, hvorefter kontakttallet estimeres baseret på de efterfølgende 30 dage.

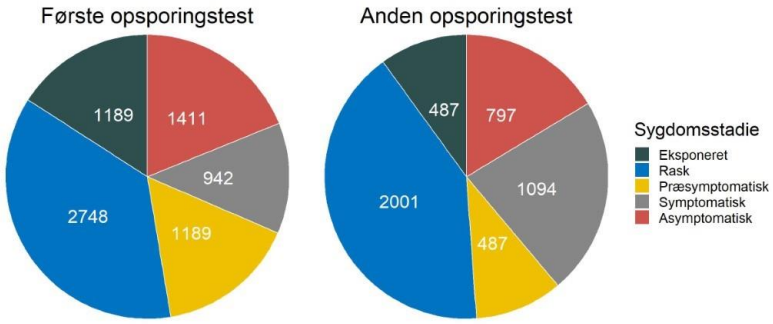
De to parametre, som betyder mest for effekten af kontaktopsporingen, er den gennemsnitlige ventetid fra en smittet får (milde) symptomer til at denne går i isolation og samtidig bestiller en test, samt andelen af kontakter som personen reducerer i perioden fra der bestilles en test til der foreligger et testsvar – det antages, at nære kontakter som opspores opretholder samme grad af isolation som andre, der venter på testsvar, hvilket vil sige, at nære kontakter går i isolation fra de bliver notificeret og indtil de får svar på deres første test.



Figur 2: Kontakttallets afhængighed af den gennemsnitlige tid fra at primærtilfældet har symptomdebut til der bestilles en test og personen går i en grad af isolation. For hver parameterværdi er der foretaget 40 simulationer, og boxplottet viser median, de indre kvartiler samt minimum og maksimum af disse.

På figur 2 ses en klar effekt af tiden fra symptomdebut til isolation og samtidig bestilling af test. For hver dag den gennemsnitlige person går hurtigere i (delvis) isolation estimeres det, at kontakttallet reduceres med 0,04 (når referencen er et kontakttal omkring 1).

Modellen viser også, at omkring 25% af alle test positive, er fundet gennem kontaktopsporing. Det er her antaget, at der udføres kontaktopsporing for alle tilfælde (Se detaljer i bilag 2), samt at test af nære kontakter bestilles på de foreskrevne tidspunkter. Endvidere viser modellen, at over halvdelen af alle smittede aldrig bliver testet positiv (både falsk negative test og asymptomatiske tilfælde). Disse starter derfor nye smittekæder uden forudgående opsporing. Dette kan være årsagen til, at det kun er 25% som findes gennem kontaktopsporing.

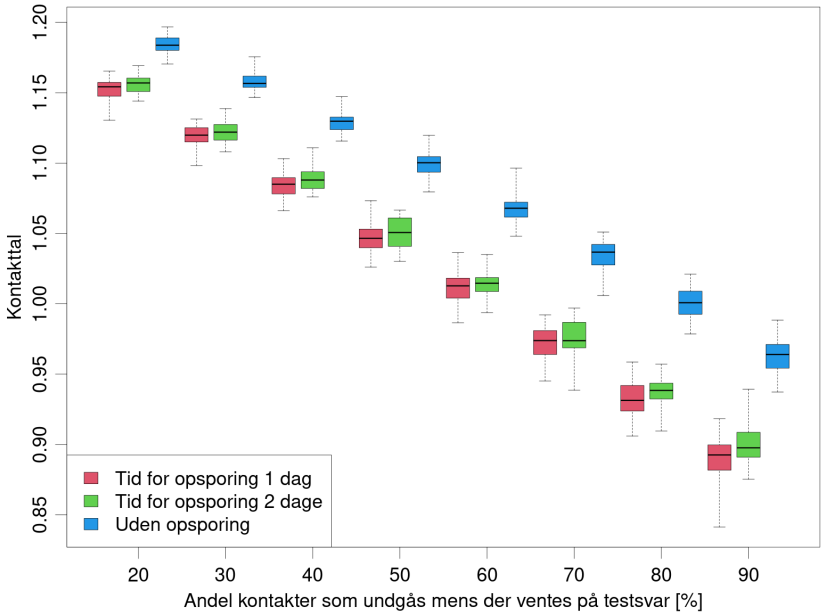


Figur 3. Antal opsporede og smittede i hvert sygdomsstadier, når de får foretaget hhv. første og anden test i opsporingsprocessen. Der er flere, som ikke kommer til anden test, bl.a. fordi de tester positiv i første test eller efter negativt testsvar vælger ikke at få taget den opfølgende test. Derudover vil der være en andel, hvor kontaktopsporingen er initieret sent, således at det kun er foreskrevet at teste personen én gang.

Figur 3 beskriver de forskellige sygdomsstadier for smittede personer, som er opsporet som nære kontakter. Det ses, at en betydelig andel af de opsporede personer, med de i modellen anvendte ventetidsfordelinger, på tidspunktet for opsporingen allerede har overstået deres infektiøse periode, når de testes første gang – en del af disse vil være smittet tidligere og ikke i forbindelse med den nærværende kontaktopsporing. I praksis vil nogle af disse teste positiv, da qPCR kan detektere virus 17 dage efter symptomdebut (Cevik et al., 2020). Desuden ses det, at personer i det præsymptomatiske stadie - hvor ca. halvdelen af smitten sker - kun udgør en lille andel af de opsporede smittede personer ved både første og anden test. Ved begge test er det således under halvdelen af dem, som er smittede, som reelt er infektiøse. Kontaktopsporingen vil derfor kunne optimeres yderligere, hvis man identificerer flere nære kontakter i den præsymptomatiske fase. Dette kan ske ved at nedbringe ventetiden fra symptomdebut til testsvar for primærtilfældet.

Personer, som tidligere er testet positiv er ikke medtaget her og bidrager derfor ikke til antallet af raske. Endvidere vil personer som modtager et positivt testresultat på deres første opsporingstest ikke få foretaget anden opsporingstest. Ovenstående diagrammer er produceret på baggrund af referenceparametrene som beskrevet i bilag 2.

Graden hvorved en smittet person isolerer sig, dvs. hvor stor en andel af ens kontakter man reducerer i perioden fra bestilling af test til testsvar, har stor betydning for kontakttallet. Referenceværdien antages at være 50% reduktion i antallet af kontakter i denne periode. Som det fremgår af figur 4 så opnås der i modellen en reduktion i kontakttallet på knap 0,04 for hver 10 procentpoint graden af isolation øges, hvis der udføres kontaktopsporing (rød og grøn). Mens reduktionen er på 0,03 når der ikke udføres kontaktopsporing (blå). Således har andelen af kontakter, der reduceres hos primærtilfældet og opsporede nære kontakter i ventetiden fra bestilling af test til testsvar, større betydning for en reduktion i kontakttallet, end en reduktion i ventetiden til opsporing af nære kontakter.



Figur 4. Kontakttallets afhængighed af andelen af kontakter et primærtilfælde og opsporede nære kontakter reducerer, i ventetiden fra der bestilles en test til at testsvar foreligger, samt betydningen af ventetiden til at en nær kontakt opspores og går i tilsvarende isolation. For hver parameterværdi er der foretaget 40 simulationer og boxplottet viser median, de indre kvartiler samt minimum og maksimum af disse.

6. Referencer

Cevik, M., Kuppalli, K., Kindrachuk, J. & Peiris, M. (2020). Virology, transmission, and pathogenesis of SARS-CoV-2. The BMJ. Lokaliseret: http://dx.doi.org/10.1136/bmj.m3862

Kretzschmar, M., Rozhnova, G., Bootsma, M., van boven, M., Wijgert, J & Bonten, M. (2020). Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. The Lancet Public Health. Lokaliseret: https://doi.org/10.1016/S2468-2667(20)30157-2

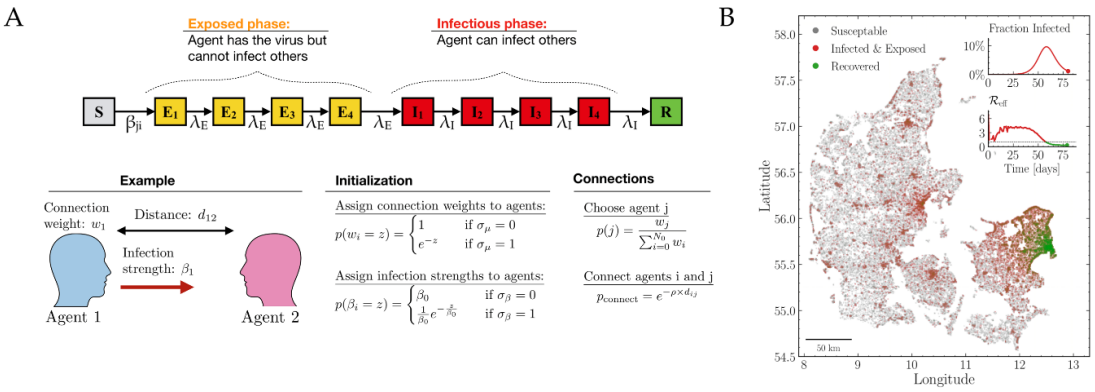
Kucirka, Lauren M., Stephen A. Lauer, Oliver Laeyendecker, et al., (2020). Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction–Based SARS-CoV-2 Tests by Time Since Exposure. Annals of Internal Medicine. Lokalseret: https://doi.org/10.7326/M20-1495

Plank, M., James, A., Lustig, A., Steyn, N., Binny, R. & Hendy, S. (2020). Potential reduction in transmission of COVID-19 by digital contact tracing systems. MedRxiv. Lokaliseret: https://doi.org/10.1101/2020.08.27.20068346

Bilag 1. Beskrivelse af den agentbaserede model fra Niels Bohr Instituttet

Bidrag og udvikling: Christian Michelsen, Emil Martiny, Tariq Halasa, Mogens H. Jensen, Troels C. Petersen og Mathias L. Heltberg

Den agentbaserede model fra NBI baseres på agenter, dvs. individer hvis karakteristika er tildelt ud fra statistiske fordelinger i befolkningen. Dette er f.eks. en aldersfordeling og en fordeling over pendlerafstande. Modellen starter med at fordele Danmarks bopæle ud i landet baseret på det danske hussalg over de sidste 15 år. Herefter placeres agenter i hver husstand baseret på deres alder og geografiske placering.



Figur 5: A) Skematisk oversigt over hvordan interaktionsnetværket i modellen ser ud. B) Eksempel på simulation af smittespredning i Danmark i modellen, gennem et simuleret tilfælde af flokimmunitet i København.

Et afgørende element i modellen er opbygningen af alle personers interaktionsnetværk. Dette genereres ved, at hver agent har et netværk, de interagerer med. Dette opdeles i tre dele: 1) kontakter i hjemmet, 2) kontakter på arbejdet, 3) kontakter i kategorien andre kontakter. Der er ikke nogen geografisk afhængighed af antallet af kontakter på arbejdet, men i den kategori der kaldes "andre", vil der generelt være flere kontakter for dem der bor i tæt befolkede områder i forhold til dem der bor på landet. Måden hvorpå netværket dannes er vist i Figur 5A.

Ud fra data fra HOPE-projektet har vi estimeret, hvor mange personer hver agent vil interagere med, og i denne model vil alle agenter have mellem 3 og 15 daglige kontakter.

Når modellen simuleres vil alle inficerede agenter gennemgå et forløb, hvor de er i en latent periode, hvor de ikke smitter, hvorefter de vil rykke over i en infektiøs periode, hvor de kan smitte agenter i deres netværk. Denne model simuleres ud fra det der kaldes Gillespie algoritmen, således at netværket opdateres instantant for alle smittebegivenheder. En samling af de væsentligste parametre er vist herunder (Tabel 2).

Tabel 2: Parametre i modellen.

Parameter	**Værdi interval for middelværdien**	**Reference**
Antal kontakter per dag	3-15	HOPE projektet
Latent tid (dage)	3-5	Litteratur se referenceliste i bilag 5
Infektiøs tid (dage)	4-8	Litteratur se referenceliste i bilag 5
Andel af kontakter i "andre" (%)	30-80	HOPE projektet
Typisk afstand mellem kontakter (km)	5-20	Trafik data
Andel afstandsuafhængige kontakter (%)	3-5	Trafik data
Tid fra symptom til test (Dage)	0-2	Fordeling fra spørgeskemaundersøgelse i foråret 2020 (ikke offentliggjort)
Sandsynlighed for at få symptomer og blive testet (%)	20-60 %	Prævalensundersøgelsen
Sandsynlighed for at kontakte husstand (%)	100%	Antagelse
Sandsynlighed for at kontakte kollegaer (%)	40-80	Antagelse
Sandsynlighed for at kontakte andre (%)	0-75	Antagelse

Bilag 2. Beskrivelse af den agentbaserede model fra DTU

Bidrag og udvikling: Freja Terp Petersen, Jacob Bahnsen Schmidt, Kasper Telkamp Nielsen, Rebekka Quistgaard-Leth, Kaare Græsbøll og Lasse Engbo Christiansen

Den agentbaserede model fra DTU baseres på en befolkningstabel, hvor hver række i tabellen svarer til en agent - eller et individ – og hver kolonne indeholder data, der beskriver den pågældende agent, herunder aldersgrupper med 5 års-intervaller, bopælskommune, netværks-ID og forskellige smitteparametre.

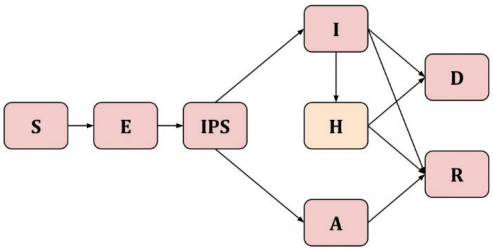
I sygdomsmodellen bæres smitten fremad ved, at agenter der deler netværks-ID, f.eks. husholdnings-ID, skole/job-ID eller omgangskreds-ID, kan smitte hinanden. Hver dag får alle agenter udregnet deres sandsynlighed for at blive smittet på baggrund af antal infektiøse i deres forskellige netværk og på baggrund af deres individuelle antal nære kontakter, som de er blevet tildelt baseret på en fordeling fra totalt antal kontakter inden for 1m i HOPE projektet.

Der er 7 forskellige netværkstyper, som en agent kan være en del af:

- Husholdning (alle agenter har en husholdning)
- Daginstitution (børn mellem 0 og 4 år)
- Grundskole (børn mellem 5 og 14 år)
- Ungdomsuddannelse (unge mellem 15-24 år samt voksne på erhvervsuddannelser)
- Arbejdsplads med kontorinddelinger (voksne op til 65 år)
- Omgangskreds (alle agenter har en omgangskreds)
- Kommune (alle agenter har en kommune)

Agenterne er blevet tildelt netværk baseret på data fra Danmarks Statistik (husholdninger og arbejdspladser), Undervisningsministeriet (grundskoler og ungdomsuddannelser) samt Institution.dk (daginstitutioner).[3] Det antages i modellen, at den gennemsnitlige kontorstørrelse og den gennemsnitlige omgangskreds uden for skole og arbejde er på 8 personer.

[3] FAM122N: https://www.statistikbanken.dk/FAM122N
FAM133N: https://statistikbanken.dk/FAM133N
FAM55N: https://statistikbanken.dk/FAM55N
PEND100: https://www.statistikbanken.dk/PEND100
ERHV6: https://www.statistikbanken.dk/ERHV6
UVM (Normering grundskoler): https://uddannelsesstatistik.dk/Pages/Reports/1577.aspx
UVM (Normering gymnasier): https://uddannelsesstatistik.dk/Pages/Reports/1851.aspx
UVM (Normering erhvervsuddannelse): https://uddannelsesstatistik.dk/Pages/Reports/1850.aspx
Daginstitutioner: https://www.institutioner.dk/



Figur 6. Flowdynamisk diagram af bevægelse gennem sygdomsstadier.

Agenter i modellen kan være i et af følgende sygdomsstadier: Modtagelig (S), Eksponeret (E), Præ-symptomatisk (IPS), Symptomatisk (I), Asymptomatisk (A), Rask (R) eller Død (D). Agenter, som befinder sig i det præ-symptomatiske, symptomatiske eller asymptomatiske stadie, er infektiøse og kan således viderebringe smitte til agenter, som befinder sig i det modtagelige stadie. Bliver en modtagelig agent inficeret, overgår de til at være eksponeret. Dette sygdomsstadie repræsenterer den latente periode, hvor den inficerede agent endnu ikke er infektiøs. Agenterne kan bevæge sig gennem sygdomsstadierne, som vist på det flowdynamiske diagram, figur 6. Modellen antager, at 2/3 af agenterne bliver symptomatiske og at 1/3 forbliver asymptomatiske ved infektiøs tilstand. En andel symptomatiske agenter får et behandlingsbehov i løbet af deres sygdomsforløb og bliver indlagt på et Hospital (H). Sandsynligheden for indlæggelse blandt symptomatiske agenter er opdelt efter regioner og 10-års aldersgrupper baseret på data over indlæggelser i Danmark i september-oktober 2020.

Når en agent skifter til et nyt sygdomsstadie, tildeles de den ventetid, som de skal opholde sig i stadiet. Ventetiden i de forskellige stadier er beskrevet ved gamma-fordelinger med parametre, som vist i tabel 3. Modellen simuleres i diskret tid. Hvert tids-skridt svarer til en halv dag.

Tabel 3. Parametre og kvartiler for varighed af de enkelte stadier.

	Parametre		*Kvartiler*			
Stadier	**Shape**	**Periode [Dage]**	**Nedre kvartil [Dage]**	**Median [Dage]**	**Øvre kvartil [Dage]**	**Referencer**
Eksponeret (E)	3	3	2	3	4	Litteratur se referenceliste i bilag 5
Præsymptomatisk (IPS)	5	1,25	1	2	2	Litteratur se referenceliste i bilag 5
Symptomatisk (I)	4	7	5	7	9	Litteratur se referenceliste i bilag 5
Asymptomatisk (A)	4	7	5	7	9	Litteratur se referenceliste i bilag 5

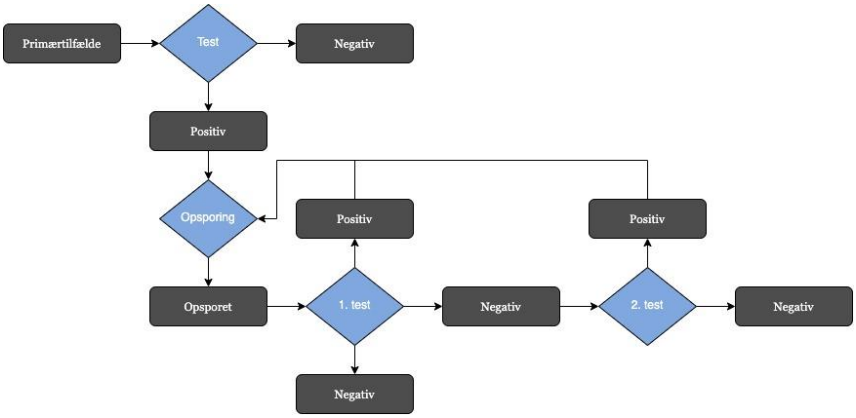
Hospitaliseret (H) Under 60 år	2	3	2	3	5	Linelisten SSI
Hospitaliseret (H) 60 år og derover	2	5	3	5	7	Linelisten SSI
Ventetider						
timeSymptoOrderTest	5	1	1	1	1	Antagelser - afventer STPS data
timeOrderToTest	2	2	1	2	3	Antagelser - afventer STPS data
timeTestToResult	6	1,5	2	2	2	Ventetider fra samfundssporet
traceDelay	5	1	1	2	3	Antagelser – afventer STPS data

Sandsynligheden for, at en modtagelig agent bliver inficeret af en infektiøs agent og overgår til at være eksponeret i et givent netværk stiger med antallet af infektiøse agenter i netværket, de infektiøse agenter i netværkets smitsomhed, samt antallet af kontakter som både de modtagelige og infektiøse agenter har i netværket. Raten hvormed en modtagelig agent bliver inficeret er summen af smitterater fra de enkelte netværk, som agenten deltager i. Test og opsporing er indført i modellen ved følgende regler:

- Når en agent får symptomer, er der en sandsynlighed (pTestGivenSymptoms =80%) for, at de bestiller en test efter en gammafordelt ventetid (timeSympToOrderTest). Hvis der er bestilt en test, vil personen reducere sine kontakter til 50% (undtagen i husholdninger, hvor kontakter reduceres til 70%).
- Der er en gammafordelt ventetid fra testen bestilles, til testen udføres (timeOrderToTest).
- Der er en gammafordelt ventetid fra testen udføres, til der kommer svar (timeTestToResult).
- Hvis der kommer positivt svar, vil agenten isolere sig yderligere; kontakter reduceres til 10% (husholdning: 50%). Derudover påbegyndes opsporing af netværk under følgende regler:
    - I skoleklasser, ungdomsskoleklasser, institutioner og i husholdninger opspores alle personer (i husholdninger foregår det dobbelt så hurtigt som i de øvrige netværk).
    - På kontorer (arbejdspladser) og i omgangskredse opspores et antal nære kontakter givet ved fordeling af kontakter under 1m i data fra HOPE projektet.
    - Personer, som tidligere er testet positiv, får ikke tildelt yderligere test.
    - Der opspores med en gammafordelt forsinkelse (traceDelay) fra den positive test.
    - Ved opsporing efter en person testes positiv tildeles de opsporede personer testtider relativt til 48 timer før den positive fik symptomer - eller blev testet i et asymptomatisk tilfælde. Hvis muligt, gives test på dag 4 og dag 6, ellers dag 5 og 7, og ellers én test hurtigt muligt.
    - Personer, som er i et igangværende opsporingsforløb, får kun tildelt test, hvis de venter på mindre end to testsvar.
    - Den opsporede person har samme ventetider på testsvar, som symptomatiske personer.

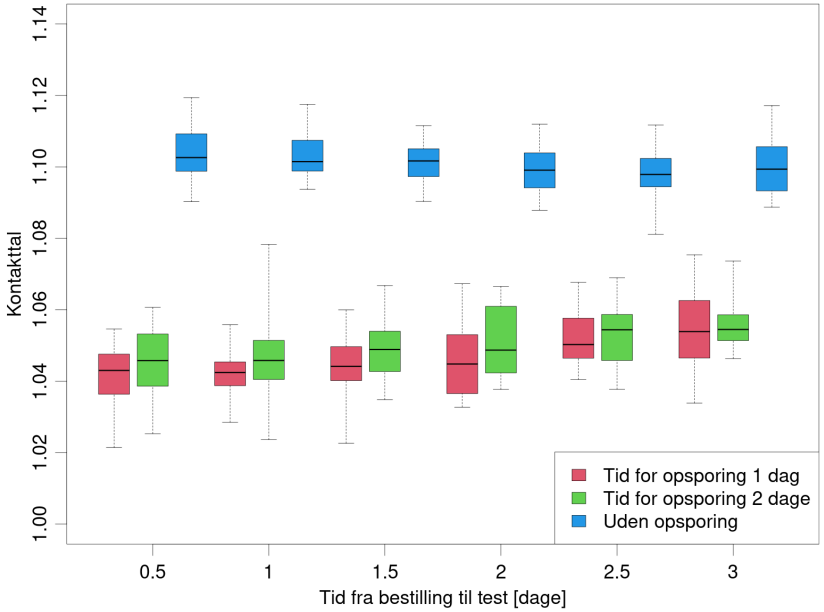
- Mens der ventes på test og testsvar, isoleres den opsporede person på samme måde som en symptomatisk, der venter på svar.
- Hvis en opsporet person får negativt svar på den første test, vil der være en sandsynlighed for (pNoShow2ndTest =40%) at de ikke tager test nummer 2.
- Efter et negativt svar på test nummer 1, vil isolationen brydes. Hvis der fås et positivt svar, inden test nummer 2 er taget, annulleres test nummer 2, og personens egne netværk opspores.

- For alle tests – om det er en opsporet person eller ej – antages der en sandsynlighed på 20% for en falsk negativ test (Kucirka et al., 2020).



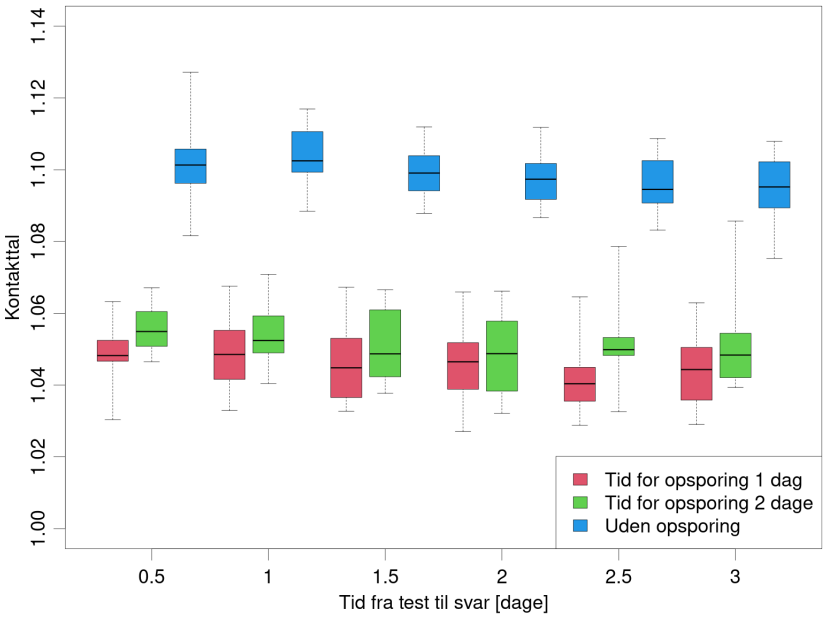
Figur 7. Diagram, der viser test og opsporing i den agentbaserede model fra DTU.

Yderligere resultater



Figur 8. Kontakttallets afhængighed af ventetiden på at få taget en test hos primærtilfældet, samt betydningen af hvor lang tid der går før nære kontakter opspores og går i tilsvarende isolation. For hver parameterværdi er der foretaget 40 simulationer og boxplottet viser median, de indre kvartiler samt minimum og maksimum af disse.

Af figur 8 fremgår det, at der ikke er nævneværdig forskel på reduktionen i kontakttallet, hvorvidt man reducerer ventetiden til at primærtilfældet testes, i forhold til at reducere ventetiden til opsporing af nære kontakter.



Figur 9. Kontakttallets afhængighed af tiden fra der testes til at der foreligger et testsvar, samt betydningen af hvor lang tid der går indtil nære kontakter opspores og går i tilsvarende isolation. For hver parameterværdi er der foretaget 40 simulationer og boxplottet viser median, de indre kvartiler samt minimum og maksimum af disse.

Af figur 9 fremgår det, at der ikke er nævneværdig forskel på reduktionen i kontakttallet, hvorvidt man reducerer ventetiden fra at primærtilfældet og opsporede kontakter testes til der foreligger et testsvar, i forhold til at reducere ventetiden til opsporing af nære kontakter. En årsag kan være, at ventetiden til testsvar gør, at en masse opsporede og modtagelige kontakter er isoleret i længere tid og derfor ikke bliver smittet. Det er ikke undersøgt om dette kun ses når kontakttallet er nær 1.

Bilag 3. Regneeksempel

Følgende er et illustrativt regneeksempel på den agentbaserede model fra Niels Bohr Instituttet beskrevet i bilag 1. Udregningerne er baseret på modellens underliggende antagelser, nemlig at perioden for eksposition (E (T_E), hvor den latente fase er en gammafordeling med middelværdi på 4.7 dage, og perioden for den smitsomme fase er en gammafordeling med middelværdi på 7 dage, samt en antagelse om, at 40% af cases findes uden kontaktopsporing. Det antages, at for de COVID-19 tilfælde der findes uafhængigt af kontaktopsporingen, er de smittet uniformt fordelt i den smitsomme periode (I). Vi udregner nu tiden man er asymptomatisk men smitsom ved at trække tal fra fordelingen af tider for hele perioden, man er smitsom og tester en andel p, på et uniformt tilfældigt tidspunkt. Det giver en fordeling og en gennemsnitlig eksponeringstid (se figur 10A).

Vi kigger nu på et sekundært tilfælde, der blev smittet på et uniformt tilfældigt tidspunkt i den smitsomme periode for primærtilfældet. Denne person kan enten findes tilfældigt, eller ved at primærtilfældet testes, og at sekundærtilfældet opspores efter en tidsperiode (d for delay). Denne ventetid, er tiden fra at primærtilfældet testes til at sekundærtilfældet kontaktes, og afspejler således både ventetid til test samt ventetid til opsporing. Igen antages det, at sekundærtilfældet går i isolation øjeblikkeligt. Ved igen at trække tal tilfældigt fra de relevante fordelinger fås en eksponeringsperiode, hvori sekundærtilfældet måske opspores, forhåbentligt inden smitten er ført videre.

Resultat

I figur 10B vises det gennemsnitlige antal dage en kontakt er eksponeret for smitte, som en funktion af den samlede ventetid til test og opsporing. Herudfra estimeres effekten af kontaktopsporing på det effektive kontakttal, Rt. Det antages, at en given andel (fc) af alle smittetilfælde, findes via kontaktopsporing, og derved reduceres smitten, idet eksponeringsperioden for opsporede kontakter reduceres. Herved fås et simpelt estimat af effekten af kontaktopsporing på kontakttallet Rt. Dette vises i figur 10C. Farverne på graferne viser, hvor stor en andel af smitten der kan reduceres, såfremt eksponeringsperioden reduceres, som følge af kontaktopsporing. Hvis det f.eks. antages, at der er 2000 nye smittede med COVID-19 per dag (ca. 1000 fundne smittede + et mørketal), så svarer 0.05 grafen (orange) til at 100 smittede bliver fundet gennem kontaktopsporing dagligt.

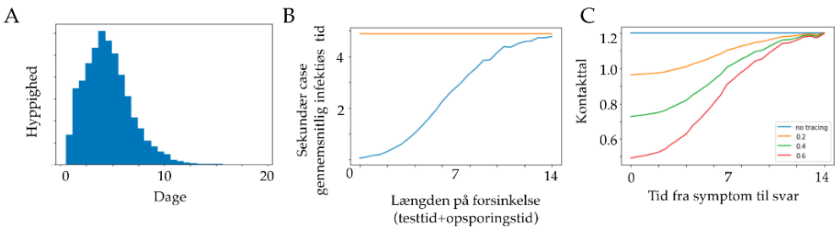
En væsentlig begrænsning er, at disse udregninger ikke medtager effekten af, at flere COVID-19 tilfælde bliver fundet pga. kontaktopsporing, men er udelukkende baseret på effekten ved at forkorte eksponeringsperioden for kontakter.

I modellen indgår 4 mulige eksponeringsperioder. 1: Kontakter opspores ikke, hvorved eksponeringsperioden ikke afkortes (blå graf), 2: 20% af kontakter opspores (gul graf), 3: 40% af kontakter opspores (grøn graf) og 4): hvis 80% af kontakter opspores (rød graf).

Af regneeksemplet fremgår det, at givet antagelserne i eksemplet vil kontakttallet kunne reduceres med ca. 50%, såfremt man opsporer 50% af alle kontakter inden for ca. 3 dage.

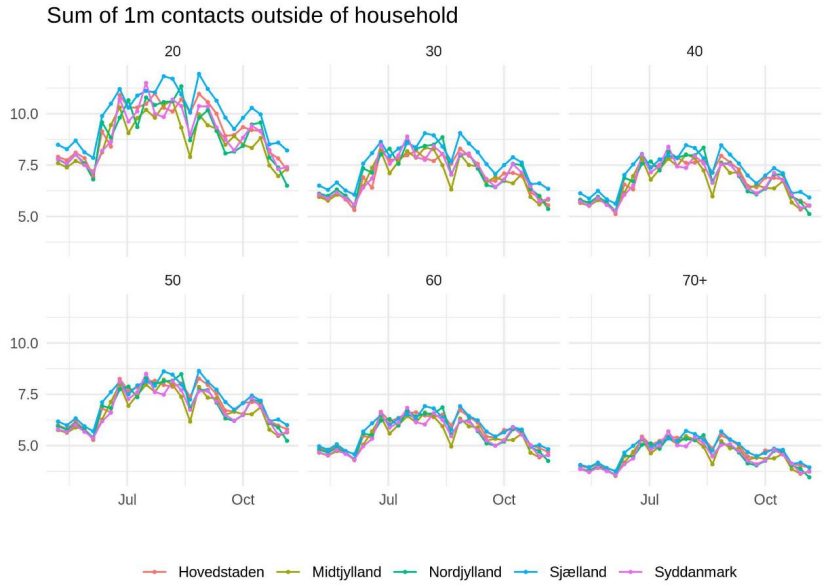
Bemærk at alle kurverne i figur 10C er meget flade i intervallet mellem dag 0 og 3. Dette betyder, at der kun opnås en lille gevinst ved at afkorte den samlede ventetid fra symptomer til der foreligger et testsvar inden for denne periode, men at der til gengæld er en stor gevinst ved at øge andelen af opsporede kontakter.

Det skal i øvrigt bemærkes, at det i eksemplet antages, at opsporede kontakter går i isolation, indtil de får svar på deres test.



Figur 10:A) Fordeling af eksponeringstiden, gennemsnit = 4.9 dage. B) Gennemsnitlig eksponeringstid for sekundære tilfælde (blå), som funktion af den samlede ventetid til test og opsporing. Den orange graf viser gennemsnittet i ventetiden til test og opsporing for primærtilfældet som reference. C) Det effektive kontakttal Rt efter kontaktopsporing som funktion af ventetiden fra symptomer til testsvar hvor udgangspunktet er et kontakttal på 1.2, inden der iværksættes opsporing. Farverne indikerer hvor stor en andel af kontakter der opspores, hvorved eksponeringstiden reduceres.

Bilag 4. Udvikling i antal kontakter fra HOPE projektet



Figur 11.Kilde: Hope-projektet (12.11.2020). Estimating Local Protective Behavior in Denmark with dynamic MRP. https://github.com/mariefly/HOPE/raw/master/HOPE_report_2020-11-12.pdf

Bilag 5. Beskrivelse af parametre brugt i rapporten

Modellerne i rapporten bygger på en række parametre. Estimaterne, som parametre er baseret på er udvalgt af den relevante institution, der har udarbejdet modellerne. Begrundelsen for valg af estimaterne er beskrevet nærmere i dette bilag.

Overordnet set er parametre om sygdomsforløb primært baseret på international litteratur på emnet, men også på data fra den danske befolkning. Estimater over befolkningens adfærd i forbindelse med covid-19 bygger på en række danske undersøgelser fra i år, samt på data over danskernes rejsemønstre.

Estimater for latensperiode, inkubationsperiode og infektiøs periode fra litteraturen:

Særligt relevant for simuleringerne over effekten af kontaktopsporing er estimaterne bag sygdomsforløbet, herunder hvor lang tid der går fra eksponering til, at vedkommende kan smitte, og derefter til, at vedkommende vises symptomer. Estimaterne i modellen er blandt andet baseret på andre forskeres data, som er offentliggjort i international litteratur om covid-19.

For at finde de bedste estimat på *latensperioden* har modelgruppen trianguleret distributioner fra nedenstående kilder. Estimatet er 3,6 dage med et interval på mellem 3-5 dage.

- Read et al. (2020). Novel coronavirus 2019-nCoV: Early estimation of epidemiological parameters and epidemic predictions. *Preprint*.
- Li et al. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N. Engl. J. Med.*
- Li et al. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science.*
- Milne and Xie (2020). The Effectiveness of Social Distancing in Mitigating COVID-19 Spread: a modelling analysis. *Preprint*.

For at finde det bedste estimat af *inkubationsperioden*, har Ekspertgruppen gennemgået nedenstående litteratur. Estimatet er 5 dage med et interval på mellem 3-7 dage.

- Lauer et al. (2020). The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Ann. Int. Med.*
- Li et al. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N. Engl. J. Med.*
- Anderson et al. (2020). How will country-based mitigation measures influence the course of the COVID-19 epidemic. *The Lancet.*
- Linton et al. (2020). Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data. *J. Clin. Med.*
- Liu et al. (2020). Transmission dynamics of 2019 novel coronavirus (2019-nCoV). *bioRχiv*.
- Shen et al. (2020). Modelling the epidemic trend of the 2019 novel coronavirus outbreak in China. *bioRχiv*.

- Backer et al. (2020). Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Euro Surveill*.
- Gostic et al. (2020). Estimated effectiveness of symptom and risk screening to prevent the spread of COVID-19. *eLife*
- Hellewell et al. (2020). Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*.
- Milne and Xie (2020). The Effectiveness of Social Distancing in Mitigating COVID-19 Spread: a modelling analysis. *Preprint*.

For estimatet af *den infektiøse periode*, hvor det bedste estimat er 5 dage, mens det bedste interval er mellem 3-7 dage, har Ekspertgruppen gennemgået følgende artikler:

- Read et al. (2020). Novel coronavirus 2019-nCoV: Early estimation of epidemiological parameters and epidemic predictions. *Preprint*.
- Prem et al (2020). The effect of control strategies that reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China. *Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working and Jit, Mark and Klepac, Petra, The Effect of Control Strategies that Reduce Social Mixing on Outcomes of the COVID-19 Epidemic in Wuhan, China*.
- Milne and Xie (2020). The Effectiveness of Social Distancing in Mitigating COVID-19 Spread: a modelling analysis. *Preprint*.

HOPE rapporter og data:

En del af estimaterne i modellerne om befolkningens adfærd, herunder kontaktmønstre, bygger på både data og rapporter for Hope-projektet (https://hope-project.dk/#/).

HOPE-projektet udsender løbende spørgeskemaer til tilfældigt udvalgte personer i Danmark vedrørende både deres tillid til myndighederne, og til deres adfærdsmønstre, herunder hvor mange de ses med i forskellige kontaktkategorier, hvor meget afstand de holder fra andre mennesker etc. Denne information samles i rapporter, der løbende offentliggøres.

Udover HOPE-rapporten, der henvises til i Bilag 4 (*https://github.com/marie-fly/HOPE/raw/master/HOPE_report_2020-11-12.pdf*), oversender HOPE-projektet løbende anonymiserede data om befolkningens adfærd under covid-19 til Ekspertgruppen, der anvender det i deres modeller. Ekspertgruppen har også adgang til HOPE-projektets rapporter, der sammenskriver data.

Trafik data:

Antagelser om befolkningens adfærd bygges ligeledes på trafikdata, hvorudfra man kan bestemme danskernes rejsemønstre. Efter aftale med Trafik-, Bygge- og Boligstyrelsen får Ekspertgruppen løbende adgang anonymiserede data over danskernes bevægelse rundt i landet. Data er bl.a. brugt til at bestemme den typiske afstand mellem kontakter og afstanden mellem afstandsuafhængige kontakter. Data bygger på 5 forskellige kilder:

- Overblik over rejsende, der bruger rejsekort, som kommer fra Rejsekort og Rejseplanen A/S
- Overblik over biltrafik på Øresunds- og Storebæltsbroen fra Sund og Bælt A/S

- Overblik over flytrafik (antal passagerer) til og fra Københavns Lufthavn og Billund Lufthavn
- Overblik over biltrafikken på Statsvejsnettet og cykeltrafikken (samlet ud fra tællestandere) leveret af Vejdirektoratet.
- Overblik og færgetrafik på 5 rederier, der dækker over 17 færgeruter. Data er leveret af Danske Rederier.

Estimater for ventetider til test

Estimater for ventetider til test og svar på test er taget fra TCDKs hjemmeside (https://tcdk.ssi.dk/vente-og-svartider).

Data fra SSIs Linelisten

Linelisten på SSI indeholder informationer om de covid-19 podninger, der tages en given dag. Data fra Linelisten er bl.a. brugt til at modellere risikoen for at blive hospitaliseret i løbet af et covid-19-forløb for personer over og under 60 år.

Spørgeskemaundersøgelse blandt covid-19 syge lavet af SSI i foråret:

I foråret 2020 foretog SSI en telefonisk spørgeskemaundersøgelse blandt en række personer, der fik konstateret covid-19. Spørgsmålene undersøgte deltagernes sygdomsforløb, herunder symptomer, hvorvidt nære kontakter i husttanden var smittet og lignende.

Data fra spørgeskemaundersøgelsen blev i modellerne brugt til at estimere tiden fra symptomdebut til tests i dage.

Den nationale prævalensundersøgelse for covid-19:

SSI iværksatte i maj en undersøgelse af, hvor udbredt covid-19 var blandt danskerne. Undersøgelsen bestemmer seroprævalencen blandt et repræsentativt udsnit af danskerne fra maj og til i dag. Informationer fra prævalensundersøgelsen har været anvendt i modellerne til at estimere sandsynligheden for at få symptomer og blive testet.

Bilag 6. Medlemmer af ekspertgruppen

Ekspertgruppen ledes af læge Camilla Holten Møller og overlæge Robert Leo Skov, Infektionsberedskabet, Statens Serum Institut.

Danmarks Tekniske Universitet, Institut for Matematik og Computer Science

- Kaare Græsbøll, ph.d., MSc, Seniorforsker, Sektion for dynamiske systemer
- Lasse Engbo Christiansen, ph.d., MSc Eng, lektor, Sektion for dynamiske systemer
- Sune Lehmann, Professor, Afdelingen for Kognitive Systemer
- Uffe Høgsbro Thygesen, Civilingeniør, ph.d., lektor, Sektion for dynamiske systemer

Københavns Universitet, Det Sundhedsvidenskabelige Fakultet, Institut for Veterinær- og Husdyrvidenskab,

- Carsten Thure Kirkeby, Seniorforsker, ph.d., MSc. Sektion for Animal Welfare and Disease Control
- Matt Denwood, BVMS, ph.d., Sektion for Animal Welfare and Disease Control

Københavns Universitet, Institut for Folkesundhedsvidenskab

- Theis Lange, Vice Institutleder, Lektor i Biostatistik, ph.d., Biostatistisk Afdeling

Københavns Universitet, Niels Bohr Instituttet

- Troels Christian Petersen, Lektor, Eksperimentel subatomar fysik

Roskilde Universitets Center, Institut for Naturvidenskab og Miljø

- Viggo Andreasen, Lektor, Matematik og Fysik

Region Hovedstaden

- Anders Perner, Professor, Overlæge, Intensivafdelingen, Rigshospitalet

Danmarks Statistik

- Laust Hvas Mortensen, Chefkonsulent, professor, ph.d., Metode og Analyse

Statens Serum Institut

- Mathias Heltberg, Postdoc ENS Paris samt Statens Serum Institut. Infektionsberedskabet
- Frederik Plesner Lyngse, Postdoc, Økonomisk Institut, Københavns Universitet samt Statens Serum Institut, Infektionsberedskabet
- Peter Michael Bager, Seniorforsker, ph.d., Infektionsberedskabet, Epidemiologisk Forskning, Statens Serum Institut
- Robert Skov, Overlæge, Infektionsberedskabet, Statens Serum Institut
- Camilla Holten Møller, Læge, PhD, Infektionsberedskabet, Statens Serum Institut

C *SSI Notat*

The following pages contain the report from Statens Serum Institut, the Danish CDC:

> Ekspertgruppen for matematisk modellering, *"Scenarier for udviklingen i den engelske virusvariant af SARS-COV-2 (cluster B.1.1.7)"* (Statens Serum Institut, 2021).

The report is from January 2 2021 and is a summary of the estimated spread of the "alpha" variant of COVID-19 (B.1.1.7) in Denmark. The report is in Danish and is based on two models, one from DTU and our agent based model from NBI.

STATENS SERUM INSTITUT

d. 2. januar 2021

Notatet er opdateret d. 22. januar 2021 med en præcisering af formuleringer vedrørende udviklingen i forholdet mellem Cluster B.1.1.7 og øvrige virusvarianter.

Scenarier for udviklingen i den engelske virusvariant af SARS-COV-2 (cluster B.1.1.7)

Ekspertgruppen for matematisk modellering, der ledes fra SSI, bringer i dette notat en række estimater for den forventede udbredelse af cluster B.1.1.7 i den kommende periode, dels ved logistisk regression af udviklingen i forekomsten af varianten, og dels ud fra simuleringer af spredningen af varianten i en agentbaseret model.

Sammenfatning

- Den observerede udvikling i forekomsten af cluster B.1.1.7 i Danmark, svarer til en ugentlig vækstrate for forholdet mellem cluster B.1.1.7 og de øvrige virusvarianter på 72% (95% CI: [37, 115] %).
- Med udgangspunkt i den aktuelle situation hvor 2,3% af virusvarianterne i den rutinemæssige helgenomsekventering tilhører cluster B.1.1.7, estimeres det, at varianten vil udgøre halvdelen af de cirkulerende virusstammer i Danmark om 40-50 dage såfremt ovennævnte stigning fortsætter.
- Det nuværende niveau af restriktioner forventes ikke at være tilstrækkeligt til at få kontakttallet for cluster B.1.1.7 under 1. Derfor vil denne vokse eksponentielt upåagtet at det samlede kontakttal (for alle virusvarianter) kan være under 1 indtil cluster B.1.1.7 overtager om omkring en måned.
- Forekomsten af cluster B.1.1.7 er højest i Region Nordjylland, og udviklingen i forekomsten er ca. fire uger foran Region Hovedstaden.
- Det er på baggrund af engelske data estimeret at kontakttallet er ca. 1,5 gange højere for den nye virusvariant i forhold til andre virusvarianter.
- Den reduktion i smittetal og indlæggelser, der kan opnås i den kommende måned vil give et lavere udgangspunkt for den forøgede smitte og stigende kontakttal, som vi må forvente.

Disse beregninger er behæftet med usikkerheder af forskellige grunde. I perioden op til jul var der stor efterspørgsel på tryghedstest, og i samme periode er der udført et stigende antal antigen test. Derimod så vi i juledagene, at kun ganske få har ladet sig teste. Disse ændringer i testdynamikker gør det svært at følge udviklingen i covid-19, idet de vanlige indikatorer såsom incidenser, positivprocenter og kontakttallet påvirkes af den ændrede fordeling af covid-19-positive blandt de testede. Et lignende mønster forventes i dagene op til og efter nytår. Desuden har vi endnu ikke set effekten af de sidst indførte tiltag, herunder lukning af detailhandlen og liberale erhverv. Samlet set giver dette usikkerhed omkring det aktuelle kontakttal. Analysen er baseret på 76 isolater med cluster B.1.1.7 fordelt på de fem regioner. Den lille stikprøve giver relativt store statistiske usikkerheder. Der vil derfor være behov for at løbende at opdatere estimaterne og lave nye analyser.

Logistisk regression for spredningen af cluster B.1.1.7

Som det fremgår af nedenstående tabel, er der stor forskel på, hvornår man har fundet cluster B.1.1.7 i de enkelte regioner.

Tabel 1. Forekomst af cluster B.1.1.7 i de fem regioner baseret på helgenomsekventering af stikprøver af SARS-CoV-2 positive isolater.

Uge	Hovedstaden		Midtjylland		Nordjylland		Sjælland		Syddanmark	
	B.1.1.7	Total	B.1.1.7	Total	B.1.1.7	Total	B.1.1.7	Total	B.1.1.7	Total
45	0	656	0	283	0	238	0	181	0	200
46	4	420	0	327	0	305	0	132	0	168
47	0	588	0	297	0	240	0	143	0	241
48	3	679	0	291	0	169	0	165	0	195
49	0	825	0	332	3	64	0	246	0	208
50	2	892	0	360	7	92	0	214	1	431
51	3	753	0	524	9	254	3	310	4	354
52	8	774	5	221	12	169	10	193	1	225

Ud fra udbredelsen af cluster B.1.1.7 i Danmark samt andelen af nye isolater i overvågningen som er relateret til clusteret, anvendes logistisk regression til at estimere den forventede udbredelse af cluster B.1.1.7. Da fokus er på spredningen af virusvarianten, og ikke på introduktioner af denne, er det kun regioner, hvor der er detekteret isolater tilhørende cluster B.1.1.7 i mindst fire uger – dvs. Region Hovedstaden og Region Nordjylland, der er medtaget i denne første analyse.

Der er lavet logistisk regression med uge og region som forklarende variable. Der er også testet en interaktion, men den er ikke signifikant.

Tabel 2. Estimater for logistisk regression af andelen af cluster B.1.1.7. Referencen repræsenterer Region Hovedstaden.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-32.812	5.679	-5.778	0.000
Uge	0.540	0.112	4.844	0.000
Region Nordjylland	2.221	0.311	7.133	0.000

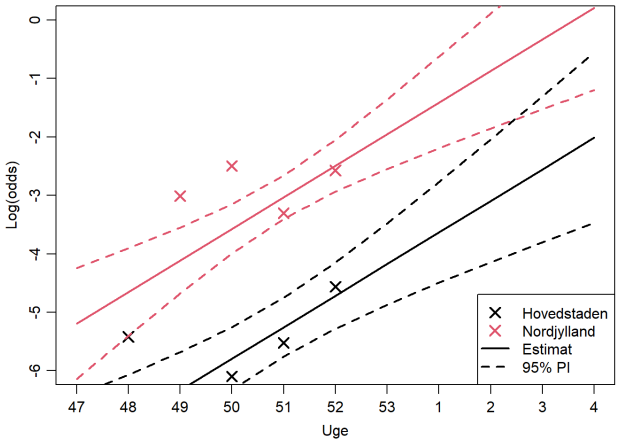
Det ses, at log(odds) for at detektere cluster B.1.1.7 er 2.2 højere i Region Nordjylland end i Region Hovedstaden. Det svarer til odds på 9.2. Det mest interessante er den tidslige udvikling, hvoraf det ses at log(odds) øges med 0.54 for hver uge. Dette svarer til at cluster B.1.1.7 har en ugentlig vækstrate i odds (forholdet mellem antal cluster B.1.1.7 og øvrige virusvarianter) på 72% (95% CI: [37, 115] %), hvilket med den nuværende lave andel af cluster B.1.1.7 svarer til den samme stigning i andelen af cluster B.1.1.7 blandt alle positive prøver. Usikkerheden på estimatet er endnu ganske stort og estimatet er følsomt over for hvilke uger der medtages. Uanset usikkerheder, svarer det fundne estimat til de der er rapporteret fra England for denne virusvariant og det tyder på, at cluster B.1.1.7 har samme forøgede transmissionsrate i Danmark som i England.

Det ses, at log(odds) for at detektere cluster B.1.1.7 er 2.2 højere i Region Nordjylland end i Region Hovedstaden. Det svarer til odds på 9,2, dvs. at sandsynligheden for at detektere cluster B.1.1.7 her er 9,2 gange højere. Det svarer også til at Region Nordjylland er fire uger foran Region Hovedstaden i andelen af cluster B.1.1.7

Det forventes, at usikkerhederne vil blive reduceret væsentligt når der er data for 1-2 uger mere. Men givet at B.1.1.7 er så meget mere smitsom end hidtidige varianter vil det kræve længerevarende restriktioner at sænke smittetallet.

De seneste estimater af kontakttallet er lige under 1,0. Dette er dog påvirket af den ændrede testaktivitet og adfærd hen over jul og nytår, og vi har endnu ikke et overblik over konsekvenserne af sammenkomster i forbindelse med jul og nytår. Endvidere har vi endnu ikke set effekten af nedlukningen af de liberale erhverv og detailhandlen omkring jul. Derfor er det forventningen, at en fastholdelse af de nuværende restriktioner vil give et fald i kontakttallet, hvis man kigger på de virusvarianter som vi har set før introduktionen af cluster B.1.1.7. I England har man estimeret, at deres reference kontakttal var 0,8 for andre virusvarianter og 1,2 for cluster B.1.1.7. Det observerede kontakttal er et vægtet gennemsnit af virusvarianterne i populationen.

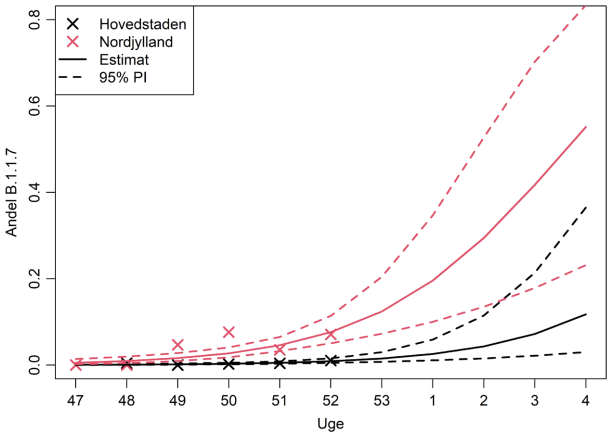
Figur 1 viser en fremskrivning af log(odds) for B.1.1.7 mod andre virusvarianter baseret på ovenstående logistiske regression. Estimatet er, at cluster B.1.1.7 allerede i uge 4 vil udgøre halvdelen af alle positive test i Region Nordjylland. Dette er dog behæftet med stor usikkerhed på baggrund af de nuværende data.



Figur 1. Log(odds) for at detektere cluster B.1.1.7 i hhv. Region Hovedstaden og Region Nordjylland

Ved sammenligning med England er vi nu, hvor de var i starten af november, hvor South East havde log(odds) på -2 svarende til Nordjylland og både London og East of England havde log(odds) omkring -4 svarende til Hovedstaden[1]

Figur 2 viser den samme fremskrivning som i figur 1. Blot er der transformeret tilbage til andelen af positive test, som tilhører cluster B.1.1.7.



[1] 2020_12_23_Transmissibility_and_severity_of_VOC_202012_01_in_England.pdf (cmmid.github.io)

Figur 2. Udviklingen i forekomsten af cluster B.1.1.7 i de kommende uger. Fremskrivningen viser, at halvdelen at isolaterne i Region Nordjylland vil være cluster B.1.1.7 omkring uge 4.

Det skal bemærkes, at udviklingen i Hovedstaden er ca. 4 uger efter udviklingen i Nordjylland. Det er endnu for tidligt at udtale sig om niveauet i de andre tre regioner, men særlig Region Sjælland synes at have oplevet en hurtig stigning, om end det er baseret på meget lidt data. De næste par uger vil forbedre estimatet af niveauet i alle regioner.
Hen over julen har der været et nyt toppunkt i antal indlagte og der er endnu kun set små fald. Det er først i uge 1, at vi kan forvente at se eventuelle indlæggelser som følge af smitte i julen. Alt andet lige må dette forventes at give en yderligere kortvarig pukkel i antal nye indlæggelser.
På nuværende tidspunkt er prognosen, at vi har omkring en måned før det samlede kontakttal for alle virusvarianter hurtigt vil stige på grund af øget udbredelse af cluster B.1.1.7. Hvis restriktionerne skærpes i den kommende tid, vil det give en reduktion i smittetal og indlæggelser og dermed et lavere udgangspunkt for den forøgede smitte og stigende kontakttal, som vi må forvente.

Et første estimat af kontakttallet for cluster B.1.1.7 for perioden uge 47 til 52 og baseret på observationer fra Region Hovedstaden og Region Nordjylland er 1.5 (95% CI [1,2 ; 1,7]) - dette er estimeret vha. Poisson regression med offset lig med 0.7*log(antal sekventerede). Det gennemsnitlige kontakttal (baseret på SSIs publicerede kontakttal 2020-12-29) for perioden er 1,1. Da kontakttallet for cluster B.1.1.7 er så meget højere må det selv med de nuværende restriktioner forventes, at det vedbliver med at være over 1 og dermed forventes cluster B.1.1.7 at vokse eksponentielt, hvis det nuværende niveau af restriktioner fastholdes.

Simulering af spredningen af cluster B.1.1.7 i en agentbaseret model

Agentbaserede modeller
Spredningen af cluster B.1.1.7 er simuleret i en agentbaseret model, som er udviklet af Niels Bohr Instituttet, Københavns Universitet (NBI). En agentbaseret model simulerer et antal agenter (individer i en population) og deres interaktioner med andre agenter, svarende til de interaktioner som en befolkning normal vis har. Hver agent repræsenterer således en person, som er knyttet til en lokation i Danmark, svarende til deres bopæl. Agenterne indgår i flere forskellige netværk, f.eks. husstand, job og skole hvor de har kontakt til andre personer. Derudover har de kontakt til tilfældige personer i samfundet i den tid, hvor personen ikke er hjemme, på job eller i skole. Hvis en agent bliver smittet med SARS-CoV-2, er forløbet for den enkelte agent beskrevet således, at agenten først er eksponeret (E) og derefter infektiøs (I), hvorefter agenten ikke længere er smitsom og betragtes som rask (R). De gennemsnitlige tider i hvert sygdomsstadie kan findes i bilag 1. Hver kontakt, som en agent eksponeres for, tildeles en sandsynlighed for at blive smittet af en anden agent, såfremt denne er smitsom. For en detaljeret beskrivelse af den agentbaserede model, herunder de inkluderede parametre, henvises til bilag 1.

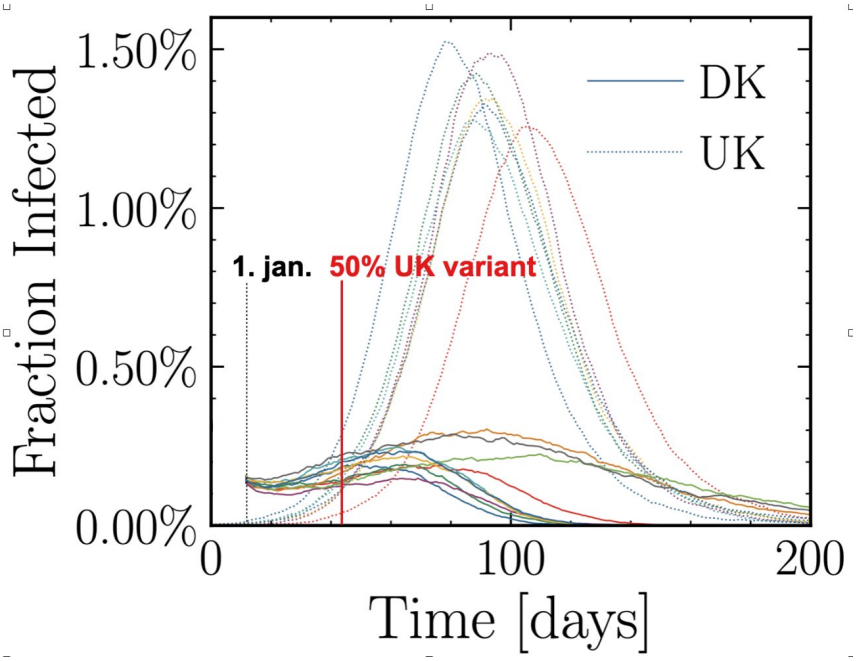
Forbehold

Mens en agentbaseret model kan medtage mere detaljerede dynamikker i en epidemi, så kræver en præcis simulation input fra data, som ofte ikke er tilgængelige eller forefindes, fx hvem en person mødes med i løbet af en dag. Derfor kan en sådan model have unøjagtigheder eller bygge på antagelser, som ikke er retvisende. Det er ikke muligt at kvantificere den nøjagtige størrelse eller effekt af disse potentielle fejlkilder. Da datagrundlaget for disse simuleringer er sparsomt, fordi vi endnu har få datapunkter for cluster B.1.1.7, vil resultatet være behæftet med væsentlig usikkerhed.

Resultater
I det følgende er udviklingen simuleret i en model, hvor udgangspunktet er 1/10 af Danmarks befolkning, og hvor cluster B.1.1.7 fra starten udgør omkring 5% af de cirkulerende virusvarianter. Epidemien simuleres ud fra et kontakttal på omkring 1,0, samt en antagelse om, at cluster B.1.1.7 smitter 50% mere, som rapporteret fra England[2]

Figur 3 viser, hvordan en epidemi vil udvikle sig i tid, forudsat at det simulerede scenarie ikke ændres. Der opdeles i hhv. de nuværende virusvarianter (DK, fulde linjer) og det engelske cluster B.1.1.7 (UK, stiplede linjer). Simulationen er gentaget flere gange (forskellige farver) for at se, hvor store variationer der forekommer. Som det kan ses, så udfases DK-versionen af smitten, mens UK-versionen B.1.1.7 giver ophav til en eksponentiel vækst, idet kontakttallet for denne er væsentligt over 1.
Af figuren fremgår det, at cluster B.1.1.7 ca. 35-40 dage fra simulationens start ("1. jan.") udgør omkring 50% af de cirkulerende virusvarianter. Da simulationen er startet med en større andel UK-varianter (5%) end det aktuelle landsgennemsnit (2.3%), så bliver estimatet 40-50 dage til at halvdelen af de sekventerede varianter tilhører cluster B.1.1.7. I de viste simulationer er de første smittet med cluster B.1.1.7 varianten placeret i Hovedstadsområdet. I andre scenarier, hvor cluster B.1.1.7 varianten i starten udvikler sig i et tyndere befolket område tager udviklingen lidt længere tid, op til 60 dage.

[2]2020_12_23_Transmissibility_and_severity_of_VOC_202012_01_in_England.pdf (cmmid.github.io)



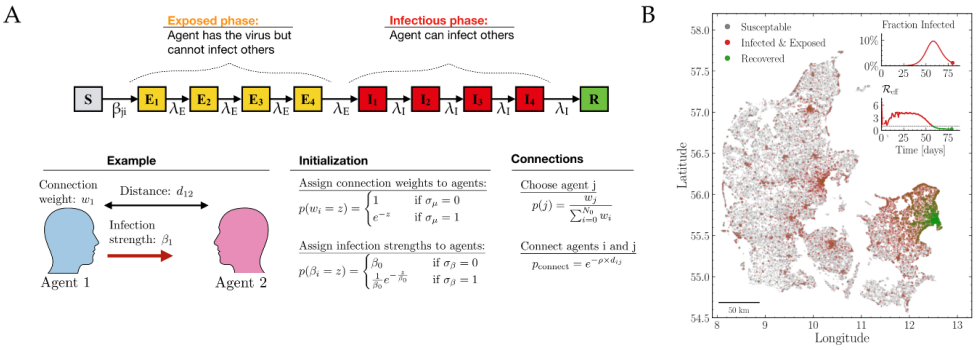
Figur 3. Den forventede udvikling i cluster B.1.1.7 sammenholdt med udviklingen i øvrige virusvarianter, simuleret i en agentbaseret model. Ud fra simulationerne estimeres det, at B.1.1.7 varianten vil være dominerende efter 40-50 dage.

Bilag 1. Beskrivelse af den agentbaserede model

Den nedenstående modelbeskrivelse er et uddrag fra ekspertrapporten "effekten af kontaktopsporing" der er publiceret d. 16. december 2020

Bidrag og udvikling: Christian Michelsen, Emil Martiny, Tariq Halasa, Mogens H. Jensen, Troels C. Petersen og Mathias L. Heltberg

Den agentbaserede model baseres på agenter, dvs. individer hvis karakteristika er tildelt ud fra statistiske fordelinger i befolkningen. Dette er f.eks. en aldersfordeling og en fordeling over pendlerafstande. Modellen starter med at fordele Danmarks bopæle ud i landet baseret på det danske hussalg over de sidste 15 år. Herefter placeres agenter i hver husstand baseret på deres alder og geografiske placering.



Figur 5: A) Skematisk oversigt over hvordan interaktionsnetværket i modellen ser ud. B) Eksempel på simulation af smittespredning i Danmark i modellen, gennem et simuleret tilfælde af flokimmunitet i København.

Et afgørende element i modellen er opbygningen af alle personers interaktionsnetværk. Dette genereres ved, at hver agent har et netværk, de interagerer med. Dette opdeles i tre dele: 1) kontakter i hjemmet, 2) kontakter på arbejdet, 3) kontakter i kategorien andre kontakter. Der er ikke nogen geografisk afhængighed af antallet af kontakter på arbejdet, men i den kategori der kaldes "andre", vil der generelt være flere kontakter for dem der bor i tæt befolkede områder i forhold til dem der bor på landet. Måden hvorpå netværket dannes er vist i Figur 5A.

Ud fra data fra HOPE-projektet har vi estimeret, hvor mange personer hver agent vil interagere med, og i denne model vil alle agenter have mellem 3 og 15 daglige kontakter.

Når modellen simuleres vil alle inficerede agenter gennemgå et forløb, hvor de er i en latent periode, hvor de ikke smitter, hvorefter de vil rykke over i en infektiøs periode, hvor de kan smitte agenter i deres netværk. Denne model simuleres ud fra det der kaldes Gillespie algoritmen, således at netværket opdateres instantant for alle smittebegivenheder. En samling af de væsentligste parametre er vist herunder (Tabel 2).

Tabel 2: Parametre i den agentbaserede model

Parameter	**Værdi interval for middelværdien**	**Reference**
Antal kontakter per dag	3-15	HOPE projektet
Latent tid (dage)	3-5	Litteratur se referenceliste i bilag 5
Infektiøs tid (dage)	4-8	Litteratur se referenceliste i bilag 5
Andel af kontakter i "andre" (%)	30-80	HOPE projektet
Typisk afstand mellem kontakter (km)	5-20	Trafik data
Andel afstandsuafhængige kontakter (%)	3-5	Trafik data
Tid fra symptom til test (Dage)	0-2	Fordeling fra spørgeskemaundersøgelse i foråret 2020 (ikke offentliggjort)
Sandsynlighed for at få symptomer og blive testet (%)	20-60 %	Prævalensundersøgelsen
Sandsynlighed for at kontakte husstand (%)	100%	Antagelse
Sandsynlighed for at kontakte kollegaer (%)	40-80	Antagelse
Sandsynlighed for at kontakte andre (%)	0-75	Antagelse

This document was typeset using LaTeX and modified version of the `tufte-style-thesis` class.
The style is heavily inspired by the works of Edward R. Tufte and Robert Bringhurst.