

UNIVERSITY OF  
COPENHAGEN



PH.D. THESIS  
by  
*Christian Michelsen*

## Biological Data Science

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

*Submitted: November 14, 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 14, 2022.

*Til kvinderne i mit liv*



# *Table of Contents*

## FRONT

Preface i

Acknowledgements iii

Abstract v

Dansk Abstract vii

Publications ix

## MAIN

1 Introduction 1

- 1.1 Ancient DNA and Bayesian Statistics 2
- 1.2 Anesthesiology – a Machine Learning Approach 4
- 1.3 COVID-19 and Agent Based Models 4
- 1.4 Diffusion Models and Bayesian Model Comparison 4

Bibliography 5

2 Paper I 7

3 Paper II 45

4 Paper III 85

5 Paper IV 95

## APPENDIX

A Kap København 109

B SSI Ekspertrapport 143

C SSI Notat 173



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 taxonomic 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*”.
- 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*.”
- 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*”. 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*”.



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*

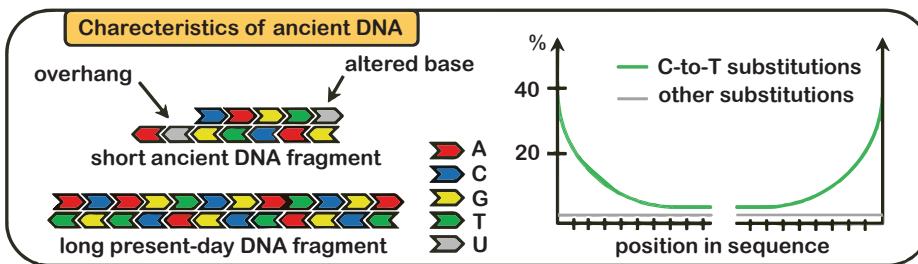
Previously, the only way to study ancient animals, plants, and other species was by studying their fossils. This changed in the middle of the 1980's when the first DNA was recovered from almost 5000 years old ancient mummies, showing that it was indeed possible to extract and sequence ancient DNA (Pääbo, 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 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; Renaud et al., 2019). 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 is a problem since XXX. Deamination is a process in which cytosine (C) in the single-stranded overhangs in the end of the DNA molecules is often hydrolyzed 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 one of the main reasons behind nucleotide misincorporations in ancient DNA (Briggs et al., 2007). Figure 1 illustrates both fragmentation and deamination of ancient DNA.



Currently, a handful 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<sup>1</sup> to identify ancient DNA damage. One of the key features of this method is the beta-binomial model which allows the uncertainty to be 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 for this type of data is a binomial distribution. The mean and variance of the binomial distribution is given by:

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

**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 primarily happening at the end of the reads. Modified from (Peyrégne and Prüfer, 2020).

<sup>1</sup> for the forward strand and the G-to-A deamination pattern for the reverse strand

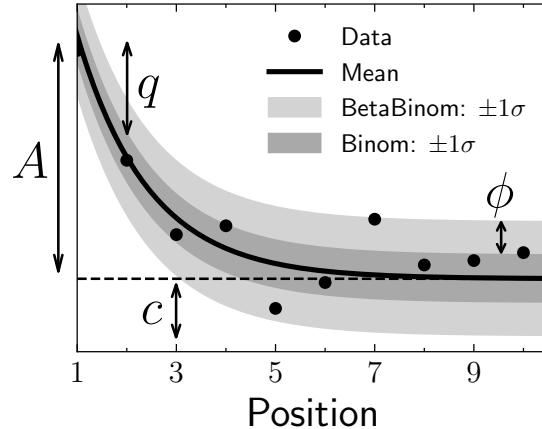
**Figure 2.**

Illustration of the damage model. The figure shows data points as circles and the damage,  $f(x)$ , 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.

## 1.2 Anesthesiology – a Machine Learning Approach

asdadasd

## 1.3 COVID-19 and Agent Based Models

asdadasdads

## 1.4 Diffusion Models and Bayesian Model Comparison

asdadas

# Bibliography

- 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. URL: <https://peerj.com/articles/11845> (visited on 2022).
- 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. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1976210/> (visited on 2019).
- 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. URL: <http://cshperspectives.cshlp.org/content/5/7/a012567> (visited on 2022).
- 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. DOI: 10.1093/bioinformatics/btt193. pmid: 23613487. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3694634/> (visited on 2019).
- 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. URL: <https://www.nature.com/articles/nature08976> (visited on 2022).
- 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.
- 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. URL: <https://www.nature.com/articles/nature12323> (visited on 2022).

- 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. URL: <https://www.nature.com/articles/314644a0> (visited on 2022).
- (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. URL: <https://www.sciencedirect.com/science/article/pii/0305440385900020> (visited on 2022).
- 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. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.202000081> (visited on 2022).
- 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. URL: [https://doi.org/10.1007/978-1-4939-9176-1\\_17](https://doi.org/10.1007/978-1-4939-9176-1_17) (visited on 2022).
- 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. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6020069/> (visited on 2022).
- 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. URL: <https://www.nature.com/articles/s41586-021-03224-9> (visited on 2022).

## **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”.

asdasdsdad

## metaDMG:

# <sup>2</sup> An Ancient DNA Damage Toolkit

**Christian Stentoft Michelsen** <sup>1</sup> <sup>†</sup>  **Mikkel Winther Pedersen** <sup>2</sup> <sup>†</sup>  **Antonio**

**Fernandez-Guerra** <sup>2</sup>  **Lei Zhao** <sup>2</sup>, **Troels C. Petersen** <sup>1</sup>  **Thorfinn Sand**  
**Korneliussen** <sup>2</sup> 

 **For correspondence:**

[christianmichelsen@gmail.com](mailto:christianmichelsen@gmail.com)

(CM); [mwpedersen@sund.ku.dk](mailto:mwpedersen@sund.ku.dk)

(MW)

<sup>1</sup> Niels Bohr Institute, University of Copenhagen; <sup>2</sup> Globe Institute, University of Copenhagen

8

<sup>†</sup>Authors contributed equally.

### Abstract

**Present address:** Niels Bohr

Institute, University of  
Copenhagen, Blegdamsvej 17,  
2100 Copenhagen, Denmark

**Data availability:** Data is  
available on [Zenodo](#) or the  
[Github](#) repository.

**Funding:** This work was  
supported by Carlsberg  
Foundation Young Researcher  
Fellowship awarded by the  
Carlsberg Foundation  
[CF19-0712], and the  
Lundbeck Foundation Centre  
for Disease Evolution:  
[R302-2018-2155 to L.Z]. The  
funders had no role in the  
decision to publish.

**Competing interests:** The  
author declare no competing  
interests.

<sup>10</sup> **1. Motivation** Under favourable conditions DNA molecules can survive for more than two million years (Kjaer et al in press). Such genetic remains can give unique insights to past assemblages, populations and evolution of species. However, DNA is degraded over time, and are therefore found in ultra low concentrations making it highly prone to contamination from modern DNA sources. Despite strict precautions implemented in the field (Llamas et al., 2017), DNA from modern sources does appear in the final output data. One authenticity criteria used in all ancient DNA studies are the high nucleotide mis-incorporation rates that can be observed as a result of chemical post-mortem DNA damage, in fact misincorporation patterns have become instrumental to authenticate ancient sequences. To date this has primarily been possible for single organisms (Jónsson et al., 2013) and recently for assemblies (Borry et al., 2021), but these methods have not been designed, nor can they be computationally upscaled to calculate the thousands of taxonomic species that occur in just one metagenome.

<sup>22</sup> **2. Methods** We present metaDMG, a novel framework that takes advantage of the information already contained in the alignment files to compute and statically evaluate post-mortem DNA damage, thus bypassing the need for classifying and splitting reads into individual organisms and realigning these to parse data to mapDamage2.0 (Jónsson et al., 2013). It

uses a Bayesian approach that combines a geometric damage profile with a beta-binomial model to fit the entire model to the misincorporations which drastically improve the damage estimates compared to previous methods.

**30 3. Results** Using a two-tier simulation setup, we find `metaDMG` to not only be a factor of 10

faster than previous methods but it is also more accurate and able to evaluate even

**32** complex metagenomes with tens of thousands of species. Even with very few number of

reads, down to even below 1000 reads. BLABLA, more results here.

**34 4. Conclusion** `metaDMG` includes state-of-the-art statistical methods for computing nucleotide

misincorporation and fragmentation patterns of even highly complex samples along with

**36** re-implementation of the current statistics used within the field such as `PMDtools`

(Skoglund et al., 2014). This suite of programs is freely available and consists of

**38** computational parts implemented as multi threaded C++ programs as well as

computationally optimized modern python libraries, and an interactive dashboard for

**40** displaying the results. `metaDMG` is furthermore flexible, compatible with custom databases,

can output nucleotide misincorporation and fragmentation patterns at different taxonomic

**42** ranks as well as per reference ID.

**46 keywords:** ancient DNA, damage estimation, DNA damage, lowest common ancestor, `metaDMG`,

**48** metadamage, metagenomics, statistics.

## 1 | INTRODUCTION

Throughout the life of an organism, it contaminates its surrounding environment with cells or

**48** tissue and hence its DNA contained within. As the cell leaves its host, DNA repair mechanisms

stops and the DNA is now subjected to chemical and mechanical degradation, resulting in frag-

**50** mented molecules and chemical damages, characteristic for ancient DNA (Briggs et al., 2007; Dab-

ney, Meyer, and Pääbo, 2013). Ancient DNA has been shown to be able to survive in the envi-

**52** ronment for thousands and even up to two million years (Kjaer et al in review), and have been

widely used to study past organisms and organism composition (Cappellini et al., 2018). Partic-

**54** ularly misincorporations of cytosines on thymines as a result of deamination has been found to

independently authenticate ancient DNA origin (Dabney, Meyer, and Pääbo, 2013; Ginolhac et al.,

56 2011). Postmortem damage with regards to DNA is characterized by the four Briggs parameters  
 (Briggs et al., 2007). A damaged dna fragment tend to be short, and is likely to be single stranded  
 58 at the termini of the fragment. There is an high proportion of C→T substitutions at the single  
 stranded part  $\epsilon_{ss}$ , a somewhat higher C→ T at the double stranded part  $\epsilon_{ds}$ . The length of the sin-  
 60 gle stranded part (*overhang*) follows a geometric distribution  $\lambda$ , and finally there might be breaks at  
 the backbone in the double stranded part  $v$ . It is possible to estimate these four Briggs parameters  
 62 Jónsson et al., 2013 but these four parameters are rarely used directly for asserting "ancientness",  
 and researchers working with ancient DNA tend to simply use the empirical C→T on the first po-  
 64 sition of the fragment together with other supporting summary statistic of the experiment. This  
 ancient DNA (aDNA) authenticity approach, were initially performed on single individual sources  
 66 such as hair, bones, teeth and later on ancient environmental samples such as soil sediments CITE  
 SOMETHING. While this is a relatively fast process for single individuals it becomes increasingly  
 68 demanding, iterative and time consuming as the samples and the diversity within increases, as in  
 the case for metagenomes from ancient soil, sediments, dental calculus, coprolites and other an-  
 70 cient environmental sources. It has therefore been practice to estimate damage for only the key  
 taxa of interest in a metagenome, as a metagenomic sample easily includes thousands of different  
 72 taxonomic entities, that would make a complete estimate an impossible task.

We have devised a novel test statistic in `metaDMG` which takes into account all relevant infor-  
 74 mation in single scalar. For these reasons, we present here `metaDMG`, a tool that enables fast and  
 accurate DNA damage estimation of whole metagenomes within hours. `metaDMG` is designed and  
 76 upscales equally for the increasingly large datasets that are generated in the field of ancient envi-  
 ronmental DNA, but can also with advantage be used to estimate DNA damage of single genomes  
 78 and samples with low complexity, it can even compute an global damage estimate for a given sam-  
 ple. `metaDMG` is compatible with the NCBI taxonomy and can use `ngSLCA` to perform a naïve last  
 80 common ancestor of the aligned reads to get precise damage estimates for the reads classified to  
 different taxonomic nodes. In addition, it is also designed to be used with custom taxonomies and  
 82 metagenomic assembled genomes.

After defining the method and notation used throughout this paper, we show through multi-  
 84 ple sets of simulations that `metaDMG` not only improves on existing methods in the case of single-  
 genome damage estimation but also work for metagenomic samples. Finally, we apply our method  
 86 on a representative mix of nine different metagenomic samples to show the real life performance

of metaDMG.

## 2 | METHODS & MATERIALS

Perhaps the most basic bioinformatic analyses is the difference between two nucleotide sequences.

This assumes that we have a haploid representation of our target organisms and larger differences can be interpreted as larger genetic differences. Obtaining a haploid representation is none trivial, firstly our target organism might not be haploid and we need to construct a consensus genome, secondly data from modern day sequencers are essentially a sampling with replacement process and we need to infer the relative location of each of the possible millions or even billions of short DNA fragments, this is the process which is called mapping or alignment. Thirdly, and the focus for this manuscript, is the quantification of the presence of postmortem damage (PMD) in DNA. PMD mainly manifests as an excess of cytosine to thymine substitutions at the termini of fragments that has been prepared for sequencing. A priori we can not directly observe these actual biochemical changes but we can align each fragment and consider the difference between reference and read as possible PMD, and it is even possible to use the excess of C to T at the single fragment level to separate modern from ancient (data with PMD) (Skoglund et al., 2014). Expanding from the single read all reads for a sequencing experiment and genome to tabulate the overall substitution or mismatch rates to obtain a statistic of the damage (Borry et al., 2021) or even estimate the four Briggs parameters that is traditionally used to characterize the damage signal (Jónsson et al., 2013).

We have devised a general ancient DNA damage toolkit with a special emphasis in a metagenomic setting which implements and expands existing relevant methods but also expands with several state of the art novel methodologies. At the most basic level we have reimplemented the approach given in (Skoglund et al., 2014) which allows for the extracting and separation of highly damaged DNA reads. Secondly under the assumption of vast amounts of data we have defined a full multinomial regression model building on the method in (Cabanski et al., 2012), we show that this will give superior and stable results if it is possible to obtain high depth and coverage data. However in standard ancient DNA context it is generally not possible to obtain vast amounts of data and we propose two novel tests statistics that is especially suited for this scenario. To our knowledge there are no currently available methods that is geared towards damage analysis in a metagenomic setting and existing approaches are essentially based on remapping against the sin-

gle target organism and does not take into account any possible issues with regards to reads being well assigned or specified. Our solution called `metaDMG` (pronounced metadamage), estimates the damage patterns in metagenomic samples in a three step approach. First, the lowest common ancestor (LCA) for each read (mapped to a multi-species reference database) is computed and the mismatch matrix for each leaf node (e.g. taxonomic ID or contig, depending on the database used) is computed based on the mapped reads. Second, `metaDMG` fits a damage model to each leaf node to compute the ancient damage estimates. Finally, the results are visualized in the `metaDMG` dashboard, which is a state of the art graphical user interface that allows for fast and user-friendly interaction with the results for further downstream analysis and visualization.

## 2.1 | Lowest Common Ancestor and Mismatch matrices

For environmental DNA (eDNA) studies we routinely apply a competitive alignment approach where we consider all possible alignments for a given read. Each read is mapped against a multi species reference databases (e.g. nucleotide or RefSeq from NCBI or custom downloaded). A single read might map to a highly conserved gene that is shared across higher taxonomic ranks such as class or even domains. This read will not provide relevant information due to the generality, whereas a read that maps solely to a single species or species from a genus would be indicative of the read being well classified. We seek to obtain the pattern or signal of damage which is done by the tabulation of the cycle specific mismatch rates between our reference and observed sequence for all well classified reads.

In details we compute the lowest common ancestor (lca) for all alignments for each read, this is done using (Wang et al., 2022) and if a read is well classified or properly assigned based on a user defined threshold (species, genus or family) we tabulate the mismatches for each cycle, if a read is not well assigned it is discarded. Pending on the run mode we allow for the construction of these mismatch tables on three different levels. Either we obtain a basic single global mismatch matrix, which could be relevant in a standard single genome aDNA study and similar to the tabulation used in (Jónsson et al., 2013). Secondly we can obtain per reference counts or if a taxonomy database has been supplied we allow for the aggregation from leaf nodes to the internal taxonomic ranks towards the root.

To suit as many users as possible, `metaDMG` takes as input an alignment file (.bam, .sam, or .sam.gz), where Each read is hereafter allowed an equal chance to map against the multiple refer-

<sup>146</sup> ences. One read can therefore attract multiple alignments, and we thus first seek to find the lowest common ancestor (LCA) among the alignments based on the tree structure from the databases and  
<sup>148</sup> a user defined read-reference similarity interval (Wang et al., 2022). Note that `metaDMG` is not limited to the NCBI database and allow for custom databases as well.

<sup>150</sup> Regardless of runmode or weighing scheme used in the possible aggregation we obtain the nucleotide substitution frequencies across reads which provides us with the position dependent  
<sup>152</sup> mismatch matrices,  $\underline{M}(x)$ , with  $x$  denoting the position in the read, starting from 1. At a specific position,  $M_{\text{ref} \rightarrow \text{obs}}(x)$  describes the number of nucleotides that was mapped to a reference base  $B_{\text{ref}}$   
<sup>154</sup> but observed to be  $B_{\text{obs}}$ , where  $B \in \{A, C, G, T\}$ . The number of C to T transitions, e.g., is denoted as  $M_{C \rightarrow T}(x)$ .

<sup>156</sup> When calculating the mismatch matrix, two different approaches can be taken. Either all alignments of the read will be counted, which we will refer to as weight-type 0, or the counts will be  
<sup>158</sup> normalized by the number of alignments of each read; weight-type 1 (default).

## 2.2 | Damage Estimation

<sup>160</sup> The damage pattern observed in aDNA has several features which are well characterized. By modelling these, one can construct observables sensitive to aDNA signal. We model the damage patterns seen in ancient DNA by looking exclusively at the  $C \rightarrow T$  transitions in the forward direction (5') and the  $G \rightarrow A$  transitions in the reverse direction (3'). For each LCA, we denote the number of  
<sup>162</sup> transitions  $k(x)$  as:

$$k(x) = \begin{cases} M_{C \rightarrow T}(x) & \text{for } x > 0 \quad (\text{forward}) \\ M_{G \rightarrow A}(x) & \text{for } x < 0 \quad (\text{reverse}) \end{cases} \quad (1)$$

<sup>166</sup> and the number of the reference counts  $N(x)$ :

$$N(x) = \begin{cases} \sum_{i \in B} M_{C \rightarrow i}(x) & \text{for } x > 0 \quad (\text{forward}) \\ \sum_{i \in B} M_{G \rightarrow i}(x) & \text{for } x < 0 \quad (\text{reverse}), \end{cases} \quad (2)$$

<sup>170</sup> where the sum is over all four bases. The damage frequency is thus  $f(x) = k(x)/N(x)$ . A natural choice of likelihood model would be the binomial distribution. However, we found that a binomial  
<sup>172</sup> likelihood lacks the flexibility needed to deal with the large amount of variance (overdispersion) we found in the data due to bad references and misalignments.

<sup>174</sup> To accommodate overdispersion, we instead apply a beta-binomial distribution,  $\mathcal{P}_{\text{BetaBinomial}}$ , which  
<sup>175</sup> Note that we do not parameterize the beta distribution in terms of the common  $(\alpha, \beta)$  parameterization, but instead using the more intuitive  $(\mu, \phi)$  parameterization. One can re-parameterize  $(\alpha, \beta) \rightarrow (\mu, \phi)$  using the following two equa-

<sup>176</sup> <sup>177</sup> treats the probability,  $p$ , as a random variable following a beta distribution<sup>1</sup> with mean  $\mu$  and concentration  $\phi$ :  $p \sim \text{Beta}(\mu, \phi)$ . The beta-binomial distribution has the following probability density function:

$$\mathcal{P}_{\text{BetaBinomial}}(k | N, \mu, \phi) = \binom{N}{k} \frac{B(k + \mu\phi, N - k + \phi(1 - \mu))}{B(\mu\phi, \phi(1 - \mu))}, \quad (3)$$

<sup>178</sup> where  $B$  is defined as the beta function:

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \quad (4)$$

<sup>179</sup> <sup>180</sup> <sup>181</sup> <sup>182</sup> with  $\Gamma(\cdot)$  being the gamma function (Cepeda-Cuervo and Cifuentes-Amado, 2017).

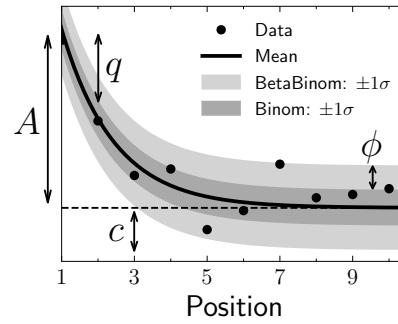
<sup>183</sup> The close resemblance to a binomial model is most easily seen by comparing the mean and variance of a random variable  $k$  following a beta-binomial distribution,  $k \sim \mathcal{P}_{\text{BetaBinomial}}(N, \mu, \phi)$ :

$$\begin{aligned} \mathbb{E}[k] &= N\mu \\ \mathbb{V}[k] &= N\mu(1-\mu)\frac{\phi+N}{\phi+1}. \end{aligned} \quad (5)$$

<sup>183</sup> The expected value of  $k$  is similar to that of a binomial distribution and the variance of the beta-binomial distribution reduces to a binomial distribution as  $\phi \rightarrow \infty$ . The beta-binomial distribution can thus be seen as a generalization of the binomial distribution.

<sup>184</sup> Note that both equation (3) and (5) relate to damage at a specific base position, i.e. for a single <sup>185</sup> <sup>186</sup> <sup>187</sup> <sup>188</sup> <sup>189</sup> <sup>190</sup> <sup>191</sup> <sup>192</sup> <sup>193</sup> <sup>194</sup> <sup>195</sup> <sup>196</sup> <sup>197</sup> <sup>198</sup> <sup>199</sup> <sup>200</sup> <sup>201</sup> <sup>202</sup> <sup>203</sup> <sup>204</sup> <sup>205</sup> <sup>206</sup> <sup>207</sup> <sup>208</sup> <sup>209</sup> <sup>210</sup> <sup>211</sup> <sup>212</sup> <sup>213</sup> <sup>214</sup> <sup>215</sup> <sup>216</sup> <sup>217</sup> <sup>218</sup> <sup>219</sup> <sup>220</sup> <sup>221</sup> <sup>222</sup> <sup>223</sup> <sup>224</sup> <sup>225</sup> <sup>226</sup> <sup>227</sup> <sup>228</sup> <sup>229</sup> <sup>230</sup> <sup>231</sup> <sup>232</sup> <sup>233</sup> <sup>234</sup> <sup>235</sup> <sup>236</sup> <sup>237</sup> <sup>238</sup> <sup>239</sup> <sup>240</sup> <sup>241</sup> <sup>242</sup> <sup>243</sup> <sup>244</sup> <sup>245</sup> <sup>246</sup> <sup>247</sup> <sup>248</sup> <sup>249</sup> <sup>250</sup> <sup>251</sup> <sup>252</sup> <sup>253</sup> <sup>254</sup> <sup>255</sup> <sup>256</sup> <sup>257</sup> <sup>258</sup> <sup>259</sup> <sup>260</sup> <sup>261</sup> <sup>262</sup> <sup>263</sup> <sup>264</sup> <sup>265</sup> <sup>266</sup> <sup>267</sup> <sup>268</sup> <sup>269</sup> <sup>270</sup> <sup>271</sup> <sup>272</sup> <sup>273</sup> <sup>274</sup> <sup>275</sup> <sup>276</sup> <sup>277</sup> <sup>278</sup> <sup>279</sup> <sup>280</sup> <sup>281</sup> <sup>282</sup> <sup>283</sup> <sup>284</sup> <sup>285</sup> <sup>286</sup> <sup>287</sup> <sup>288</sup> <sup>289</sup> <sup>290</sup> <sup>291</sup> <sup>292</sup> <sup>293</sup> <sup>294</sup> <sup>295</sup> <sup>296</sup> <sup>297</sup> <sup>298</sup> <sup>299</sup> <sup>300</sup> <sup>301</sup> <sup>302</sup> <sup>303</sup> <sup>304</sup> <sup>305</sup> <sup>306</sup> <sup>307</sup> <sup>308</sup> <sup>309</sup> <sup>310</sup> <sup>311</sup> <sup>312</sup> <sup>313</sup> <sup>314</sup> <sup>315</sup> <sup>316</sup> <sup>317</sup> <sup>318</sup> <sup>319</sup> <sup>320</sup> <sup>321</sup> <sup>322</sup> <sup>323</sup> <sup>324</sup> <sup>325</sup> <sup>326</sup> <sup>327</sup> <sup>328</sup> <sup>329</sup> <sup>330</sup> <sup>331</sup> <sup>332</sup> <sup>333</sup> <sup>334</sup> <sup>335</sup> <sup>336</sup> <sup>337</sup> <sup>338</sup> <sup>339</sup> <sup>340</sup> <sup>341</sup> <sup>342</sup> <sup>343</sup> <sup>344</sup> <sup>345</sup> <sup>346</sup> <sup>347</sup> <sup>348</sup> <sup>349</sup> <sup>350</sup> <sup>351</sup> <sup>352</sup> <sup>353</sup> <sup>354</sup> <sup>355</sup> <sup>356</sup> <sup>357</sup> <sup>358</sup> <sup>359</sup> <sup>360</sup> <sup>361</sup> <sup>362</sup> <sup>363</sup> <sup>364</sup> <sup>365</sup> <sup>366</sup> <sup>367</sup> <sup>368</sup> <sup>369</sup> <sup>370</sup> <sup>371</sup> <sup>372</sup> <sup>373</sup> <sup>374</sup> <sup>375</sup> <sup>376</sup> <sup>377</sup> <sup>378</sup> <sup>379</sup> <sup>380</sup> <sup>381</sup> <sup>382</sup> <sup>383</sup> <sup>384</sup> <sup>385</sup> <sup>386</sup> <sup>387</sup> <sup>388</sup> <sup>389</sup> <sup>390</sup> <sup>391</sup> <sup>392</sup> <sup>393</sup> <sup>394</sup> <sup>395</sup> <sup>396</sup> <sup>397</sup> <sup>398</sup> <sup>399</sup> <sup>400</sup> <sup>401</sup> <sup>402</sup> <sup>403</sup> <sup>404</sup> <sup>405</sup> <sup>406</sup> <sup>407</sup> <sup>408</sup> <sup>409</sup> <sup>410</sup> <sup>411</sup> <sup>412</sup> <sup>413</sup> <sup>414</sup> <sup>415</sup> <sup>416</sup> <sup>417</sup> <sup>418</sup> <sup>419</sup> <sup>420</sup> <sup>421</sup> <sup>422</sup> <sup>423</sup> <sup>424</sup> <sup>425</sup> <sup>426</sup> <sup>427</sup> <sup>428</sup> <sup>429</sup> <sup>430</sup> <sup>431</sup> <sup>432</sup> <sup>433</sup> <sup>434</sup> <sup>435</sup> <sup>436</sup> <sup>437</sup> <sup>438</sup> <sup>439</sup> <sup>440</sup> <sup>441</sup> <sup>442</sup> <sup>443</sup> <sup>444</sup> <sup>445</sup> <sup>446</sup> <sup>447</sup> <sup>448</sup> <sup>449</sup> <sup>450</sup> <sup>451</sup> <sup>452</sup> <sup>453</sup> <sup>454</sup> <sup>455</sup> <sup>456</sup> <sup>457</sup> <sup>458</sup> <sup>459</sup> <sup>460</sup> <sup>461</sup> <sup>462</sup> <sup>463</sup> <sup>464</sup> <sup>465</sup> <sup>466</sup> <sup>467</sup> <sup>468</sup> <sup>469</sup> <sup>470</sup> <sup>471</sup> <sup>472</sup> <sup>473</sup> <sup>474</sup> <sup>475</sup> <sup>476</sup> <sup>477</sup> <sup>478</sup> <sup>479</sup> <sup>480</sup> <sup>481</sup> <sup>482</sup> <sup>483</sup> <sup>484</sup> <sup>485</sup> <sup>486</sup> <sup>487</sup> <sup>488</sup> <sup>489</sup> <sup>490</sup> <sup>491</sup> <sup>492</sup> <sup>493</sup> <sup>494</sup> <sup>495</sup> <sup>496</sup> <sup>497</sup> <sup>498</sup> <sup>499</sup> <sup>500</sup> <sup>501</sup> <sup>502</sup> <sup>503</sup> <sup>504</sup> <sup>505</sup> <sup>506</sup> <sup>507</sup> <sup>508</sup> <sup>509</sup> <sup>510</sup> <sup>511</sup> <sup>512</sup> <sup>513</sup> <sup>514</sup> <sup>515</sup> <sup>516</sup> <sup>517</sup> <sup>518</sup> <sup>519</sup> <sup>520</sup> <sup>521</sup> <sup>522</sup> <sup>523</sup> <sup>524</sup> <sup>525</sup> <sup>526</sup> <sup>527</sup> <sup>528</sup> <sup>529</sup> <sup>530</sup> <sup>531</sup> <sup>532</sup> <sup>533</sup> <sup>534</sup> <sup>535</sup> <sup>536</sup> <sup>537</sup> <sup>538</sup> <sup>539</sup> <sup>540</sup> <sup>541</sup> <sup>542</sup> <sup>543</sup> <sup>544</sup> <sup>545</sup> <sup>546</sup> <sup>547</sup> <sup>548</sup> <sup>549</sup> <sup>550</sup> <sup>551</sup> <sup>552</sup> <sup>553</sup> <sup>554</sup> <sup>555</sup> <sup>556</sup> <sup>557</sup> <sup>558</sup> <sup>559</sup> <sup>560</sup> <sup>561</sup> <sup>562</sup> <sup>563</sup> <sup>564</sup> <sup>565</sup> <sup>566</sup> <sup>567</sup> <sup>568</sup> <sup>569</sup> <sup>570</sup> <sup>571</sup> <sup>572</sup> <sup>573</sup> <sup>574</sup> <sup>575</sup> <sup>576</sup> <sup>577</sup> <sup>578</sup> <sup>579</sup> <sup>580</sup> <sup>581</sup> <sup>582</sup> <sup>583</sup> <sup>584</sup> <sup>585</sup> <sup>586</sup> <sup>587</sup> <sup>588</sup> <sup>589</sup> <sup>590</sup> <sup>591</sup> <sup>592</sup> <sup>593</sup> <sup>594</sup> <sup>595</sup> <sup>596</sup> <sup>597</sup> <sup>598</sup> <sup>599</sup> <sup>600</sup> <sup>601</sup> <sup>602</sup> <sup>603</sup> <sup>604</sup> <sup>605</sup> <sup>606</sup> <sup>607</sup> <sup>608</sup> <sup>609</sup> <sup>610</sup> <sup>611</sup> <sup>612</sup> <sup>613</sup> <sup>614</sup> <sup>615</sup> <sup>616</sup> <sup>617</sup> <sup>618</sup> <sup>619</sup> <sup>620</sup> <sup>621</sup> <sup>622</sup> <sup>623</sup> <sup>624</sup> <sup>625</sup> <sup>626</sup> <sup>627</sup> <sup>628</sup> <sup>629</sup> <sup>630</sup> <sup>631</sup> <sup>632</sup> <sup>633</sup> <sup>634</sup> <sup>635</sup> <sup>636</sup> <sup>637</sup> <sup>638</sup> <sup>639</sup> <sup>640</sup> <sup>641</sup> <sup>642</sup> <sup>643</sup> <sup>644</sup> <sup>645</sup> <sup>646</sup> <sup>647</sup> <sup>648</sup> <sup>649</sup> <sup>650</sup> <sup>651</sup> <sup>652</sup> <sup>653</sup> <sup>654</sup> <sup>655</sup> <sup>656</sup> <sup>657</sup> <sup>658</sup> <sup>659</sup> <sup>660</sup> <sup>661</sup> <sup>662</sup> <sup>663</sup> <sup>664</sup> <sup>665</sup> <sup>666</sup> <sup>667</sup> <sup>668</sup> <sup>669</sup> <sup>670</sup> <sup>671</sup> <sup>672</sup> <sup>673</sup> <sup>674</sup> <sup>675</sup> <sup>676</sup> <sup>677</sup> <sup>678</sup> <sup>679</sup> <sup>680</sup> <sup>681</sup> <sup>682</sup> <sup>683</sup> <sup>684</sup> <sup>685</sup> <sup>686</sup> <sup>687</sup> <sup>688</sup> <sup>689</sup> <sup>690</sup> <sup>691</sup> <sup>692</sup> <sup>693</sup> <sup>694</sup> <sup>695</sup> <sup>696</sup> <sup>697</sup> <sup>698</sup> <sup>699</sup> <sup>700</sup> <sup>701</sup> <sup>702</sup> <sup>703</sup> <sup>704</sup> <sup>705</sup> <sup>706</sup> <sup>707</sup> <sup>708</sup> <sup>709</sup> <sup>710</sup> <sup>711</sup> <sup>712</sup> <sup>713</sup> <sup>714</sup> <sup>715</sup> <sup>716</sup> <sup>717</sup> <sup>718</sup> <sup>719</sup> <sup>720</sup> <sup>721</sup> <sup>722</sup> <sup>723</sup> <sup>724</sup> <sup>725</sup> <sup>726</sup> <sup>727</sup> <sup>728</sup> <sup>729</sup> <sup>730</sup> <sup>731</sup> <sup>732</sup> <sup>733</sup> <sup>734</sup> <sup>735</sup> <sup>736</sup> <sup>737</sup> <sup>738</sup> <sup>739</sup> <sup>740</sup> <sup>741</sup> <sup>742</sup> <sup>743</sup> <sup>744</sup> <sup>745</sup> <sup>746</sup> <sup>747</sup> <sup>748</sup> <sup>749</sup> <sup>750</sup> <sup>751</sup> <sup>752</sup> <sup>753</sup> <sup>754</sup> <sup>755</sup> <sup>756</sup> <sup>757</sup> <sup>758</sup> <sup>759</sup> <sup>760</sup> <sup>761</sup> <sup>762</sup> <sup>763</sup> <sup>764</sup> <sup>765</sup> <sup>766</sup> <sup>767</sup> <sup>768</sup> <sup>769</sup> <sup>770</sup> <sup>771</sup> <sup>772</sup> <sup>773</sup> <sup>774</sup> <sup>775</sup> <sup>776</sup> <sup>777</sup> <sup>778</sup> <sup>779</sup> <sup>780</sup> <sup>781</sup> <sup>782</sup> <sup>783</sup> <sup>784</sup> <sup>785</sup> <sup>786</sup> <sup>787</sup> <sup>788</sup> <sup>789</sup> <sup>790</sup> <sup>791</sup> <sup>792</sup> <sup>793</sup> <sup>794</sup> <sup>795</sup> <sup>796</sup> <sup>797</sup> <sup>798</sup> <sup>799</sup> <sup>800</sup> <sup>801</sup> <sup>802</sup> <sup>803</sup> <sup>804</sup> <sup>805</sup> <sup>806</sup> <sup>807</sup> <sup>808</sup> <sup>809</sup> <sup>810}</sup> <sup>811</sup> <sup>812</sup> <sup>813</sup> <sup>814</sup> <sup>815</sup> <sup>816</sup> <sup>817</sup> <sup>818</sup> <sup>819</sup> <sup>820</sup> <sup>821</sup> <sup>822</sup> <sup>823</sup> <sup>824</sup> <sup>825</sup> <sup>826</sup> <sup>827</sup> <sup>828</sup> <sup>829</sup> <sup>830</sup> <sup>831</sup> <sup>832</sup> <sup>833</sup> <sup>834</sup> <sup>835</sup> <sup>836</sup> <sup>837</sup> <sup>838</sup> <sup>839</sup> <sup>840</sup> <sup>841</sup> <sup>842</sup> <sup>843</sup> <sup>844</sup> <sup>845</sup> <sup>846</sup> <sup>847</sup> <sup>848</sup> <sup>849</sup> <sup>850</sup> <sup>851</sup> <sup>852</sup> <sup>853</sup> <sup>854</sup> <sup>855</sup> <sup>856</sup> <sup>857</sup> <sup>858</sup> <sup>859</sup> <sup>860</sup> <sup>861</sup> <sup>862</sup> <sup>863</sup> <sup>864</sup> <sup>865</sup> <sup>866</sup> <sup>867</sup> <sup>868</sup> <sup>869</sup> <sup>870</sup> <sup>871</sup> <sup>872</sup> <sup>873</sup> <sup>874</sup> <sup>875</sup> <sup>876</sup> <sup>877</sup> <sup>878</sup> <sup>879</sup> <sup>880</sup> <sup>881</sup> <sup>882</sup> <sup>883</sup> <sup>884</sup> <sup>885</sup> <sup>886</sup> <sup>887</sup> <sup>888</sup> <sup>889</sup> <sup>890</sup> <sup>891</sup> <sup>892</sup> <sup>893</sup> <sup>894</sup> <sup>895</sup> <sup>896</sup> <sup>897</sup> <sup>898</sup> <sup>899</sup> <sup>900</sup> <sup>901</sup> <sup>902</sup> <sup>903</sup> <sup>904</sup> <sup>905</sup> <sup>906</sup> <sup>907</sup> <sup>908</sup> <sup>909</sup> <sup>910</sup> <sup>911</sup> <sup>912</sup> <sup>913</sup> <sup>914</sup> <sup>915</sup> <sup>916</sup> <sup>917</sup> <sup>918</sup> <sup>919</sup> <sup>920</sup> <sup>921</sup> <sup>922</sup> <sup>923</sup> <sup>924</sup> <sup>925</sup> <sup>926</sup> <sup>927</sup> <sup>928</sup> <sup>929</sup> <sup>930</sup> <sup>931</sup> <sup>932</sup> <sup>933</sup> <sup>934</sup> <sup>935</sup> <sup>936</sup> <sup>937</sup> <sup>938</sup> <sup>939</sup> <sup>940</sup> <sup>941</sup> <sup>942</sup> <sup>943</sup> <sup>944</sup> <sup>945</sup> <sup>946</sup> <sup>947</sup> <sup>948</sup> <sup>949</sup> <sup>950</sup> <sup>951</sup> <sup>952</sup> <sup>953</sup> <sup>954</sup> <sup>955</sup> <sup>956</sup> <sup>957</sup> <sup>958</sup> <sup>959</sup> <sup>960</sup> <sup>961</sup> <sup>962</sup> <sup>963</sup> <sup>964</sup> <sup>965</sup> <sup>966</sup> <sup>967</sup> <sup>968</sup> <sup>969</sup> <sup>970</sup> <sup>971</sup> <sup>972</sup> <sup>973</sup> <sup>974</sup> <sup>975</sup> <sup>976</sup> <sup>977</sup> <sup>978</sup> <sup>979</sup> <sup>980</sup> <sup>981</sup> <sup>982</sup> <sup>983</sup> <sup>984</sup> <sup>985</sup> <sup>986</sup> <sup>987</sup> <sup>988</sup> <sup>989</sup> <sup>990</sup> <sup>991</sup> <sup>992</sup> <sup>993</sup> <sup>994</sup> <sup>995</sup> <sup>996</sup> <sup>997</sup> <sup>998</sup> <sup>999</sup> <sup>9999</sup>

<sup>2</sup> Parameterized as  $(\mu, \phi)$



**Figure 1.** Illustration of the damage model. The figure shows data points as circles and the damage,  $f(x)$ , 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 additional uncertainty of the beta-binomial model, compared to the binomial model, is related to  $\phi$ , see equation (5).

and  $c$ . In addition to this, we assume an exponential prior on  $\phi$  with the requirement of  $\phi > 2$  to avoid too much focus on 0-or-1 probabilities (McElreath, 2020). The final model is thus:

$$\begin{aligned}
 \text{[A prior]} & A \sim \text{Beta}(0.1, 10) \\
 \text{[q prior]} & q \sim \text{Beta}(0.2, 5) \\
 \text{[c prior]} & c \sim \text{Beta}(0.1, 10) \\
 \text{[\phi prior]} & \phi \sim 2 + \text{Exponential}(1000) \\
 \text{[likelihood]} & k_i \sim P_{\text{BetaBinomial}}(N_i, \tilde{f}(x_i; A, q, c), \phi),
 \end{aligned} \tag{7}$$

where  $i$  is an index running over all positions.

We define the damage due to deamination,  $D$ , as the background-subtracted damage frequency at the first position:  $D \equiv \tilde{f}(|x| = 1) - c$ . As such,  $D$  is the damage related to ancientness. Using the properties of the beta-binomial distribution, eq. (5), we find the mean and variance of the damage,

$D$ :

$$\begin{aligned}
 \mathbb{E}[D] & \equiv \bar{D} = A \\
 \mathbb{V}[D] & \equiv \sigma_D^2 = \frac{A(1-A)}{N} \frac{\phi + N}{(\phi + 1)}.
 \end{aligned} \tag{8}$$

Since  $D$  estimates the overexpression of damage due to ancientness, not only the mean is relevant but also the certainty of  $D > 0$ . We quantify this through the significance  $Z = \bar{D}/\sigma_D$

<sup>220</sup> which is thus the number of standard deviations ("sigmas") away from zero. Assuming a Gaussian  
<sup>221</sup> distribution of  $D$ ,  $Z > 2$  would indicate a probability of  $D$  being larger than zero, i.e. containing  
<sup>222</sup> ancient damage, with more than 97.7% probability. These two values allows us to not only quantify  
<sup>223</sup> the amount of ancient damage (ie.  $\bar{D}$ ) but also the certainty of this damage ( $Z$ ) without even having  
<sup>224</sup> to run multiple models and comparing these.

We perform the Bayesian inference of the parameters models using Hamiltonian Monte Carlo  
<sup>226</sup> (HMC) sampling which is a particular of Monte Carlo Markov Chain (MCMC) algorithm (Betancourt,  
<sup>227</sup> 2018). Specifically, we use the NUTS implementation in NumPyro (Phan, Pradhan, and Jankowiak,  
<sup>228</sup> 2019), a Python package which uses JAX (Bradbury et al., 2018) under the hood for automatic dif-  
<sup>229</sup> ferentiation and JIT compilation. We treat each leaf node of the LCA as being independent and  
<sup>230</sup> generate 1000 MCMC samples after an initial 500 samples as warm up.

Since running the full Bayesian model is computationally expensive, we also allow for a faster,  
<sup>232</sup> approximate method by just fitting the maximum a posteriori probability (MAP) estimate. We use  
<sup>233</sup> iMinuit (Dembinski et al., 2021) for the MAP optimization with Numba acceleration (Lam, Pitrou,  
<sup>234</sup> and Seibert, 2015) for even faster run times. On a Macbook M1 Pro model from 2021, the timings  
<sup>235</sup> for running the full Bayesian model is  $1.41 \pm 0.04$  s/fit and for the MAP it is  $4.34 \pm 0.07$  ms/fit, showing  
<sup>236</sup> more than a 2 order increase in performance (around 300x) for the approximate model. Both  
models allow for easy parallelisation to decrease the computation time.

### <sup>238</sup> 2.3 | Visualisation

We provide an interactive dashboard to properly visualise the results from the modelling phase,  
<sup>240</sup> see <https://metadmg.onrender.com/> for an example. The dashboard allows for filtering, styling and  
<sup>241</sup> variable selection, visualizing the mismatch matrix related to a specific leaf node, and exporting of  
<sup>242</sup> both fit results and plots. By filtering, we include both filtering by sample, by specific cuts in the fit  
<sup>243</sup> results (e.g. requiring  $D$  to be above a certain threshold), and even by taxonomic level (e.g. only  
<sup>244</sup> looking tax IDs that are part of the Mammalia class). We greatly believe that a visual overview of  
<sup>245</sup> the fit results increase understanding of the data at hand. The dashboard is implemented with  
<sup>246</sup> Plotly plots and incorporated into a Dash dashboard (Plotly, 2015).

### 3 | SIMULATION STUDY

248 We conducted two sets of simulations, one to gauge the performance of the damage model itself  
250 and one to determine the performance of the full `metaDMG` pipeline, i.e. both LCA and damage  
252 model.

#### 3.1 | Single-genome Simulations

252 The first set of simulations was performed by taking a single, representative genome and adding  
254 deamination and sequencing noise to it followed by a mapping step and finally damage estima-  
256 tion using `metaDMG`. The deamination was applied using NGSNGS (XXX, ref here) which is a recent  
258 implementation of the original Briggs model similar to Gargammel (Neukamm, Peltzer, and Nieselt,  
n.d.) but with better performance and more accurate deamination patterns. In this step we vary  
260 the simulated amount of damage added (in particular the single-stranded DNA deamination,  $\delta_{ss}$   
262 in the original Briggs model (Briggs et al., 2007)), the number of reads, and the fragment length  
264 distribution.

266 We chose five different, representative genomes, in each of these varying the three simulation  
268 parameters. These genomes where the homo sapiens, the betula, and three microbial organisms  
270 with respectively low, median, and high amount of GC-content. For each of these simulations,  
272 we performed 100 independent runs to measure the variability of the parameter estimations and  
274 quantify the robustness of the estimates. We simulated eight different sets of damage (approxi-  
276 mately 0%, 1%, 2%, 5%, 10%, 15%, 20%, 30%), 13 sets of different number of reads (10, 25, 50, 100, 250,  
278 500, 1,000, 2,500, 5,000, 10,000, 25,000, 50,000, 100,000), three sets of different fragment length distri-  
280 butions (samples from a lognormal distributions with mean 35, 60, and 90, each with a standard  
282 deviation of 10), and five different genomes, each simulation set repeated 100 times.

284 In addition to this, we also create 1000 repetitions of the non-damaged simulations for Homo  
286 Sapiens to be able to gauge the risk of finding false positives. Finally, to show that the damage esti-  
288 mates that `metaDMG` provides are independent of the contig size, we artificially create three different  
290 genomes by sampling 1.000, 10.000 or 100.000 different basepairs from a uniform categorical dis-  
292 tribution of {A, C, G, T}.

294 To be able to compare our estimates to a known value, we generate 1.000.000 reads from  
296 NGSNGS without any added sequencing noise for each of other sets of simulation parameters.

<sup>276</sup> The difference in damage frequency at position 1 and 15 is then the value to compare to:

$$D_{\text{known}} = \frac{f(x=1) - f(x=15)}{2} + \frac{f(x=-1) - f(x=-15)}{2}, \quad (9)$$

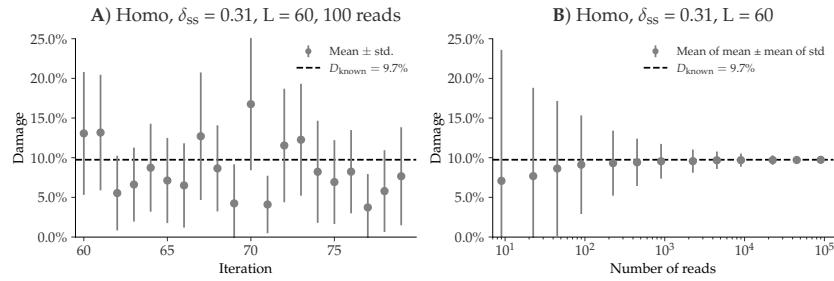
<sup>278</sup> where we take the average of the C to T damage frequency difference and the G to A damage frequency difference.

<sup>280</sup> The fastq files were simulated with NGSNGS using the above mentioned simulation parameters, all with the same quality scores profiles as used in ART (Huang et al., 2012), based on the Illumina <sup>282</sup> HiSeq 2500 (150 bp). The mapping was performed using Bowtie-2 with the `-no-unal` flag (Langmead and Salzberg, 2012).

### <sup>284</sup> 3.2 | Metagenomic Simulations

<sup>286</sup> While the previously mentioned simulation study is perfectly aimed at quantifying the performance of the damage model in the case of single-reference genomics it does lack the complexity related to metagenomic samples. Therefore, we also conduct a more advanced simulation study to determine the accuracy of the full metaDMG pipeline.

<sup>288</sup> The previously mentioned simulation study quantifies the damage model's performance for single-reference genomics, but it does not address the complexity of metagenomic samples. Therefore, we also conducted a more advanced simulation study to determine the accuracy of the full metaDMG pipeline. Based on an ancient metagenome, we created a synthetic dataset that reproduces the composition, fragment length distribution, and damage patterns for each genome. We selected X metagenomes (Supp table XXX) covering several environmental conditions and ages. First, we mapped the reads of each metagenome with bowtie2 against a database that contained the GTDB r202 (Parks et al., 2018) species cluster reference sequences, all organelles from NCBI RefSeq (NCBI Resource Coordinators, 2018), and the reference sequences from CheckV (Nayfach et al., 2021). We used bam-filter 1.0.11 with the flag `--read-length-freqs` to get the mapped read length distribution for each genome and their abundance. The genomes with an observed-to-expected coverage ratio greater than 0.75 were kept. The filtered BAM files were processed by metaDMG to obtain the misincorporation matrices. The abundance tables, fragment length distribution, and misincorporation matrices were used in aMGSIM-smk v0.0.1 (Fernandez-Guerra, 2022), a Snakemake workflow (Mölder et al., 2021) that facilitates the generation of many synthetic ancient metagenomes. The data used and generated by the workflow can be obtained from Figshare link



**Figure 2.** Overview of the single-genome simulations based on the homo sapiens genome with the Briggs parameter  $\delta_{ss} = 0.065$  and a fragment length distribution with mean 60. **A)** This plot shows the estimated damage ( $D$ ) of 10 simulations with 100 simulated reads. The grey points shows the mean damage (with its standard deviation as errorbars). The known damage ( $D_{known}$ ) is shown as a dashed line, see eq. (9). **B)** This plot shows the average damage as a function of the number of reads. The grey points show the average of the individual means (with the average of the standard deviations as errors).

(XXX). We then performed taxonomic profiling using the same parameters used for the synthetic  
306 reads generated by aMGSIM-smk.

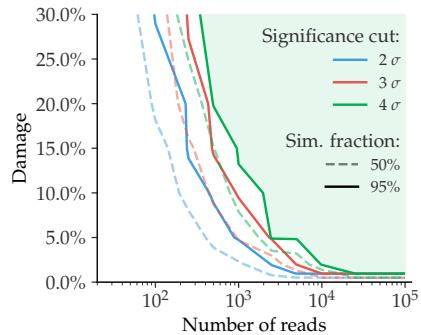
## 4 | RESULTS

308 The accuracy of all methods in metaDMG was tested in various simulation scenarios. In general we  
find that metaDMG yields accurate, precise damage estimates even in extreme low-coverage data.

### 310 4.1 | Single-genome Simulations

The results of the single-genome simulations can be seen in Figure 2. The left part of the figure  
312 shows metaDMG damage estimates based on the homo sapiens genome with the Briggs parameter  
 $\delta_{ss} = 0.31$  and a fragment length distribution with mean 60, each of the 10 simulations generated  
314 with 100 simulated reads for 10 representative simulations. When the damage estimates are low,  
the distribution of  $D$  is highly skewed (restricted to positive values) leading to errorbars sometimes  
316 going into negative damage, which of course represents un-physical values. The right hand side of  
the figure visualizes the average amount of damage across a varying number of reads. This shows  
318 that the damage estimates converge to the known value with more data, and that one needs more  
than 100 reads to even get strictly positive damage estimates (when including uncertainties).

320 Across more than 5 different species, 3 different fragment length distributions, and 3 different

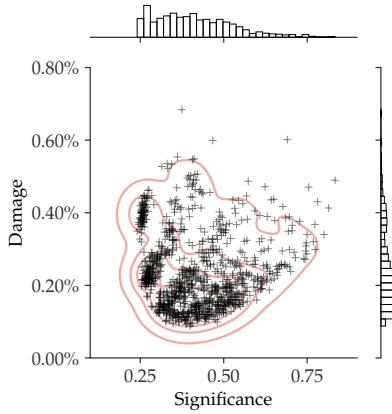


**Figure 3.** Given a specific significance cut, the solid contour line shows the relationship between the amount of damage and the number of reads required to be able to correctly infer damage in 95% of the species. The dashed line shows the similar value for a simulation fraction of 50%. The green part of the figure shows the “good” region of number of reads and estimated damage, given than one wants to be more than 95% certain of correctly identifying damage with more than  $4\sigma$  confidence.

contig length distributions, each with 100 simulations for 104 different sets of simulation parameters, the only difference we note in the damage estimates is between species with low, median, and high GC-levels. In general, species with higher GC-levels exhibit lower variations in their damage estimates compared to species with lower GC-levels, leading to high-GC species requiring fewer reads to establish damage estimates.

Based on the single-genome simulations, we can compute the relationship between the amount of damage in a species and the number of reads required to correctly infer that the given species is damaged, see Figure 3. If we want to find damage with a significance of more than 2 (solid blue line) in a sample with around 5% expected damage, it requires about 1000 reads to be 95% certain that we will find results this good. Said in other words: given 100 different samples, each with 1000 reads and around 5% damage, one would expect to find damage (with a  $Z > 2$ ) in 95 of the total 100 samples, on average. If we loose the requirement such that it is okay to only find it in every second sample, it would be enough with only around 250 reads in each sample (dashed blue line).

Finally, to quantify the risk of incorrectly assigning damage to a non-damaged species, we created 1000 independent simulations for a varying number of reads, where none of them had any artificial ancient damage applied, only sequencing noise. Figure 4 shows the damage ( $D$ ) as a function of the significance ( $Z$ ) for the case of 1000 simulated reads. Even though the estimated damage is larger than zero, the damage is non-significant since the significance is less than one. When looking



**Figure 4.** This figure shows the inferred damage estimates of 1000 independent simulations, each with 1000 reads and no artificial ancient damage applied, with the inferred damage shown on the y-axis and the significance on the x-axis. Each simulation is shown as a single cross and the red lines show the kernel density estimate (KDE) of the damage estimates. The marginal distributions are shown as histograms next to the scatter plot.

at all the figures across the different number of reads, see Figure bayesian\_zero\_damage\_plots.pdf,

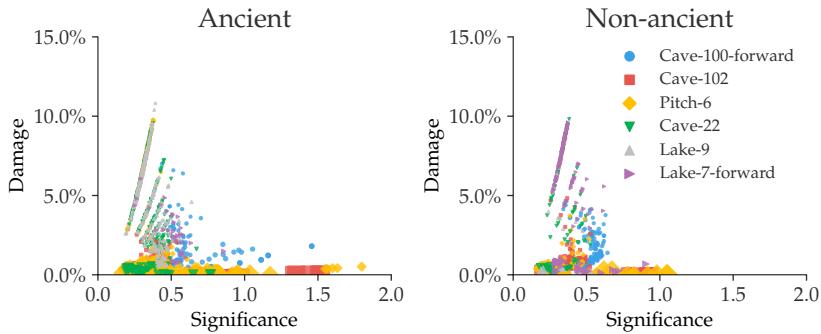
340 we note that a loose cut requiring that  $D > 1\%$  and  $Z > 2$  would filter out all of non-damaged points.

#### 342 4.2 | Metagenomic Simulations

With the full metagenomic simulation pipeline we can further probe the performance of metaDMG.

344 By looking at the six different samples at different steps in the pipeline we are able to show that  
 345 metaDMG provides relevant, accurate damage estimates. First of all, we run metaDMG on the six samples  
 346 after fragmentation with FragSim. Since no deamination has yet been added at this step in the pipeline,  
 this is also a test of the risk of getting false positives. The results can be seen in Figure 5  
 348 where we see the damage estimates for both the species that we simulate to be ancient and the  
 species that we do not add deamination to. We see that the damage estimates are quite similar,  
 350 as expected, and that our previously established loose cut of  $D > 1\%$  and  $Z > 2$  still filters out all  
 of non-damaged points.

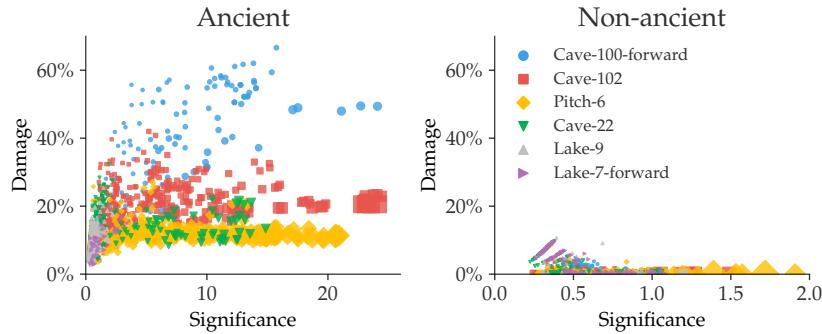
352 In comparison we can look at Figure 6 which shows the same plot, but after the deamination  
 (deamSim) and sequencing errors (ART) has been added. Here we see a clear difference between



**Figure 5.** Estimated amount of damage as a function of significance using the fragSim data. The left figure shows the damage of the species that we simulated to be ancient (however with no deamination added yet) and the right figure shows the same for the species that are not going to have deamination added.

354 the ancient and the non-ancient ones, as expected. The non-ancient species would still not pass  
 355 the loose cut, however, we note that a large number of the ancient samples would. By looking at  
 356 Figure 6 we see that not all of the samples show similar amount of damage. These observations  
 357 are summarised in Table 1 where we see that Cave-100-forward, Cave-102, Pitch-6 all have more  
 358 than 60% of their ancient species labelled as damaged according to the loose cut, Cave-22 (18%)  
 359 and Lake-7-forward (12%) a bit lower, while Lake-9 (0.5%) does not show any clear signs of damage.  
 360 However, once we condition on the requirement of having more than 100 reads, the fraction of  
 361 ancient species correctly identified as ancient increases to more than 90% for most the samples.  
 362 To better understand the damage estimates, we can look a them individually. Figure 7 shows  
 363 the *Stenotrophomonas maltophilia* species from the Pitch-6 sample. We see that none of the  
 364 fragmentation-only files were estimated to have damage and that most of the deamination and  
 365 final files including sequencing errors have damage – at a simulation size of 1 million, the signif-  
 366 icance of both are  $Z \approx 1.9$ , so this one of the few fits with more than 100 reads that does not  
 367 pass the loose cut. Furthermore, we notice that the error bars decrease with simulation size, as  
 368 expected.

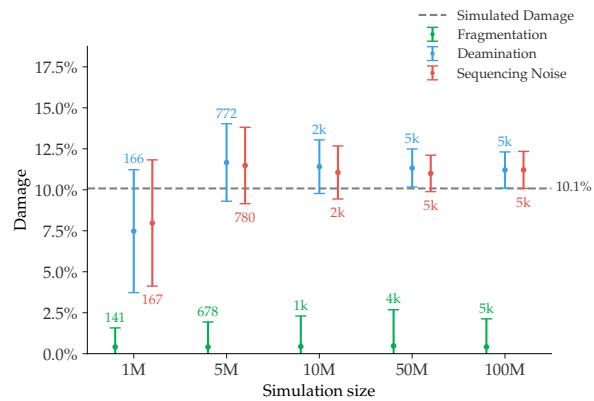
The rest of the metagenomic simulation results are shown in Figure XXX.



**Figure 6.** Estimated amount of damage as a function of significance using the ART data. The left figure shows the damage of the species that we simulated to be ancient and the right figure shows the same for the species that have not had deamination added.

**Table 1.** Number of ancient species for each of the six simulated samples. The first column is the total number of species, the second column is the total number of species that would pass the loose cut of  $D > 1\%$  and  $Z > 2$ , the third column is the number of species with more than 100 reads, and the final column is the number of species with more than 100 reads that also do pass the cut.

Sample	Total	Pass	+100 Reads	+100 Reads and Pass		
Cave-100-forward	135	107	79.3%	88	87	98.9%
Cave-102	500	326	65.2%	309	285	92.2%
Pitch-6	415	260	62.7%	274	260	94.9%
Cave-22	393	71	18.1%	73	69	94.5%
Lake-9	410	2	0.5%	8	0	0%
Lake-7-forward	32	4	12.5%	6	4	66.7%



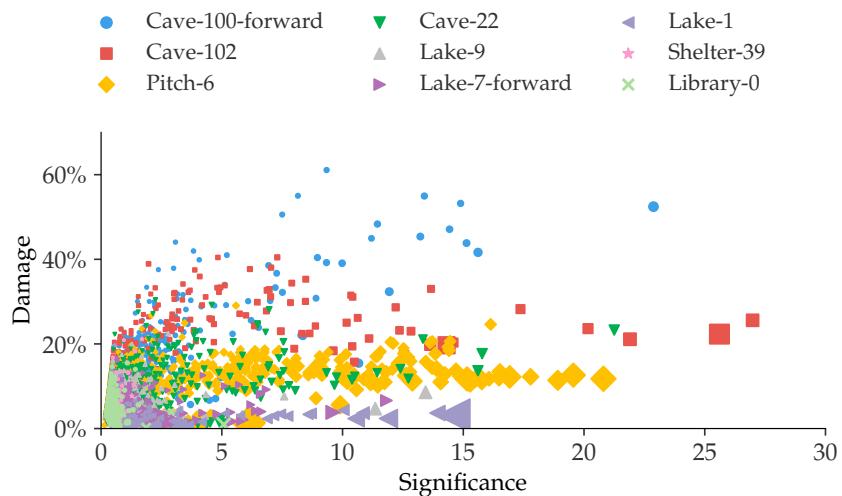
**Figure 7.** Damage estimates of the *Stenotrophomonas maltophilia* species from the Pitch-6 sample. Damage is shown as a function of simulation size, with the fragmentation files in green, the deamination files in blue and the final files including sequencing errors in red. All errors are  $1\sigma$  error bars (standard deviation). The number of reads for each fit is shown as text and since this was a species simulated to have ancient damage, the simulated amount of damage is shown as a dashed grey line.

### 370 4.3 | Real Data

The results from running the full metaDMG pipeline on real data can be seen in Figure 8. The figures  
 372 shows Blablabla, real life data here. We find that the loose cut ( $D > 1\%$ ,  $Z > 2$ ) accepts only one of  
 374 the fits from the control test Library-0, which would not have been accepted by more conservative  
 376 cut ( $D > 2\%$ ,  $Z > 3$ , more than 100 reads).

### 4.4 | Bayesian vs. MAP

376 Due to increased computational burden of running the full Bayesian model compared to faster,  
 378 approximate MAP model, in samples with several thousand species, the MAP model is often the  
 380 most realistic model to use due to time constraints. In this case, it is of course important to know  
 382 that the damage estimates are indeed trustworthy. Figure 9 compares the estimated damage  
 384 between the Bayesian model and the MAP model and the estimated significances for species with  
 $D > 1\%$ ,  $Z > 2$  and more than 100 reads. The figure shows that the vast majority of species map  
 386 1:1 between the Bayesian and the MAP model. One should note, though, that the few species with  
 388 the highest mismatch, all are based on forward-only fits, i.e. with no information from the reverse  
 390 strand, which thus leads to less data to base the fits on. For the comparison with no cuts, see



**Figure 8.** Estimated amount of damage as a function of significance using the real data.

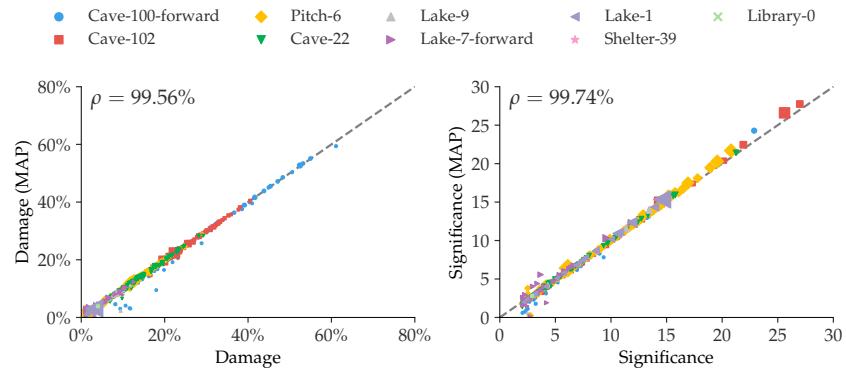
Figure 1 in appendix.

#### 386 4.5 | Existing Methods

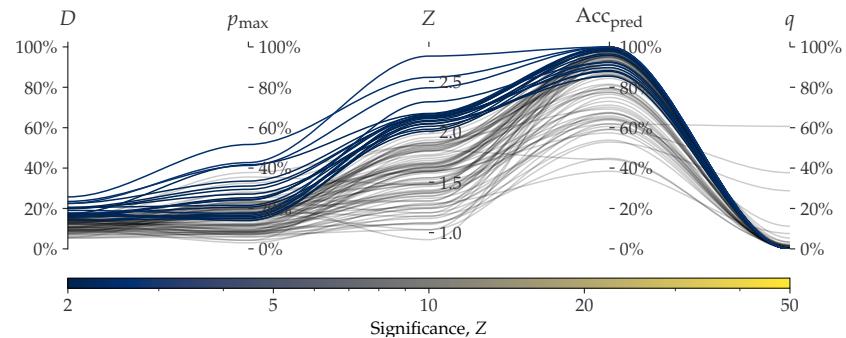
We have also compared `metaDMG` to existing methods such as PyDamage (Borry et al., 2021). Since  
 388 PyDamage does not include the LCA step, this comparison is based on the non-LCA mode (local-  
 mode) of `metaDMG`. This mode iterates through the different assigned species for all mapped reads  
 390 and estimates the damage for each. In general, we find that `metaDMG` is more conservative, accurate  
 and precise in its damage estimates.

392 On example of this is can be found in Figure 10, which shows both the `metaDMG` and PyDamage  
 results of the 100 *Homo Sapiens* single-genome simulations with 100 reads and 15% added artificial  
 394 damage (and a fragment length distribution with mean 60).

To compare the computational performance, we use the Pitch-6 sample which has 11.433  
 396 unique taxa. When using only a single core, PyDamage took 1105 s to compute all fits, while `metaDMG`  
 took 88 s, a factor of 12.6x faster. Out of the 88 s, `metaDMG` spent 53 s on the actual fits, the rest was for  
 398 loading and reading the alignment file and computing the mismatch matrices. This makes `metaDMG`  
 more than 20x faster than PyDamage for the fit computation. For the rest of the timings, see Ta-  
 400 ble 2. PyDamage requires the alignment file to be sorted by chromosome position and be supplied  
 with an index file, allowing it to iterate fast through the alignment file, at the expense of computa-



**Figure 9.** Comparison between the full Bayesian model and the fast, approximate, MAP model for the estimated damage and significance. The figure shows data after a loose cut of  $D > 1\%$ ,  $Z > 2$  and more than 100 reads. The dashed, grey line shows the 1:1 ratio and the correlation,  $\rho$ , is shown in the upper right corner.



**Figure 10.** Parallel Coordinates plot comparing metaDMG and PyDamage for the Homo Sapiens single-genome simulation with 100 reads and 15% added artificial damage. The different axis shows the five different variables: metaDMG-damage ( $D$ , by metaDMG), PyDamage-damage ( $p_{\max}$ , by PyDamage), significance ( $Z$ , by metaDMG), predicted accuracy ( $Acc_{pred}$ , by PyDamage), and the p-value ( $q$ , by PyDamage). Each of the 100 simulations are plotted as single lines showing the values of the different dimensions. Simulations with  $D > 1\%$  and  $Z > 2$ , i.e. damaged according to the loose metaDMG cut, are shown in color proportional to their significance. Non-damaged simulations are shown in semi-transparent black lines.

**Table 2.** Computational performance of PyDamage and metaDMG. The table contains the times it takes to run either PyDamage or metaDMG on the full Pitch-6 sample containing 11.433 species. The timings are shown for both single-processing case (1 core) and multi-processing (2 and 4 cores). The timings were performed on a Macbook M1 Pro model from 2021. “12.6x” means that metaDMG was 12.6 times faster than PyDamage for that particular test.

Cores	Pitch-6		Pydamage		metaDMG	
	Total	Fits	Total	Fits		
1	1105 s	1102 s	88 s	12.6x	53 s	20.8x
2	592 s	590 s	66 s	9.0x	25 s	23.6x
4	398 s	397 s	54 s	7.4x	14s	28.4x

402 tional load before running the actual damage estimation. metaDMG on the other hand requires the  
 403 reads to be sorted by name to minimize the time it takes to run the LCA, which however, is not  
 404 tested in this comparison.

## 5 | DISCUSSION

405 Preliminary work indicates that the computational performance of the models can be even fur-  
 406 ther optimized by using Julia (Bezanson et al., 2017), which shows around 7x optimization for the  
 407 Bayesian model (~ 0.2 s/fit) and 4x for the MAP model (~ 1.1 ms/fit).

### 5.1 | Acknowledgment

410 Acknowledgements here

### 5.2 | Data availability

412 Source code is hosted at GitHub: <https://github.com/metaDMG-dev>. Sequencing data can be found  
 413 at: <https://somewhere.com> XXX.

## REFERENCES

- <sup>414</sup> Betancourt, Michael (2018). "A Conceptual Introduction to Hamiltonian Monte Carlo". In: *arXiv:1701.02434 [stat]*. arXiv: 1701.02434.
- <sup>416</sup> Bezanson, Jeff et al. (2017). "Julia: A fresh approach to numerical computing". In: *SIAM review* 59.1. Publisher: SIAM, pp. 65–98. URL: <https://julialang.org/>.
- <sup>418</sup> Borry, Maxime et al. (2021). "PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA de novo assembly". en. In: *PeerJ* 9. Publisher: PeerJ Inc., e11845. ISSN: 2167-8359. DOI: [10.7717/peerj.11845](https://doi.org/10.7717/peerj.11845). URL: <https://peerj.com/articles/11845> (visited on 2022).
- <sup>420</sup> Bradbury, James et al. (2018). *JAX: composable transformations of Python NumPy programs*. URL: <http://github.com/google/jax>.
- <sup>422</sup> Briggs, Adrian W. et al. (2007). "Patterns of damage in genomic DNA sequences from a Neandertal". en. In: *Proceedings of the National Academy of Sciences* 104.37. Publisher: National Academy of Sciences Section: Biological Sciences, pp. 14616–14621. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.0704665104](https://doi.org/10.1073/pnas.0704665104). URL: <https://www.pnas.org/content/104/37/14616>.
- <sup>424</sup> Cabanski, Christopher R. et al. (2012). "ReQON: a Bioconductor package for recalibrating quality scores from next-generation sequencing data". In: *BMC Bioinformatics* 13.1, p. 221. ISSN: 1471-2105. DOI: [10.1186/1471-2105-13-221](https://doi.org/10.1186/1471-2105-13-221). URL: <https://doi.org/10.1186/1471-2105-13-221> (visited on 2022).
- <sup>426</sup> Cappellini, Enrico et al. (2018). "Ancient Biomolecules and Evolutionary Inference". In: *Annual Review of Biochemistry* 87.1. \_eprint: <https://doi.org/10.1146/annurev-biochem-062917-012002>, pp. 1029–1060. DOI: [10.1146/annurev-biochem-062917-012002](https://doi.org/10.1146/annurev-biochem-062917-012002). URL: <https://doi.org/10.1146/annurev-biochem-062917-012002> (visited on 2022).
- <sup>428</sup> Cepeda-Cuervo, Edilberto and MARÍA VICTORIA Cifuentes-Amado (2017). "Double Generalized Beta-Binomial and Negative Binomial Regression Models". en. In: *Revista Colombiana de Estadística* 40.1. Publisher: Universidad Nacional de Colombia, pp. 141–163. ISSN: 0120-1751. DOI: [10.15446/rce.v40n1.61779](https://doi.org/10.15446/rce.v40n1.61779).
- <sup>430</sup> 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](https://doi.org/10.1101/cshperspect.a012567). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3685887/> (visited on 2022).
- <sup>432</sup> Michelsen & Pedersen *et al.* 2022 | metaDMG: An Ancient DNA Damage Toolkit
- <sup>434</sup> bioRxiv | 21 of 36

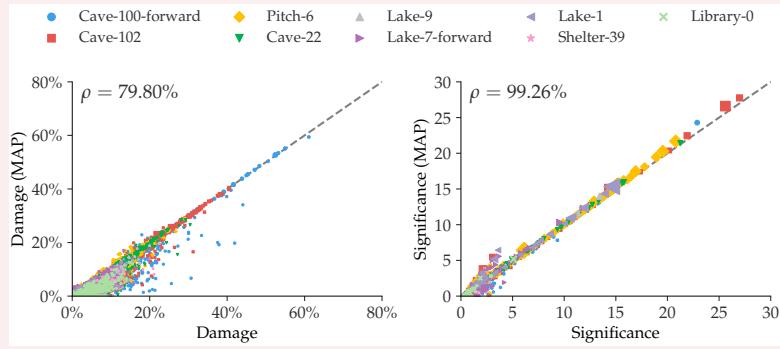
- Dembinski, Hans et al. (2021). *scikit-hep/iminuit*: v2.8.2. DOI: [10.5281/ZENODO.3949207](https://doi.org/10.5281/ZENODO.3949207). (Visited on  
 444 2021).
- Fernandez-Guerra, Antonio (2022). *genomewalker/aMGSim-smk*: v0.0.1. DOI: [10.5281/zenodo.7298422](https://doi.org/10.5281/zenodo.7298422).  
 446 URL: <https://doi.org/10.5281/zenodo.7298422>.
- Ginolhac, Aurelien et al. (2011). "mapDamage: testing for damage patterns in ancient DNA se-  
 448 quences". In: *Bioinformatics* 27.15, pp. 2153–2155. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btr347](https://doi.org/10.1093/bioinformatics/btr347). URL: <https://doi.org/10.1093/bioinformatics/btr347> (visited on 2022).
- 450 Huang, Weichun et al. (2012). "ART: a next-generation sequencing read simulator". eng. In: *Bioinfor-  
 matics (Oxford, England)* 28.4, pp. 593–594. ISSN: 1367-4811. DOI: [10.1093/bioinformatics/btr708](https://doi.org/10.1093/bioinformatics/btr708).
- 452 Jónsson, Hákon et al. (2013). "mapDamage2.0: fast approximate Bayesian estimates of ancient DNA  
 damage parameters". en. In: *Bioinformatics* 29.13, pp. 1682–1684. ISSN: 1367-4803, 1460-2059.  
 454 DOI: [10.1093/bioinformatics/btt193](https://doi.org/10.1093/bioinformatics/btt193).
- Lam, Siu Kwan, Antoine Pitrou, and Stanley Seibert (2015). "Numba: a LLVM-based Python JIT com-  
 456 piler". 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.  
 458 DOI: [10.1145/2833157.2833162](https://doi.org/10.1145/2833157.2833162). URL: <https://github.com/numba/numba>.
- Langmead, Ben and Steven L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2". en. In:  
 460 *Nature Methods* 9.4. Number: 4 Publisher: Nature Publishing Group, pp. 357–359. ISSN: 1548-  
 7105. DOI: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923). URL: <https://www.nature.com/articles/nmeth.1923> (visited on  
 462 2022).
- Llamas, Bastien et al. (2017). "From the field to the laboratory: Controlling DNA contamination in  
 464 human ancient DNA research in the high-throughput sequencing era". en. In: *STAR: Science &  
 Technology of Archaeological Research* 3.1, pp. 1–14. ISSN: 2054-8923. DOI: [10.1080/20548923.2016.1258824](https://doi.org/10.1080/20548923.2016.1258824). URL: <https://www.tandfonline.com/doi/full/10.1080/20548923.2016.1258824> (visited  
 466 on 2022).
- 468 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-  
 470 13991-9.
- Mölder, Felix et al. (2021). *Sustainable data analysis with Snakemake*. en. Tech. rep. 10:33. Type: arti-  
 472 cle. F1000Research. DOI: [10.12688/f1000research.29032.2](https://doi.org/10.12688/f1000research.29032.2). URL: <https://f1000research.com/articles/10-33> (visited on 2022).

- 474 Nayfach, Stephen et al. (2021). "CheckV assesses the quality and completeness of metagenome-  
assembled viral genomes". en. In: *Nature Biotechnology* 39.5. Number: 5 Publisher: Nature Pub-  
lishing Group, pp. 578–585. ISSN: 1546-1696. DOI: [10.1038/s41587-020-00774-7](https://doi.org/10.1038/s41587-020-00774-7). URL: <https://www.nature.com/articles/s41587-020-00774-7> (visited on 2022).
- 476 NCBI Resource Coordinators (2018). "Database resources of the National Center for Biotechnology  
Information". In: *Nucleic Acids Research* 46.D1, pp. D8–D13. ISSN: 0305-1048. DOI: [10.1093/nar/gkx1095](https://doi.org/10.1093/nar/gkx1095). URL: <https://doi.org/10.1093/nar/gkx1095> (visited on 2022).
- 480 Neukamm, Judith, Alexander Peltzer, and Kay Nieselt (n.d.). "DamageProfiler: Fast damage pattern  
calculation for ancient DNA". en. In: (), p. 10.
- 482 Parks, Donovan H. et al. (2018). "A standardized bacterial taxonomy based on genome phylogeny  
substantially revises the tree of life". en. In: *Nature Biotechnology* 36.10. Number: 10 Publisher:  
Nature Publishing Group, pp. 996–1004. ISSN: 1546-1696. DOI: [10.1038/nbt.4229](https://doi.org/10.1038/nbt.4229). URL: <https://www.nature.com/articles/nbt.4229> (visited on 2022).
- 484 Phan, Du, Neeraj Pradhan, and Martin Jankowiak (2019). "Composable Effects for Flexible and Accel-  
erated Probabilistic Programming in NumPyro". In: *arXiv:1912.11554 [cs, stat]*. arXiv: 1912.11554.
- 486 Plotly, Technologies (2015). *Collaborative data science*. Place: Montreal, QC Publisher: Plotly Tech-  
nologies Inc. URL: <https://plot.ly>.
- 488 Skoglund, Pontus et al. (2014). "Separating endogenous ancient DNA from modern day contami-  
nation in a Siberian Neandertal". In: *Proceedings of the National Academy of Sciences* 111.6. Pub-  
lisher: Proceedings of the National Academy of Sciences, pp. 2229–2234. DOI: [10.1073/pnas.1318934111](https://doi.org/10.1073/pnas.1318934111). URL: <https://www.pnas.org/doi/10.1073/pnas.1318934111> (visited on 2022).
- 490 Wang, Yucheng et al. (2022). "ngsLCA—A toolkit for fast and flexible lowest common ancestor in-  
ference and taxonomic profiling of metagenomic data". en. In: *Methods in Ecology and Evolution* n/a.n/a (). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14006>. ISSN:  
492 2041-210X. DOI: [10.1111/2041-210X.14006](https://doi.org/10.1111/2041-210X.14006). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14006> (visited on 2022).
- 494

## 500 Appendix 1

## 502 A | EXAMPLE FIGURE

This is an example of including a figure in the appendix.



504 Appendix 1—figure 1. Comparison between the full Bayesian model and the fast, approximate, MAP  
506 model for the estimated damage and significance. The dashed, grey line shows the 1:1 ratio.

## Appendix 2

### B | EXAMPLE TABLE

This is an example of including a table in the appendix.

**Appendix 2—table 1.** An example table.

Speed (mph)	Driver	Car	Engine	Date
407.447	Craig Breedlove	Spirit of America	GE J47	8/5/63
413.199	Tom Green	Wingfoot Express	WE J46	10/2/64
434.22	Art Arfons	Green Monster	GE J79	10/5/64
468.719	Craig Breedlove	Spirit of America	GE J79	10/13/64
526.277	Craig Breedlove	Spirit of America	GE J79	10/15/65
536.712	Art Arfons	Green Monster	GE J79	10/27/65
555.127	Craig Breedlove	Spirit of America, Sonic 1	GE J79	11/2/65
576.553	Art Arfons	Green Monster	GE J79	11/7/65
600.601	Craig Breedlove	Spirit of America, Sonic 1	GE J79	11/15/65
622.407	Gary Gabelich	Blue Flame	Rocket	10/23/70
633.468	Richard Noble	Thrust 2	RR RG 146	10/4/83
763.035	Andy Green	Thrust SSC	RR Spey	10/15/97

## Appendix 3

### C | LIKELIHOOD CALCULATION

#### C.1 | Full multinomial logistic Regression models

514 Postmortem damages will have impacts on the NGS (next generation sequencing) reads.

516 A common phenomenon is the calling error rates increases from nucleotide C to T due to  
518 the cytosine deamination process. Unawareness of this will lead to inaccurate inferences.

520 Evidences show that the magnitude of such changes are related to the positions the site is  
522 within a read (the fraction of the ancient DNA). Here we present 3 slightly different ways to  
unveil the relationship between the calling error rates and the mismatching reference/read  
pairs as well as the site positions within a read. The methods are based on the multinomial  
logistic regressions.

#### 524 Data Description

526 We perform the regression based on the summary statistic of the mismatch matrix which  
528 is a table which contains the counts of reads of different reference/read categories (in total  
16) and positions on the forward/reversed strand (15 positions on each direction). Table 1  
and Table 2 give an example of the data format we use for the inference.

Ref.	Read Counts								
	A					C			
	Read	A	C	G	T	A	C	G	T
1	12794053	8325	28769	16073	10404	8045811	8020	2092619	
2	13480290	6812	21107	12102	9151	8260185	6531	1145605	
3	12760253	6131	18859	10327	7772	8385423	5899	914709	
4	12995572	5240	17671	8940	7880	8345892	5252	767237	
5	12930102	4601	17021	8188	8374	8474964	5161	703283	
6	12879355	4684	16435	7536	8726	8571141	4811	643607	
7	12684349	4557	15298	7394	8835	8727254	4762	586674	
8	12585563	4454	15497	7236	8898	8888173	5058	527691	
9	12468622	4309	14704	6942	8948	9076851	4673	481170	
10	12491183	4437	14567	6912	9103	9237982	4702	443329	
11	12430899	4296	14083	6515	9313	9364121	4609	404431	
12	12419506	4226	13985	6503	9342	9357468	4367	371475	
13	12469412	4147	13851	6375	9586	9386737	4588	345390	
14	12549936	4045	13650	6246	9673	9324488	4628	322294	
15	12566555	4174	13499	6213	9735	9305820	4518	301360	
-1	11599167	8800	16164	14851	90888	9613102	10843	19810	
-2	11985637	8769	14044	12040	28799	9561124	7184	18424	
-3	12941743	7805	13861	12001	24988	9400151	6368	15466	
-4	12808985	7141	12885	9889	23067	9509723	5421	14901	
-5	12869585	6954	12100	9428	22349	9464831	5789	13987	
-6	12784911	6440	12080	8735	20556	9566794	6544	14021	
-7	12878349	5946	12311	8225	19480	9566359	6478	16419	
-8	12719722	9521	12156	8131	19226	9725468	6709	23434	
-9	12652860	5634	11940	7671	18035	9762224	6321	31667	
-10	12566817	5448	11850	7178	17353	9701382	6306	37831	
-11	12702498	5309	12092	7568	16121	9526031	6035	43215	
-12	12731940	5207	11933	6856	15637	9533858	5557	47650	
-13	12697647	4989	12199	7153	15072	9508117	5434	51614	
-14	12689924	4944	11891	6816	15050	9525285	5237	bioRxiv preprint doi: https://doi.org/10.1101/272727; this version posted April 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.	
-15	12660634	4746	11753	6732	14815	9561359	5184	59633	

530                   **Appendix 3—table 1.** The read counts per position given the reference nucleotides are A or C of an  
532                   ancient human data. The negative position indices are the position on the reversed strand. In the  
534                   manuscript, the elements (the values of a specific nucleotide read counts per position given the  
                      reference nucleotide is A or C) in this table are denoted as  $o_{A \rightarrow i,p}$  or  $o_{C \rightarrow i,p}$ .

Ref.	Read Counts								
	G					T			
	Read	A	C	G	T	A	C	G	T
1	16389	8976	9639767	86584	11733	15878	8351	11718463	
2	17614	6483	9510149	26655	10761	13958	7011	11974947	
3	15164	5949	9488917	23374	9509	13767	6046	12839015	
4	14844	5186	9566468	21960	8170	12509	5585	12721790	
5	14005	5612	9497118	20468	7186	11991	5233	12795244	
6	13671	6195	9622572	19096	6948	11683	4790	12686645	
7	16648	6394	9609855	18594	6203	12122	4780	12794172	
8	23659	6405	9768666	17341	6131	11847	4758	12626614	
9	31680	6139	9785449	17034	5998	12040	4469	12579260	
10	38484	5982	9700857	16235	5487	11546	4175	12513653	
11	44665	5722	9536341	15284	5651	12044	4176	12646627	
12	48949	5371	9547134	14569	5449	11663	4060	12684645	
13	53076	5234	9543953	14090	5262	11785	4046	12631297	
14	57343	5186	9551477	13855	5257	11768	4006	12624840	
15	61236	5137	9583481	13667	5122	11733	3947	12612416	
-1	2078554	7947	8096447	11847	15732	28461	8551	12890628	
-2	1138478	6656	8232666	10760	12299	20759	6999	13446882	
-3	921712	5970	8399013	8643	10514	18226	6564	12718084	
-4	775038	5720	8319235	8416	9415	17800	5388	12977322	
-5	710955	5499	8462058	8926	8526	17088	4911	12886576	
-6	647761	5052	8545455	9193	7640	16351	4879	12852322	
-7	593854	4872	8693834	9318	7600	15523	5048	12664576	
-8	535542	7828	8889921	9399	7163	18704	4718	12510123	
-9	486549	4696	9075263	9522	7109	14547	4611	12409220	
-10	448895	4622	9226758	9432	6816	14567	4668	12438344	
-11	409027	4654	9352528	9544	6575	14019	4611	12388650	
-12	376069	4637	9344701	9419	6511	13874	4486	12390148	
-13	350609	4655	9384853	9885	6197	13877	4327	12432024	
-14	326760					9337266	9889	5986	13928
	4595								4403
-15	305014					9310617	10065	5919	13442
	4541								4232

**Appendix 3—table 2.** The read counts per position given the reference nucleotides are G or T of the same human data as in Table 1. The negative position indices are the position on the reversed strand. In the manuscript, the elements (the values of a specific nucleotide read counts per position given the reference nucleotide is G or T) in this table are denoted as  $o_{G \rightarrow i,p}$  or  $o_{T \rightarrow i,p}$ .

The terminology used here might not be standard. The term full regression here is to distinguish itself from the folded regression discussed later, which simply means inferring the coefficients of forward strand and reversed strand separately. Full regression includes both unconditional regression and conditional regression. The unconditional regression's objective is to infer the probability of observing a read of nucleotide  $j$  and its reference is  $i$  at position  $p$ , i.e.,  $P(\text{Obs} : i \rightarrow j | \text{Pos} : p)$ ; while the unconditional regression's target is to estimate the probability of observing a read of nucleotide  $j$  given its reference is  $i$  at position  $p$ , i.e.,  $P(\text{Obs} : j | \text{Ref} : i, \text{Pos} : p)$ . Their relationship is as follows:

$$P(\text{Obs} : j | \text{Ref} : i, \text{Pos} : p) = \frac{P(\text{Obs} : i \rightarrow j | \text{Pos} : p)}{\sum_j P(\text{Obs} : i \rightarrow j | \text{Pos} : p)}.$$

So in fact, unconditional regression can give us more detailed inferred results (extra information the nucleotide composition per position of the reference, which may be related to the prepared libraries).

#### Unconditional Regression likelihood

$$\begin{aligned} l_1 &= \sum_p \sum_{i,j \in \{A,C,G,T\}} o_{i \rightarrow j,p} \log P_{i,j|p} \\ &= \sum_p \left[ o_p \log P_{TT|p} + \sum_{(i,j) \neq TT} o_{i \rightarrow j,p} \log \frac{P_{ij|p}}{P_{TT|p}} \right], \end{aligned} \quad (10)$$

where  $P_{ij|p} = P(\text{Obs} : i \rightarrow j | \text{Pos} : p)$ , and  $o_p = \sum_{i,j \in \{A,C,G,T\}} o_{i \rightarrow j,p}$ .

$$\log \frac{P_{ij|p}}{P_{TT|p}} = \sum_{n=0}^{\text{order}} \alpha_{i,j,p,n} p^n \quad (11)$$

$$\begin{aligned} l_1 &= \sum_p \left\{ -o_p \log \left[ 1 + \sum_{(i,j) \neq TT} \exp \left( \sum_{n=0}^{\text{order}} \alpha_{i,j,p,n} p^n \right) \right] + \sum_{(i,j) \neq TT} o_{i \rightarrow j,p} \sum_{n=0}^{\text{order}} \alpha_{i,j,p,n} p^n \right\} \\ &= l_{1,5'} + l_{1,3'}. \end{aligned} \quad (12)$$

568

The number of inferred parameters for the full conditional regression is 30 (order + 1).

$$\frac{\partial l_1}{\partial \alpha_{i,j,p,n}} = -o_p \frac{p^n \exp \left( \sum_{n=0}^{\text{order}} \alpha_{i,j,p,n} p^n \right)}{1 + \sum_{(i,j) \neq \text{TT}} \exp \left( \sum_{n=0}^{\text{order}} \alpha_{i,j,p,n} p^n \right)} + o_{i \rightarrow j,p} p^n. \quad (13)$$

572

### Conditional Regression likelihood

$$\begin{aligned} l_2 &= \sum_{i \in \{\text{A,C,G,T}\}} \sum_p \sum_{j \in \{\text{A,C,G,T}\}} o_{i \rightarrow j,p} \log P_{j|p,i} \\ &= \sum_{i \in \{\text{A,C,G,T}\}} \sum_p \left[ o_{i,p} \log P_{\text{T}|p,i} + \sum_{j \neq \text{T}} o_{i \rightarrow j,p} \log \frac{P_{j|p,i}}{P_{\text{T}|p,i}} \right], \end{aligned} \quad (14)$$

where  $P_{j|p,i} = P(\text{Obs : } j | \text{Ref : } i, \text{Pos : } p)$ , and  $o_{i,p} = \sum_{j \in \{\text{A,C,G,T}\}} o_{i \rightarrow j,p}$ .

578

$$\log \frac{P_{j|p,i}}{P_{\text{T}|p,i}} = \sum_{n=0}^{\text{order}} \beta_{i,j,p,n} p^n \quad (15)$$

582

$$\begin{aligned} l_2 &= \sum_{i \in \{\text{A,C,G,T}\}} \sum_p \left\{ -o_{i,p} \log \left[ 1 + \sum_{j \neq \text{T}} \exp \left( \sum_{n=0}^{\text{order}} \beta_{i,j,p,n} p^n \right) \right] + \sum_{j \neq \text{T}} o_{i \rightarrow j,p} \sum_{n=0}^{\text{order}} \beta_{i,j,p,n} p^n \right\} \\ &= l_{2,\text{A},5'} + l_{2,\text{C},5'} + l_{2,\text{G},5'} + l_{2,\text{T},5'} + l_{2,\text{A},3'} + l_{2,\text{C},3'} + l_{2,\text{G},3'} + l_{2,\text{T},3'}. \end{aligned} \quad (16)$$

586

The number of inferred parameters for the full unconditional regression is 24 (order + 1).

$$\frac{\partial l_2}{\partial \beta_{i,j,p,n}} = -o_{i,p} \frac{p^n \exp \left( \sum_{n=0}^{\text{order}} \beta_{i,j,p,n} p^n \right)}{1 + \sum_{j \neq \text{T}} \exp \left( \sum_{n=0}^{\text{order}} \beta_{i,j,p,n} p^n \right)} + o_{i \rightarrow j,p} p^n. \quad (17)$$

590

### Folded Regression

The folded regressions use the same log-likelihood function as the full regression (i.e., Equation ) but are conducted based on the assumptions that,

592

$$\alpha_{i,j,p,n} = \alpha_{c(i),c(j),-p,n}, \quad (18)$$

594

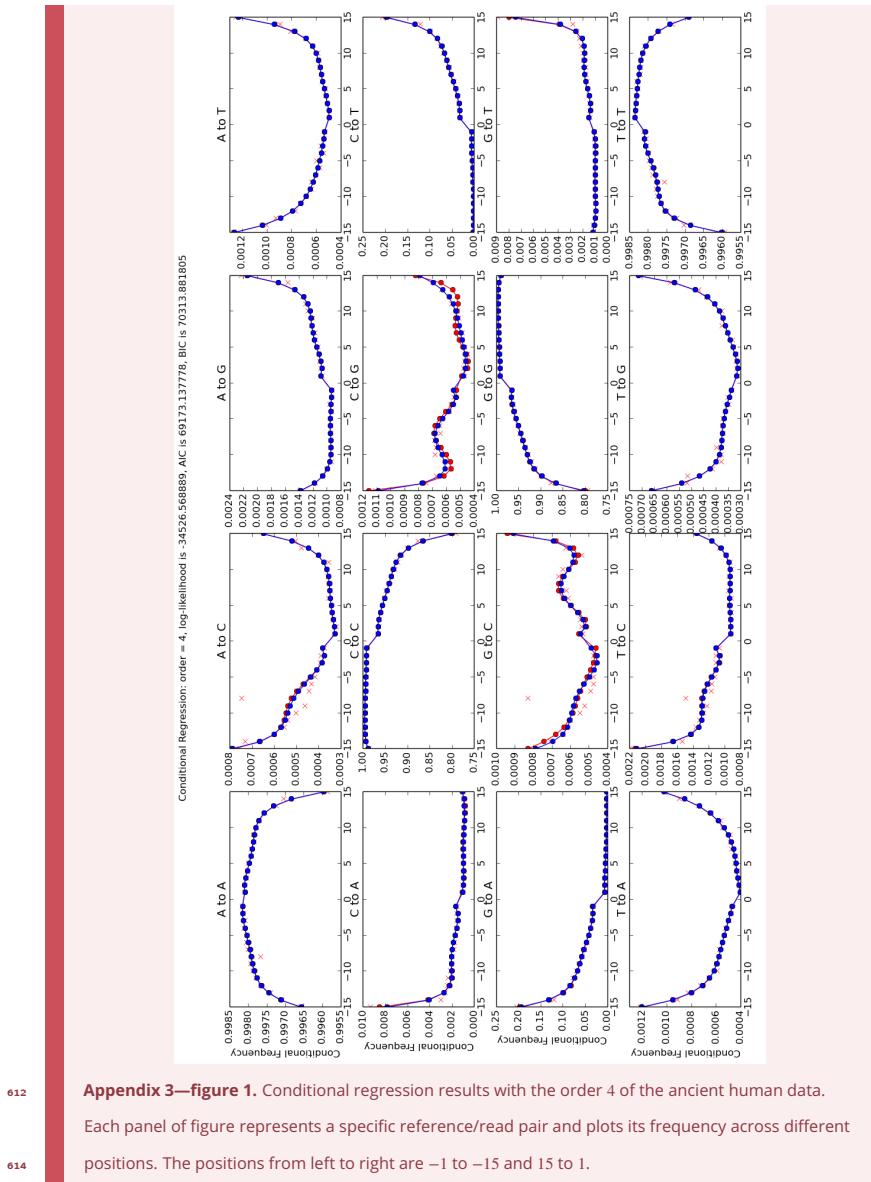
$$\beta_{i,j,p,n} = \beta_{c(i),c(j),-p,n}, \quad (19)$$

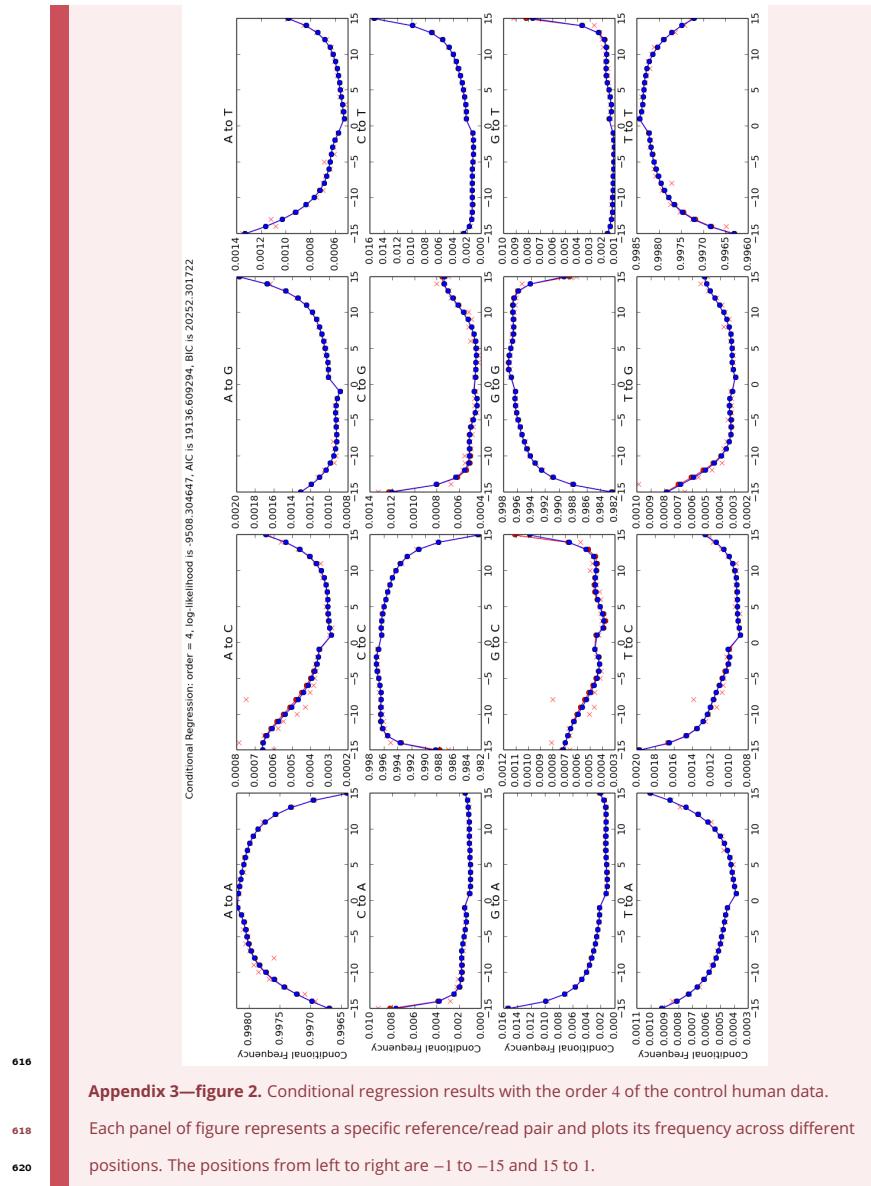
where  $c(i)$  means the complimentary nucleotide of the nucleotide  $i$ , e.g.,  $c(\text{A}) = \text{T}$  and  $c(\text{G}) = \text{C}$ . Our data (both taxon and human data) and some models studies seem to support this assumption.

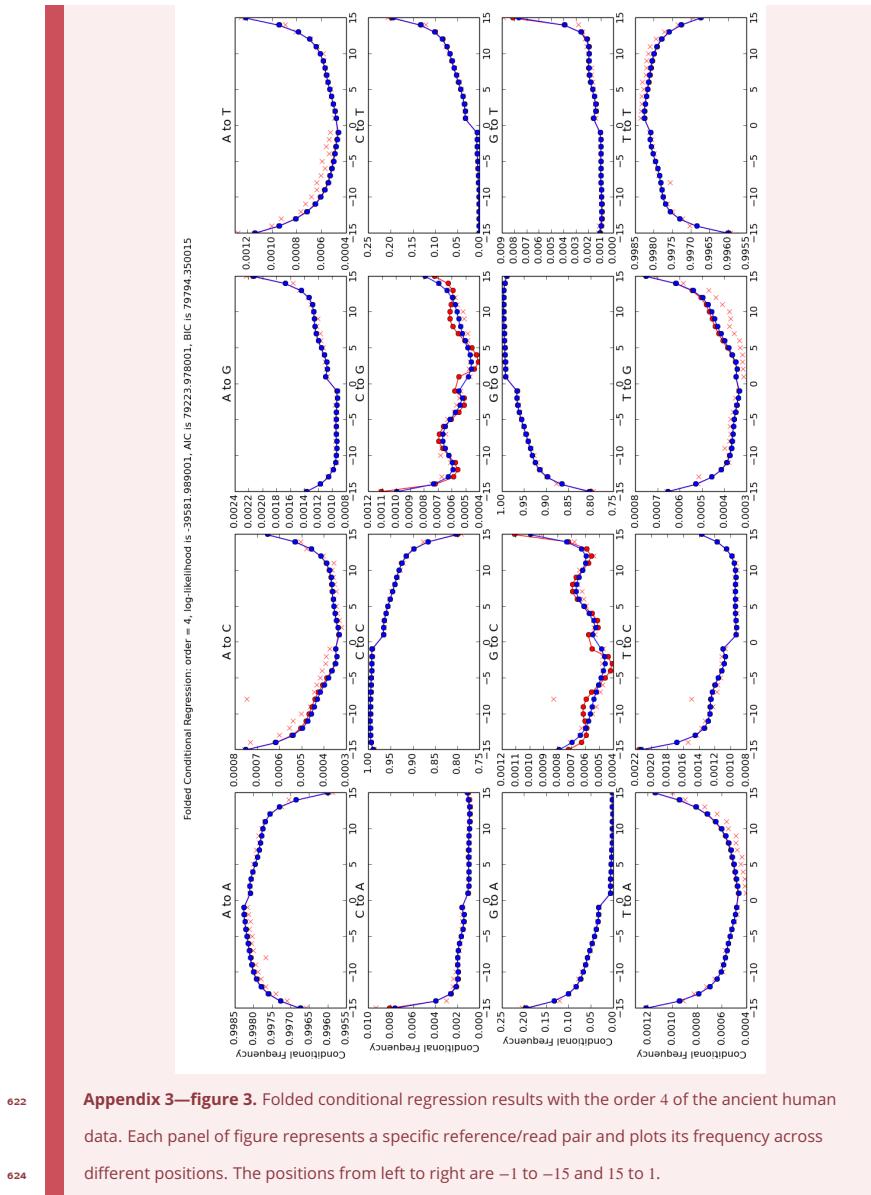
598 By doing the folded regression, we halve the number of inferred parameters. Hence The  
599 number of inferred parameters for the folded unconditional regression is 15 (order + 1), and  
600 that of folded conditional regression is 12 (order + 1).

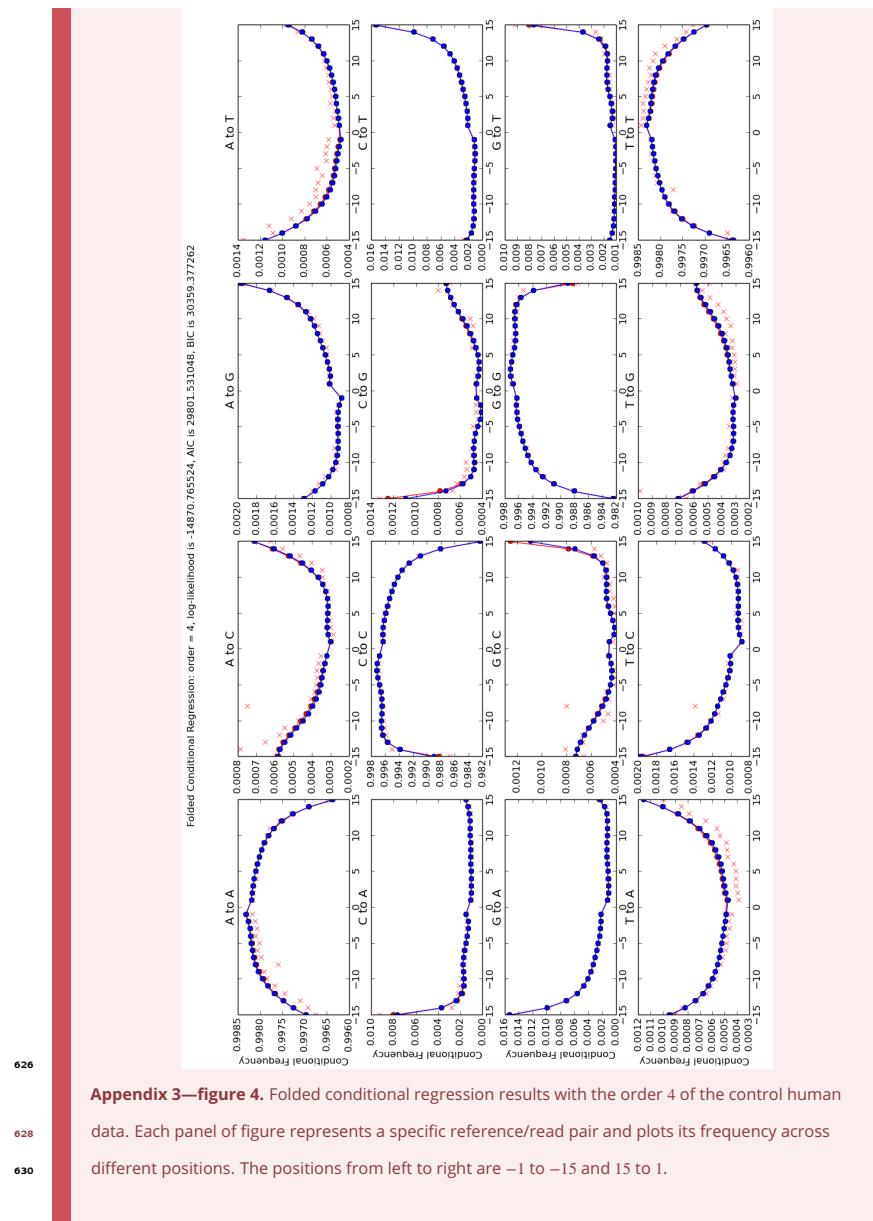
#### Results for multinomial logistic regression

602 Currently, the optimization of the likelihood functions are based on the C++ library of gsl and  
603 use the function *gsl\_multimin\_fminimizer\_nmsimplex2*. with the initial searching point is set to  
604 be the results of logistic regression. We here present here 4 figures pertaining to showcase  
605 the performance of our model. The regression methods are based on the summary statistic  
606 of the counts of mismatches and the optimization is therefore in the scale of miliseconds.  
607 Fig. 1 and Fig. 2 are the conditional regression results of the ancient and control human  
608 data correspondingly. And Fig. 3 and Fig. 4 are the folded conditional regression results of  
609 the same data as above. Our codes can also do the unconditional regression, but I have not  
610 generated the results for now.









### **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.”.

Manuscript (Title Page, Abstract, Body, References, Figure  
Legends)

1  
2  
3  
4  
5

6                   **Preoperative prediction of medical morbidity after fast-**  
7                   **track hip and knee arthroplasty - a machine learning**  
8                   **based approach.**

9  
10  
11  
12  
13

14                   **2. Author information:**

15 Christian Michelsen, M.Sci., Research Fellow, The Niels Bohr Institute, University of Copenhagen,  
16 Blegdamsvej 17 2100 Copenhagen, Denmark

17 Christoffer C Jørgensen, M.D., Senior Researcher, Section of Surgical Pathophysiology and Centre for  
18 Fast-track Hip and Knee Replacement, 7621, Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen,  
19 Denmark

20 Mathias Heltberg, M.Sci, Research Fellow, The Niels Bohr Institute, University of Copenhagen,  
21 Blegdamsvej 17 2100 Copenhagen, Denmark

22 Mogens H. Jensen, D.Sci., Prof., The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17  
23 2100 Copenhagen, Denmark

24 Alessandra Lucchetti, M.Sci., Research Fellow., The Niels Bohr Institute, University of Copenhagen,  
25 Blegdamsvej 17 2100 Copenhagen, Denmark

26 Pelle B Petersen, M.D., Ph.D, Section of Surgical Pathophysiology and Centre for Fast-track Hip and  
27 Knee Replacement, 7621, Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen, Denmark

28 Troels Petersen, M.Sci, Ass.Prof., The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17  
29 2100 Copenhagen, Denmark

30 Henrik Kehlet, M.D., Ph.D., Prof. Section of Surgical Pathophysiology and Centre for Fast-track Hip and  
31 Knee Replacement, 7621, Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen, Denmark

32  
33  
34

35                   \*This is a joint first-authorship between CM and CJ

36  
37  
38

39                   **3. Corresponding author:**

40 Dr. Christoffer Calov Jørgensen  
41 Section for Surgical Pathophysiology 7621  
42 Rigshospitalet, Blegdamsvej 9,  
43 DK-2100 Copenhagen, Denmark  
44 Phone +45 3545 4616 Fax: +45 3545 6543  
45 E-mail: christoffer.calov.joergensen@regionh.dk

46  
47  
48

49                   **4. Clinical Trial Number:** The Centre for Fast-track Hip and Knee Replacement Database was registered  
50 as a study registry on ClinicalTrials.gov:NCT01515670

51  
52  
53

54                   **5. Prior presentations:** Not applicable

55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4     **6. Acknowledgements:** The members of the Centre for Fast-track Hip and Knee Replacement Database  
5     collaborative group all contributed by implementing the fast-track protocol at their respective departments  
6     and reviewing the final manuscript.  
7  
8     Frank Madsen M.D. Consultant, Department of Orthopedics, Aarhus University Hospital, Aarhus,  
9     Denmark  
10     Torben B. Hansen M.D., Ph.D., Prof. Department of Orthopedics, Holstebro Hospital, Holstebro, Denmark  
11     Kirill Gromov, M.D., Ph.D., Ass.Prof. Department of Orthopedics Hvidovre Hospital, Hvidovre Denmark  
12     Thomas Jakobsen, M.D., Ph.D., DM.Sci., Ass. Prof. Department of Orthopedics, Aalborg University  
13     Hospital, Farsø, Denmark  
14     Claus Varnum, M.D., Ph.D., Ass. Prof. Department of Orthopedic Surgery, Lillebaelt Hospital - Vejle,  
15     University Hospital of Southern Denmark, Denmark  
16     Soren Overgaard, M.D., DM.Sci., Prof, Department of Orthopedics, Bispebjerg Hospital, Copenhagen,  
17     Denmark  
18     Mikkel Rathsach, M.D., Ph.D., Ass. Prof. Department of Orthopedics, Gentofte Hospital, Gentofte,  
19     Denmark  
20     Lars Hansen, M.D., Consultant, Department of Orthopedics, Sydvestjysk Hospital, Grindsted, Denmark  
21  
22     **7. Word and Element Counts:**  
23     Abstract: 300/300 Introduction: 466/500 Discussion:1278/1500 Figures:3 Tables:2 Appendices:2  
24     Supplementary Digital Files:4  
25     **8. Abbreviated title:** Machine learning models in joint replacement  
26     **9. Summary Statement:** Not applicable.  
27     **10. Funding:** The study received funding from the Lundbeck Foundation, Denmark, as well as from  
28     institutional and departmental sources.  
29     **11. Conflict of interest:** Prof. Kehlet is a board member of "Rapid Recovery", by Zimmer Biomet. Mr.  
30     Heltberg is sponsored by a grant from the Lundbeck Foundation, independently of the present study.  
31     Dr. Petersen is an advisory member of Sanofi outside of the present study.  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7

## Abstract

8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36

**Background:** Introduction of machine-learning models has potentially improved prediction of postoperative hospitalization and morbidity after hip and knee replacement. However, few studies include enhanced recovery programs, and most rely on administrative coding with limited follow-up and information on perioperative care. Thus, benefits of machine-learning models for prediction of postoperative morbidity in enhanced recovery hip and knee replacement remain uncertain.

**Methods:** Multicenter cohort study from 2014-2017 in enhanced recovery total hip and knee replacement. Prospective recording of comorbidity and prescriptions. Information on length of stay and readmissions through the Danish National Patient Registry and medical records. Data was split into training (n:18013) and test sets (n:3913). A machine-learning model with 33 variables was used for predicting “medical” morbidity with a length of stay of >4 days or 90-days readmission and compared to a full logistic regression model. In addition, a machine-learning model excluding age, an age-only model and parsimonious machine-learning and logistic regression models using the ten most important variables were evaluated. Model performances were evaluated using several metrics, including precision, operating receiver (AUC) and precision recall curves (AUPRC). Variable importance was analyzed using Shapley Additive Explanations values.

**Results:** With 782 (20%) “risk-patients”, precision, AUC and AUPRC were 13.6%, 76.3% and 15.5% for the full and 12.8%, 75.9% and 17.1% for the parsimonious machine-learning models vs. 12.5%, 74.5% and 15.7% for the full logistic regression model. The machine-learning model excluding age and the Age-only model performed worse. Of the top ten variables, eight were shared between the full machine-learning and logistic regression models, and the importance of specific prescribed drugs varied considerably with age.

**Conclusion:** A machine-learning algorithm using preoperative characteristics and prescriptions likely improves identification of patients in high-risk of medical complications after fast-track hip and knee replacement. Such algorithms could help identify patients who benefit from intensified perioperative care.

53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1

2

3

4

## INTRODUCTION

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

Prediction of postoperative morbidity and requirement for hospitalization is important for planning of health care resources. With regard to the common surgical procedures of primary total hip and knee arthroplasty, the introduction of enhanced recovery or fast-track programs has led to a significant reduction of postoperative length of stay (length of stay) as well as morbidity and mortality.<sup>1-3</sup> However, despite such progress, a fraction of patients still have postoperative complications leading to prolonged length of stay or readmissions.<sup>1,3,4</sup>

Consequently, in order to prioritize perioperative care, many efforts have been published to preoperatively predict length of stay and morbidity using traditional risk factors such as age, preoperative cardio-pulmonary disease, anemia, diabetes, frailty, etc.<sup>4-8</sup> These efforts have been based on traditional statistical methods, most often multiple regression analyses, and essentially concluding that it is "better to be young and healthy than old and sick".

Consequently, despite being statistically significant, conventional risk-stratification based on such studies has had a relatively limited clinically relevant ability to predict and reduce potentially preventable morbidity and length of stay.<sup>4-8</sup>

More recently, machine-learning methods have been introduced with success in several areas of healthcare and where preliminary data suggest them to improve surgical risk prediction compared to traditional risk calculation in certain anesthetic and surgical conditions.<sup>9,10</sup> This is also the case in total hip replacement, total knee replacement and uni-compartmental knee replacement, where several publications on machine-learning algorithms for prediction of length of stay,<sup>11,12</sup> complications,<sup>13</sup> disability,<sup>14</sup> potential outpatient setup,<sup>15</sup> readmissions<sup>16</sup> or payment models,<sup>17,18</sup> have shown promising predictive value compared to conventional statistical methods.<sup>19</sup>

However, few papers have included enhanced recovery programs, and most are based on large database cohorts with the presence of risk factors and complications often relying on administrative coding with limited information on perioperative care, follow-up and discharge destination. In our previous study of 9512 total hip and knee replacements within an enhanced recovery protocol and including the above information, we did not find advantages of machine-learning methods compared to logistic regression in predicting a length of stay > 2 days.<sup>20</sup> However, this may have been due to data imbalance, lack of details on medication and the chosen outcome of length of stay of >2 days.<sup>20</sup> Thus, machine-learning models remain promising and could provide an improved basis for identifying a potential "high-risk" surgical

1  
2  
3  
4 population who may benefit from more extensive preoperative evaluation and postoperative  
5 medical care.  
6  
7 Consequently, we analyzed whether an improved machine-learning model was better for  
8 preoperative prediction of medical complications resulting in prolonged length of stay and  
9 readmissions compared to a traditional logistic regression model, in a large consecutive cohort  
10 of patients undergoing fast-track total hip and knee replacement within a national public health-  
11 care system.<sup>1</sup> In addition to well-defined patient-reported preoperative risk-factors, we also  
12 included information on dispensed reimbursed prescriptions 6 months prior to surgery using a  
13 nationwide registry.<sup>21</sup>  
14  
15  
16  
17  
18  
19  
20

## 21 Method

22

23  
24 Reporting of the study is done in accordance with the Transparent reporting of multivariable  
25 prediction model for individual prognosis or diagnosis (TRIPOD) statement<sup>22</sup> and the Clinical AI  
26 Research (CAIR) checklist proposal.<sup>23</sup>  
27  
28 The study is based on the Centre for Fast-track Hip and Knee Replacement database which is a  
29 prospective database on preoperative patient characteristics and enrolling consecutive patients  
30 from 7 departments between 2010 and 2017. The database is registered on ClinicalTrials.gov  
31 as a study registry (NCT01515670). Permission to review and store information from medical  
32 records without informed consent was acquired from Center for Regional Development (R-  
33 20073405) and the Danish Data Protection Agency (RH-2007-30-0623). Patients completed a  
34 preoperative questionnaire with nurse assistance if needed. Additional information on  
35 reimbursed prescriptions 6 months prior to surgery was acquired using the Danish National  
36 Database of Reimbursed Prescriptions (DNDRP) which records all dispensed prescriptions with  
37 reimbursement in Denmark.<sup>21</sup> Finally, data were combined with the Danish National Patient  
38 Registry (DNPR) for information on length of stay (counted as postoperative nights spent in  
39 hospital), 90-days readmissions with overnight stay and mortality. In case of length of stay >4  
40 days or readmission, patient discharge summaries were reviewed for information on  
41 postoperative morbidity and in case of insufficient information, the entire medical records were  
42 reviewed. Readmissions were only included if considered related to the surgical procedure, thus  
43 excluding planned procedures like cancer workouts, cataract surgery, etc. Readmissions due to  
44 urinary tract infection or dizziness after day 30 were also considered unrelated to the surgical  
45 procedure. In case of postoperative mortality the entire medical record, including potential  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 readmissions, was reviewed to identify cause of death. Evaluation of discharge and medical  
5 records was performed by PP supervised by CJ. In case of disagreement, records were  
6 conferred with HK. Subsequently, causes of length of stay >4, readmissions or mortality were  
7 classified as "medical" when related to perioperative care (renal failure, falls, pain, thrombosis,  
8 anemia, venous thromboembolism or infection etc.) and "surgical" if related to surgical  
9 technique (prosthetic infection, revision surgery, periprosthetic fracture, hip dislocation, etc.).<sup>1</sup> In  
10 case of a length of stay 4-6 days with a standard discharge summary describing a successful  
11 postoperative course, it was assumed that no clinically relevant postoperative complications had  
12 occurred. If length of stay was >6 days but with standard discharge summary, the entire medical  
13 record was evaluated to confirm that no relevant complications had occurred.  
14  
15 For the present study, only cases between 2014 and 2017 were used to provide the most up-to  
16 date data. All patients had elective unilateral total hip and knee replacement in dedicated  
17 arthroplasty departments with similar fast-track protocols, including multimodal opioid sparing  
18 analgesia with high-dose (125mg) methylprednisolone, preference for spinal anesthesia, only in-  
19 hospital thromboprophylaxis when length of stay ≤5 days, early mobilization, functional  
20 discharge criteria and discharge to own home.<sup>1</sup> There was no selection criteria for the fast-track  
21 protocol as it is considered standard of care, but we excluded patients with previous major hip  
22 or knee surgery within 90-days of their total hip or total knee replacement and total hip  
23 replacement due to severe congenital joint disorder or cancer.  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39

## Outcomes

40 The primary outcome was to compare prediction quality when using a machine-learning model  
41 to predict the occurrence of "medical" complications resulting in a length of stay >4 days or  
42 readmission compared to a traditional logistic regression model (outcome A). Secondarily, we  
43 investigated how inclusion of cases with a length of stay >4 days but no reported "medical"  
44 complication as a positive outcome influenced the model (outcome B). For both outcomes, we  
45 also investigated whether a parsimonious model including only the top ten variables would  
46 perform equally well as the full model, and whether the effect of age per se would compare to  
47 the full machine-learning model. All figures and tables in the main text and Appendix are based  
48 on outcome A; the corresponding figures for outcome B are reported in the Supplemental Digital  
49 Content.  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## Statistical Analysis

Data was initially trimmed by removing 156 patients (1.7%) who were outliers with regards to weight (<30 kg or >250 kg) and height (<100 cm or >210 cm) or where these data were missing. To reduce the risk of overfitting, the dataset was subsequently split into a training set consisting of 18.013 (82.2%) procedures from 2014-2016 and a test set of 3913 (17.8%) procedures from 2017.

As a reference model, we used classical logistic regression using all 33 input variables (table 1). Cases of missing values in the logistic regression model were handled by imputing missing values with the median of present values. All variables were then normalized.

In addition, we used Boosted Decision Trees (LightGBM)<sup>24</sup> for the machine-learning models, as such methods work well with categorical data and missing values. We tried using both normal cross entropy and FocalLoss<sup>25</sup> as the objective function for the machine-learning model. The reason for testing FocalLoss was to allow the machine-learning model to focus more on the (few) positives.

The full machine-learning model was trained and hyperparameter optimized using state of the art machine-learning methods. The models were trained on the training data and then used for making predictions on the unseen test data (see supplementary for details). The classification threshold was calibrated such that no more than 20% of the total number of patients were predicted as positive by the model (a positive predictive fraction (PPF) of 20%). We also included results for PPF values of 25% and 30%. Furthermore, we trained two parsimonious models using machine-learning and logistic regression with only the 10 most important features. Finally, we specifically explored the influence of increasing age, by constructing a model based only on age (Age), and a machine-learning model based on all variables except for age.

To investigate the importance of the included variables, we computed the SHapley Additive exPlanations (SHAP) values, which provide estimates on which variables contribute most to the risk score predictions.<sup>26,27</sup> Finally, we investigated a potential relation between reimbursed prescribed cardiac drugs, anticoagulants, psychotropics and pulmonary drugs and age, as the relation between polypharmacy and postoperative outcomes have mainly been found in older patients.<sup>28</sup>

For evaluating model performance, we computed the number of true positives, false positives, false negatives, true negatives, sensitivity (true positive rate), precision (positive predictive value). Since the data was quite imbalanced (about a 1:20 positive:negative ratio) we also computed the Matthews Correlation Coefficient (MCC) which is independent of class

1  
2  
3  
4 imbalance.<sup>29,30</sup> The MCC ranges between -1 (the 100% wrong classifier), 0 (the random  
5 classifier), and +1 (the perfect classifier). Finally, we computed the area under the receiver  
6 operating characteristic curve (AUC) and the area under the precision recall curve (AUPRC). To  
7 evaluate the statistical difference between the classifiers, we applied a Bayesian metric  
8 comparison P(sensitivity),<sup>31</sup> which is the probability that a model will perform better than the  
9 machine-learning model relative to the sensitivity. Thus, for two equally performing models  
10 P(sensitivity) is ≈ 50%.  
11  
12  
13  
14  
15

## 16 Results

17

18 Median age in the 3913 patients was 70 years (IQR 62-76), 59% were female and 58% had  
19 total hip replacement (table 1). Details on prescribed drug types are shown in Appendix 1.  
20 Median length of stay was 2 (IQR: 1-2) days with 7.6% 90-days readmissions and outcome A  
21 occurring in 182 (4.7%) patients. When applying any model with a positive prediction fraction of  
22 20% to the 3913 patients, 782 qualified as “risk-patients”. The results are summarized in figure  
23 1 and table 2. When considering risk scores from the full machine-learning (figure 1a) and full  
24 logistic regression model leading to this risk-patient selection, 106 and 98 had outcome A,  
25 respectively. Correspondingly, the sensitivity and precision were 58.2% and 13.6% for the full  
26 machine-learning and 53.8% and 12.5% for the full logistic regression model, respectively. The  
27 full machine-learning model was superior (figure 1b) on essentially all parameters compared to  
28 any of the other models, although the differences were minor (table 2). The results were similar  
29 when using positive prediction fractions of 25% and 30%, but with the sensitivity for the full  
30 machine-learning model increasing to 64.4% and 69.2% and precision decreasing to 12.0% and  
31 10.7%, respectively (Appendix table 2).  
32 Both the machine-learning model excluding age and age-only model had significantly lower  
33 sensitivity than the full machine-learning model (figure 1b). Despite age being the single most  
34 important variable (figure 2), the machine-learning model excluding age performed as well as  
35 the age-only model.  
36 When evaluating feature importance, we found a strong correlation between the full machine-  
37 learning and full logistic regression model, with age and use of walking aids being the most  
38 important variables in both (figure 2a). From the combined importance of variables outside the  
39 top ten, the machine-learning approach extracted more information with fewer variables than  
40 logistic regression (figure 1b).  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 For the full machine-learning model, there was a clear signal that increasing age, number of  
5 reimbursed prescriptions, and presence of comorbidity, all contributed to an increased risk  
6 score. In contrast, a recent date of surgery and an increased hemoglobin level seemed to  
7 reduce the calculated risk (figure 2b). Individual analysis of the SHAP interaction values for  
8 types of anticoagulant prescriptions revealed that prescriptions on vitamin-K antagonists (VKA)  
9 or adenosine diphosphate (ADP) antagonists increased, while acetylic salicylic acid and direct  
10 oral anticoagulants (DOAC) reduced the risk score of the full machine-learning model,  
11 regardless of age (figure 3a). The SHAP analysis of prescribed cardiac drugs revealed that  
12 prescriptions on  $\text{Ca}^{2+}$ -antagonists and betablockers in combination with one or two other  
13 antihypertensives increased the risk-score, as did prescriptions on nitrates, other  
14 antihypertensives and antiarrhythmics. For the remaining cardiac drugs, prescriptions either  
15 reduced or had minor influence, and with limited relation with age (figure 3b). Preoperative  
16 psychotropic prescriptions increased the risk-score except for antipsychotics (0.6%). For users  
17 of selective serotonin inhibitors there was a clear age-related distinction with the risk score  
18 being increased in elderly patients but decreased in those < 60 years (figure 3c). Finally, the risk  
19 score increased with prescriptions on inhalation steroid and  $\beta$ -blockers, and more accentuated  
20 in the younger patients (figure 3d).

21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65 The results including patients with a length of stay >4 days, but no reported postoperative  
complications (outcome B) were similar as for outcome A. In general, we found that the full  
machine-learning model was superior to the others, although the difference were smaller than  
for outcome A. (Supplemental Digital Content table S1 listing outcome parameters and  
Supplemental Digital Content 2 figure S1a-b showing distributions and ROC curves for outcome  
B). While the ten most important variables for the full machine-learning model remained  
unchanged, familiar disposition for venous thromboembolism replaced gender as one of the top  
ten important variables in the full logistic regression model (Supplemental Digital Content figure  
S2a-b showing SHAP values for outcome B). Furthermore, the SHAP analysis on specific  
prescribed drugs demonstrated that the machine-learning model found no benefits from  
information on prescriptions on respiratory drugs, why all SHAP values were zero. In addition,  
the reduced risk with acetylsalicylic acid and DOAC prescriptions, as well as the influence of  
practically all cardiac drugs except for nitrates, other antihypertensives and antiarrhythmics, was  
attenuated (Supplemental Digital Content 4 figure S3a-d showing SHAP-values of prescriptions  
of specific drugs for outcome B).

1

2

3

4

## Discussion

5

6

7

We found that using a machine-learning algorithm including all 33 available variables and a parsimonious machine-learning-algorithm encompassing only the 10 most important predictors improved prediction of patients at increased risk of having a length of stay >4 days or readmissions due to medical complications compared to traditional logistic regression models. In contrast, when also including patients having a length of stay >4 days but without a well-defined complication as an outcome, the parsimonious machine-learning model was slightly worse than a traditional logistic regression model including all variables. We also found that although age was the single most important predictor of both outcome A and B, it was less suited for prediction of postoperative medical complications after fast-track total hip and knee replacement on its own. Finally, we demonstrated how the chosen classification threshold of the machine-learning algorithm influenced model performance through an increase in sensitivity at the cost of decreased precision.

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

A previous systematic review also found that machine-learning algorithms may provide better prediction of postoperative outcomes in THA and TKA.<sup>32</sup> However, the authors concluded that such models performed best at predicting postoperative complications, pain and patient reported outcomes and were less accurate at predicting readmissions and reoperations.<sup>32</sup> That machine-learning algorithms may improve prediction of complications after THA and TKA compared to traditional logistic regression was also found by Shah *et al.* who used an automated machine-learning framework to predict selected major complications after THA.<sup>13</sup> However, theirs was a retrospective study based on diagnostic and administrative coding and the selected complications occurred only in 0.61% of patients, potentially limiting clinical relevance. In contrast, we aimed at identifying a cohort which would comprise 20% of patients in which we found about 60% of all medical complications. This we believe, is within the means of the Danish socialized healthcare system to allocate additional resources for intensified perioperative care and with both patient-related and economic benefits due to potentially avoided complications and costs.

In contrast to many other machine-learning studies,<sup>33</sup> our dataset included not only preoperative data but also only one paraclinical variable, which was preoperative hemoglobin. Although the inclusion of other laboratory tests such as preoperative albumin, sodium and alkaline phosphatase has been found to be of importance in machine-learning algorithms for home discharge in UKA<sup>12</sup> and spine surgery,<sup>9</sup> they are not standard in our fast-track protocols and not easy to interpret from a pathophysiological point of view. As there is a need to prioritize the

1  
2  
3  
4 limited health-care resources, most decisions on which patients may benefit from more  
5 extensive postoperative care will likely need to be conducted preoperatively. Thus, although  
6 postoperative information such as duration of surgery, perioperative blood length of stays or  
7 postoperative hemoglobin have been included in other studies<sup>33</sup>, we decided against the use of  
8 peri- and postoperative data. The same approach has been used by Ramkumar *et al.* who used  
9 U.S. National Inpatient Sample data including 15 preoperative variables, to predict length of  
10 stay, patient charges and disposition after both TKA<sup>34</sup> and THA.<sup>18</sup> However, these studies were  
11 not conducted in a socialized health care system, and the main focus was on the need for  
12 differentiated payment bundles and without specific information on the reason for increased  
13 length of stay or non-home discharge.<sup>34</sup> Wei *et al.* used an artificial neural network model to  
14 predict same-day discharge after TKA, based on the NSQUIP database from 2018 and found  
15 that six of the ten most important variables were the same compared with logistic regression,  
16 similar to our findings.<sup>35</sup> However, patients with one-day length of stay were intentionally  
17 excluded due to variations in in-patient vs. out-patient registration.<sup>35</sup>  
18 Age has traditionally been a major factor when predicting surgical outcomes which is why we  
19 choose to specifically evaluate its effect on our risk-prediction. That age is important for risk-  
20 prediction was further illustrated by the machine-learning model without age being comparable  
21 to the age-only model. Note that, although elderly patients had increased risk of postoperative  
22 complications, likely related to decline of physical reserves,<sup>36</sup> the use of chronological age alone  
23 as a selection criteria for being a “risk-patient” was inferior compared to both machine-learning  
24 and logistic regression models incorporating comorbidity and functional status.  
25 We used the SHAP values for estimation of feature importance, thus providing a better  
26 understanding of the otherwise “black-box” machine-learning model. The SHAP values showed  
27 which variables contribute most to the risk-score predictions.  
28 Our inclusion of specific data on reimbursed prescriptions 6 months prior to surgery based upon  
29 the unique Danish registries, unsurprisingly found increased risk-scores with increased number  
30 of prescriptions and with the majority being in elderly patients. Similarly, a Canadian study in  
31 elective non-cardiac surgery found decreased survival and increased length of stay and  
32 readmissions and costs in patients >65 years with polypharmacy.<sup>28</sup> However, this is a complex  
33 relationship where some patients benefit from their treatments, while other may suffer from  
34 undesirable side-effects. Consequently, the authors cautioned against altering perioperative  
35 practices based on current evidence.<sup>28</sup> However, the information from the included prescriptions  
36 with SHAP analysis may provide inspiration for new hypothesis-generating studies such as  
37 investigation of the potential differences in risk-profile between having preoperative prescribed  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 VKA and DOAKs. Also, the age-related differences in risk from SSRI's seen in our study could  
5 guide further studies on "deprescription".  
6  
7 Another requirement for machine-learning-algorithms to be clinically useful is user friendliness  
8 and not depending on excessive additional data collection by the attending clinicians. In this  
9 context, it was a bit disappointing that the parsimonious machine-learning algorithm with only  
10 the ten most important variables was slightly worse at predicting outcome B than the full logistic  
11 regression model. A reason for this could be that when including a length of stay >4 days but  
12 without described medical complications, the combination of all variables provide information  
13 not available by merely including the ten most important ones. This highlights the need for as  
14 much detailed, and preferably non-binary, data as possible to fulfill the true potential of  
15 machine-learning algorithms.  
16  
17 Our study has some limitations. First, one of the strengths of machine learning compared to  
18 logistic regression is the analysis of multilevel continuous data, whereas we included only a  
19 limited number of, often binary, preoperative variables. This could have limited the full  
20 realization of our machine learning model. As previously mentioned, we excluded intraoperative  
21 information, including type of anesthesia, surgical approach etc. all of which may influence  
22 postoperative outcomes. The observational design of this study means that we cannot exclude  
23 unmeasured confounding or confounding by indication. Also, despite that the DNDRP has a  
24 near complete registration of dispensed medicine in Denmark, some types or drugs, especially  
25 benzodiazepines, are exempt from general reimbursement and thus not sufficiently captured.<sup>21</sup>  
26 Furthermore, it is doubtful whether the patients used all types of drugs at the time of surgery  
27 (e.g. heparin which is rarely for long-term use). Finally, classification of a complication being  
28 "medical" depended on review of the discharge records which can also introduce bias. However,  
29 we believe our approach to be superior to depending only on diagnostic codes which often are  
30 inaccurate<sup>37</sup> and provide limited details on whether the complication may be attributed to a  
31 medical or surgical adverse event. The strengths of our study include the use of national  
32 registries with high degree of completion (>99% of all somatic admissions in case of the  
33 DNDRP),<sup>38</sup> prospective recording of comorbidity, extensive information on prescription patterns  
34 6 months prior to surgery and similar established enhanced recovery protocols in all  
35 departments.  
36  
37 In summary, our results suggest that machine-learning-algorithms likely provide clinically  
38 relevant improved predictions for defining patients in high-risk of medical complications after  
39 fast-track THA and TKA compared to a logistic regression model. Future studies could benefit  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 from using such algorithms to find a manageable population of patients who benefit from  
5 intensified perioperative care.  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## References

- Petersen PB, Kehlet H, Jorgensen CC, Lundbeck Foundation Centre for Fast-track H, Knee Replacement Collaborative G: Improvement in fast-track hip and knee arthroplasty: a prospective multicentre study of 36,935 procedures from 2010 to 2017. *Sci Rep* 2020; 10: 21233
  - Khan SK, Malviya A, Muller SD, Carluke I, Partington PF, Emmerson KP, Reed MR: Reduced short-term complications and mortality following Enhanced Recovery primary hip and knee arthroplasty: results from 6,000 consecutive procedures. *Acta Orthop.* 2014; 85: 26-31
  - Partridge T, Jameson S, Baker P, Deehan D, Mason J, Reed MR: Ten-Year Trends in Medical Complications Following 540,623 Primary Total Hip Replacements from a National Database. *J Bone Joint Surg Am* 2018; 100: 360-367
  - Jorgensen CC, Gromov K, Petersen PB, Kehlet H: Influence of day of surgery and prediction of LOS > 2 days after fast-track hip and knee replacement. *Acta Orthop* 2021; 92: 170-175
  - Jorgensen CC, Petersen MA, Kehlet H: Preoperative prediction of potentially preventable morbidity after fast-track hip and knee arthroplasty: a detailed descriptive cohort study. *BMJ Open*. 2016; 6: e009813
  - Johns WL, Layon D, Golladay GJ, Kates SL, Scott M, Patel NK: Preoperative Risk Factor Screening Protocols in Total Joint Arthroplasty: A Systematic Review. *J Arthroplasty* 2020; 35: 3353-3363
  - Adhia AH, Feinglass JM, Suleiman LI: What Are the Risk Factors for 48 or More-Hour Stay and Nonhome Discharge After Total Knee Arthroplasty? Results From 151 Illinois Hospitals, 2016-2018. *J Arthroplasty* 2020; 35: 1466-1473 e1
  - Shah A, Memon M, Kay J, Wood TJ, Tushinski DM, Khanna V, McMaster Arthroplasty Collective g: Preoperative Patient Factors Affecting Length of Stay following Total Knee Arthroplasty: A Systematic Review and Meta-Analysis. *J Arthroplasty* 2019; 34: 2124-2165 e1
  - Li Q, Zhong H, Girardi FP, Poeran J, Wilson LA, Memtsoudis SG, Liu J: Machine Learning Approaches to Define Candidates for Ambulatory Single Level Laminectomy Surgery. *Global Spine J* 2021; 2192568220979835
  - Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR: Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission. *Ann Surg* 2020; 272: 1133-1139
  - Li H, Jiao J, Zhang S, Tang H, Qu X, Yue B: Construction and Comparison of Predictive Models for Length of Stay after Total Knee Arthroplasty: Regression Model and Machine Learning Analysis Based on 1,826 Cases in a Single Singapore Center. *J Knee Surg* 2022; 35: 7-14
  - Lu Y, Khazi ZM, Agarwalla A, Forsythe B, Taunton MJ: Development of a Machine Learning Algorithm to Predict Nonroutine Discharge Following Unicompartmental Knee Arthroplasty. *J Arthroplasty* 2021; 36: 1568-1576
  - Shah AA, Devana SK, Lee C, Kianian R, van der Schaaf M, SooHoo NF: Development of a Novel, Potentially Universal Machine Learning Algorithm for Prediction of Complications After Total Hip Arthroplasty. *J Arthroplasty* 2021; 36: 1655-1662 e1
  - Sniderman J, Stark RB, Schwartz CE, Imam H, Finkelstein JA, Nousiainen MT: Patient Factors That Matter in Predicting Hip Arthroplasty Outcomes: A Machine-Learning Approach. *J Arthroplasty* 2021; 36: 2024-2032
  - Kugelman DN, Teo G, Huang S, Doran MG, Singh V, Long WJ: A Novel Machine Learning Predictive Tool Assessing Outpatient or Inpatient Designation for Medicare Patients Undergoing Total Hip Arthroplasty. *Arthroplast Today* 2021; 8: 194-199

- 1  
2  
3  
4     16. Mohammadi R, Jain S, Namin AT, Scholem Heller M, Palacholla R, Kamarthi S, Wallace  
5     B: Predicting Unplanned Readmissions Following a Hip or Knee Arthroplasty: Retrospective  
6     Observational Study. JMIR Med Inform 2020; 8: e19761  
7     17. Ramkumar PN, Navarro SM, Haeberle HS, Karnuta JM, Mont MA, Iannotti JP, Patterson  
8     BM, Krebs VE: Development and Validation of a Machine Learning Algorithm After Primary  
9     Total Hip Arthroplasty: Applications to Length of Stay and Payment Models. J Arthroplasty 2019;  
10    34: 632-637  
11    18. Ramkumar PN, Karnuta JM, Navarro SM, Haeberle HS, Iorio R, Mont MA, Patterson  
12    BM, Krebs VE: Preoperative Prediction of Value Metrics and a Patient-Specific Payment Model  
13    for Primary Total Hip Arthroplasty: Development and Validation of a Deep Learning Model. J  
14    Arthroplasty 2019; 34: 2228-2234 e1  
15    19. Haeberle HS, Helm JM, Navarro SM, Karnuta JM, Schaffer JL, Callaghan JJ, Mont MA,  
16    Kamath AF, Krebs VE, Ramkumar PN: Artificial Intelligence and Machine Learning in Lower  
17    Extremity Arthroplasty: A Review. J Arthroplasty 2019; 34: 2201-2203  
18    20. Johannesson KB, Kehlet H, Petersen PB, Aasvang EK, Sørensen HBD, Jørgensen  
19    CC: Machine learning classifiers do not improve prediction of hospitalization > 2 days after fast-  
20    track hip and knee arthroplasty compared with a classical statistical risk model. Acta Orthop  
21    2022; 93: 117-123  
22    21. Johannesson SA, Horvath-Puhó E, Ehrenstein V, Schmidt M, Pedersen L, Sorensen  
23    HT: Existing data sources for clinical epidemiology: The Danish National Database of  
24    Reimbursed Prescriptions. Clin.Epidemiol. 2012; 4: 303-313  
25    22. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers  
26    AJ, Ransohoff DF, Collins GS: Transparent Reporting of a multivariable prediction model for  
27    Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;  
28    162: W1-73  
29    23. Olczak J, Pavlopoulos J, Prijs J, Ijpmma FFA, Doornberg JN, Lundstrom C, Hedlund J,  
30    Gordon M: Presenting artificial intelligence, deep learning, and machine learning studies to  
31    clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical  
32    AI Research (CAIR) checklist proposal. Acta Orthop 2021; 92: 513-525  
33    24. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T: LightGBM: a highly  
34    efficient gradient boosting decision tree, Proceedings of the 31st International Conference on  
35    Neural Information Processing Systems. Red Hook, NY, USA, Curran Associates Inc, 2017, pp  
36    3149-57  
37    25. Lin T-Y, Goyal P, Girshick R, He K, Dollár P: Focal Loss for Dense Object Detection.  
38    <http://arxiv.org/abs/1708.02002>, ArXiv170802002 Cs 2018  
39    26. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J,  
40    Bansal N, Lee SI: From Local Explanations to Global Understanding with Explainable AI for  
41    Trees. Nat Mach Intell 2020; 2: 56-67  
42    27. Lundberg SMLSI: A Unified Approach to Interpreting Model Predictions. Edited by  
43    Guyon I. Adv Neural Inf Process Syst [Internet], Curran Associates, Inc., 2017  
44    28. McIsaac DL, Wong CA, Bryson GL, van Walraven C: Association of Polypharmacy with  
45    Survival, Complications, and Healthcare Resource Use after Elective Noncardiac Surgery: A  
46    Population-based Cohort Study. Anesthesiology 2018; 128: 1140-1150  
47    29. Chicco D: Ten quick tips for machine learning in computational biology. BioData Mining  
48    2017; 10: 35 (2017)  
49    30. Chicco D, Totsch N, Jurman G: The Matthews correlation coefficient (MCC) is more  
50    reliable than balanced accuracy, bookmaker informedness, and markedness in two-class  
51    confusion matrix evaluation. BioData Mining 2021; 14: 13 (2021)  
52    31. Totsch N, Hoffmann D: Classifier uncertainty: evidence, potential impact, and  
53    probabilistic treatment. PeerJ Comput Sci 2021; 7: e398  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- 1  
2  
3  
4       32. Lopez CD, Gazgalis A, Boddapati V, Shah RP, Cooper HJ, Geller JA: Artificial Learning  
5 and Machine Learning Decision Guidance Applications in Total Hip and Knee Arthroplasty: A  
6 Systematic Review. *Arthroplast Today* 2021; 11: 103-112  
7       33. Han C, Liu J, Wu Y, Chong Y, Chai X, Weng X: To Predict the Length of Hospital Stay  
8 After Total Knee Arthroplasty in an Orthopedic Center in China: The Use of Machine Learning  
9 Algorithms. *Front Surg* 2021; 8: 606038  
10      34. Ramkumar PN, Karnuta JM, Navarro SM, Haeberle HS, Scuderi GR, Mont MA, Krebs  
11 VE, Patterson BM: Deep Learning Preoperatively Predicts Value Metrics for Primary Total Knee  
12 Arthroplasty: Development and Validation of an Artificial Neural Network Model. *J Arthroplasty*  
13 2019; 34: 2220-2227 e1  
14      35. Wei C, Quan T, Wang KY, Gu A, Fassihi SC, Kahlenberg CA, Malahias MA, Liu J,  
15 Thakkar S, Gonzalez Della Valle A, Sculco PK: Artificial neural network prediction of same-day  
16 discharge following primary total knee arthroplasty based on preoperative and intraoperative  
17 variables. *Bone Joint J* 2021; 103-B: 1358-1366  
18      36. Griffiths R, Beech F, Brown A, Dhesi J, Foo I, Goodall J, Harrop-Griffiths W, Jameson J,  
19 Love N, Pappenheim K, White S: Peri-operative care of the elderly. *Anaesthesia* 2014; 69 Suppl  
20 1: 81-98  
21      37. Bedard NA, Pugely AJ, McHugh MA, Lux NR, Bozic KJ, Callaghan JJ: Big Data and  
22 Total Hip Arthroplasty: How Do Large Databases Compare? *J Arthroplasty* 2018; 33: 41-45.e3  
23      38. Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sorensen HT: The  
24 Danish National Patient Registry: a review of content, data quality, and research potential. *Clin  
25 Epidemiol* 2015; 7: 449-90  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1

2

3

4

## Figure legends

5

6

7

### Figure 1a-b

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

### Figure 2a-b

2a) The overall importance of the 10 most important variables measured by the SHAP-values for the full machine-learning and full logistic regression models on outcome A (LOS >4 days or readmission due to "medical" morbidity). Only the importance of prescribed anticholesterols and gender differ between the models. The contributions of the remaining variables are summed in the bottom bar.

2b) The SHAP-values for the full machine-learning model on outcome A, where positive increase and negative values decrease the risk score. The color is related to the value of the variable with blue being lowest and red highest and each dot represents a patient.

### Figure 3a-d

SHAP scatter-plot on the contributions to the full machine-learning model on outcome A (LOS >4 days or readmission due to "medical" morbidity), for individual types of prescribed anticoagulants, cardiac drugs, psychotropics and respiratory drugs stratified by age.

3a) Prescribed anticoagulants

VKA: vitamin K antagonists ASA: acetylsalicylic acid DOAC: direct oral anticoagulant ADP: Adenosine diphosphate ACE: angiotensin converting enzyme

3b) Prescribed cardiac drugs

ACE: angiotensin converting enzyme AHT: antihypertensive. Other AHT were defined as AHT different from diuretics ANG-II/ACE inhibitors or Ca<sup>2+</sup>antagonists. IHD: Ischemic heart disease

3c) Prescribed psychotropics

SSRI: Selective serotonin inhibitor SNRI: Serotonin and norepinephrine reuptake inhibitor NaRI: Norepinephrine reuptake inhibitor NaSSA: Norepinephrine and specific serotonergic antidepressants. AD: antidepressants BZ: Benzodiazepines (likely underreported due to limited general reimbursement in Denmark). ADHD: Attention-deficit/hyperactivity disorder

3d) Prescribed respiratory drugs

SABA: Short-acting beta agonist LABA: long-acting beta agonist LAMA: Long-acting muscarinic antagonist.

table1

Table 1. patient demographics with and without outcome A (length of stay >4 days or readmissions due to "medical" morbidity) in the combined test and training dataset.

Preoperative characteristics n (%) unless otherwise specified	+outcome A (n:1180)	-outcome A (n:20837)
mean age (SD)	75.0 (68.0-81.0)	69.0 (62.0-75.0)
mean number of reimbursed prescriptions <sup>1</sup> (SD)	3.0 (1.0-4.0)	2.0 (0.0-3.0)
female gender	755 (64.0)	12133 (58.2)
Total hip replacement	636 (53.9)	11542 (55.4)
mean weight in kg (SD)	78.0 (67.0-91.0)	81 (70.0-93.0)
mean height in cm (SD)	168 (162.0-175.0)	170.0 (164.0-178.0)
mean body mass index (SD)	27.3 (23.9-31.2)	27.5 (24.6-31.1)
regular use of walking aid	552 (46.8)	4398 (21.5)
missing	29 (2.5)	359 (1.7)
living alone	578 (49.0)	6717 (32.2)
with others	571 (48.4)	13869 (66.6)
institution	24 (2.0)	113 (0.5)
missing	7 (0.6)	138 (0.7)
hemoglobin	8.2 (7.7-8.8)	8.6 (8.1-9.2)
missing	11 (0.9)	314 (1.5)
>2 units of alcohol/day	79 (6.7)	1589 (7.6)
missing	10 (0.8)	174 (0.8)
active smoker	130 (11.0)	2751 (13.2)
missing	11 (0.9)	141 (0.7)
cardiac disease	306 (25.9)	2750 (13.2)
missing	8 (0.8)	153 (0.7)
hypercholesterolemia	467 (39.6)	6062 (29.1)
missing	8 (0.7)	120 (0.6)
hypertension	738 (62.5)	10141 (48.7)
missing	64 (5.4)	663 (3.2)
pulmonary disease	182 (15.4)	1841 (8.8)
missing	5 (0.4)	96 (0.5)
previous cerebral attack	165 (14.0)	1086 (5.2)
missing	25 (2.1)	282 (1.4)
previous VTE	133 (11.3)	1481 (7.1)
missing	26 (2.2)	325 (1.6)
malignancy (undefined)	557 (47.2)	8843 (42.4)
previous radically treated malignancy	127 (10.8)	2065 (9.9)
missing	14 (1.2)	162 (0.8)
chronic kidney disease	50 (4.2)	273 (1.3)
missing	35 (3.0)	292 (1.4)
family member with VTE	155 (13.1)	2510 (12.0)
missing	1190 (16.1)	2569 (12.3)
regular snoring	266 (22.5)	5522 (26.5)
uncertain about snoring	208 (17.6)	3781 (18.1)
missing	259 (21.9)	3309 (15.9)
not feeling rested	468 (39.7)	9340 (44.8)

uncertain about being rested	48 (4.1)	809 (3.9)
missing	105 (8.9)	1230 (5.9)
psychiatric disorder	156 (13.2)	1590 (7.6)
missing	62 (5.3)	703 (3.4)

Characteristic based on combination of questionnaire and DNDRP

diabetes

diet treated diabetes <sup>2</sup>	29 (2.5)	274 (1.3)
oral antidiabetics	137 (11.6)	1448 (6.9)
insulin treated diabetes <sup>3</sup>	60 (5.1)	413 (2.0)
missing	7 (0.6)	98 (0.5)

SD: standard deviation VTE: venous thromboembolic event DNDRP: Danish National Database of Reimbursed Prescriptions.

<sup>1</sup>Antirheumatica, steroids, anticoagulants, cardiac, cholesterol lowering, respiratory and psychotropic drugs.

<sup>2</sup>Reported diabetes but no registered prescriptions <sup>3</sup>+/- oral antidiabetics

table2

Table 2: Performance of the six different models with a predefined positive prediction fraction of 20% for outcome A

Positive prediction fraction 20%	TP	FP	FN	TN	sensitivity	precision	MCC	AUROC	AUPRC	P (sensitivity)
Full machine-learning model	106	676	76	3055	58.2%	13.6%	21.1%	76.3%	15.5%	-
Full logistic regression model	98	684	84	3047	53.8%	12.5%	18.7%	74.5%	15.7%	19.7%
Parsimonious machine-learning model	100	682	82	3049	54.9%	12.8%	19.3%	75.9%	17.3%	26.1%
Parsimonious logistic regression model	95	687	87	3045	52.2%	12.1%	17.8%	73.7%	13.6%	12.4%
machine-learning model excluding age	88	694	94	3037	48.4%	11.3%	15.7%	72.3%	13.6%	3.1%
Age-only model	87	676	95	3055	47.8%	11.4%	15.8%	69.7%	12.1%	2.3%

TP: true positives FP: false positives FN: false negatives TN: true negatives MCC: Matthews correlation coefficient  
AUC: area under the ROC curve AUPRC: area under the precision recall curve P(sensitivity): probability that the  
model performs better than the machine-learning model relative to sensitivity.

Appendix1 (to appear in print publication)

## Appendix table 1

Details on specific drugs with reimbursed prescriptions 6 months preoperatively.  
Numbers are n (%)

		+Outcome A	-Outcome A
1	<b>Anticoagulants</b>		
2	none	679 (57.5)	15844 (76.0)
3	VKA	106 (9.0)	750 (3.6)
4	Heparin & Acetylsalicylic acid	0 (0.0)	7 (0.0)
5	DOAC	48 (4.1)	659 (3.2)
6	Acetylsalicylic acid	205 (17.4)	2492 (12.0)
7	Dipyradimol	5 (0.4)	29 (0.1)
8	ADP-antagonist	75 (6.4)	569 (2.7)
9	Acetylsalicylic acid & Dipyradimol	17 (1.4)	168 (0.8)
10	VKA & Acetylsalicylic acid	10 (0.8)	78 (0.4)
11	DOAC & Acetylsalicylic acid	6 (0.5)	41 (0.2)
12	VKA & ADP-antagonist	4 (0.3)	11 (0.1)
13	DOAC & ADP-antagonist	3 (0.3)	14 (0.1)
14	VKA & Heparin	1 (0.1)	21 (0.1)
15	DOAC & Acetylsalicylic acid & ADP-antagonist	1 (0.1)	3 (0.0)
16	Acetylsalicylic acid & ADP-antagonist	18 (1.5)	132 (0.6)
17	Acetylsalicylic acid & ADP-antagonist & Heparin	1 (0.1)	12 (0.1)
18	Acetylsalicylic acid & ADP-antagonist & Dipyradimol	1 (0.1)	7 (0.0)
19	<b>Cardiac prescriptions</b>		
20	none	321 (27.2)	9200 (44.2)
21	diuretics	77 (6.5)	1184 (5.7)
22	angiotensin-II/ACE-inhibitors	132 (11.2)	2683 (12.9)
23	Ca <sup>2+</sup> antagonists	55 (4.7)	773 (3.7)
24	β-blocker	29 (2.5)	559 (2.7)
25	nitrates	1 (0.1)	18 (0.1)
26	other antihypertensives	0 (0.0)	12 (0.1)
27	other types of medication for IHD	2 (0.2)	21 (0.1)
28	2 antihypertensives	177 (15.0)	2696 (12.9)
29	β-blocker & 1 antihypertensive <sup>1</sup>	92 (8.1)	1069 (5.1)
30	3 antihypertensives	50 (4.2)	548 (2.6)
31	β-blocker & 2 antihypertensives <sup>1</sup>	95 (8.1)	975 (4.7)
32	β-blocker & 3 antihypertensives <sup>1</sup>	25 (2.1)	265 (1.3)
33	4 antihypertensives	2 (0.2)	18 (0.1)
34	β-blocker & 4 antihypertensives	2 (0.2)	19 (0.1)
35	other antihypertensive & antihypertensives <sup>1</sup>	9 (0.8)	87 (0.4)
36	nitrates & any hypertensive	49 (4.2)	331 (1.6)
37	other drugs for IHD & any antihypertensive and/or nitrate	5 (0.4)	15 (0.1)
38	other antiarrhythmics & any antihypertensives	57 (4.8)	364 (1.7)
39	<b>Anticholesterols</b>		
40	none	708 (60.0)	14719 (70.6)
41	statins	457 (38.7)	5866 (28.2)
42	other anti-lipids	7 (0.6)	135 (0.6)
43	Statins +other anti-lipids	8 (0.7)	117 (0.6)

1	<u>Systemic steroids</u>	123 (10.4)	1149 (5.5)
2	<u>Antirheumatics</u>		
3	none	1143 (96.9)	20388 (97.8)
4	disease-modifying antirheumatic drugs	37 (3.1)	446 (2.1)
5	other antirheumatics	0 (0.0)	3 (0.0)
6	<u>Respiratory prescriptions</u>		
7	none	1000 (84.7)	18754 (90.0)
8	SABA	13 (1.1)	276 (1.3)
9	LABA or LAMA	19 (1.6)	217 (1.0)
10	inhalation steroid only	8 (0.7)	211 (1.0)
11	SABA & Ipratropium (+/- others)	6 (0.5)	18 (0.1)
12	LABA & steroid	45 (3.8)	474 (2.3)
13	LABA & LAMA & steroid	19 (1.6)	122 (0.6)
14	LAMA & steroid	0 (0.0)	11 (0.1)
15	LABA & LAMA	7 (0.6)	80 (0.4)
16	other pulmonary drugs	3 (0.3)	32 (0.2)
17	other pulmonary drugs & steroid	9 (0.8)	98 (0.5)
18	SABA & LABA or LAMA	6 (0.5)	96 (0.5)
19	SABA & LABA or LAMA & steroid	45 (3.8)	448 (2.2)
20	<u>Psychotropic prescriptions</u>		
21	none	952 (80.7)	18657 (89.5)
22	SSRI/SNRI/NaRI	100 (8.5)	1164 (5.6)
23	other antidepressants	1 (0.1)	17 (0.1)
24	antipsychotics	8 (0.7)	116 (0.6)
25	benzodiazepines <sup>2</sup>	0 (0.0)	7 (0.0)
26	anti-cholinergics or memantine	6 (0.5)	27 (0.1)
27	anti-ADHD drugs	1 (0.1)	10 (0.0)
28	NaSSA	25 (2.1)	184 (0.9)
29	other psychotropics	28 (2.4)	182 (0.9)
30	SSRI + other antidepressants	4 (0.3)	6 (0.0)
31	SSRI + NaSSA	8 (0.7)	94 (0.5)
32	SRRI + antipsychotics	11 (0.9)	87 (0.4)
33	SRRI + other psychotropics	7 (0.6)	84 (0.4)
34	benzodiazepines + any psychotropic	3 (0.3)	12 (0.1)
35	antipsychotics + any psychotropic	20 (1.7)	149 (0.7)
36	anti-ADHD + any psychotropic	0 (0.0)	14 (0.1)
37	NaSSA + any psychotropic	4 (0.3)	18 (0.1)
38	other psychotropics + any specified psychotropic	2 (0.2)	9 (0.0)

VKA: vitamin K antagonists DOAC: direct oral anticoagulant ADP: Adenosine diphosphate ACE: angiotensin converting enzyme IHD: Ischemic heart disease SABA: Short-acting beta agonist LABA: long-acting beta agonist LAMA: Long-acting muscarinic antagonist SSRI: Selective serotonin inhibitor SNRI: Serotonin and norepinephrine reuptake inhibitor NaRI: Norepinephrine reuptake inhibitor NaSSA: Norepinephrine and specific serotonergic antidepressants

<sup>1</sup>either diuretics, ACE/ANG-II inhibitors or Ca<sup>2+</sup>antagonists <sup>2</sup>likely underreported due to limited general reimbursement for benzodiazepines in Denmark

54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

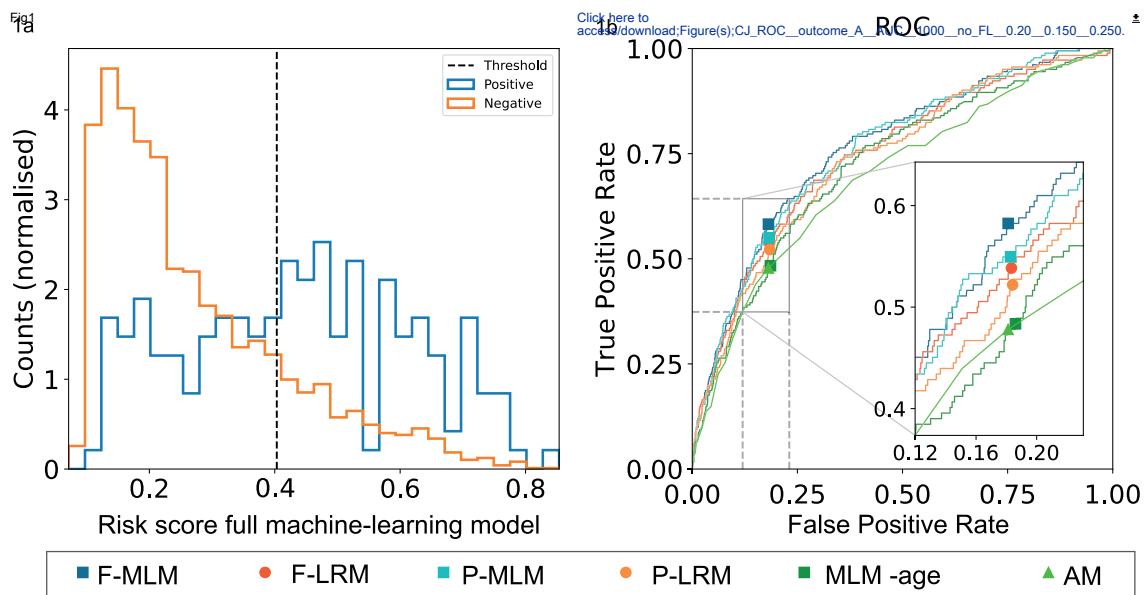
Appendix (to appear in print publication)

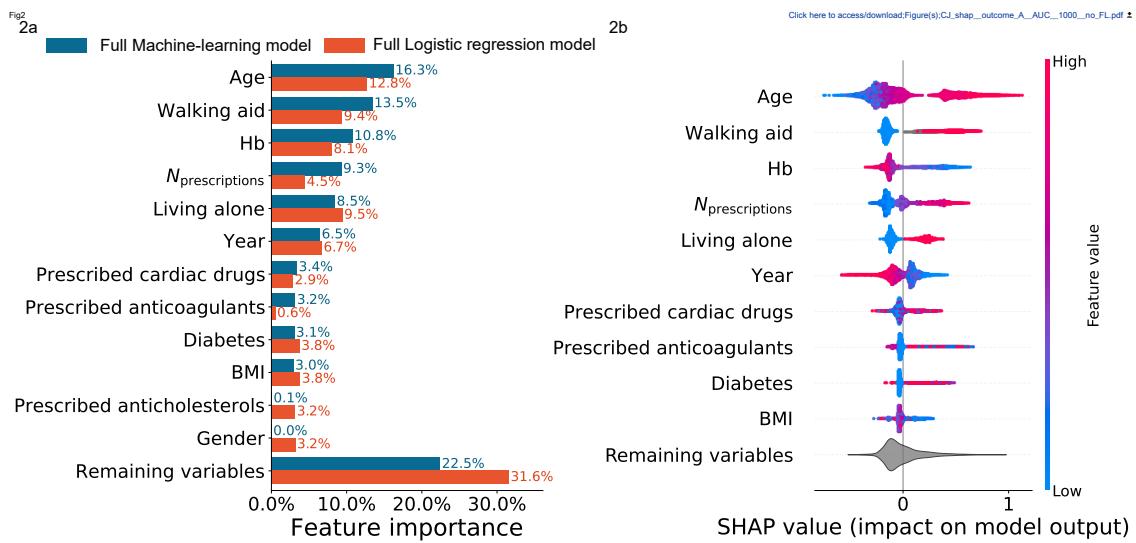
## Appendix table 2

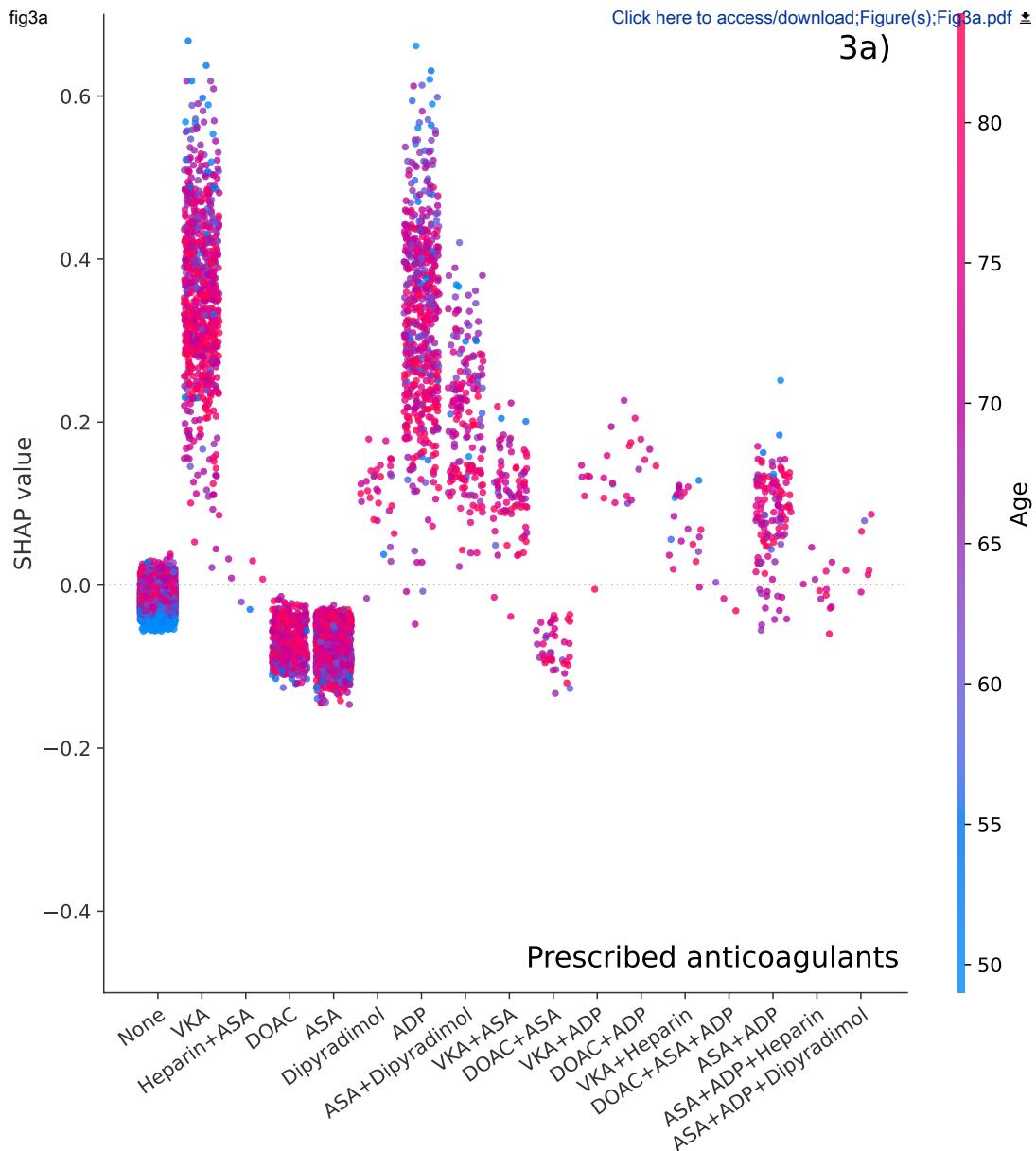
Performance of the six different models with a predefined positive prediction fraction of 25% and 30% for outcome A (LOS >4 days or readmission due to "medical" morbidity.

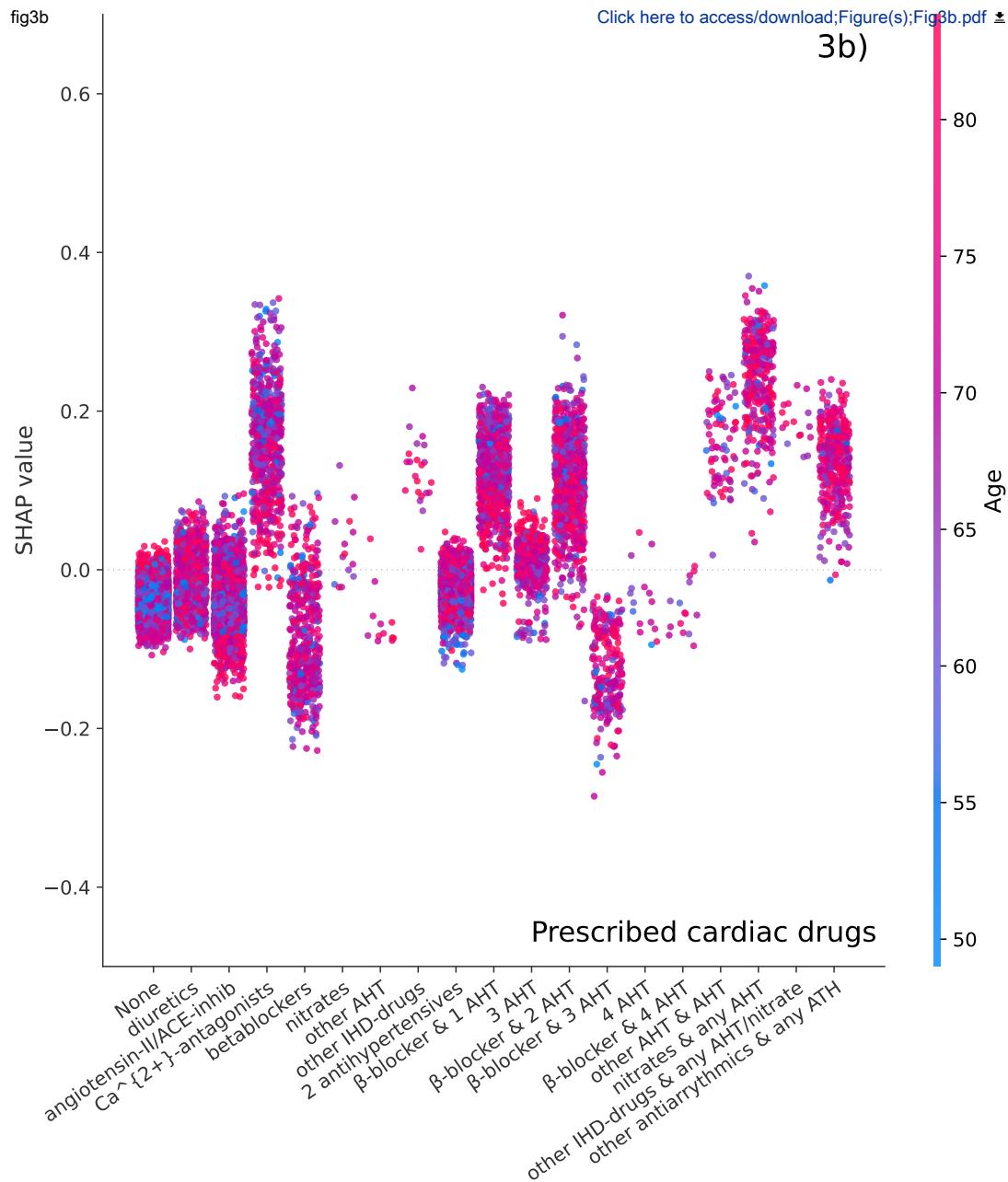
Positive prediction fraction 25%	TP	FP	FN	TN	sensitivity	precision	MCC	AUC	AUPRC	P (sensitivity)
Full machine-learning model	117	861	65	2870	64.3%	12.0%	20.0%	76.3%	15.5%	-
Full logistic regression model	110	868	72	2863	60.4%	11.2%	18.1%	74.5%	15.7%	23.1%
Parsimonious machine-learning model	115	863	67	2868	63.2%	11.8%	19.5%	75.9%	17.3%	41.2%
Parsimonious logistic regression model	106	872	76	2859	58.2%	10.8%	17.0%	73.4%	15.5%	11.8%
machine-learning model excluding age	106	872	76	2859	58.2%	10.8%	17.0%	72.3%	13.6%	11.8%
Age-model	94	824	88	2907	51.6%	10.2%	14.7%	69.7%	12.2%	0.7%
Positive prediction fraction 30%	TP	FP	FN	TN	sensitivity	precision	MCC	AUC	AUPRC	P (sensitivity)
Full machine-learning model	126	1047	56	2684	69.2%	10.7%	18.9%	76.3%	15.5%	-
Full logistic regression model	120	1053	62	2678	65.9%	10.2%	17.3%	74.5%	15.7%	25.2%
Parsimonious machine-learning model	124	1049	58	2682	68.1%	10.6%	18.4%	75.9%	17.3%	40.8%
Parsimonious logistic regression model	115	1058	67	2673	63.2%	9.8%	16.0%	73.7%	15.5%	11.1%
machine-learning model excluding age	116	1057	66	2674	63.7%	9.9%	16.3%	72.3%	13.6%	13.8%
Age-model	100	955	82	2776	54.9%	9.5%	13.9%	69.7%	12.2%	0.2%

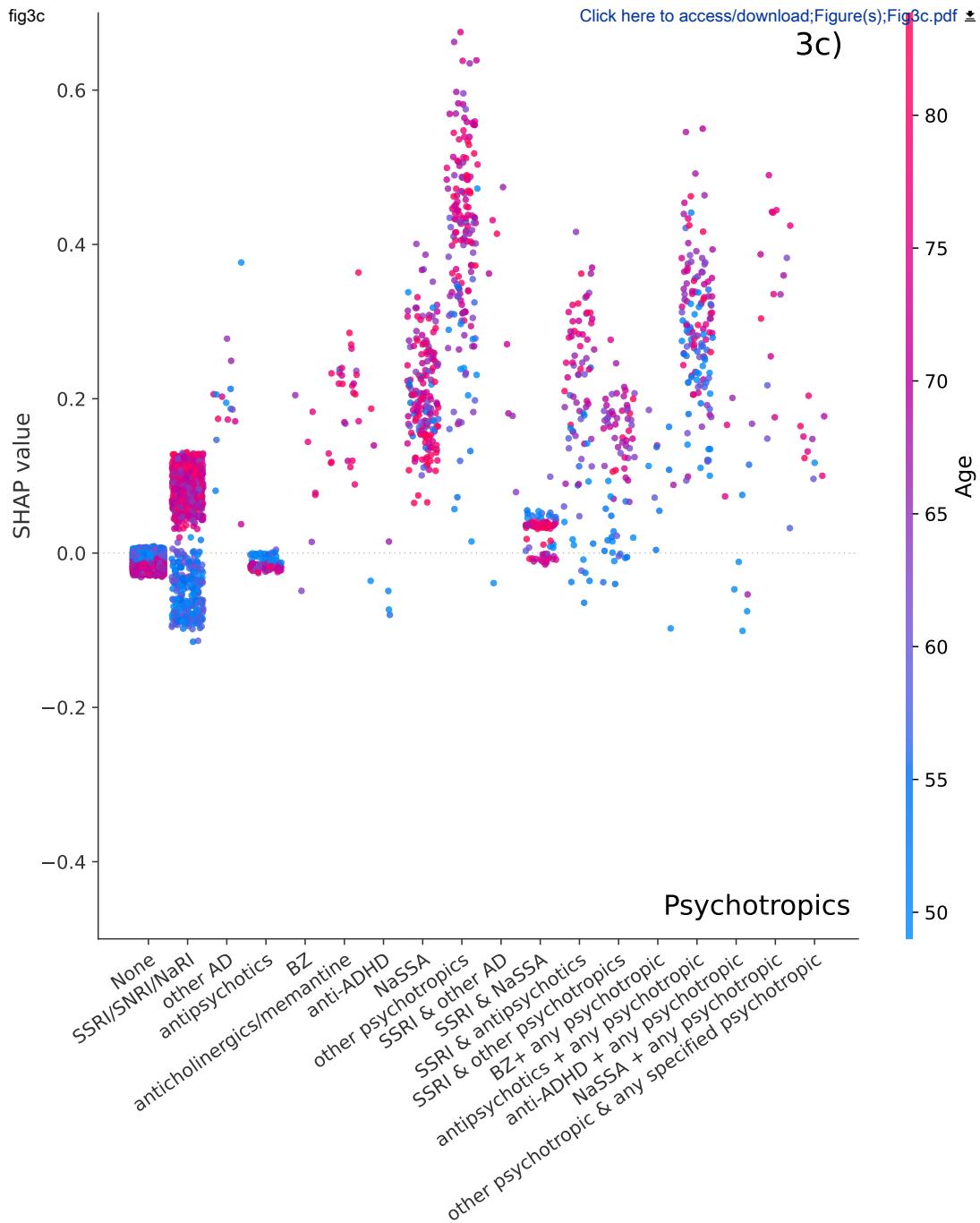
TP: true positives FP: false positives FN: false negatives TN: true negatives MCC: Matthews correlation coefficient AUC: area under the ROC curve AUPRC: area under the precision recall curve P(sensitivity): probability that the model performs better than the machine-learning model relative to sensitivity.

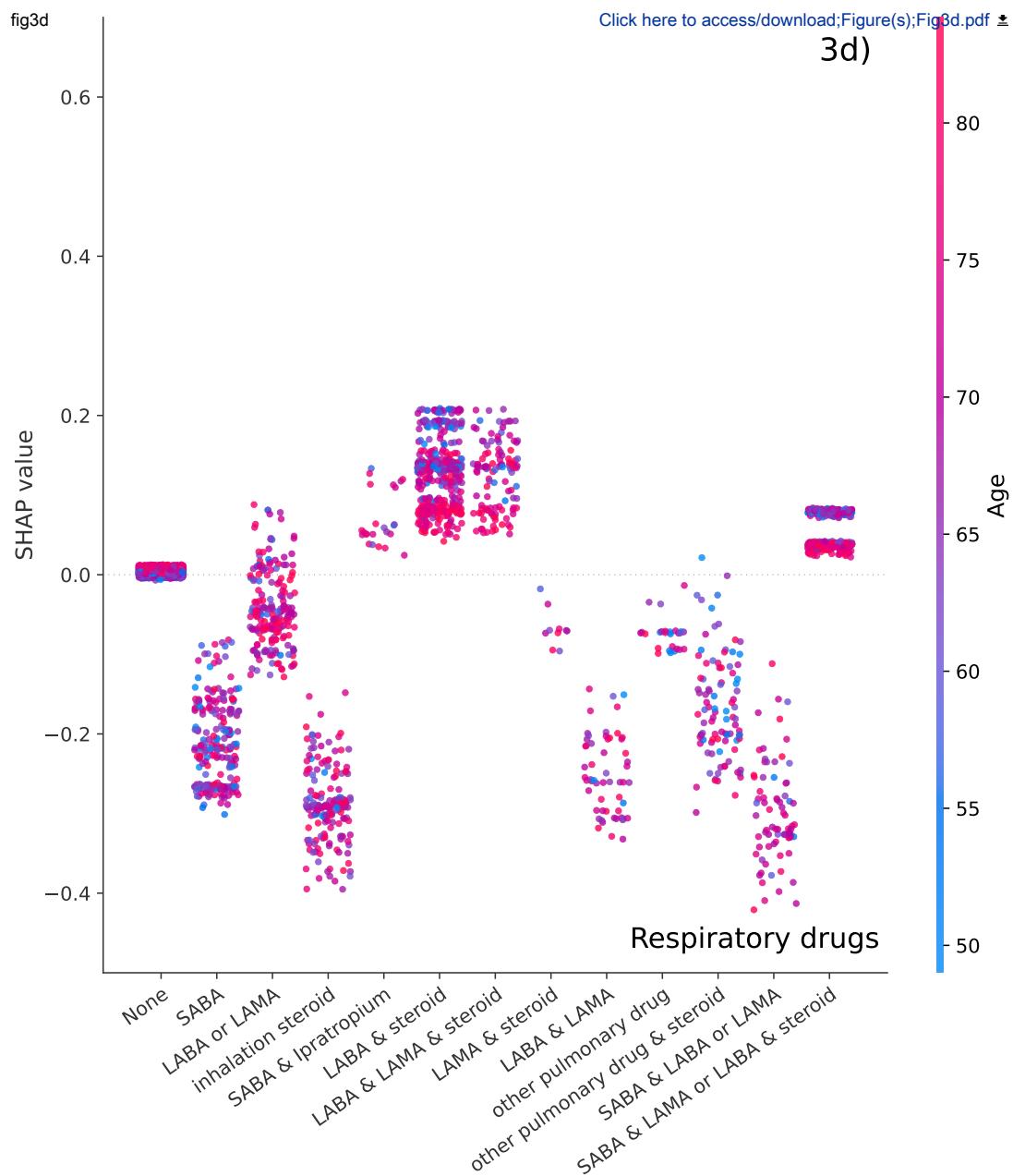












## Supplemental Digital Content 1

Table S1: Performance of different models for Outcome B (Los &gt;4 days or readmissions due to "medical" morbidity or LOS &gt;4 days but without recorded morbidity)

Positive prediction fraction 20%	TP	FP	FN	TN	sensitivity	precision	MCC	AUC	AUPRC	P (sensitivity)
Full machine-learning model	121	661	108	3023	52.8%	15.5%	20.5%	75.3%	17.1%	-
Full logistic regression model	115	667	114	3017	50.2%	14.7%	18.9%	74.1%	16.7%	28.3%
Parsimonious machine-learning model	111	671	118	3013	48.4%	14.2%	17.8%	74.4%	16.8%	17.2%
Parsimonious logistic regression model	109	673	120	3011	47.6%	13.9%	17.2%	73.1%	16.8%	12.9%
machine-learning model excluding age	110	672	119	3012	48.0%	14.1%	17.5%	72.8%	16.9%	15.1%
Age-model	102	661	127	3023	44.5%	13.4%	15.8%	68.7%	13.4%	3.8%
Positive prediction fraction 25%	TP	FP	FN	TN	sensitivity	precision	MCC	AUC	AUPRC	P (sensitivity)
Full machine-learning model	140	838	89	2846	61.1%	14.3%	20.8%	75.3%	17.1%	-
Full logistic regression model	136	842	93	2842	59.4%	13.9%	19.8%	74.1%	16.7%	35.3
Parsimonious machine-learning model	134	844	95	2840	58.5%	13.7%	19.3%	74.4%	16.8%	28.3
Parsimonious logistic regression model	125	853	104	2831	54.6%	12.8%	17.0%	73.1%	16.8%	7.8
machine-learning model excluding age	121	857	108	2827	52.8%	12.4%	16.0%	72.8%	16.9%	3.6
Age-model	113	805	116	2879	49.3%	12.3%	15.2%	68.7%	13.4%	0.5
Positive prediction fraction 30%	TP	FP	FN	TN	sensitivity	precision	MCC	AUC	AUPRC	P (sensitivity)
Full machine-learning model	153	1020	76	2664	66.8%	13.0%	20.0%	75.3%	17.1%	-
Full logistic regression model	147	1026	82	2658	64.2%	12.5%	18.6%	74.1%	16.7%	27.9

Parsimonious machine-learning model	147	1026	82	2658	64.2%	12.5%	18.6%	74.4%	16.8%	27.7
Parsimonious logistic regression model	145	1028	84	2656	63.3%	12.4%	18.1%	73.1%	16.8%	21.6
machine-learning model excluding age	140	1033	89	2651	61.1%	11.9%	17.0%	72.8%	16.9%	10.2
Age-model	122	933	107	2751	53.3%	11.6%	14.8%	69.8%	13.4%	0.1

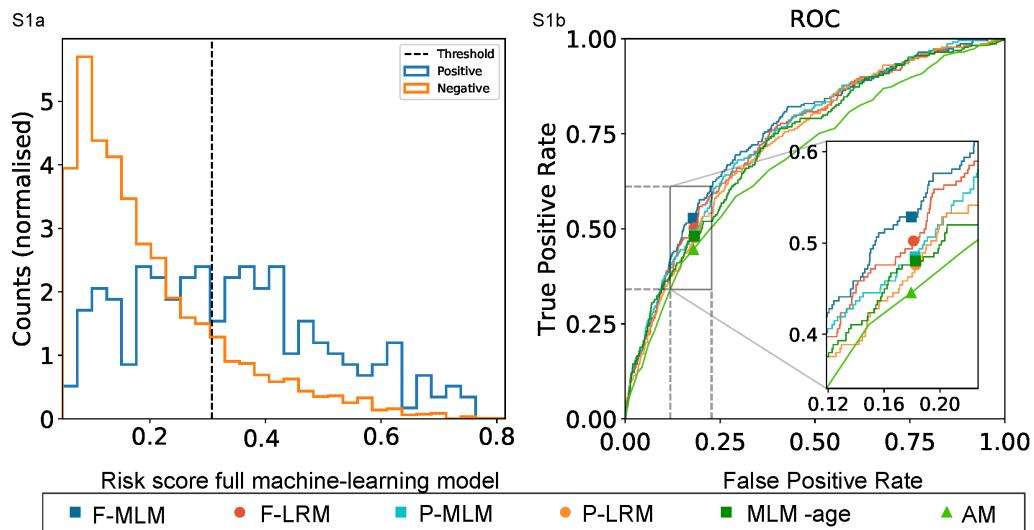
TP: true positives FP: false positives FN: false negatives TN: true negatives MCC: Matthews correlation coefficient AURC: area under the ROC curve AUPRC: area under the precision recall curve P(sensitivity): probability that the model performs better than the machine-learning model relative to sensitivity.

SDC2

[Click here to access/download Supplemental Digital Content SDC2.pdf](#)

## Supplemental Digital Content 2

Figure S1a-b



S1a) Distribution of full machine learning model risk scores for patients +/- outcome B(LOS >4 days or readmissions due to "medical" morbidity or LOS >4 days with no recorded morbidity). The dashed line marks the classification threshold of 20% positive prediction fraction.

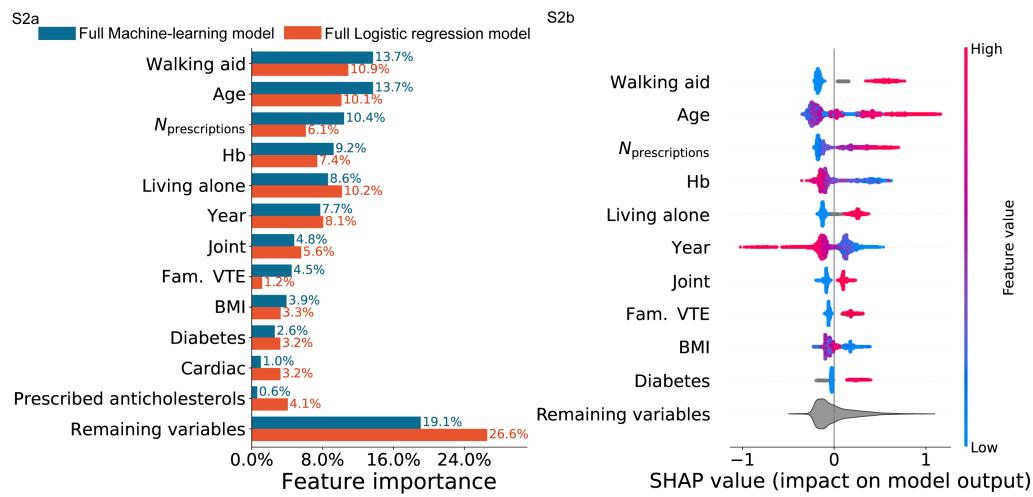
S1b) Receiver operating curves (ROC) for the full machine learning model (F-MLM), full logistic regression model (F-LRM), parsimonious machine learning model (P-MLM), parsimonious logistic regression model (P-LRM), machine learning excluding age (MLM -age) and the age-only model (AM).

SDC3

[Click here to access/download:Supplemental Digital Content;SDC3.pdf](#)

## Supplemental Digital Content 3

Figure S2a-b



S2a) The overall importance of the 10 most important variables measured by the SHAP-values for the full machine-learning and full logistic regression models for outcome B (LOS >4 days or readmissions due to "medical" morbidity or LOS >4 days with no recorded morbidity).

Only the importance of prescribed anti-coagulants and familiar disposition for venous thromboembolism differed between the models. The contributions of the remaining variables are summed in the bottom bar.

S2b) The SHAP-values for the full machine-learning model where values increase while negative values decrease the risk score. The color is related to the value of the variable with blue being lowest and red highest and each dot represents a patient.

## Supplemental Digital Content 4

### Figure S3a-d

SHAP scatter-plot on the contributions to the full machine-learning model on outcome B (LOS >4 days or readmission due to "medical" morbidity), for individual types of prescribed anticoagulants, cardiac drugs, psychotropics and respiratory drugs stratified by age.

#### Legend:

##### 3a) Prescribed anticoagulants

VKA: vitamin K antagonists ASA: acetylsalicylic acid DOAC: direct oral anticoagulant ADP: Adenosine diphosphate ACE: angiotensin converting enzyme

##### 3b) Prescribed cardiac drugs

ACE: angiotensin converting enzyme AHT: antihypertensive. Other AHT were defined as AHT different from diuretics ANG-II/ACE inhibitors or Ca<sup>2+</sup>-antagonists. IHD: Ischemic heart disease

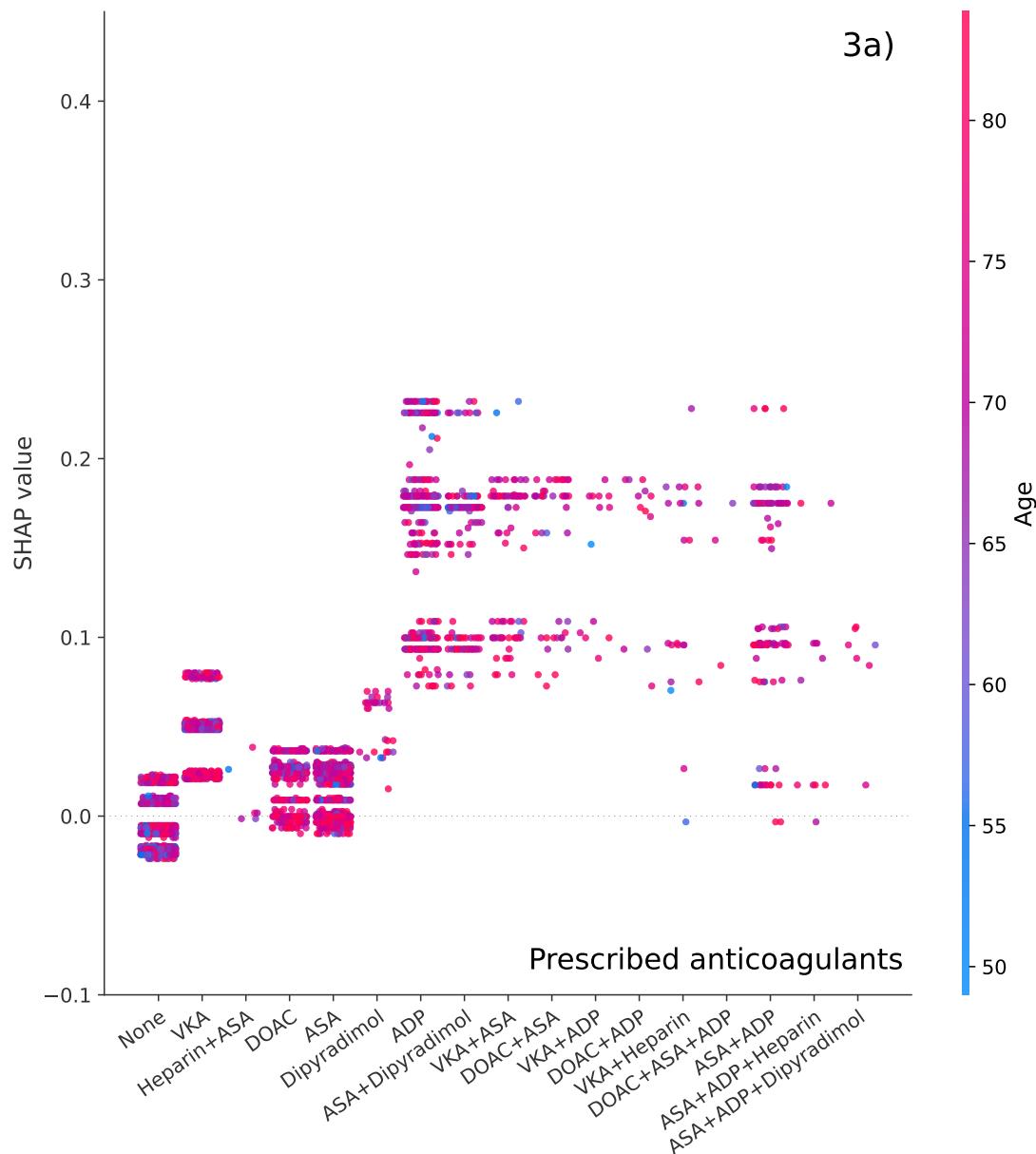
##### 3c) Prescribed psychotropics

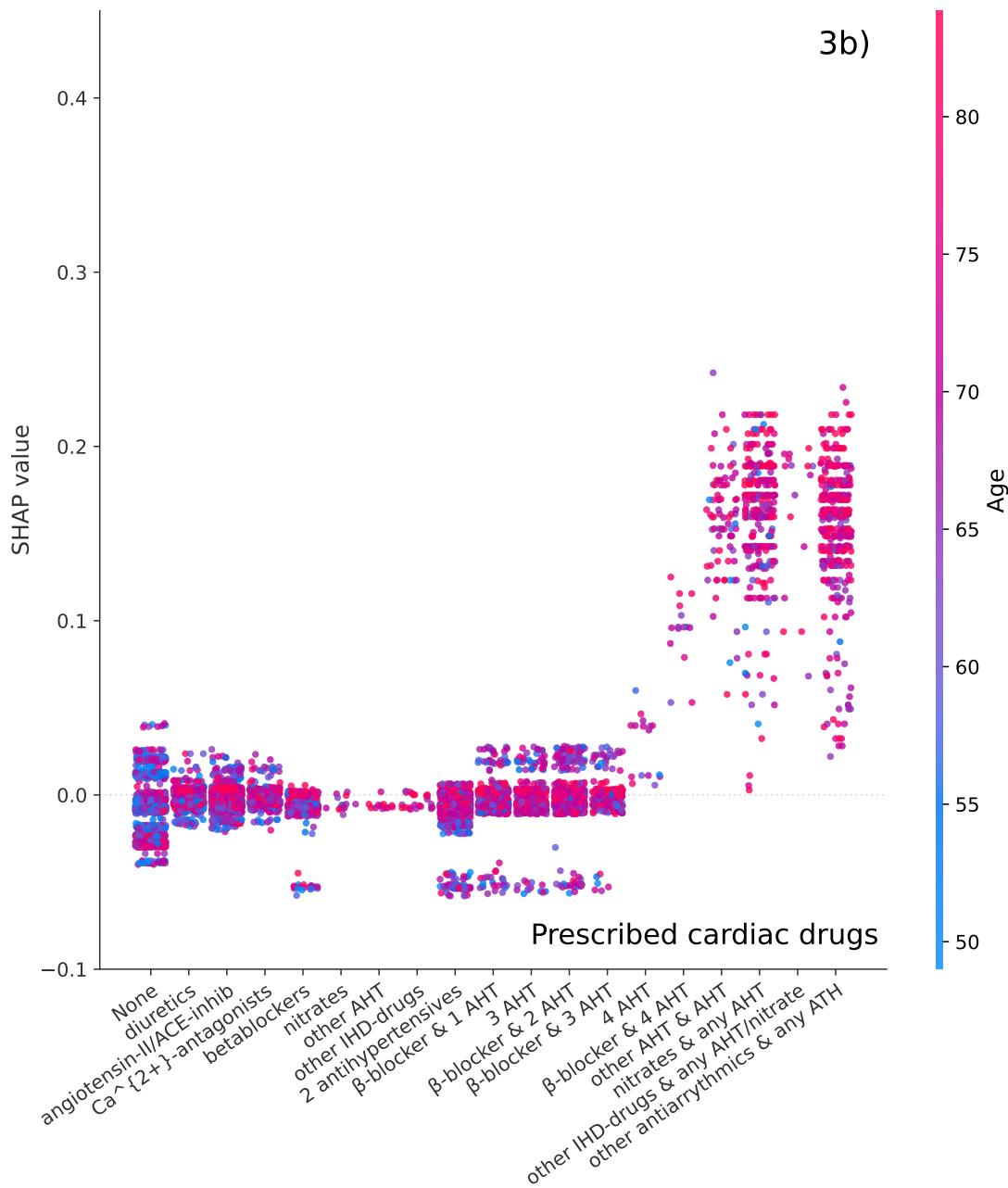
SSRI: Selective serotonin inhibitor SNRI: Serotonin and norepinephrine reuptake inhibitor NaRI: Norepinephrine reuptake inhibitor NaSSA: Norepinephrine and specific serotonergic antidepressants. AD: antidepressants BZ: Benzodiazepines (likely underreported due to limited general reimbursement in Denmark). ADHD: Attention-deficit/hyperactivity disorder

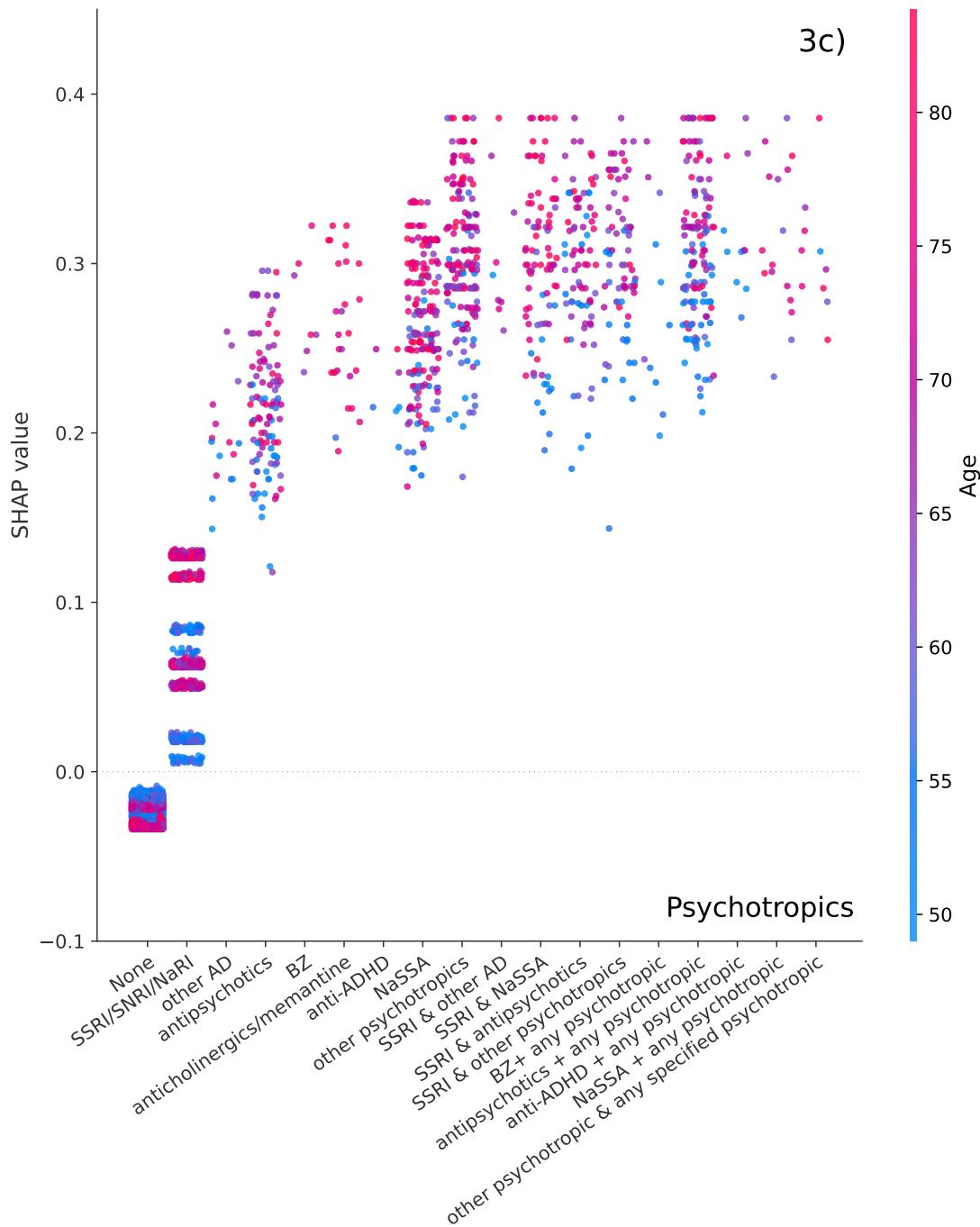
##### 3d) Prescribed respiratory drugs

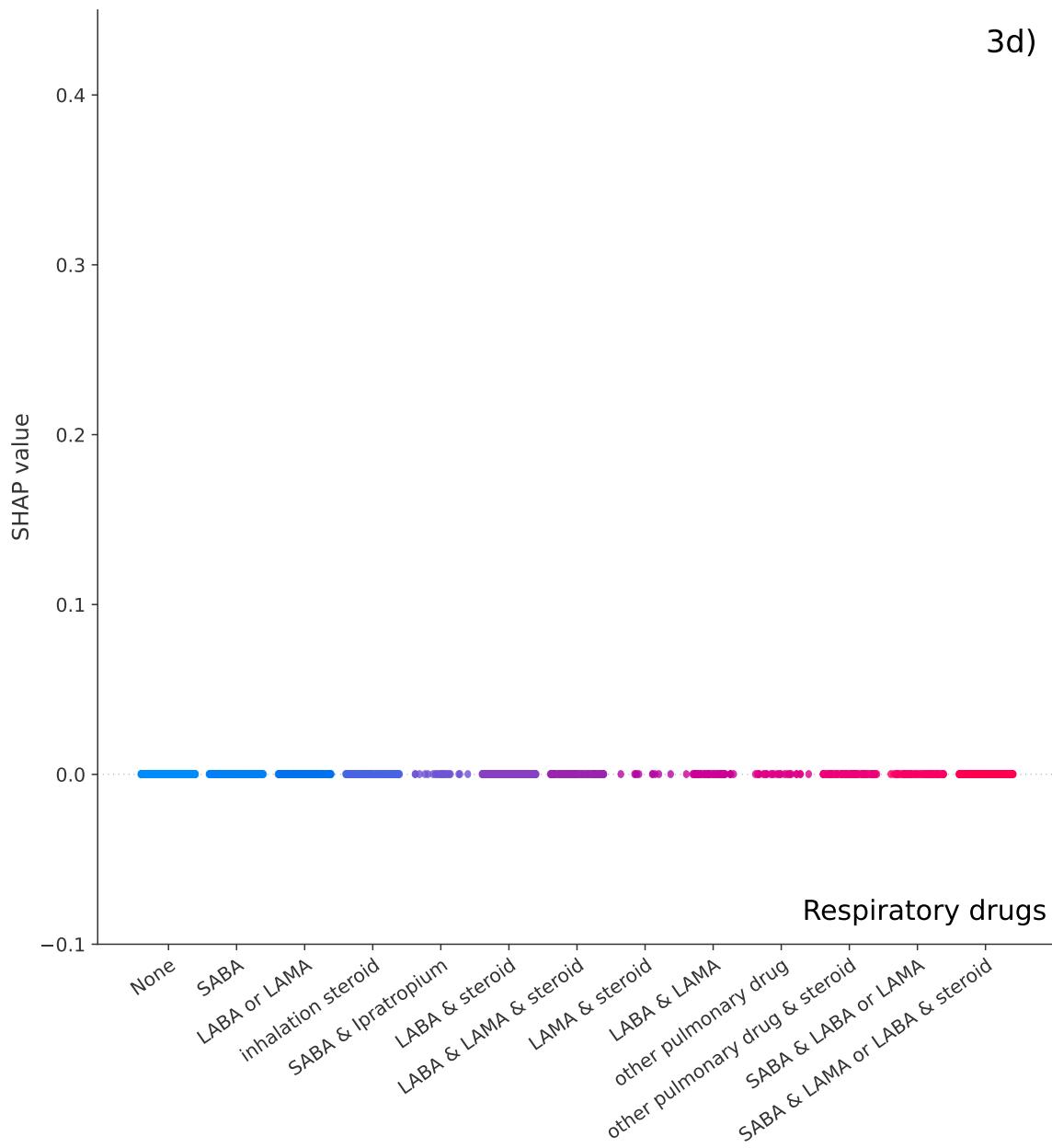
The model found no additional information from this variable why all values equal 0.

SABA: Short-acting beta agonist LABA: long-acting beta agonist LAMA: Long-acting muscarinic antagonist.











## **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](https://doi.org/10.1098/rsos.220018).

## ROYAL SOCIETY OPEN SCIENCE

[royalsocietypublishing.org/journal/rsos](http://royalsocietypublishing.org/journal/rsos)

Research



**Cite this article:** Heltberg ML, Michelsen C, Martiny ES, Christensen LE, Jensen MH, Halasa T, Petersen TC. 2022 Spatial heterogeneity affects predictions from early-curve fitting of pandemic outbreaks: a case study using population data from Denmark. *R. Soc. Open Sci.* **9**: 220018. <https://doi.org/10.1098/rsos.220018>

Received: 18 January 2022

Accepted: 16 August 2022

### Subject Category:

Mathematics

### Subject Areas:

mathematical modelling/biophysics/  
computational biology

### Keywords:

pandemics, agent-based modelling,  
spatial heterogeneity, fitting, COVID-19

### Author for correspondence:

Mathias L. Heltberg

e-mail: [heltberg@nbi.ku.dk](mailto:heltberg@nbi.ku.dk)

# Spatial heterogeneity affects predictions from early-curve fitting of pandemic outbreaks: a case study using population data from Denmark

Mathias L. Heltberg<sup>1,2,3,†</sup>, Christian Michelsen<sup>1,†</sup>,  
Emil S. Martiny<sup>1,†</sup>, Lasse Engbo Christensen<sup>4</sup>, Mogens  
H. Jensen<sup>1</sup>, Tariq Halasa<sup>5</sup> and Troels C. Petersen<sup>1</sup>

<sup>1</sup>Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, Copenhagen E 2100, Denmark

<sup>2</sup>Laboratoire de Physique, Ecole Normale Supérieure, Rue Lhomond 15, Paris 07505, France

<sup>3</sup>Infektionsberedskab, Statens Serum Institut, Artillerivej, Copenhagen S 2300, Denmark

<sup>4</sup>DTU Compute, Section for Dynamical Systems, Department of Applied Mathematics and

Computer Science, Technical University of Denmark, Anker Engelunds Vej 101A, Kongens Lyngby 2800, Denmark

<sup>5</sup>Animal Welfare and Disease Control, University of Copenhagen, Grønnegårdsvej 8,

Fredensborg C 1870, Denmark

MLH, 0000-0002-9699-4075; LEC, 0000-0001-5019-1931

The modelling of pandemics has become a critical aspect in modern society. Even though artificial intelligence can help the forecast, the implementation of ordinary differential equations which estimate the time development in the number of susceptible, (exposed), infected and recovered (SIR/SEIR) individuals is still important in order to understand the stage of the pandemic. These models are based on simplified assumptions which constitute approximations, but to what extent this are erroneous is not understood since many factors can affect the development. In this paper, we introduce an agent-based model including spatial clustering and heterogeneities in connectivity and infection strength. Based on Danish population data, we estimate how this impacts the early prediction of a pandemic and compare this to the long-term development. Our results show that early phase SEIR model predictions overestimate the peak number of infected and the equilibrium level by at least a factor of two. These results are robust to variations of parameters influencing connection distances and independent of the distribution of infection rates.

<sup>†</sup>These authors contributed equally.

## 1. Introduction

Over the past years, the pathogen now known as SARS-CoV-2 has spread dramatically, risen in several waves, paralyzing societies, resulting in a large number of deaths and severe economic damage worldwide [1,2]. Mathematical models have estimated the reproduction number and guided the authorities in an attempt to minimize the damage caused by the virus [3–6]. Even though modern algorithms using machine learning have helped the process [7,8], the majority of models used to predict the size of the pandemic (or a rising wave of the disease) have been variants of the SIR/SEIR model. The SIR model was originally proposed in 1927, in the seminal work of Kermack and McKendrick, who successfully described the evolution of a pandemic, using a mean field approximation where all individuals are described as one population [9]. In the investigations of the SARS-CoV-2 pandemic, the mathematical models have varied in complexity including simple deterministic compartmental models [6,10], meta-population compartmental models [11–13], individual based models without including spatial specifications [4,14,15] and spatio-temporal agent-based models [16].

One aspect in the modelling is the ability to predict the infection peak height and the number of individuals who will be infected based on the early rise in the number of infected (before governmental interference). Earlier work has pointed out the importance of including heterogeneity when modelling the spread of infectious disease such as contact patterns between individuals [17], population mixing assumptions [18], heterogeneities caused by super-spreaders [15], and the spatial dependency of COVID-19 [19,20]. These mathematical models have not combined heterogeneous elements nor quantified how much the early SIR/SEIR predictions might be biased.

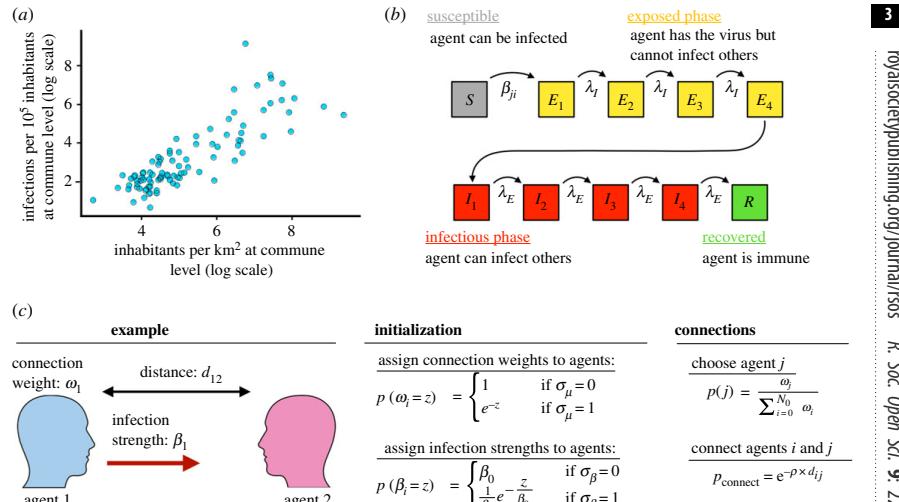
In this paper, we include geographical distributions based on an entire population, using population data of Denmark. When the SIR model was originally formulated, 95 years ago, data was not available to investigate the effects of geographical and demographic differences among the population, which might be one of the reasons why fundamental properties for diseases, such as the basic reproduction number ( $R_0$ ), can vary significantly between different regions [21]. However, with modern collection of data, these geographical aspects might be accounted for. Our main goal of this work is therefore to investigate the importance of heterogeneities in a geographically distributed population on the spread of a pandemic. We find that the heterogeneity arising from spatial inhomogeneities causes an increase in the early stage of the pandemic, affecting the initial forecast and highlighting the importance of early intervention in order to minimize the effects of the pandemic.

### 1.1. Construction of the model

In order to investigate the effect of a geographically distributed population, we extracted the number of infected per commune (from the Danish Serum Institute [22]) and divided this number with the number of inhabitants in each commune to obtain the number of infected per individual in each commune. This number we then plotted against the number of inhabitants in that specific commune (extracted from statistics Denmark [23]). Doing so, we found a strong correlation between the population density and the number of infections per inhabitant as seen in figure 1a. This observation has been made for many other countries [24–29] and underlines the aspect of disease spreading that has been observed since ancient times; that densely populated regions often have larger pandemics than the rural areas. Note that in the very early stage of a pandemic, before the exponential growth rate is reached, micro outbreaks will guide its evolution and these events can likely take place in regions with low density [30].

### 1.2. Disease simulation

To simulate evolution of the disease, we assigned each individual (agent) to a state (predominantly initialized in state  $S$ ) and assigned four states to the exposed phase and four states to the infectious phase, in order to achieve an Erlang distribution (which is related to the Gamma distribution) of time in each phase [31]. Once in the exposed phase, the infected agent has a rate to move into another state, where the rate is fixed based on experimental data in order to achieve a mean time in the exposed phase of approximately 4 days (table 1). Each agent in the Infectious phase can infect other agents that have a connection to this agent in the network. This definition of agents in discrete states is naturally a simplification of the real pandemic, and we stress that this mathematical model aims at describing the spread of the disease in a simple way that does not capture all aspects of the real disease. We do not believe that this impacts our main conclusions in any way, as we are aware that one should always be careful when making these kinds of simplifications. To investigate the effect of



**Figure 1.** (a) Population density ( $x$ -axis) and the number of infections per  $10^5$  inhabitants ( $y$ -axis) for each commune in Denmark. (b) Illustration of the modified susceptible-exposed-infected-removed (SEIR) model used. It consists of 10 consecutive states ( $S$ ,  $E_{1-4}$ ,  $I_{1-4}$  and  $R$ ), with transition rates governed by  $\beta$ ,  $\lambda_E$  and  $\lambda_I$ , respectively. (c) Illustration of how the spatial network is generated and heterogeneities in individuals included.

infection heterogeneities, we assigned an infection strength to each connection in the network, so some agents were more infectious than others. In order to control the degree of this heterogeneity, we assigned a boolean parameter  $\sigma_\beta$  that if switched on generated an exponential distribution in infection strengths, keeping the mean field reproduction number fixed. The reproduction number between the ABM and the SIR model is related through the parameter  $\bar{\beta} = \beta(\mu/2N_0)$ . All transitions between states and infection of other individuals were done using the Gillespie algorithm [32]. This is schematized in figure 1b.

### 1.3. Network creation

In order to construct the underlying network, we created a set-up whereby two agents were chosen at random but based on their individual connectivity weight each iteration and connected with some probability based on their spatial position. To include the possibility of highly connected individuals independent of their spatial position, we assigned a boolean parameter  $\sigma_\mu$  that, if switched on, generated an exponential distribution in weights for the individuals, keeping the mean field reproduction number fixed similar to the heterogeneity in infection strengths. To include the spatial position in the network, we introduced a parameter  $\rho$ , so the probability of connecting two chosen agents decayed exponentially with the distance between them:  $p_{\text{connect}} = e^{-\rho \times d_{ij}}$ . In order to allow some long-distance connections we introduced another parameter  $\varepsilon \in [0; 1]$ , that determines the fraction of distance-independent contacts. To construct the network of spatially distributed contacts, we chose the parameters using data based on:

- The geographical location of people in Denmark (from Boligsiden [33])
- The average number of contacts per individual per day of 11 (from HOPE [34]). Given an average infectious period of 4 days, we approximate the average number of effective contacts to be  $\mu = 40$
- The average commuting distance  $\rho = 0.1 \text{ km}^{-1}$  and the fraction of long-distance commutes  $\epsilon_\rho = 4\%$  (from statistics Denmark [23])

This is schematized in figure 1c and further described in the Methods section. All 10 parameters in this model are defined and outlined in table 1. We note that in order to keep the parameters space low, this model does not include the effects of temporal changes such as seasonality and holidays. While all agents

**Table 1.** Overview of the 10 parameters applied in this study, their typical value, and the ranges we have considered. The first six parameters are standard SEIR parameters, whereas the last four parameters define the heterogeneity in the model. These four parameters do not affect the SEIR model.

variable	description	value	range	units
$N_0$ :	population size	$5.8 \times 10^6$	$10^5 - 10^7$	—
$N_{\text{init}}$ :	number of individuals initially infected	100	$1 - 10^4$	—
$\mu$ :	average number of network contacts	40	10–100	—
$\beta$ :	typical infection strength	0.01	0.001–0.1	$\text{d}^{-1}$
$\lambda_E$ :	rate to move through $\frac{1}{4}$ of latency period	1	0.5–4	$\text{d}^{-1}$
$\lambda_I$ :	rate to move through $\frac{1}{4}$ of infectious period	1	0.5–4	$\text{d}^{-1}$
$\sigma_\mu$ :	population clustering spread	0	0–1	—
$\sigma_\beta$ :	interaction strength spread	0	0–1	—
$\rho$ :	typical acceptance distance	0.1	0–0.5	$\text{km}^{-1}$
$\epsilon_p$ :	fraction of distance-independent contacts	0.04	0–1	—

have been assigned parameters to their infection network that are derived from statistics of Denmark for both employees and students, we have not divided each agent into specific occupations.

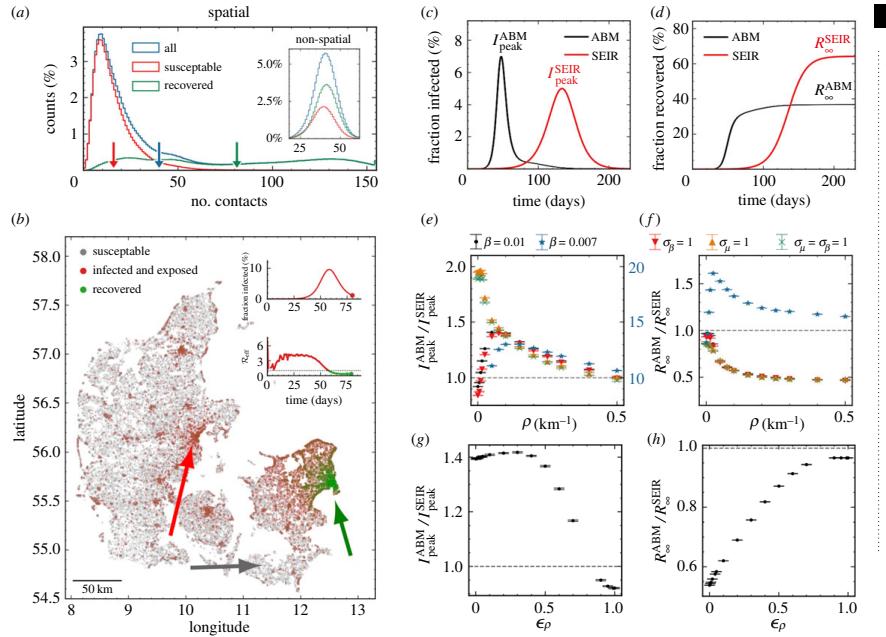
Before including heterogeneity, we compared the ABM to the corresponding SEIR model as a test, and found them to agree within 5% for all parameter configurations tested. Here, we also tested the effect of the number of individuals initially infected (see electronic supplementary material). This concludes that the SEIR and ABM model are calibrated to have the same reproduction number in the absence of heterogeneities. Next, we will introduce heterogeneities into the system, while keeping the sum of contacts and infection strengths constant, to study how this affects the evolution of the pandemic.

## 2. Results

### 2.1. Geographical distributions in a population and large variances in numbers of contacts leads to increased infection levels

Having introduced heterogeneity, the distributions of connections in this network were created automatically through the population clustering, see figure 2a. This naturally leads to individuals living in densely populated areas having higher number of connections. In an example simulation with 100 initially infected individuals,  $N_{\text{init}} = 100$ , we observed a spatial difference in areas affected by the disease (figure 2b), as expected. Note that we also show the effective reproduction number ( $\mathcal{R}_{\text{eff}}$ ) as a function of time in the lower part of the inserted panel. One region reached local endemic steady state (green arrow, figure 2b) while other regions of similar density were highly infected (red arrow, figure 2b) and yet other districts were almost unaffected (grey arrow, figure 2b). To quantify the effect of population clustering, we compared the ABM result to the reference SEIR model of similar parameters. Generally, we observed that the epidemic developed faster with a higher infection peak  $I_{\text{peak}}$  but also subsided quicker, leading to a lower number of infected once reaching endemic steady state,  $R_\infty$  (figure 2c,d).

In order to explore how population clustering affects the epidemic, we chose a reference value of infection rates,  $\beta = 0.01$ , and an alternative value of  $\beta = 0.007$ . In the absence of spatial dependence ( $\rho = 0 \text{ km}^{-1}$ ), these correspond to initial reproduction numbers  $\mathcal{R}_0 \approx 1.7$  and 1.1, respectively. Here, we define the reproduction number as the average number of agents each infectious agent will infect in the first part of the disease. Increasing the spatial dependence (i.e. increasing  $\rho$ ) leads to a significant rise in the infection peak for the ABM,  $I_{\text{peak}}^{\text{ABM}}$ , compared to the (unaffected) SEIR model,  $I_{\text{peak}}^{\text{SEIR}}$  for both the reference value and the alternative lower value of  $\beta$  (black and blue points, figure 2e). We introduced heterogeneity in infection strengths ( $\sigma_\beta = 1$ , see figure 1b), thus making some individuals much more infectious than others (i.e. including *super shedders*). We found no significant impact from this effect (red points in figure 2e). Similarly, we introduced heterogeneity in connection weights ( $\sigma_\mu = 1$ , see figure 1b), thus making some individuals much more likely to form contacts than others



**Figure 2.** (a) Histograms showing the number of susceptible (red) and recovered (green) individuals at the end of an epidemic with  $\rho = 0.1 \text{ km}^{-1}$ . The distribution before the epidemic is shown in blue. The arrows show the mean of each distribution. The inset shows the same for  $\rho = 0 \text{ km}^{-1}$ . (b) Visualization of the spatial position of individuals during the infection and which state they are in. Green arrow: largest city in Denmark (Copenhagen); mostly recovered. Red arrow: Second largest city in Denmark (Aarhus); mostly infected. Grey arrow: low-population area; mostly susceptible (i.e. have not been infected). (c) Number of infected individuals as a function of time. Data shown for the spatially distributed network ( $\rho = 0.1 \text{ km}^{-1}$ ). Simulation was repeated 10 times. (d) Cumulative sum of individuals who have had the disease as a function of time (with  $\rho = 0.1 \text{ km}^{-1}$ ). (e) Relative difference in maximal number of infected,  $I_{\text{peak}}$ , between deterministic (SEIR) and ABM as a function of  $\rho$ , and shown for different parameters. Note the data for  $\beta = 0.007$  are shown in blue with a factor 10 scaling (right y-axis). (f) Relative difference in total number of infected at the end of the epidemic,  $R_\infty$ , between deterministic (SEIR) and ABM as a function of  $\rho$ . Colours similar to (e). (g) Same as (e), but as a function of  $\epsilon_\rho$ . (h) Same as (f), but as a function of  $\epsilon_\rho$ .

(i.e. including *super connecters*). This leads to a significant effect for  $\rho = 0 \text{ km}^{-1}$ , which converges towards the other curves for  $\rho > 0.1 \text{ km}^{-1}$  (orange (only super connecters) and green (super connecters and super shedders) points in figure 2e). The total number of individuals that have been in the infectious state, when there are not enough susceptible agents for the disease to keep infecting new individuals, is termed  $R_\infty$ , and this converged towards half of the SEIR model prediction as a function of  $\rho$  except for  $\beta = 0.007$  where the endemic steady state level is larger than the one obtained by the SEIR model (figure 2f). We note that in reality, individuals can lose immunity and therefore new waves can emerge. But for a completely susceptible population,  $R_\infty$  describes the fraction of the population that will get the disease during a specific wave. Fixing  $\rho = 0.1 \text{ km}^{-1}$  and increasing the fraction of distance-independent contacts,  $\epsilon_\rho$ , we found that  $I_{\text{peak}}^{\text{ABM}}$  is almost unaffected for  $\epsilon_\rho < 0.5$  (figure 2g), while  $R_\infty^{\text{ABM}}$  increases linearly towards the SIER model  $R_\infty^{\text{SEIR}}$ , as expected (figure 2h).

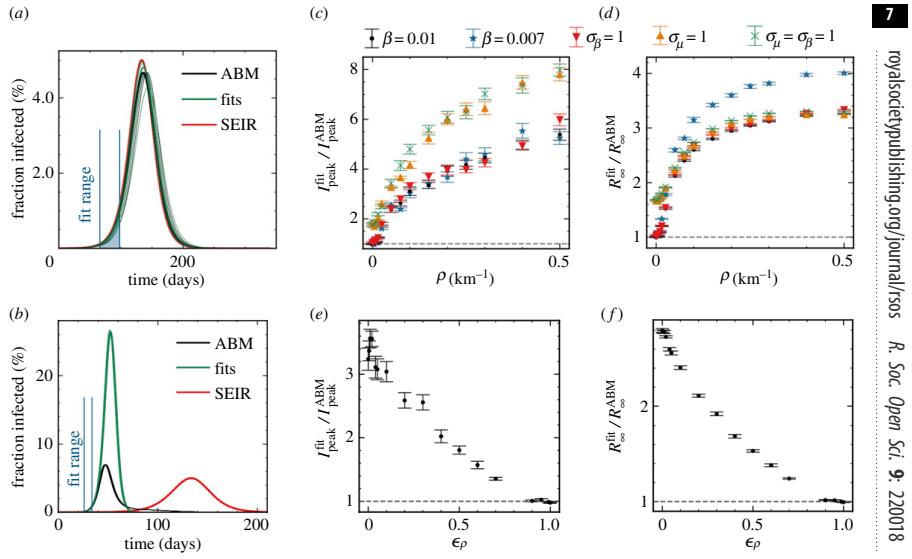
## 2.2. Fitting early infection curves leads to significant bias in estimating the size of the pandemic

Next, we consider how these heterogeneities bias the traditional SEIR model predictions, especially the predictions based on fits to the number of infected (i.e. the curve to be flattened) in the beginning of the epidemic (see Methods). Without spatial dependence, the predicted curves fitted the number of infected

individuals very well (figure 3a). Introducing spatial dependence ( $\rho = 0.1 \text{ km}^{-1}$ ) leads to a severe overestimation of the epidemic based on the number of early infection cases (figure 3b). This result can be interpreted by the fact that in societies where population density and thus individual contact number varies significantly, the early phase will be driven by people with many contacts (*super connectors*). This typically happens in cities where the population density is high. Increasing the spatial dependence  $\rho$ , we found that the SEIR model predictions overestimated the infection peak height  $I_{\text{peak}}$  and the total number of infected  $R_{\infty}$  significantly even for very small spatial heterogeneities (figure 3c, d). We observed this general trend for all tested combinations of parameters and heterogeneities. In particular, we found that if long-distance connections  $\epsilon_{\rho}$  are below 10%, the bias in the estimated infection peak height,  $I_{\text{peak}}$ , was constant within statistical uncertainty (figure 3e). For the total number of infected,  $R_{\infty}$ , we observed an almost linear regression to the SEIR model as  $\epsilon_{\rho}$  approaches one. However, even when  $\epsilon_{\rho} = 0.25$ , the prediction bias was still a factor of two (figure 3f). We concluded from these curves a general trend; if one fits an SEIR model to infection numbers during the beginning of an epidemic, and use these estimates to predict the characteristics of the epidemic at a national level, one overestimates the number of infected by at least a factor of two.

### 3. Discussion

In summary, this work outlines that the degree of population clustering in Denmark creates a discrepancy between the early predictions made by the SEIR models and the underlying agent-based interactions. It results in a significant overestimation of the impact of the disease, both in terms of maximal number of simultaneously infected (by a factor of 3) and the endemic steady state level (by a factor of 2.5). Such discrepancies have been observed for earlier pandemics, for instance, the 1918 Spanish flu, where the predicted number of individuals that would get the disease within a season turned out to be higher than the actual outcome [35]. The present results can be an important element in explaining these mismatches, even though other elements, such as for instance social distancing and the population behaviour, play a vital part. When facing a rising pandemic, societies are faced with the task of laying out strategies to minimize the consequences, including the importance of *flattening the curve*. While this is truly crucial to avoid overpopulated hospitals, the understanding of the pandemic should be taken seriously enough that we might specify to a higher degree of certainty which curve to be flattened. Our results highlight an important element in the prediction of infection levels and quantify the effect of density heterogeneities. We are aware that these results are not directly applicable to the pandemic of SARS-CoV-2 as a whole, since numerous mutations have increased the infection rates compared to the early estimates and created a strong heterogeneity in the infection worldwide. Furthermore, the actual evolution of the pandemic was highly affected by the different governmental interventions, that are not included in this work. However, this study emphasizes the abnormally large reproduction rates in the beginning of a pandemic, due to individuals with more connections than the rest of the population and attempts to quantify this bias, when countries should estimate the severity of a disease based on the data collected in the early phase. This also underlines the benefits by making lockdowns early in the pandemic, when a population is highly susceptible (for instance to a new mutation) and therefore can be driven by *super connectors*. Since people living in city-clusters are more likely to have many contacts, or infection events, they are on average more likely to be affected in the early stage of the pandemic (if they do not implement social distancing). By removing contacts from these individuals, through some level of interaction in order to reduce the number of social contacts, one can avoid the worst peak while affecting the lowest number of people. While our work describes some fundamental aspects of the disease spreading, this model does not consider asymptomatic individuals, which has been an important aspect of the SARS-CoV-2 pandemic [36,37]. Effectively, asymptotic individuals would correspond to a very heterogeneous distribution of time the agents spend in the infectious state. While agents with symptoms would predominantly isolate themselves and thereby significantly reduce their ability to infect other agents, asymptomatic agents would have a long time in the infectious state, thereby infecting more individuals. In this work, we have not considered the observation that individuals lose their immunity to SARS-CoV-2 which was first studied in the Brazilian city of Manaus. For this model, the temporal decline of immunity would lead to more pandemic ‘waves’, but for a fixed disease transmissibility this would not alter the maximal height of the peak number of infected, since this occurred for all the initially susceptible population. Finally, we note that this work does not include a vast range of divisions for the population, including age, socio-economic status etc. We have not included this directly, since we wanted to estimate as cleanly as possible how the heterogeneity in the



**Figure 3.** (a) Number of infected individuals for the ABM in black, the SEIR model in red and the SEIR fits to the ABM data in green. Blue lines show the interval where parameters are fitted (also shown below the curves). Here,  $\rho = 0 \text{ km}^{-1}$ . (b) Same as (a) but with population clustering ( $\rho = 0.1 \text{ km}^{-1}$ ). (c) Relative difference in maximal number of infected,  $I_{\text{peak}}$ , between the fit and the ABM for different values of  $\rho$ . Simulations repeated 10 times for each data-point. (d) Relative difference in total number of infected at the end of the epidemic,  $R_{\infty}$ , between the fit and the ABM for different values of  $\rho$ . (e) Same as (c), but as a function of  $\epsilon_{\rho}$ . (f) Same as (d), but as a function of  $\epsilon_{\rho}$ .

contact pattern, arising from a geographically distributed population, could affect the evolution of a disease. We are aware that for instance the distribution of age has an enormous impact on the health risk and that this risk is vital in the prediction of hospitalizations in modern society. However, our aim was to understand the bias in the prediction of a disease, based on the data that comes during the early periods of a disease, independently of the mortality of this disease. Mathematical predictions of disease progression have been heavily criticized [38,39] and it is important to improve the theoretical foundations of the mathematical descriptions, in order to increase the confidence in the predictions. Our work highlights the importance of estimating the spatial clustering and connectivity skewness in the population in order to correct the predictions based on SEIR models, by quantifying their biases from not including spatial clustering. We hope that this work could serve as an input to the modelling and prediction of future pandemics and the importance of avoiding super-spreaders in high-density areas.

### 3.1. Methods

#### 3.1.1. Construction of spatial network

We initialized  $N_0$  agents on a network generating a total of  $\mu \times N_0$  links between two agents, with an assigned interaction strength  $\beta_{ij}$  for each link. The average contact number,  $\mu$ , was fixed to 20, based on results from the Danish HOPE project, gathering data on population behaviour since April 2020 [34]. In order to include a realistic, geographical distribution of the population, we randomly selected agent locations from a two-dimensional kernel density estimate we had generated based on housing sales in Denmark 2007–2019 (data given with permission from Boligsiden, [33]). We note that in this distribution, we do not take specific geographical elements such as roads or environment into account (which has been previously studied for other diseases [40]) as we assume that this effect is small in a country like Denmark, where all parts are connected and natural obstacles such as mountains and rivers are not present. To connect the agents, we used a hit and miss method, where two random agents are first suggested and then connected with probability,  $p(d) = e^{-\rho d_{ij}}$ . Here,  $d_{ij}$  is the distance between agents and

$\rho$  is a parameter with units of inverse distance. We choose  $\rho = 0.1 \text{ km}^{-1}$  (i.e. 10 km) which is the average distance travelled by labour force (statistics Denmark [23]). To allow some long-distance interactions, we introduced a parameter  $\epsilon_\rho = 4\%$  representing the fraction of distance-independent connections. This value was based on the fraction of workers travelling longer than 50 km to work (statistics Denmark [23]).

### 3.1.2. Fits and predictions

We defined an early phase to be the period of time when between 0.1% and 1% of the population were infected (blue lines figure 3a). We then fitted  $\beta$  and a time delay,  $\tau$ , to the SEIR model with a  $\chi^2$ -fit (assuming Poissonian statistics) and kept  $\lambda_E$  and  $\lambda_I$  fixed to the true numbers (used in the simulation). The initial number of infected,  $N_{\text{init}}$ , was also fixed to the true numbers. The fit parameters were then inserted into the SEIR model, and  $I_{\text{peak}}^{\text{fit}}$  and  $R_{\infty}^{\text{fit}}$  were extracted from the fitted model and compared to the  $I_{\text{peak}}^{\text{ABM}}$  and  $R_{\infty}^{\text{ABM}}$  from the ABM simulation.

Data accessibility. Data and relevant code for this research work are stored in GitHub: [www.github.com/ChristianMichelsen/NetworkSIR](https://www.github.com/ChristianMichelsen/NetworkSIR) and have been archived within the Zenodo repository: <https://zenodo.org/badge/latestdoi/258223118>.

Authors' contributions. C.M.: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; E.S.M.: investigation, software, validation, visualization, writing—review and editing; L.E.C.: supervision, validation, writing—review and editing; T.C.P.: conceptualization, investigation, methodology, project administration, software, supervision, validation, visualization, writing—original draft, writing—review and editing; M.L.H.: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; M.H.J.: formal analysis, investigation, supervision, validation, writing—review and editing; T.H.: conceptualization, investigation, supervision, validation, visualization, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein. Conflict of interest declaration. We declare that we have no competing interests.

Funding. M.L.H. acknowledges the Carlsberg Foundation grant no. CF20-0621 and the Lundbeck Foundation grant no. R347-2020-2250. E.S.M. and M.H.J. acknowledge support from the Independent Research Fund Denmark grant no. 9040-00116B and Danish National Research Foundation through StemPhys Center of Excellence, grant no. DNRF116. Acknowledgements. The authors are grateful to the Danish expert group of SARS-CoV-2 modelling led by Statens Serum Institute, especially Robert L. Skov, Kåre Mølbak, Camilla Holten Møller, Viggo Andreasen, Kaare Græsbøl, Theis Lange, Carsten Kirkeby, Frederik P. Lyngse, Matt Denwood, Jonas Juul, Sune Lehman, Uffe Thygesen and Laust Hvas Mortensen. Furthermore, we thank Kim Sneppen for valuable discussions.

## References

- Chinazzi M *et al.* 2020 The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400. ([doi:10.1126/science.abb9757](https://doi.org/10.1126/science.abb9757))
- WHO: see [www.who.int/news-room/detail/27-04-2020-who-timeline-covid-19](http://www.who.int/news-room/detail/27-04-2020-who-timeline-covid-19) (accessed 29 September 2020).
- Anderson RM, Heesterbeek H, Klinkenberg D, Hollingsworth TD. 2020 How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet* **395**, 931–934. ([doi:10.1016/S0140-6736\(20\)30567-5](https://doi.org/10.1016/S0140-6736(20)30567-5))
- Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, Flasche S. 2020 Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Global Health* **8**, e488–e496. ([doi:10.1016/S2214-109X\(20\)30074-7](https://doi.org/10.1016/S2214-109X(20)30074-7))
- Keeling MJ, Hollingsworth TD, Read JM. 2020 Efficacy of contact tracing for the containment of the 2019 novel coronavirus (COVID-19). *J. Epidemiol. Commun. Health* **74**, 861–866. ([doi:10.1101/2020.02.14.20023036](https://doi.org/10.1101/2020.02.14.20023036))
- Kuniya T. 2020 Prediction of the epidemic peak of coronavirus disease in Japan, 2020. *J. Clin. Med.* **9**, 789. ([doi:10.3390/jcm9030789](https://doi.org/10.3390/jcm9030789))
- Ghafouri-Fard S, Mohammad-Rahimi H, Motie P, Minabi MA, Taheri M, Nateghinia S. 2021 Application of machine learning in the prediction of COVID-19 daily new cases: a scoping review. *Helyon* **7**, e08143. ([doi:10.1016/j.helyon.2021.e08143](https://doi.org/10.1016/j.helyon.2021.e08143))
- Fokas AS, Dikaios N, Kastis GA. 2020 Mathematical models and deep learning for predicting the number of individuals reported to be infected with SARS-CoV-2. *J. R. Soc. Interface* **17**, 20200494. ([doi:10.1098/rsif.2020.0494](https://doi.org/10.1098/rsif.2020.0494))
- Kermack WO, McKendrick AG. 1927 A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A* **115**, 700–721. ([doi:10.1098/rspa.1927.0118](https://doi.org/10.1098/rspa.1927.0118))
- Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J. 2020 Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493. ([doi:10.1126/science.abb3221](https://doi.org/10.1126/science.abb3221))
- Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, Abbott S. 2020 The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *Lancet Public Health* **5**, e261–e270. ([doi:10.1016/S2468-2667\(20\)30073-6](https://doi.org/10.1016/S2468-2667(20)30073-6))
- van Bunnik BA, Morgan AL, Bessell P, Calder-Gerke G, Zhang F, Haynes S, Lepper HC. 2020 Segmentation and shielding of the most vulnerable members of the population as elements of an exit strategy from COVID-19 lockdown. *medRxiv* ([doi:10.1101/2020.05.04.20090597](https://doi.org/10.1101/2020.05.04.20090597))
- Danon L, Brooks-Pollock E, Bailey M, Keeling MJ. 2020 A spatial model of COVID-19 transmission in England and Wales: early spread and peak timing. *medRxiv* ([doi:10.1101/2020.02.12.20022566](https://doi.org/10.1101/2020.02.12.20022566))
- Chang SL, Harding N, Zachreson C, Cliff OM, Prokopenko M. 2020 Modelling transmission and control of the COVID-19 pandemic in Australia. *Nat. Commun.* **11**, 1–13. ([doi:10.1038/s41467-020-19393-6](https://doi.org/10.1038/s41467-020-19393-6))
- Sneppen K, Nielsen BF, Taylor RJ, Simonsen L. 2021 Overdispersion in COVID-19 increases the effectiveness of limiting nonreplicative contacts for transmission control. *Proc. Natl. Acad. Sci. USA* **118**, e2016623118. ([doi:10.1073/pnas.2016623118](https://doi.org/10.1073/pnas.2016623118))
- Milne GJ, Xie S. 2020 The effectiveness of social distancing in mitigating COVID-19 spread: a

- modelling analysis. *medRxiv* (doi:10.1101/2020.03.20.20040055).
17. Bansal S, Grenfell BT, Meyers LA. 2007 When individual behaviour matters: homogeneous and network models in epidemiology. *J. R. Soc. Interface* **4**, 879–891. (doi:10.1098/rsif.2007.1100)
  18. Kong L, Wang J, Han W, Cao Z. 2016 Modeling heterogeneity in direct infectious disease transmission in a compartmental model. *Int. J. Environ. Res. Public Health* **13**, 253. (doi:10.3390/ijerph13030253)
  19. Kang D, Choi H, Kim JH, Choi J. 2020 Spatial epidemic dynamics of the COVID-19 outbreak in China. *Int. J. Infect. Dis.* **94**, 96–102. (doi:10.1016/j.ijid.2020.03.076)
  20. Giuliani D, Dickson MM, Espa G, Santi F. 2020 Modelling and predicting the spatio-temporal spread of COVID-19 in Italy. *BMC Infect. Dis.* **20**, 1–10. (doi:10.1186/s12879-020-05415-7)
  21. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. 2019 Complexity of the basic reproduction number ( $R_0$ ). *Emerg. Infect. Dis.* **25**, 1–4. (doi:10.3201/eid2501.171901)
  22. Danish Serum Institute: www.ssi.dk (accessed 29 September 2020).
  23. Statistics Denmark: www.statistikbanken.dk (accessed 29 September 2020).
  24. Wong DW, Li Y. 2020 Spreading of COVID-19: density matters. *PLoS ONE* **15**, e0242398. (doi:10.1371/journal.pone.0242398)
  25. Ganasegaran K, Jamil MFA, Chng ASH, Looi I, Peariasamy KM. 2021 Influence of population density for COVID-19 spread in Malaysia: an ecological study. *Int. J. Environ. Res. Public Health* **18**, 9866. (doi:10.3390/ijerph18189866)
  26. Kodera S, Rashed EA, Hirata A. 2020 Correlation between COVID-19 morbidity and mortality rates in Japan and local population density, temperature, and absolute humidity. *Int. J. Environ. Res. Public Health* **17**, 5477. (doi:10.3390/ijerph1715477)
  27. Bhadra A, Mukherjee A, Sarkar K. 2021 Impact of population density on COVID-19 infected and mortality rate in India. *Model. Earth Syst. Environ.* **7**, 623–629. (doi:10.1007/s40808-020-00984-7)
  28. Chen K, Li Z. 2020 The spread rate of SARS-CoV-2 is strongly associated with population density. *J. Travel Med.* **27**, taaa186. (doi:10.1093/jtm/taaa186)
  29. Martins-Filho PR. 2021 Relationship between population density and COVID-19 incidence and mortality estimates: a county-level analysis. *J. Infect. Public Health* **14**, 1087–1088. (doi:10.1016/j.jiph.2021.06.018)
  30. Hittner JB, Fasina FO, Hoogesteyn AL, Piccinini R, Maciorowski D, Kempala P, Rivas AL. 2021 Testing-related and geo-demographic indicators strongly predict COVID-19 deaths in the united states during March of 2020. *Biomed. Environ. Sci.* **34**, 734–738.
  31. Huang S, Li J, Dal C, Tie Z, Xu J, Xiong X, Lu C. 2021 Incubation period of coronavirus disease 2019: new implications for intervention and control. *Int. J. Environ. Health Res.* **32**, 1707–1715. (doi:10.1080/09603123.2021.1905781)
  32. Gillespie DT. 1977 Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361. (doi:10.1021/j100540a008)
  33. Boligsiden: www.boligsiden.dk (accessed 29 September 2020).
  34. HOPE project: www.hope-project.dk (accessed 29 September 2020).
  35. Andreasen V, Viboud C, Simonsen L. 2008 Epidemiologic characterization of the 1918 influenza pandemic summer wave in Copenhagen: implications for pandemic control strategies. *J. Infect. Dis.* **197**, 270–278. (doi:10.1086/524065)
  36. Arcede JP, Caga-Anan RL, Mentuda CQ, Mammeri Y. 2020 Accounting for symptomatic and asymptomatic in a SEIR-type model of COVID-19. *Math. Model. Nat. Phenomena* **15**, 34. (doi:10.1051/mmnp/2020021)
  37. Guan J, Zhao Y, Wei Y, Shen S, You D, Zhang R, Chen F. 2022 Transmission dynamics model and the coronavirus disease 2019 epidemic: applications and challenges. *Med. Rev.* **2**, 89–109. (doi:10.1515/mr-2021-0022)
  38. Holmdahl I, Buckee C. 2020 Wrong but useful—what COVID-19 epidemiologic models can and cannot tell us. *N. Engl. J. Med.* **383**, 303–305. (doi:10.1056/NEJMmp2016822)
  39. Wynants L, Van Calster B, Bonten MM, Collins GS, Debray TP, De Vos M, Schuit E. 2020 Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal. *BMJ* **369**, m1328. (doi:10.1136/bmj.m1328)
  40. Rivas AL, Fasina FO, Hoogesteyn AL, Konah SN, Febles JL, Perkins DJ, Smith SD. 2012 Connecting network properties of rapidly disseminating epizootics. *PLoS ONE* **7**, e39778. (doi:10.1371/journal.pone.0039778)

## **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”.

# Microscopic single molecule

## dynamics suggest underlying physical properties of the silencing foci

Susmita Sridar<sup>1</sup> , Mathias Spliid Heltberg<sup>2</sup> , Christian Stentoft

Michelsen<sup>2</sup> , Judith Mine Hattab<sup>1</sup>, Angela Taddei<sup>1</sup>

<sup>1</sup>Institut Curie, PSL University, Sorbonne Universite, CNRS, Nuclear Dynamics, Paris,

<sup>2</sup> France; <sup>2</sup>Niels Bohr Institute, University of Copenhagen

For correspondence:

[mathias.heltberg@nbi.ku.dk](mailto:mathias.heltberg@nbi.ku.dk)

(MH)

<sup>†</sup>Authors contributed equally.

### 10 Abstract

**Present address:** Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark

**Data availability:** Data availability is available on Zenodo or the [Github](#) repository.

**Funding:** This work was supported by XXX Foundation. The funders had no role in the decision to publish.

**Competing interests:** The author declare no competing interests.

In order to obtain fine-tuned regulation of protein production while maintaining cell integrity, it is of fundamental importance to living organisms to express a specific subset of the genes available in the genome. One way to achieve this is through the formation of subcompartments in the nucleus, known as foci, that can form at various locations on the DNA fibers and repress the transcriptional activity of all genes covered. In this work we investigate the physical nature of such foci, by applying single molecule microscopy in living cells. Here we study the motion of the protein SIR3. By combining various statistical methods, and combining a frequentist with a bayesian approach, we extract the diffusion properties for motion in a repair foci. In order to obtain useful information based on this, we derive similar measures for the foci itself, the motion of SIR3 outside the foci and other mutants of the cell. We reveal that the behaviour inside a repair foci is highly immobile and we compare this to theoretical expressions. Based on this we hypothesize that the repair foci is probably not a result of a second order liquid-liquid phase separation but rather a so-called Polymer Bridgng Model with numerous binding sites.

24

## 1 | INTRODUCTION

26 Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo.  
 27 Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan  
 28 bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit  
 29 mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus  
 30 et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcor-  
 31 per vestibulum turpis. Pellentesque cursus luctus mauris. Nulla malesuada porttitor diam. Donec  
 32 felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Do-  
 33 nec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac  
 34 quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ip-  
 35 sum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc  
 36 eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna.  
 37 Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam  
 38 cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis  
 39 eu massa.

## 40 2 | METHODS & MATERIALS

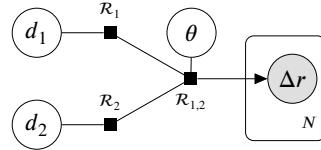
### 2.1 | Diffusion model

42 For each of the different types of data (XXX), we load in the cells and group them by cell number  
 43 and ID. For each group we compute the distance  $\Delta r$  between the subsequent observations  $\vec{x}_i$ :

$$44 \quad \Delta r_i = \|\vec{x}_{i+1} - \vec{x}_i\|. \quad (1)$$

E.g., for Wild Type 1, we find 914 groups across 43 different cells, leading to a total of  $N = 10.025$   
 45 distances. We model the diffusion distances with a Rayleigh likelihood, where the Rayleigh distri-  
 46 bution is given by:

$$48 \quad \text{Rayleigh}(x; \sigma) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}, \quad x > 0. \quad (2)$$



**Figure 1.** A graphical representation of the Bayesian model case of two diffusion components using the directed factor graph notation (Dietz, 2022). Here  $d_1$  is the diffusion coefficient,  $\mathcal{R}_1$  is the  $d$ -parameterized Rayleigh distribution and  $\mathcal{R}_{1,2}$  is the mixture model of the Rayleigh distributions with a  $\theta$  prior.

In this study, we parameterize the Rayleigh distribution in terms of the diffusion coefficient  $d$ , which  
is related to the scale parameter  $\sigma$  in eq. (2), through the XXX parameter,  $\tau$ :

$$\sigma = \sqrt{2d\tau}, \quad (3)$$

with  $\tau = 0.02$  in the current study. In the simplest form, where we assume only a single diffusion coefficient,  $d$ , the Bayesian model for this process is:

54 [d prior]	$d \sim \text{Exponential}(0.1)$
55 [transformation]	$\sigma = \sqrt{2d\tau}$
56 [likelihood]	$\Delta r_i \sim \text{Rayleigh}(\sigma).$

58 A more realistic diffusion model include more than a single diffusion coefficient. Figure 1 shows  
this for the two-component case in directed factor graph notation (Dietz, 2022). In particular, the  
60 figure shows the combination of the  $K = 2$  diffusion coefficients  $d_k$  through a mixture model  $\mathcal{R}_{1,2}$  of  
the two  $d$ -parameterized Rayleigh distributions  $\mathcal{R}_k$  with a  $\theta$ -prior. We model each of the distances  
62 as independent, indicated by the  $N$ -replications plate. In equations, the figure is similar to:

63 [d1 prior]	$d_1 \sim \text{Exponential}(0.1)$
64 [d2 prior (ordered)]	$d_2 \sim \text{Exponential}(0.1), \quad d_1 < d_2$
65 [θ prior]	$\theta_1 \sim \text{Uniform}(0, 1), \quad \bar{\theta} = [\theta_1, 1 - \theta_1]$
66 [mixture model]	$\mathcal{R}_{1,2}(d_1, d_2, \bar{\theta}) = \text{MixtureModel}([\mathcal{R}(d_1), \mathcal{R}(d_2)], \bar{\theta})$
67 [likelihood]	$\Delta r_i \sim \mathcal{R}_{1,2}(d_1, d_2, \bar{\theta}).$

## 2.2 | Model comparison

70 We can generalize the  $K = 2$  diffusion model to higher values of  $K$  by having  $d_1, \dots, d_K$  (ordered  
such that  $d_1 < d_k < d_K$  to prevent the classical label-switching problem in the case of mixture mod-

72    els (McLachlan and Peel, 2004)) diffusion coefficients and letting the mixture model's  $\bar{\theta}$ -prior be a  
 73    random variable from a flat Dirichlet distribution (such that  $\sum_k \theta_k = 1$ ). We find that including up  
 74    to three diffusion coefficients yields appropriate results. To compare the three models of differ-  
 75    ent complexity, we compute the Widely Applicable Information Criterion (WAIC) (Watanabe, 2010)  
 76    which is a generalized version of the Akaike information criterion (AIC) useful for Bayesian model  
 77    comparison (Gelman, Hwang, and Vehtari, 2014). In short, the WAIC is an approximation of the  
 78    out-of-sample performance of the model and consists of two terms, the log-pointwise-predictive-  
 79    density, lppd, and the effective number of parameters  $p_{\text{WAIC}}$ :

$$80 \quad \text{WAIC} = -2 (\text{lppd} - p_{\text{WAIC}}). \quad (6)$$

The lppd is the Bayesian version of the accuracy of the model and  $p_{\text{WAIC}}$  is a penalty term related to  
 81    the risk of over-fitting; complex models (usually) have higher values of  $p_{\text{WAIC}}$  than simple models,  
 82    (McElreath, 2020). The minus 2 factor is just a scaling included for historical reasons leading to low  
 83    WAICs being better. Given two models, A and B, we compute both the individual WAIC values,  $W_A$   
 84    and  $W_B$ , their standard deviations,  $\sigma_{W_A}$  and  $\sigma_{W_B}$ , their difference,  $\Delta_{A,B}$ , and the standard error of  
 85    their difference,  $\sigma_{\Delta_{A,B}}$ .

### 2.3 | MSD and energy

86    After choosing the optimal model, we extract the slow diffusion coefficient from the model,  $d_{\text{slow}}$ ,  
 87    and use this to compute the mean squared displacement (MSD) for the groups with a mean diffu-  
 88    sion  $D = \langle \frac{\Delta r^2}{4t} \rangle$  being slow, where slow is defined as  $D < d_{\text{slow}} + 3\sigma_{\text{slow}}$ . From the MSD, we can either  
 89    infer the full XXX (Mathias) model, based on XXX equation:  
 90    or simply approximate the DCon2\_WT1 (XXX) with half of the slope of the first three data points of

91    the MSD (Mathias, why half?).  
 92    We can compute the energy,  $U$ , in two different ways;  $U_{\text{left}}$  and  $U_{\text{right}}$ . The first method is based  
 93    on a geometric calculation depending on the fraction of the slow diffusion coefficient from the Wild

Type 1 calculation,  $\theta_{\text{slow}}^{\text{WT1}} \equiv \theta_1^{\text{WT1}}$ :

$$\begin{aligned}
 V_{\text{cap}} &= \frac{\pi h^2}{3(3r_0 - h)} \\
 V_0 &= \frac{4\pi}{3 - 2V_{\text{cap}}} \\
 V_F &= \frac{8V_0}{4\pi/3} \frac{4\pi}{3R_R^3} \\
 U_{\text{left}} &= -\log \left( \theta_{\text{slow}} \frac{V_0 - V_F}{(1 - \theta_{\text{slow}}^{\text{WT1}})V_F} \right),
 \end{aligned} \tag{8}$$

where  $r_0 = 1.0$ ,  $h = 0.85$ , and  $R_R = 0.13$ .

The other energy,  $U_{\text{right}}$ , can be calculated from the value of DCon2\_WT1 (half slope) from Wild Type 1, the Db\_focus (half slope) from the Focus files, and the fast diffusion coefficient from the delta files:  $\theta_{\text{fast}}^{\text{delta}} \equiv \theta_2^{\text{delta}}$ :

$$U_{\text{left}} = \log \left( \frac{\text{DCon2\_WT1} - \text{Db\_focus}}{\theta_{\text{fast}}^{\text{delta}} - \text{Db\_focus}} \right). \tag{9}$$

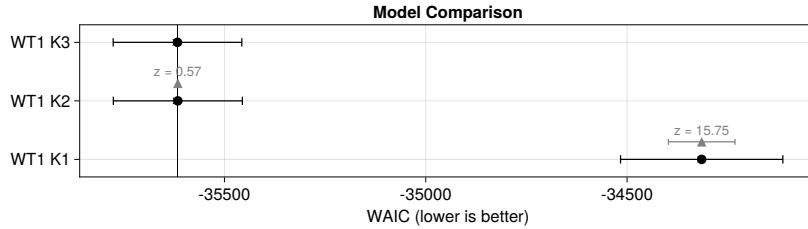
## 2.4 | Implementation

The data analysis has been carried out in Julia (Bezanson et al., 2017) and the Bayesian models are computed using the Turing.jl package (Ge, Xu, and Ghahramani, 2018). We use Hamiltonian Monte Carlo sampling (Betancourt, 2018) with the NUTS algorithm (Hoffman and Gelman, 2011). In particular, each Bayesian model have been run with 4 chains, each chain 1000 iterations long after discarding the initial 1000 samples (“warm up”).

## 3 | RESULTS

### 3.1 | Dynamics of SIR3 reveals two dominating populations of the motion

We started out by investigating the mobility of individual SIR3 molecules in vivo. Here we typically have 5-8 repair foci. To image SIR3 without changing its normal level, we generated haploid cells expressing the endogenous SIR3 fused to Halo (Figure 1A). Before we wanted to visualize the cells on a PALM microscope (see Materials and methods), we incubated the exponentially growing cells with fluorescent and fluorogenic JF647, a dye emitting light only once bound to SIR3. We were very used a low concentration of JF646 allowing for the observation of individual molecules (Ranjan et al., 2020; Figure 1— figure supplement 2). Rad52-Halo bound to JF646 (Rad52-Halo/JF646) were visualized at 20 ms time intervals (50 Hz) in 2-dimensions during 1000 frames until no signal was visible.



**Figure 2.** Comparison between diffusion models with  $K = 1$ ,  $K = 2$ , or  $K = 3$  diffusion coefficients for the Wild Type 1 data (WT1). The x-axis shows the WAIC score, where lower values indicate higher-performing models. The WAIC-score for each model is shown in black along with its uncertainty. The difference in WAIC-scores between the model and the best performing model (WT1 K3) is shown in grey with  $z$  being the number of standard deviations between them.

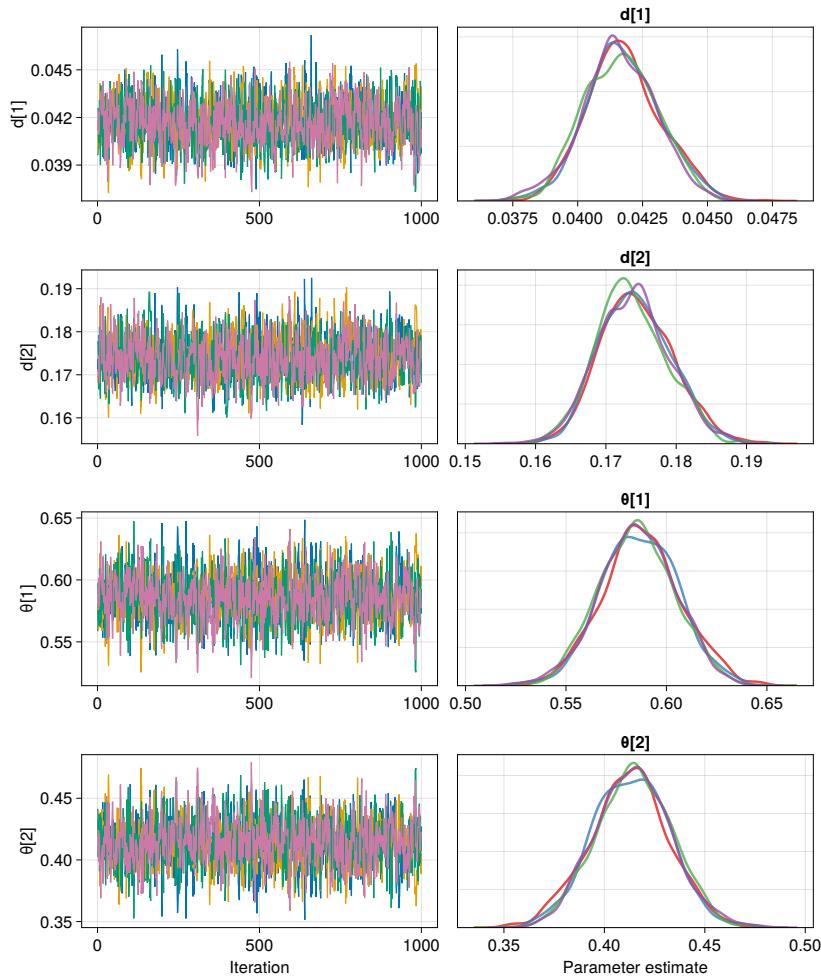
### 3.2 | Bayesian Analysis

122 Comparing the three diffusion models with 1, 2, or 3 diffusion coefficients, respectively, we find that  
 123 the model with only a single diffusion component is simply not advanced enough to fully explain  
 124 the data, see Figure 2. This figure shows that, even though the 3 component model is the best-  
 125 performing of the models, when judging by the number of standard deviations,  $z$ , that the best  
 126 model's WAIC is higher than the second best model's WAIC, it is statistically non-significant ( $z < 2$ ).  
 127 Since the performance of both the 2 and 3 component models are indistinguishable, we follow  
 128 Occam's razor and continue with the former model.

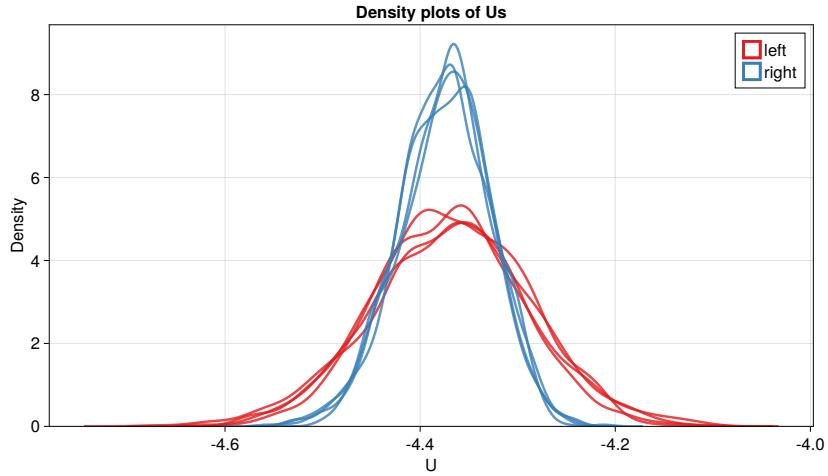
Bayesian models allow for far greater flexibility than traditional frequentist models, including  
 129 internal validation checks and diagnostic criteria to make sure that the model has not converged.  
 130 In particular, we made sure that all  $\hat{R}$ -values were less than 1.01. To fully validate the  $K = 2$  model,  
 131 we show the traceplots and posterior distributions for the different parameters in Figure 3. The  
 132 left part of the figure shows the parameter estimate as a function of MCMC iteration, i.e. traceplot,  
 133 which, for correctly sampled chains, should resemble a fuzzy caterpillar (and not a skyline which  
 134 would indicate bad mixing) (Roy, 2020). We find that the slow diffusion coefficient for WT1 data is:  
 135  $\theta_{\text{slow}}^{\text{WT1}} = 0.0417 \pm 0.0014 \text{ XXXunit}$ .

XXX Energy computation shows that, see Figure 4.

136 Final results:



**Figure 3.** Results of the  $K = 2$  diffusion model. Left) Traceplots. Right) Density plots.



**Figure 4.** Density plot of the energy,  $U$ , using either the left or the right computation approach (XXX Mathias). The energy computed using the left computation is shown in red, and in blue with the right computation. The four different MCMC chains (for each approach) are shown as individual lines.

$$\text{DCon1\_WT1} = 0.010\,29 \pm 0.001\,83$$

$$\text{DCon2\_WT1} = 0.017\,334\,3 \pm 0.000\,000\,4$$

$$\text{Db\_focus} = 0.006\,629\,0 \pm 0.000\,000\,1$$

$$U_{\text{left}} = -4.373 \pm 0.078$$

$$U_{\text{right}} = -4.375 \pm 0.047$$

$$\text{Din\_hyper\_WT} = 0.031\,51 \pm 0.002\,71 \quad (10)$$

$$\text{DCon1\_hyper\_WT} = 0.008\,14 \pm 0.002\,34$$

$$\text{DCon2\_hyper\_WT} = 0.015\,391\,1 \pm 0.000\,001\,8$$

$$\text{Db\_hyper\_focus} = 0.001\,315\,7$$

$$U_{\text{left-hyper}} = -4.364 \pm 0.204$$

$$U_{\text{right-hyper}} = -4.108 \pm 0.047$$

## <sup>140</sup> 4 | DISCUSSION

<sup>140</sup> Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac,  
<sup>142</sup> adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, conse-  
tuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique  
<sup>144</sup> senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus  
sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida  
<sup>146</sup> placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo  
ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla.  
<sup>148</sup> Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan  
eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.  
<sup>150</sup> Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo.  
Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan  
<sup>152</sup> bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit  
mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus  
<sup>154</sup> et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcor-  
per vestibulum turpis. Pellentesque cursus luctus mauris.  
<sup>156</sup> Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique,  
libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing  
<sup>158</sup> semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie  
nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi  
<sup>160</sup> blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellen-  
tesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec  
<sup>162</sup> bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu  
enim. Vestibulum pellentesque felis eu massa.

### <sup>164</sup> 4.1 | Acknowledgment

Acknowledgements here

### <sup>166</sup> 4.2 | Data availability

Source code is hosted at GitHub: <https://github.com/ChristianMichelsen/diffusion>.

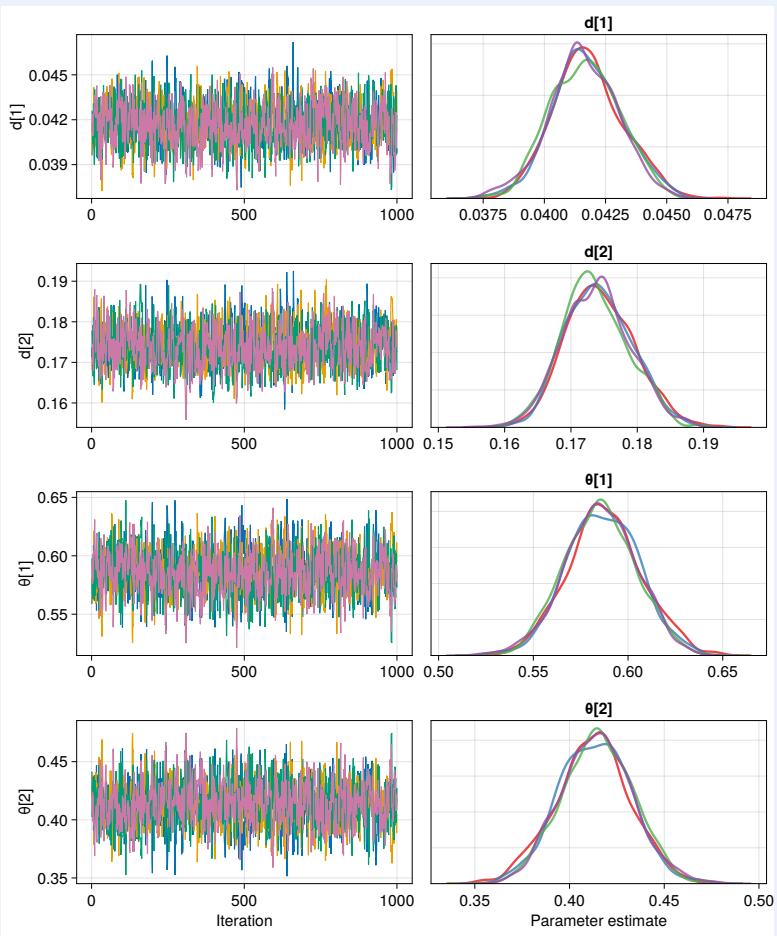
## REFERENCES

- <sup>168</sup> Betancourt, Michael (2018). "A Conceptual Introduction to Hamiltonian Monte Carlo". In: *arXiv:1701.02434 [stat]*. arXiv: 1701.02434.
- <sup>170</sup> Bezanson, Jeff et al. (2017). "Julia: A fresh approach to numerical computing". In: *SIAM review* 59.1. Publisher: SIAM, pp. 65–98. URL: <https://julialang.org/>.
- Dietz, Laura (2022). "Directed factor graph notation for generative models". In: Ge, Hong, Kai Xu, and Zoubin Ghahramani (2018). "Turing: A Language for Flexible Probabilistic Inference". en. In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. ISSN: 2640-3498. PMLR, pp. 1682–1690. URL: <https://proceedings.mlr.press/v84/ge18b.html> (visited on 2022).
- <sup>174</sup> Gelman, Andrew, Jessica Hwang, and Aki Vehtari (2014). "Understanding predictive information criteria for Bayesian models". en. In: *Statistics and Computing* 24.6, pp. 997–1016. ISSN: 1573-180 DOI: [10.1007/s11222-013-9416-2](https://doi.org/10.1007/s11222-013-9416-2). URL: <https://doi.org/10.1007/s11222-013-9416-2> (visited on 2022).
- <sup>182</sup> Hoffman, Matthew D. and Andrew Gelman (2011). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *arXiv:1111.4246 [cs, stat]*. arXiv: 1111.4246.
- <sup>184</sup> 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.
- McLachlan, Geoffrey J. and David Peel (2004). *Finite Mixture Models*. en. Google-Books-ID: c2\_fAx0DQoC. John Wiley & Sons. ISBN: 978-0-471-65406-3.
- Roy, Vivekananda (2020). "Convergence Diagnostics for Markov Chain Monte Carlo". In: *Annual Review of Statistics and Its Application* 7.1. \_eprint: <https://doi.org/10.1146/annurev-statistics-031219-041300>, pp. 387–412. DOI: [10.1146/annurev-statistics-031219-041300](https://doi.org/10.1146/annurev-statistics-031219-041300). URL: <https://doi.org/10.1146/annurev-statistics-031219-041300> (visited on 2022).
- Watanabe, Sumio (2010). "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory". In: *Journal of Machine Learning Research* 11.116, pp. 3571–3594. ISSN: 1533-7928.

## 196 Appendix 1

## A | APPENDIX FIGURE 1

198 Here an example of an appendix figure.



200 Appendix 1—figure 1. Blablabla.

## 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](https://doi.org/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



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

Late Pliocene and Early Pleistocene epochs 3.6 to 0.8 million years ago<sup>1</sup> had climates resembling those forecasted under future warming<sup>2</sup>. Palaeoclimatic records show strong polar amplification with mean annual temperatures of 11–19 °C above contemporary values<sup>3,4</sup>. The biological communities inhabiting the Arctic during this time remain poorly known because fossils are rare<sup>5</sup>. Here we report an ancient environmental DNA<sup>6</sup> (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.

Q1 Q2  
Q3 Q4

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<sup>7</sup> (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<sup>7–11</sup>.

These conditions must have driven substantial ablation of the Greenland Ice Sheet, possibly producing one of the last ice-free intervals<sup>7</sup> 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<sup>10,11</sup>. However, the approximately 3.4 Myr old Fyles Leaf bed and Beaver Pond on Ellesmere Island in Arctic Canada preserve fossils of

Q5

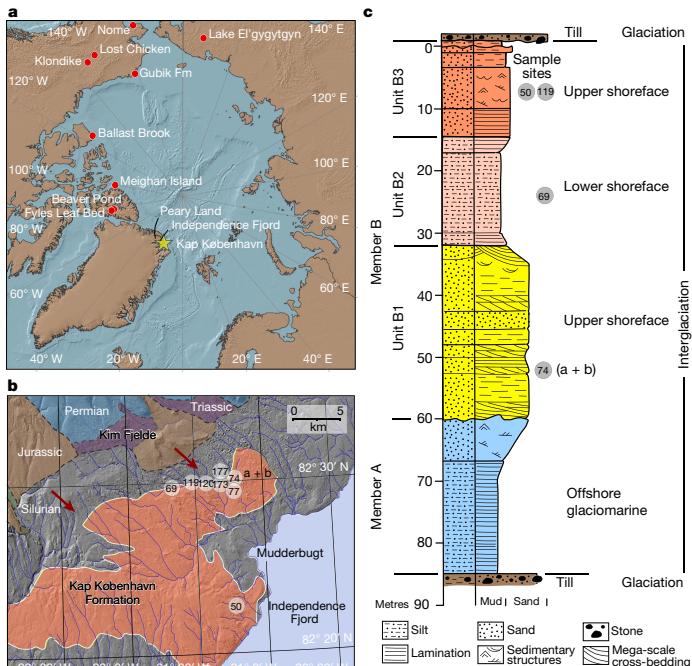
Q6

<sup>1</sup>Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark. <sup>2</sup>Department of Zoology, University of Cambridge, Cambridge, UK.

<sup>3</sup>Department of Genetics, University of Cambridge, Cambridge, UK. <sup>4</sup>Section for Evolutionary Genomics, Faculty of Health and Medical Sciences, The Globe Institute, Copenhagen K, Denmark.

<sup>5</sup>Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark. <sup>6</sup>Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, IRD, Université Gustave Eiffel, ISTerre, Grenoble, France. <sup>7</sup>Halsnaes Kommune, Frederiksverk, Denmark. <sup>8</sup>GEUS, Geological Survey of Denmark and Greenland, Copenhagen K, Denmark. <sup>9</sup>Geological Institute, Russian Academy of Sciences, Moscow, Russia. <sup>10</sup>Department of Earth Sciences, Uppsala University, Uppsala, Sweden. <sup>11</sup>Department of Earth and Environmental Sciences, Dalhousie University, Halifax, Canada. <sup>12</sup>The Arctic University Museum of Norway, UiT—The Arctic University of Norway, Tromsø, Norway. <sup>13</sup>Carlsberg Research Laboratory, Copenhagen V, Denmark. <sup>14</sup>Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, UK. <sup>15</sup>Trace and Environmental DNA (TrEnD) Laboratory, School of Molecular and Life Sciences, Curtin University, Perth, Western Australia, Australia. <sup>16</sup>University of Grenoble-Alpes, Université Savoie Mont Blanc, CNRS, LECA, Grenoble, France. <sup>17</sup>Department of Geosciences, UiT—The Arctic University of Norway, Tromsø, Norway. <sup>18</sup>Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM Université PSL, Paris, France. <sup>19</sup>School of Geography and the Environment, University of Oxford, Oxford, UK. <sup>20</sup>Center for Accelerator Mass Spectrometry, Lawrence Livermore National Laboratory, Livermore, CA, USA. <sup>21</sup>Department of Archaeology, University of Cambridge, Cambridge, UK. <sup>22</sup>MARUM, University of Bremen, Bremen, Germany. <sup>23</sup>These authors contributed equally: Kurt H. Kjær, Mikkel W. Pedersen. <sup>24</sup>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. <sup>25</sup>e-mail: kurt@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 ( $82^{\circ} 24' N$ ,  $22^{\circ} 12' W$ ) 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<sup>7,10,11,14,98</sup>.

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

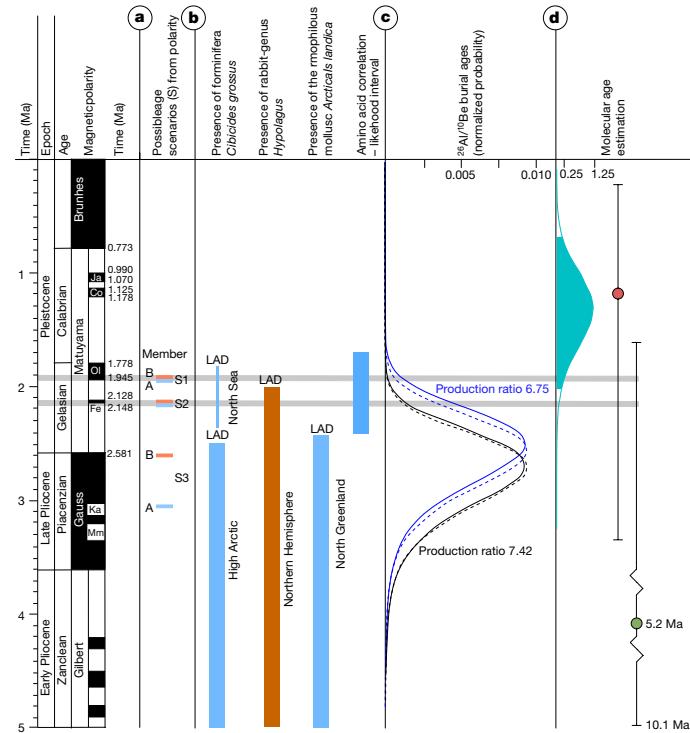
The Kap København Formation is formally subdivided into two members<sup>7</sup> (Fig. 1). The lower Member A consists of up to 50 m of laminated mud with an Arctic ostracod, foraminifera and mollusc fauna deposited in an offshore glaciomarine environment<sup>14</sup>. The overlying Member B consists 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<sup>7</sup>.

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<sup>10,15</sup>. Therefore, the taphonomy of the DNA

most probably reflects the biological communities eroded from a range of habitats, fluviably 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<sup>7,14,16–18</sup>. 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<sup>7</sup>. **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 distribution (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.

Q14

Q15

Q16

Q17

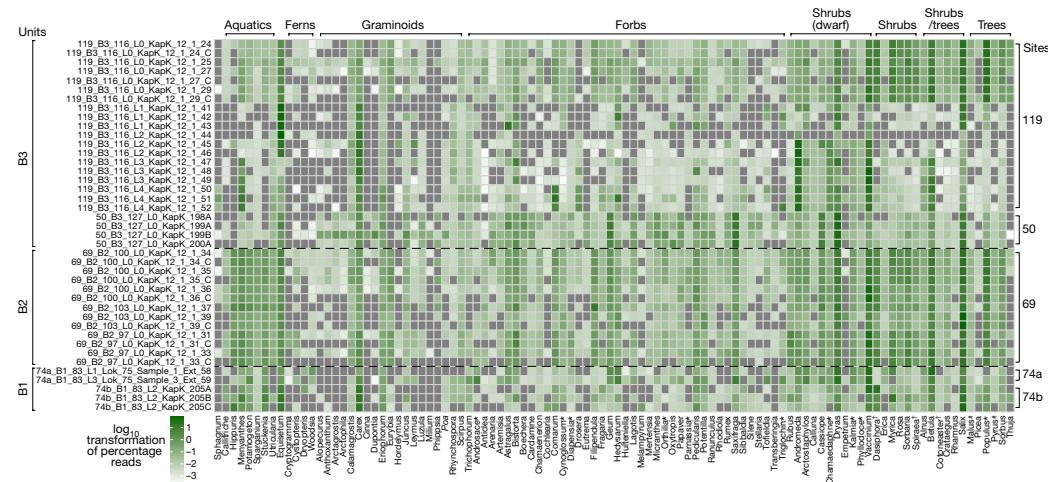
Information, section 1). Furthermore, we constrain the age using cosmogenic  $^{26}\text{Al}$ : $^{10}\text{Be}$  burial dating of Member B at four sites in this study (Supplementary Information, section 3). The recommended maximum burial age for the Kap København Formation is  $2.70 \pm 0.46$  Myr (Fig. 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<sup>7,16,18,19</sup>. 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<sup>20</sup>. 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<sup>21</sup>. 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<sup>22</sup> (MAT) of  $-17^\circ\text{C}$ , we found a thermal age of  $2.7 \text{ kyr}_{\text{DNA}}@10^\circ\text{C}$ —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<sup>23</sup>, 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<sup>24–27</sup>. 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

## Article



**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). 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<sup>28</sup> (4.9 ng ml<sup>-1</sup> DNA), the DNA adsorption capacity of smectite is 200 times greater than for quartz. We applied a sedimentary eDNA extraction protocol<sup>29</sup> 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.

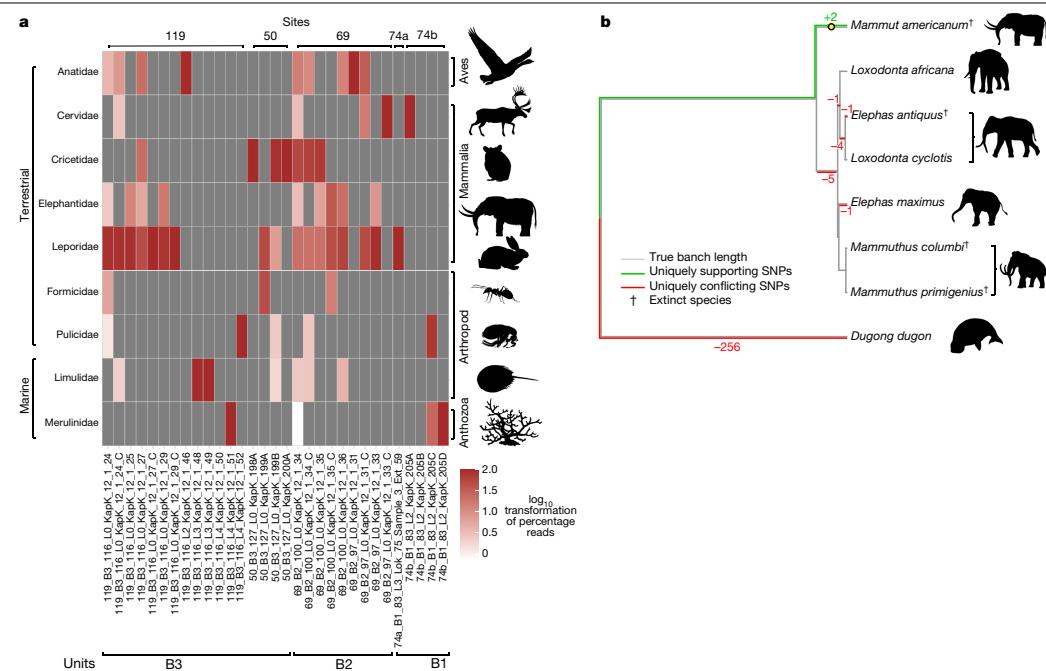
## Kap København metagenomes

We extracted DNA<sup>29</sup> 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<sup>30</sup>. 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<sup>31</sup> 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<sup>32</sup> (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<sup>33,34</sup> (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<sup>35</sup>. 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<sup>36,37</sup>. There are 175 vascular plant genera native to Greenland, excluding historically introduced species<sup>38–40</sup>. 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<sup>ss</sup> 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.<sup>41</sup> 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.<sup>42</sup>). 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<sup>10,15</sup> (Source Data 1, sheets 1 and 2). For example, 12 genera of Poaceae were identified by DNA (*Alopecurus*, *Anthoxanthum*, *Arctagrostis*, *Arctophila*, *Calamagrostis*, *Cinna*, *Dupontia*, *Hordeelymus*, *Leymus*, *Milium*, *Phippisia* and *Poa*), of these only *Hordeelymus* 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<sup>10,15</sup>, 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<sup>43</sup>. 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<sup>44</sup>, 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<sup>45</sup> 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<sup>10</sup>. Bennike<sup>45</sup> also notes a high proportion of club mosses and ferns and suggests they may be overrepresented owing to their spore wall

Q18

## Article

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 $\times$ , 57.06 $\times$  and 27.04 $\times$ , 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<sup>46</sup> 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<sup>47</sup> 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.

Most notably, we found reads in unit B2 and B3 assigned to the family Elephantidae, which includes elephants and mammoths, but taxonomically not mastodon (*Mammuthus* sp.)—which are, however, in the NCBI taxonomy, and therefore our analysis reads classified to Elephantidae or below therefore include *Mammuthus* sp. A consensus genome of our Elephantidae mitochondrial reads falls on the *Mammuthus* 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<sup>48</sup>. 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).

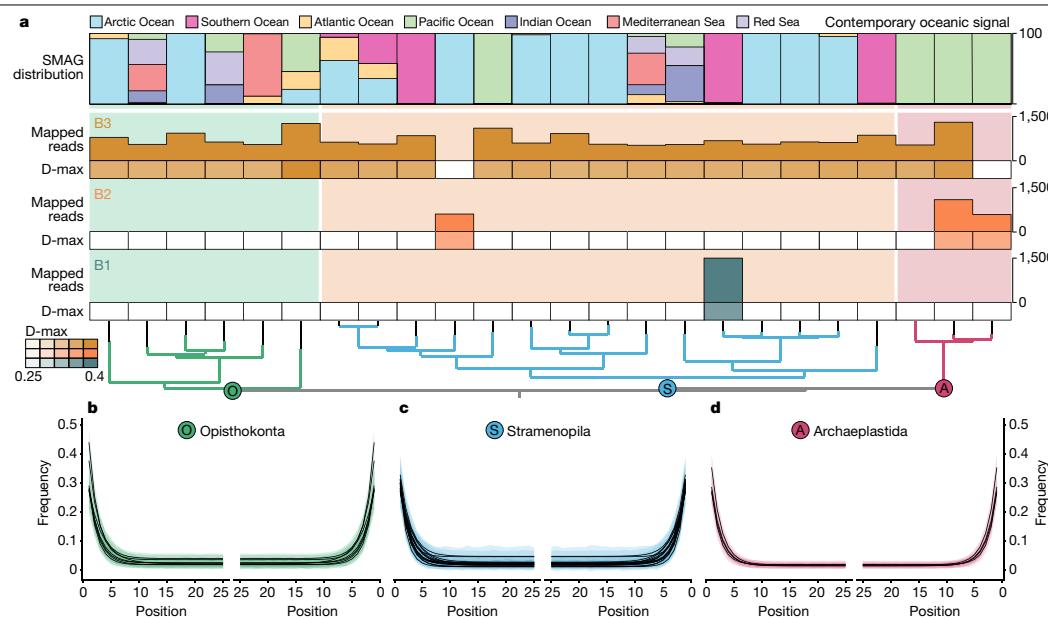
Similarly, reads assigned to the Cervidae support a basal placement on the *Rangifer* (reindeer and caribou) branch (Extended Data Fig. 3). Mitochondrial reads mapping to Leporidae (hares and rabbits) place near the base to the Eurasian hare clade (Extended Data Fig. 2), which is the only mammal found in the fossil record<sup>7</sup>. *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<sup>49</sup>. 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

Q19



**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.<sup>73</sup>. The SMAGs are ordered on the basis of

phylogenetic inference from Delmont et al.<sup>73</sup>. **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<sup>21</sup>. Our DNA results detected five times as many plant genera as previous studies using shotgun sequencing of ancient sediments<sup>29,34,50,51</sup>, which is well within the range of the richest northern boreal metabarcoding records<sup>52</sup>. 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<sup>34</sup>. 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<sup>5</sup>. 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<sup>53</sup>. 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<sup>10,15</sup>. To some degree, this is expected, as the ecological amplitudes of modern members of these genera have been modified by evolution<sup>54</sup>. Furthermore, the combination of the High Arctic photoperiod with warmer conditions and lower atmospheric CO<sub>2</sub> concentrations<sup>55</sup> 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<sup>56,57</sup> coupled with the absence of anthropogenic fire, which has had a role in some Holocene boreal habitats<sup>58</sup>, 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<sup>12,59</sup>. 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. Gross, 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. Bespovoz.* **16**, 126–153 (2019).
6. Wilberslev, 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. *Rap. 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. *Palaeontology of the Kap København Formation, a Plio-Pleistocene sequence in Peary Land, North Greenland* (Museum Tusculanum Press, 1995).
12. Rybcyznski, 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., Rybcyznski, N., Harington, C. R., White, S. C. & Tedford, R. H. A basal ursine bear (*Proartos abstrusus*) from the Pliocene High Arctic reveals Eurasian affinities and a diet rich in fermentable sugars. *Sci. Rep.* **7**, 17722 (2017).
14. Simonsen, 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 in 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 Undersøgelser i Grønland, 1990).
19. Feyling-Hanssen, R. W. *Foraminiferal Stratigraphy in the Plio-Pleistocene Kap København 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/ktlima/temaforside-klimaet-frem-til-i-dag/klimaet-i-grønland/> (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**, 305–326 (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 short-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**, 11129 (2019).
37. Bay, C. Floristic and ecological characterization of the polar desert zone of Greenland. *J. Veg. Sci.* **8**, 685–696 (1997).
38. Boermann, D. & Bay, C. *Grønlands Røddliste 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., Razibin, 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 Phyogeographical 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. Altsos, 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 cause 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).

Q27

Q29

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., Rybcynski, 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. & Straubnet, 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).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

**PhyloNorway Consortium**

Inger Greve Alsos<sup>12</sup> & Eric Coissac<sup>12,18</sup>

## Article

### Methods

#### Sampling

Q20

Sediment samples were obtained from the Kap København Formation in North Greenland ( $82^{\circ} 24' 00''\text{N}$   $22^{\circ} 12' 00''\text{W}$ ) 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  $40 \times 40 \times 40$  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  $-22^{\circ}\text{C}$ .

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  $8 \times 1$  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 ( $^{26}\text{Al}/^{10}\text{Be}$  versus normalized  $^{10}\text{Be}$ ). The  $^{26}\text{Al}$  and  $^{10}\text{Be}$  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  $^{26}\text{Al}/^{10}\text{Be}$  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  $^{26}\text{Al}/^{10}\text{Be}$  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  $^{26}\text{Al}$  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  $^{26}\text{Al}/^{10}\text{Be}$  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 Supplementary Information). However, once the marine sediments emerged above sea level, in-situ production by both nucleogenic and muogenic production could alter the  $^{26}\text{Al}/^{10}\text{Be}$ . The  $^{26}\text{Al}$  versus  $^{10}\text{Be}$  isochron plot reveals this complex burial history (Supplementary Information, section 3) and the concentration versus depth composite profiles for both  $^{26}\text{Al}$  and  $^{10}\text{Be}$  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 (Supplementary 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 (KKO6A, B and C). However, this calculation depends on the  $^{26}\text{Al}/^{10}\text{Be}$  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  $1\sigma$  confidence (Supplementary Information, section 3). Taking the midpoint between the negative and positive  $3\sigma$  confidence limits, we obtain a maximum burial age of  $2.70 \pm 0.46$  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 than 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<sup>60</sup> 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  $10^{\circ}\text{C}$ , thus estimating the thermal age in years at  $10^{\circ}\text{C}$  for each sample (Supplementary Information, section 4). The mean annual air temperature (MAT) for the Kap København sediment was taken from Funder et al. (2001)<sup>61</sup> and for the Krestovka Mammoth the MAT was calculated using temperature data from the Cerskij Weather Station (WMO no. 251230)  $68.80^{\circ}\text{N}$   $161.28^{\circ}\text{E}$ , 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 ( $6.49^{\circ}\text{C km}^{-1}$ ). We scaled the long-term temperature model of Hansen et al. (2013)<sup>62</sup> 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  $127 \text{ kJ mol}^{-1}$  (ref. <sup>60</sup>).

#### 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  $\sim 1$  g of sediment with ethanol for 10 min in a McCrone Mill. The samples were dried at  $60^{\circ}\text{C}$  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 ( $\Theta-\Theta$  geometry) and the LynxEye detector (opening  $2.71^{\circ}$ ), with  $\text{Cu } K_{\alpha1,2}$  radiation ( $1.54 \text{ \AA}$ ; 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  $5\text{--}90^{\circ}\text{2}\theta$  with a step size of  $0.1^{\circ}$  and a step time of 4 s while the sample was spun at 20 rpm. The opening of the divergence slit was  $0.3^{\circ}$  and of the anticatter slit  $3^{\circ}$ . Primary and secondary Soller slits had an opening of  $2.5^{\circ}$  and the opening of the detector window was  $2.71^{\circ}$ . For the Rietveld analysis, we used the Profex interface for the BGMIN software<sup>62,63</sup>. The instrumental parameters and peak broadening were determined by the fundamental parameters ray-tracing procedure<sup>64</sup>. 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

Q21

Q22

**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<sup>65</sup> 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<sup>-1</sup>. 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<sup>66</sup>. 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<sup>67,68</sup>. The pollen diagram was initially compiled using *Tilia* version 1.5.1<sup>69</sup> but replotted for this study using *Psimpoll* 4.10<sup>70</sup>.

#### 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<sup>71</sup> and converted to 65 dual-indexed Illumina sequencing libraries (including 13 negative extraction-and library controls)<sup>30</sup>. 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)ACTCCCATATGAAA(BHQ1)-3') and Poaceae psbA (primer: 5'-CTCACAACTTCCCTCTAGAC-3', probe 5'-(HEX) AGCTGCTGTTGAAGTT(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<sup>31</sup> 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 kmer similarity between all samples using simka<sup>32</sup> (setting heuristic count for max number of reads (-max-reads 0) and a kmer 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<sup>33</sup> 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<sup>34</sup> 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://metadmග. dev.github.io/metaDMG-core/index.html](https://metadmഗ. 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<sup>2</sup> 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<sup>73</sup> that contain 683 metagenome-assembled genomes (MAGs) and 30 single-cell genomes (SAGs) from *Tara Oceans*<sup>74</sup>, following the naming convention in Delmont et al.<sup>73</sup>, 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<sup>75</sup> against the SMAGs. We used MarkDuplicates from Picard (<https://github.com/broadinstitute/picard>) to remove 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.II.1). We used anvi'o<sup>76</sup> in manual mode to plot the mapping and damage results using the SMAGs phylogenomic tree inferred by Delmont et al.

Q23

## Article

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<sup>42</sup> and NCBI. For example, since Bennike (1990)<sup>18</sup>, *Potamogeton* has been split into *Potamogeton* and *Stuckenia*. *Polygonum* 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*<sup>39</sup>. Plant families have changed names—for instance, Gramineae is now called Poaceae and Scrophulariaceae has been re-circumscribed to exclude Plantaginaceae and Orobanchaceae<sup>77</sup>. 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<sup>38</sup>. 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<sup>79</sup>, including a single representative sequence from a recently diverged outgroup. For the *Betula* genus, we also included three chloroplast genomes from the PhyloNorway database<sup>34,80</sup>. We changed all ambiguous bases in the fasta files to N. We used MAFFT<sup>81</sup> to align each of these sets of reference sequences, and inspected multiple sequence alignments in NCBI MSAViewer to confirm quality<sup>82</sup>. 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<sup>83</sup>.

The BEAST suite<sup>48</sup> 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<sup>84</sup> to create a reference consensus fasta sequence for each set of taxa. Furthermore, we used SNPSites<sup>85</sup> 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<sup>86</sup> 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<sup>87</sup>. 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<sup>88</sup>, 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 *Mammuthus*, 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<sup>81</sup> 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. <sup>46</sup>) and repeated the previously described pathPhynder steps, with the exception that the pathPhynder tree path analysis for the *Mammuthus* SNPs was based on transitions and transversions, not restricting to only transversions due to low coverage.

***Mammuthus americanus*.** 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<sup>89</sup> 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<sup>46</sup> v1.10.4. We used two separate approaches when dating our mastodon mitogenome, as demonstrated in a recent publication<sup>90</sup>. 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<sup>90</sup>. We utilized the same BEAST parameters as Karpinski et al.<sup>90</sup> and set the age of our sample with a gamma distribution (5% quantile:  $8.72 \times 10^4$ ; Median:  $1.178 \times 10^5$ ; 95% quantile:  $5.093 \times 10^6$ ; 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. <sup>46</sup>). 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*<sup>91</sup>. 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 Phylo-Norway database generated in a recent study<sup>29</sup>, 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<sup>81</sup> and checked the quality of our alignment by eye in Seqtron<sup>92</sup> and NCBI MsaViewer. Next, we called a consensus sequence from this MSA using the BioAlign consensus function<sup>84</sup> 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<sup>93</sup>, 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<sup>94</sup>. We mapped these read sets against the consensus *Betula* chloroplast genome using BWA<sup>95</sup> 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<sup>96</sup>. 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<sup>96</sup> to make an mpileup and call a vcf file, using options for haplody 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<sup>95</sup>, 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<sup>96</sup>,

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 Seqtron<sup>92</sup> 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<sup>46</sup>. 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<sup>46</sup> 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<sup>47</sup>; 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)<sup>97</sup>. 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<sup>98</sup> 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<sup>46</sup> had placed the ancient sequence phylogenetically identically to pathPhynders<sup>88</sup> placement, which is shown in Extended Data Fig. 9. For our major results, we report the uniform ancient age prior, and the GTR+G, 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., Kleeborg, R. & Bergmann, J. The new Seifert Rietveld program BGMIN and its application to quantitative phase analysis. *Mater. Struct.* **5**, 57–66 (1998).
63. Doebelein, N. & Kleeborg, R. Profex: a graphical user interface for the Rietveld refinement program BGMIN. *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 seawater. *Limnol. Oceanogr.* **12**, 176–179 (1967).
66. Grichuk, E. D. & Zakininskaya, 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 Tiliograph* (Illinois State Museum, 1991).
70. Bennett, K. D. Manual for psimpoll and pscomb. [http://www.chrono.qub.ac.uk/psimpoll/psimpoll.html](http://www.chrono.qub.ac.uk/psimpoll/) (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. Alhos, I. G. et al. Last Glacial Maximum environmental conditions at Andeja, northern Norway: evidence for a northern ice-edge ecological 'hotspot'. *Quat. Sci. Rev.* **239**, 103634 (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. MSView: 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 Santis, 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. Fournier, M. & Holmes, E. C. Seqtron: 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. Fleying-Hansen, 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. Levens, 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 (WT200203). J.A.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 John's College, Cambridge for providing him with a stimulating environment for scientific thoughts and discussion.

Q25

**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.S. performed cartography and GIS analysis. L.S. designed and carried out palaeomagnetic analysis and interpreted the results. J.C.G. prepared and analysed eight samples for cosmogenic <sup>26</sup>Al and <sup>10</sup>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.A.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., J.S. and N.K.L., with inputs from the other authors.

Q26

**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.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

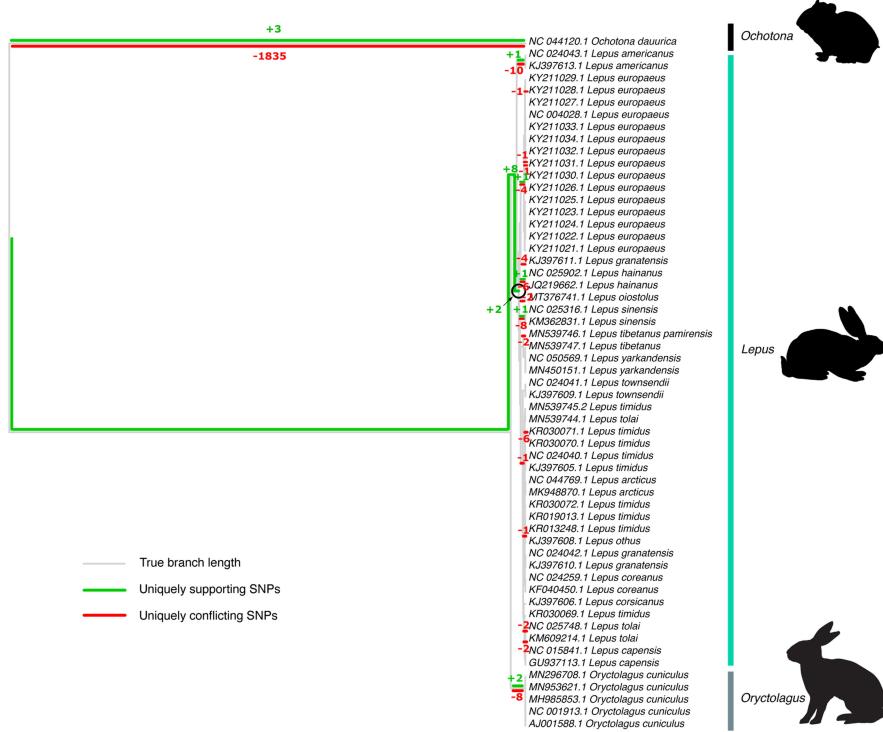
Q31

Q32



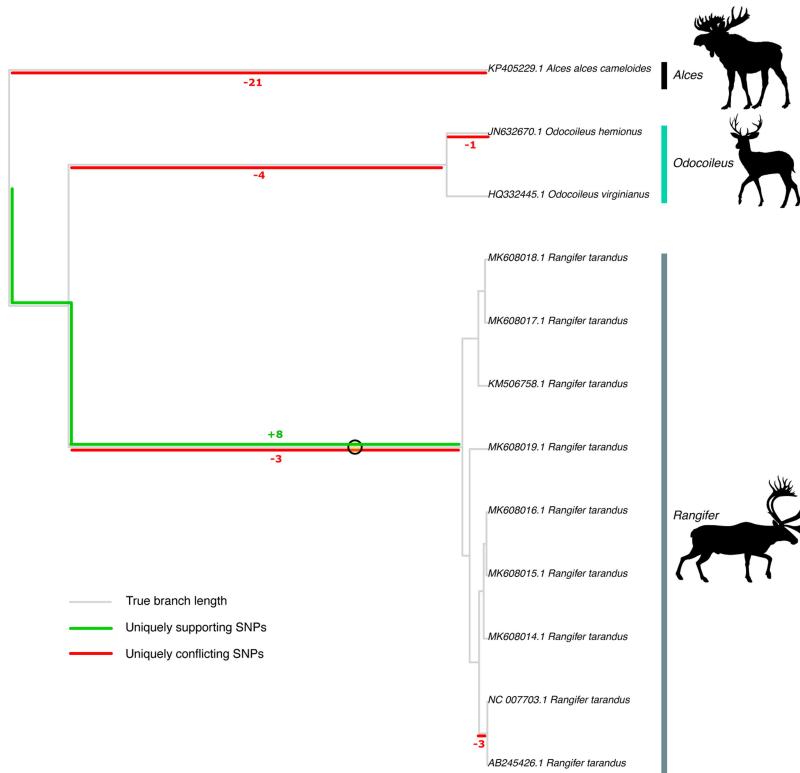
**Extended Data Fig. 1 | Setting A.** Type locality 50 indicating units in formation **b**. Overview locality 74a+b with a complete sediment sequence. **c.** Overview of locality 69. **d.** Detail of organic rich sediment in unit B3 before excavation and cleaning for ancient eDNA samples. **e.** Sampling 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

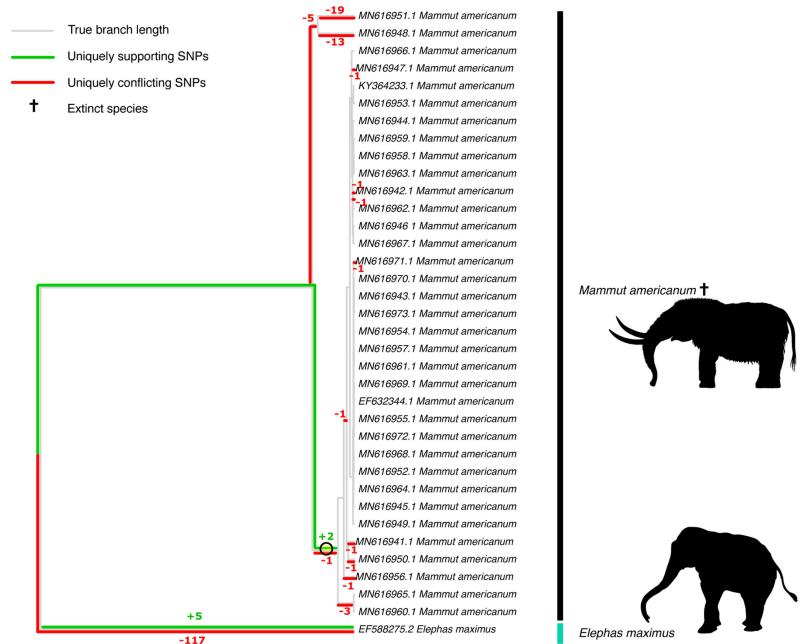
conflicting (-) SNPs in the ancient Leporidae environmental mitochondrial genome overlapping the reference SNPs assigned to the respective edge. There is a clear placement for the ancient Leporidae environmental mitochondrial genome on the edge marked +2, basal to the extant *Lepus* lineage.



**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

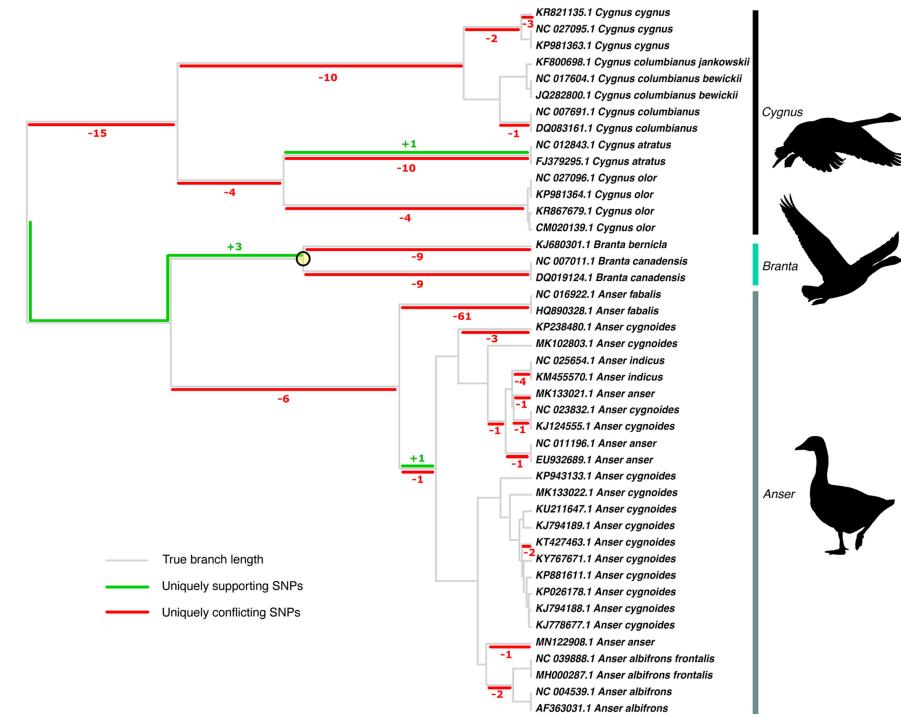
mitochondrial genome overlapping the reference SNPs assigned to the respective edge. There is a clear placement for the ancient Capreolinae environmental mitochondrial genome on the edge marked +8/-3, basal to the *Rangifer* genus.

## 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

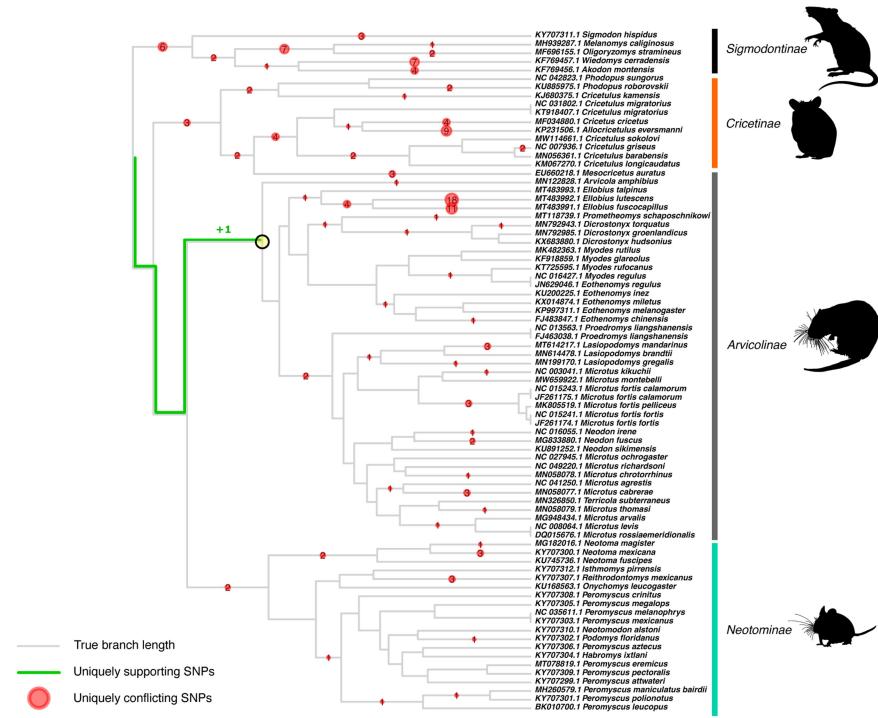
indicate conflicting (-) SNPs in the ancient Elephantidae environmental mitochondrial genome overlapping the reference SNPs assigned to the respective edge. There is a placement for the ancient Elephantidae environmental mitochondrial genome on the edge marked +2/-1, identifying it as basal to the mastodon (*Mammuthus americanus*) 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

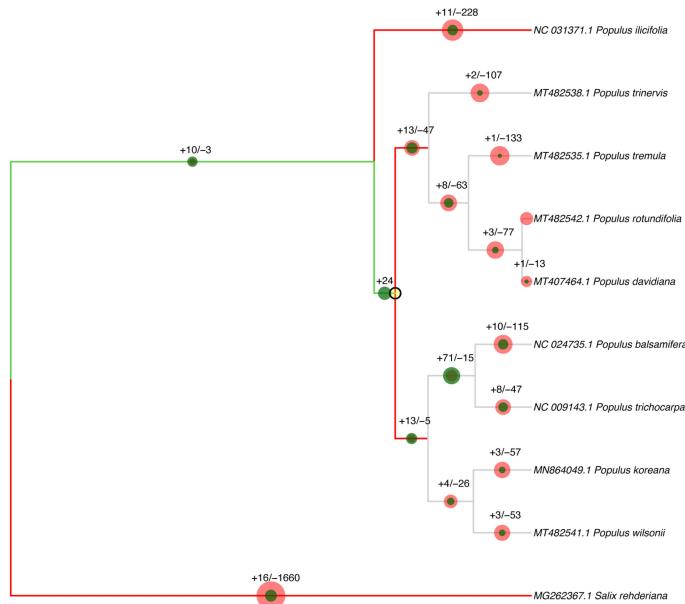
Anatidae environmental mitochondrial genome overlapping the reference SNPs assigned to the respective edge. There is a clear placement for the ancient Anatidae environmental mitochondrial genome on the edge marked +3, basal to the *Branta* genus.

## 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

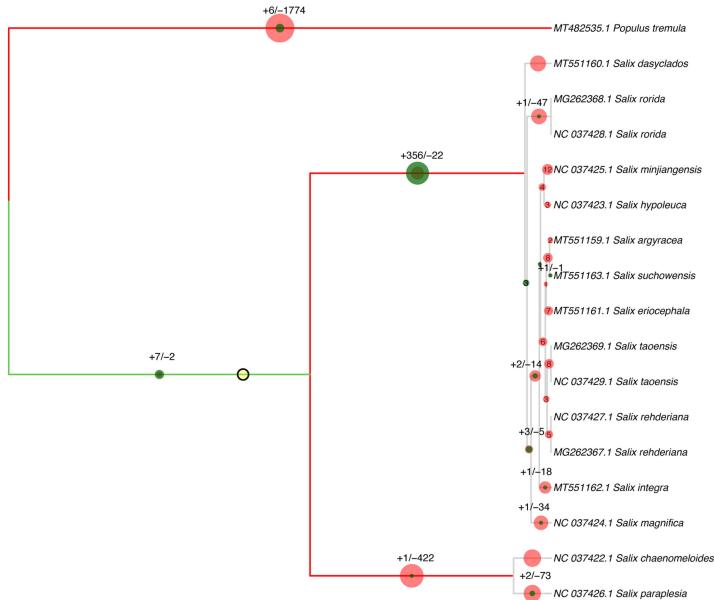
mitochondrial genome overlapping the reference SNPs assigned to the respective edge. There is a placement for the ancient Cricetidae environmental mitochondrial genome on the edge marked +1, basal to the Arvicolinae subfamily.



**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

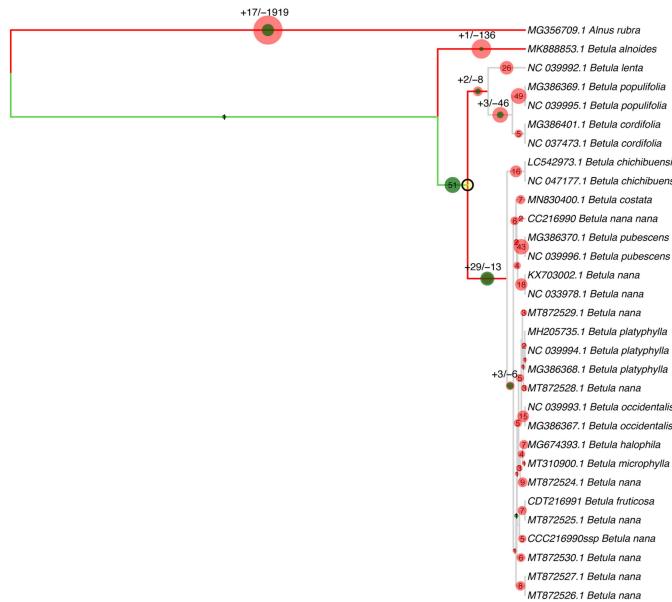
branches directly leading to these species as well. *Populus balsamifera* and *P. trichocarpa* are considered sister species. They are both distributed in North America, as far North as Alaska, are known to hybridise both among themselves and other *Populus* species and are morphologically very similar<sup>91,99,100</sup>. Previous analyses found a very recent nuclear genome divergence time of only 75000 years ago for *Populus trichocarpa* and *P. balsamifera*<sup>100</sup>, but a deep chloroplast genome divergence time of at least 6–7 Ma<sup>9</sup>, 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.

## Article



**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<sup>91</sup>, 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<sup>91</sup>, 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.

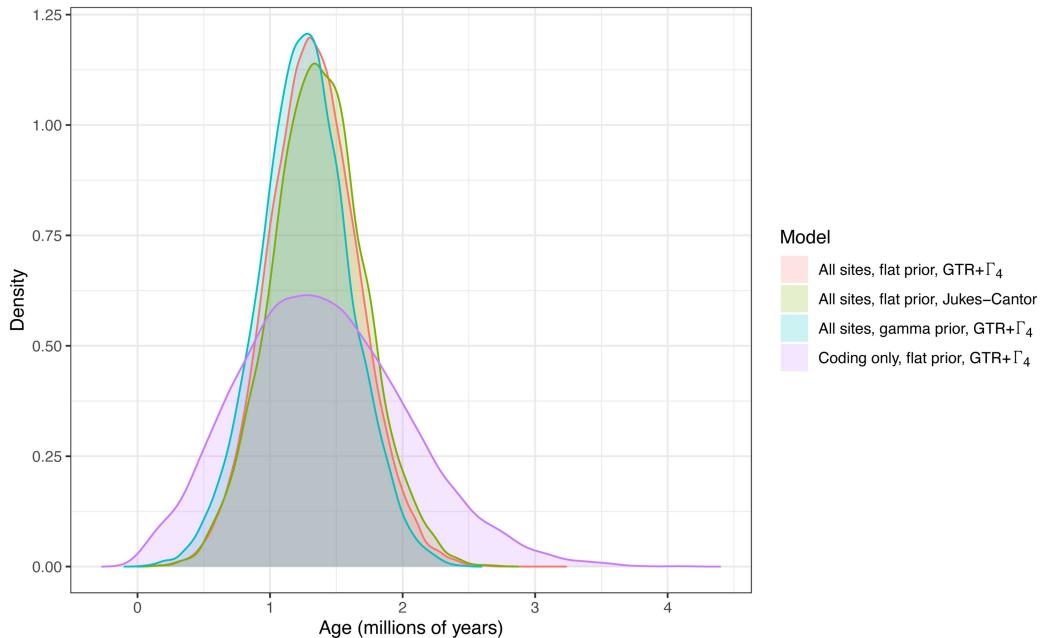


**Extended Data Fig. 9 | Phylogenetic placement results for our *Betula* chloroplast reads, using both transition and transversion SNPs, and using reads merged from all layers and sites.** Our ancient *Betula* sample was placed basal to a main *Betula* clade, based on 29 supporting (green) and

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).

## Article

Ancient *Betula* chloroplast age distribution under different BEAST models



**Extended Data Fig 10 | Molecular age distribution.** Results of running the ancient *Betula* chloroplast molecular dating analysis BEAST v1.10.4 (ref. <sup>40</sup>) 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+ $\Gamma_4$  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 temperature <sup>22</sup> (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 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 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.

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](https://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 reconstruction.

### 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?	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> 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		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern		



## **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.



# Ekspertrapport af den 10. december 2020

Effekten af kontaktopsporing



## Indhold

<b>1. Sammenfatning og konklusion .....</b>	<b>3</b>
<b>2. Formål og baggrund .....</b>	<b>4</b>
2.1 Formål og baggrund for modelgruppen.....	4
2.2 Formål med rapporten.....	4
<b>3. Opsporing og håndtering af nære kontakter i Danmark .....</b>	<b>5</b>
3.1 Forudsætninger for en effektiv kontaktsporing .....	5
3.2 Definition af en nær kontakt.....	5
3.3 Periode for smitteopsporing .....	6
3.4 Opsporing af nære kontakter .....	6
<b>4. Agentbaserede modeller.....</b>	<b>8</b>
4.1 Om agentbaserede modeller .....	8
4.2 Forbehold.....	8
<b>5. Resultater .....</b>	<b>9</b>
5.1 Resultater fra den agentbaserede model udviklet af Niels Bohr Instituttet, Københavns Universitet. ....	9
5.2 Resultater fra den agentbaserede model udviklet af DTU Compute, Danmarks Tekniske Universitet.....	10
<b>6. Referencer.....</b>	<b>13</b>
<b>Bilag 1. Beskrivelse af den agentbaserede model fra Niels Bohr Instituttet</b>	<b>14</b>
<b>Bilag 2. Beskrivelse af den agentbaserede model fra DTU .....</b>	<b>16</b>
<b>Bilag 3. Regneeksempel.....</b>	<b>22</b>
<b>Bilag 4. Udvikling i antal kontakter fra HOPE projektet .....</b>	<b>24</b>
<b>Bilag 5. Beskrivelse af parametre brugt i rapporten.....</b>	<b>25</b>
<b>Bilag 6. Medlemmer af ekspertgruppen .....</b>	<b>258</b>



## 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 Institutet (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 blyses 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. Indevarende rapport er udviklet i henhold til de tidlige 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 primært tilfælde 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 primært tilfælde. Gevinsten i form af en reduktion i kontakttallet er således større, såfremt der sikres effektiv opsporing for samtlige primært tilfælde, relativt til reduktionen i kontakttallet, som kan opnås ved at nedbringe ventetiden til test og testsvar for primært tilfældet.

Ventetiden til test og testsvar for et primært tilfælde 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 primært tilfælde 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 primært tilfældet, 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 blyses ved at estimere effekten af centrale elementer i kontaktopsporringen, såsom ventetid til test og testresultat hos primærtildældet, 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 misstanke 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<sup>1</sup>. Ved at udbyde adgang til test for flest mulige personer, vil man også finde flere asymptomatiske smittekærere.

### 3.2 Definition af en nær kontakt

*Sundhedsstyrelsen ud kom d. 23. november 2020 med opdaterede retningslinjer for smitteopsporing af nære kontakter. Indværende 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<sup>2</sup>.

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

<sup>1</sup> <https://www.ssi.dk/-/media/arkiv/dk/aktuelt/nyheder/2020/notat---covid-19-prvalensundersgelsen.pdf?la=da>

<sup>2</sup> <https://www.sst.dk/da/Udgivelser/2020/COVID-19-Smitteopsporing-af-naere-kontakter>



### 3.3 Periode for smitteopsporing

Der foretages smitteopsporing for perioden, hvor primærtildældet 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 primærtildældet
- 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 primærtildæld, der foretages kontaktopsporing for. Dette er et samlet gennemsnit for opsporede nære kontakter som STPS opsporer, og som primærtildældet 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 opspored kontakter fra sådanne udbrud, som kontakter Coronaopsporingen for at få rådgivning om hvilke dage de bør testes, samt for at få rekvizitioner 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 latentstiden, 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 primærtildældet 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, hvorfaf 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 coronaprover.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 Institutet, 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 infektøs (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 unojagtigheder eller bygge på antagelser, som ikke er retvisende. Det er ikke muligt at kvantificere den nøjagtige størrelse eller effekt af disse potentielle fejlkilder.



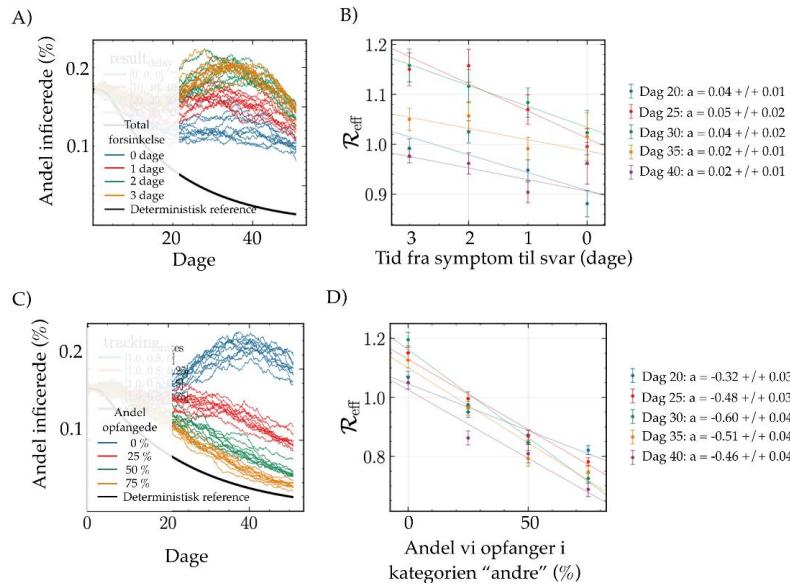
## 5. Resultater

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

Modelkørslene 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 scenario 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 ( $0.2 \times 0.3 = 0.06$ ) af reduktionen i kontakttallet, som ellers kunne opnås ved kontaktopsporing.

Ventetiden fra at et primært tilfælde 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)).



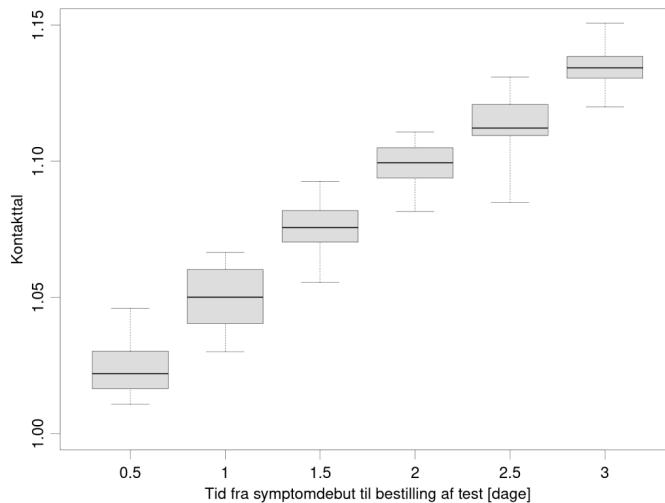


*Figur 1: A) Simuleret model, hvor hver kørsel (markeret med samme farve) gentages 10 gange for forskellige værdier af tiden fra symptom til svar. B) Verdien af kontakttallet estimeret på forskellige tidspunkter i simulationen vist i A). Den lineære sammenhæng giver en værdi for hvor mange procent kontakttallet 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 kontaktopsporningen, 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.

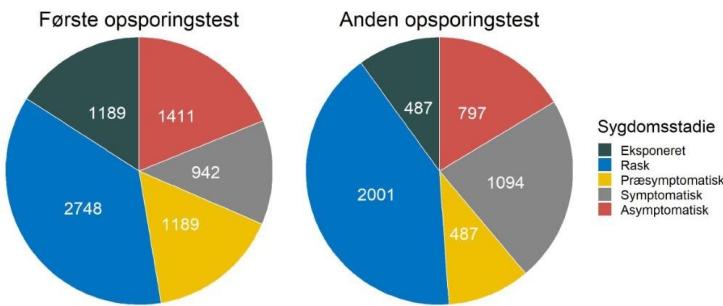


*Figur 2: Kontakttallets afhængighed af den gennemsnitlige tid fra at primærtilfældet har symptomdebut til der bestilles en test og personen går i en grad af isolation. For hver parameterværdi er der foretaget 40 simulationer, og boxplottet viser median, de indre kvartiler samt minimum og maksimum af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



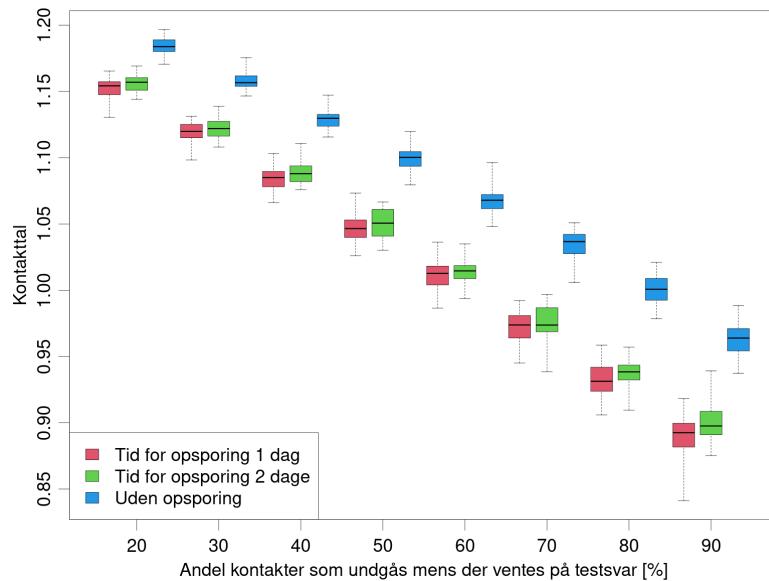
*Figur 3. Antal opsporedt og smittede i hvert sygdomsstadiet, når de får foretaget hhv. første og anden test i opsporingsprocessen. Der er flere, som ikke kommer til anden test, bl.a. fordi de tester positivt i første test eller efter negativt testsvar vælger ikke at få taget den opfølgende test. Derudover 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 opsporedt 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 positivt, da qPCR kan detektere virus 17 dage efter symptomdebut (Cevik et al., 2020). Desuden ses det, at personer i det præsymptomatiske stade - hvor ca. halvdelen af smitten sker - kun udgør en lille andel af de opsporedt 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. Kontaktopsporningen 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 primært tilfældet.

Personer, som tidligere er testet positivt er ikke medtaget her og bidrager derfor ikke til antallet af raskt. 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 procent-point 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 opsporedे 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 opsporedе 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](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*. Lokaliseret: <https://doi.org/10.7326/M20-1495>

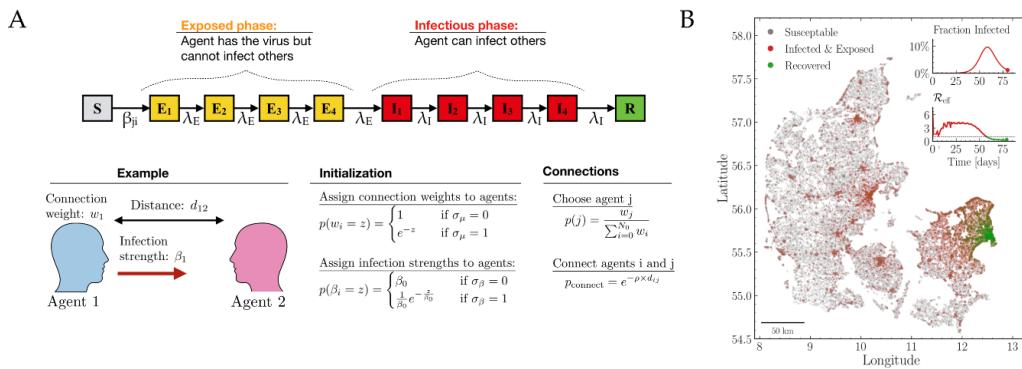
Plank, M., James, A., Lustig, A., Steyn, N., Binny, R. & Hendy, S. (2020). Potential reduction in transmission of COVID-19 by digital contact tracing systems. *MedRxiv*. Lokaliseret: <https://doi.org/10.1101/2020.08.27.20068346>



## Bilag 1. Beskrivelse af den agentbaserede model fra Niels Bohr Institutet

*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 middel-værdien	Reference
Antal kontakter per dag	3-15	HOPE projektet
Latent tid (dage)	3-5	Litteratur se referenceliste i bilag 5
Infektios 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 befolkningstabell, 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 smitteparametere.

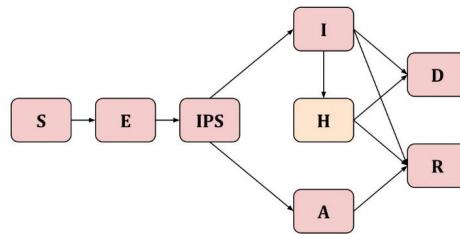
I sygdomsmodellen bæres smitten fremad ved, at agenter der deler netværks-ID, f.eks. husholdnings-ID, skole/job-ID eller omgangskreds-ID, kan smitte hinanden. Hver dag får alle agenter udregnet deres sandsynlighed for at blive smittet på baggrund af antal infektioner i deres forskellige netværk og på 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).<sup>3</sup> Det antages i modellen, at den gennemsnitlige kontorstørrelse og den gennemsnitlige omgangskreds uden for skole og arbejde er på 8 personer.

<sup>3</sup> 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 befinner sig i det præ-symptomatiske, symptomatiske eller asymptomatiske stadiet, er infektios og kan således videreførte smitte til agenter, som befinner sig i det modtagelige stadiet. Bliver en modtagelig agent inficeret, overgår de til at være eksponeret. Dette sygdomsstadi reprenterer den latente periode, hvor den inficerede agent endnu ikke er infektios. 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 infektios 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 sygdomsstadiet, 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.

Stadier	Parametre		Kvartiler			Referencer
	Shape	Periode [Dage]	Nedre kvartil [Dage]	Median [Dage]	Øvre kvartil [Dage]	
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



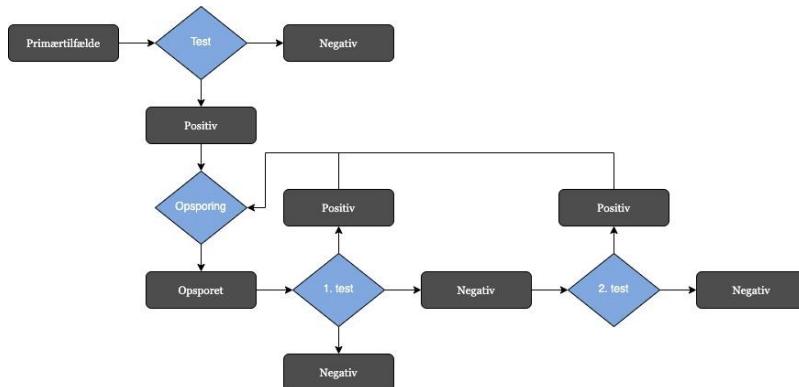
Hospitaliseret (H) Under 60 år	2	3	2	3	5	Linelisten SSI
Hospitaliseret (H) 60 år og der- over	2	5	3	5	7	Linelisten SSI
<b>Ventetider</b>						
timeSymp- ToOrderTest	5	1	1	1	1	Antagelser - af- venter STPS data
timeOrderTo- Test	2	2	1	2	3	Antagelser - af- venter STPS data
timeTestToRe- sult	6	1,5	2	2	2	Ventetider fra samfundssporet
traceDelay	5	1	1	2	3	Antagelser – af- venter STPS data

Sandsynligheden for, at en modtagelig agent bliver inficeret af en infektios agent og overgår til at være eksponeret i et givent netværk stiger med antallet af infektiose agenter i netværket, de infektiose agenter i netværkets smitsomhed, samt antallet af kontakter som både de modtagelige og infektiose 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 opsporedes person har samme ventetider på testsvar, som symptomatiske personer.



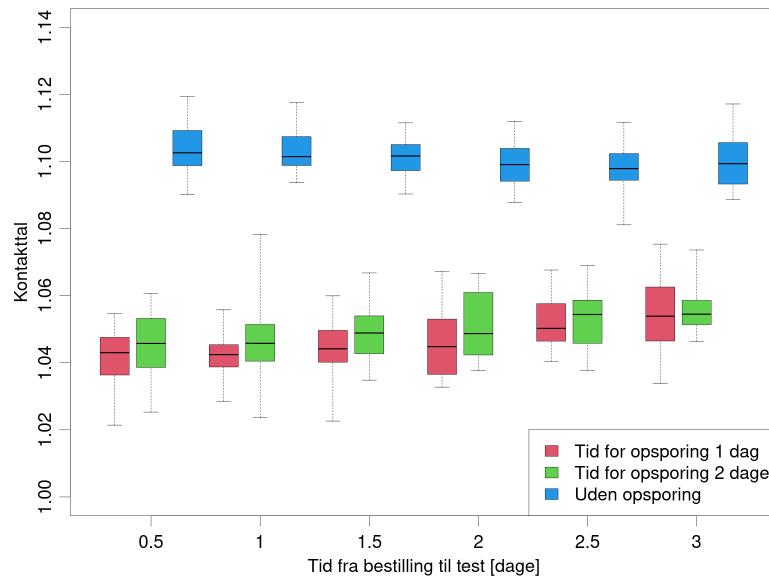
- Mens der ventes på test og testsvar, isoleres den opspored person på samme måde som en symptomatisk, der venter på svar.
- Hvis en opsporet person får negativt svar på den første test, vil der være en sandsynlighed for ( $pNoShow2ndTest = 40\%$ ) at de ikke tager test nummer 2.
- Efter et negativt 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.

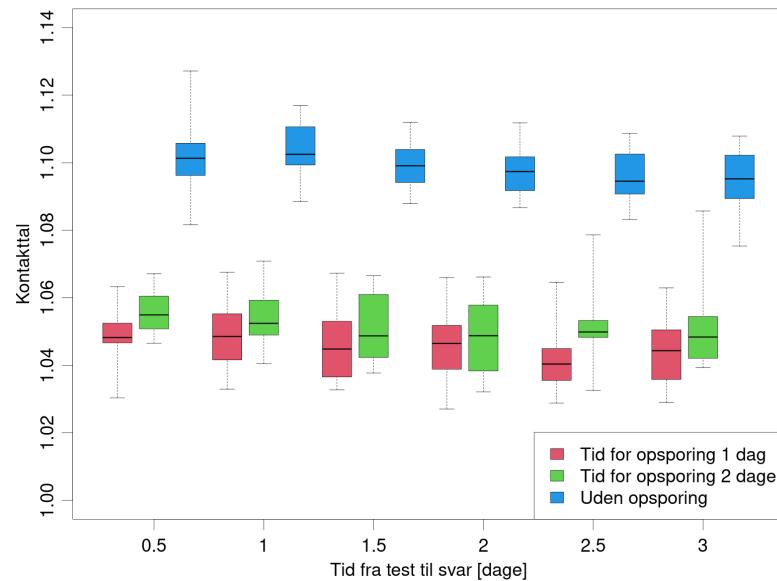


### Yderligere resultater



Figur 8. Kontakttalets 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 isolations. 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 primærtifeldet og opspored 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 opspored 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 Institutet 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 primærtildældet. Denne person kan enten findes tilfældigt, eller ved at primærtildældet testes, og at sekundærtildældet opspores efter en tidsperiode ( $d$  for delay). Denne ventetid, er tiden fra at primærtildældet testes til at sekundærtildældet kontaktes, og afspejler således både ventetid til test samt ventetid til opsporing. Ingen antages det, at sekundærtildældet går i isolation øjeblikkeligt. Ved igen at trække tal tilfældigt fra de relevante fordelinger fås en eksponeringsperiode, hvori sekundærtildældet 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,  $R_t$ . Det antages, at en given andel ( $f_c$ ) af alle smittetilfælde, findes via kontaktopsporing, og derved reduceres smitten, idet eksponeringsperioden for opspored kontakter reduceres. Herved fås et simpelt estimat af effekten af kontaktopsporing på kontakttallet  $R_t$ . 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 forlorte eksponeringsperioden for kontakter.

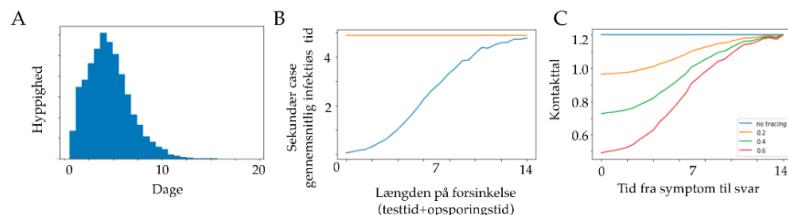
I modellen indgår 4 mulige eksponeringsperioder. 1: Kontakter opspores ikke, hvorved eksponeringsperioden ikke afkortes (blå graf), 2: 20% af kontakter opspores (gul graf), 3: 40% af kontakter opspores (grøn graf) og 4: hvis 80% af kontakter opspores (rød graf).

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 opspored kontakter.



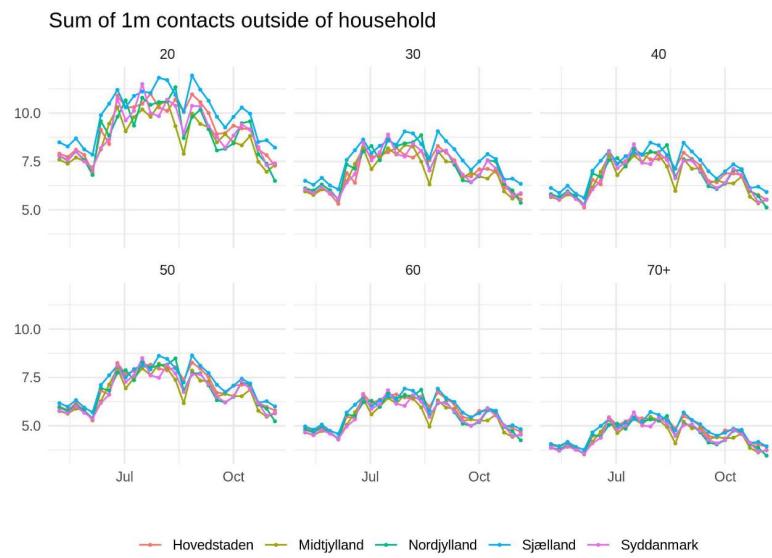
Det skal i øvrigt bemærkes, at det i eksemplet antages, at opsporede kontakter går i isolations, indtil de får svar på deres test.



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 primært tilfældet som reference. C) Det effektive kontakttal  $R_t$  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](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 infektions 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). *bioRxiv*.
- Shen et al. (2020). Modelling the epidemic trend of the 2019 novel coronavirus outbreak in China. *bioRxiv*.



- 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 infektiose periode*, hvor det bedste estimat er 5 dage, mens det bedste interval er mellem 3-7 dage, har Ekspertruppen 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 estimererne 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/mariefly/HOPE/raw/master/HOPE\\_report\\_2020-11-12.pdf](https://github.com/mariefly/HOPE/raw/master/HOPE_report_2020-11-12.pdf)), oversender HOPE-projektet løbende anonymiserede data om befolkningens adfærd under covid-19 til Ekspertruppen, der anvender det i deres modeller. Ekspertruppen har også adgang til HOPE-projektets rapporter, der sammenskriver data.

#### **Trafik data:**

Antagelser om befolkningens adfærd bygges ligeledes på trafikdata, hvorudfra man kan bestemme danskerne rejsemonstre. Efter aftale med Trafik-, Bygge- og Boligstyrelsen får Ekspertruppen løbende adgang anonymiserede data over danskerne bevægelse rundt i landet. Data er bl.a. brugt til at bestemme den typiske afstand mellem kontakter og afstanden mellem afstands- uafhængige 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 Luft-havn
- Overblik over biltrafikken på Statsvejsnettet og cykeltrafikken (samlet ud fra tællestan-dere) 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 husstanden var smittet og lignende.

Data fra spørgeskemaundersøgelsen blev i modellerne brugt til at estimere tiden fra symptomde-but til tests i dage.

**Den nationale prævalensundersøgelse for covid-19:**

SSI iværksatte i maj en undersøgelse af, hvor udbredt covid-19 var blandt danskerne. Undersø-gelsen 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.



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 estimerer 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 juliedagene, 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



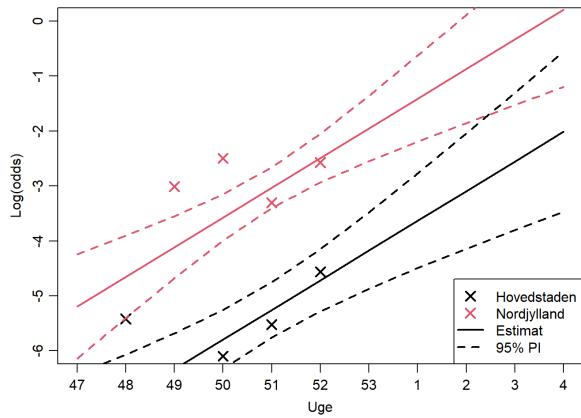
Det ses, at log(odds) for at detektere cluster B.1.1.7 er 2.2 højere i Region Nordjylland end i Region Hovedstaden. Det svarer til odds på 9.2. Det mest interessante er den tidslige udvikling, hvoraf det ses at log(odds) øges med 0.54 for hver uge. Dette svarer til at cluster B.1.1.7 har en ugentlig 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 estimerer 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 vægget gennemsnit af virusvarianterne i populationen.

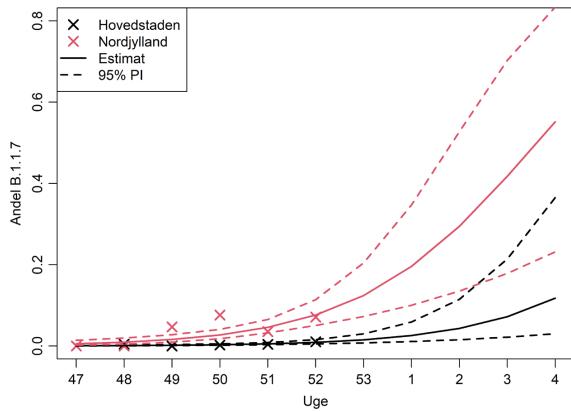
Figur 1 viser en fremskrivning af log(odds) for B.1.1.7 mod andre virusvarianter baseret på ovenstående logistiske regression. Estimatet er, at cluster B.1.1.7 allerede i uge 4 vil udgøre halvdelen af alle positive test i Region Nordjylland. Dette er dog behæftet med stor usikkerhed på baggrund af de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

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.



<sup>1</sup> 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 af 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 udtales 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  $0.7 * \log(\text{antal sekventerede})$ . 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 Institutet, 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 normalt viser. 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 sygdomsstadiet 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 fejkilder. 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*

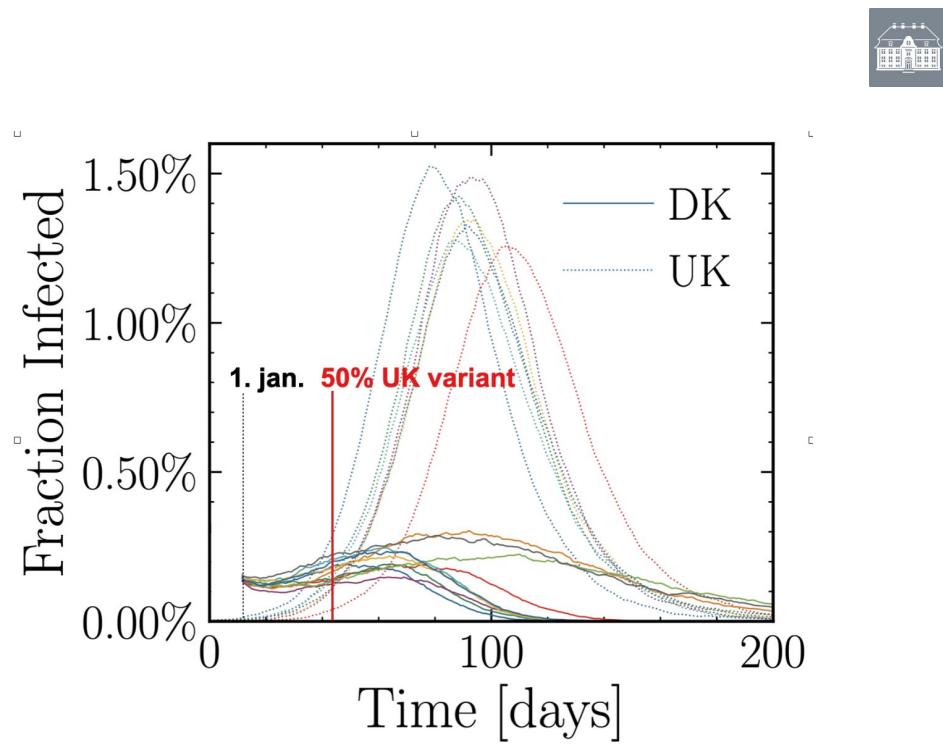
I det følgende er udviklingen simuleret i en model, hvor udgangspunktet er 1/10 af Danmarks befolkning, og hvor cluster B.1.1.7 fra starten udgør omkring 5% af de cirkulerende virusvarianter. Epidemien simuleres ud fra et kontakttal på omkring 1,0, samt en antagelse om, at cluster B.1.1.7 smitter 50%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

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, stippled 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.

---

<sup>2</sup>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.

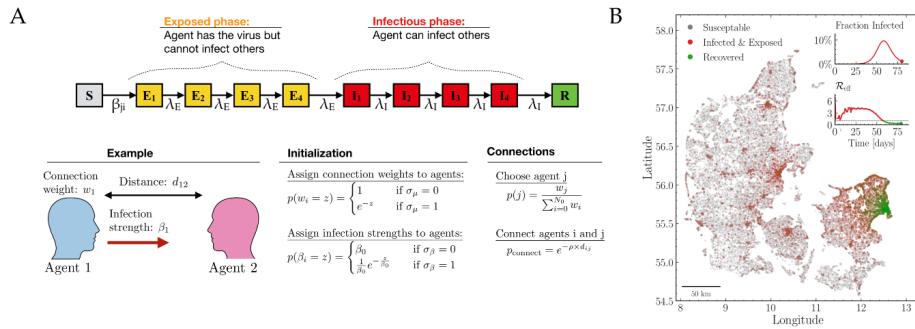


### Bilag 1. Beskrivelse af den agentbaserede model

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 afstandsufhængige 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  $\text{\LaTeX}$  and modified version of the `tufte-style-thesis` class.  
The style is heavily inspired by the works of Edward R. Tufte and Robert Bringhurst.