

UNIVERSITY OF
COPENHAGEN



PH.D. THESIS
by
Christian Michelsen

Biological Data Science

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

Submitted: 30th November 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,
30th November 2022.

Til Blob.

Contents

Preface	i
Abstract	iii
Dansk Resumé	v
Publications	vii
Acknowledgements	ix
1 Introduction	1
1.1 Ancient DNA and Bayesian Statistics	2
1.2 Anesthesiology – a Machine Learning Approach	8
1.3 COVID-19 and Agent Based Models	13
1.4 Diffusion Models and Bayesian Model Comparison	15
Bibliography	18
2 Paper I	25
3 Paper II	95
4 Paper III	133
5 Paper IV	143
APPENDIX	
A Kap København	163
B Explainable ML and Anaemia	179
C SSI Eksperttrapport	191
D SSI Notat	221

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 cross-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 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 and then I present the research in the form of four papers, each of which reflects a different aspect of the research. The introduction is written with my former self in mind, containing the background knowledge I would have liked to know when I started the projects. I hope that it will be useful for anyone interested in the research presented in this thesis.

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 worked for Statens Serum Institut, the Danish Center of Disease Control, 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.

Lastly, in the fourth paper I show how advanced Bayesian methods can be utilized to better estimate the diffusion coefficients in silencing foci in the cell nucleus with single-particle tracking experiments.

Abstract

In recent years, methods such as next generation sequencing in genomics and the use of electronic records in the health care sector has dramatically increased the amount of data in the life sciences. In the field of ancient genomics, newer lab protocols, combined with strict precautions, now allow for the sequencing of ancient environmental DNA millions of years old. In health care, electronic records have allowed for the use of modern machine learning models due to the increased amount of collected data. This has led to a need for new methods and tools to analyze and interpret this vast amount of information that seems to keep increasing in size in the coming years. This thesis focuses on the use cases and potential issues with applying modern statistical and data science related methods on biological data.

The work of this thesis is split into four parts, each with a dedicated paper supporting it. The first paper introduces a novel statistical method that we developed for analysing ancient metagenomic DNA damage. To our knowledge, no prior methods exist which are designed to cover this specific use case in genomics. We show that the work of this project, the metaDMG software, is both faster at ancient DNA damage estimation than existing methods and provides more accurate damage estimates – even at taxonomic levels down to 100 reads. As such, metaDMG is state-of-the-art for ancient DNA damage estimation for both simple and complex ancient genomic datasets.

The second paper presents a machine learning approach to predict medical complications after surgery, in particular knee and hip operations. The use of machine learning in anaesthesiology is still in its infancy, and this work is a first step towards the use of machine learning in this field. We show that modern machine learning models can be used to predict complications after surgery with higher accuracy than classical statistical methods commonly used in the field. Concretely, we find a 9.7% increase in precision and 1.6 percentage points increase in the area-under-ROC-curve metric when using a boosted decision tree compared to logistic regression. We further show how explainability methods can not only be used to better understand the “black box” of machine learning models, and thus the risk predictions themselves, but also help support the doctors in their decision making process.

The third paper describes how spatial heterogeneities affect the fitted predictions of an epidemic curve in the early phase. In collaboration with Statens Serum Institut, the Danish Center for Disease Control, we developed an agent based

model which extends on the classical SIR models often used in epidemiology. This allowed us to model the spread of disease in the Danish population and introduce complex interaction patterns between the agents in the form of heterogeneities based on geographical density. We found that fitting with classical SEIR models overestimate the peak number of infected and the total number of infected by a factor of two if only fitted on an early-stage epidemic.

All living cells share the same DNA, yet the expression of genes differ wildly between cells. The mechanisms regulating gene expressions and the silencing of specific genes are not yet fully understood, however, it is known that the heterogeneous environment in the cell nucleus is a key factor in this. In particular, the silencing and repair foci play an important role. The fourth paper presents the analysis of these foci by analysing the single molecule dynamics using Bayesian inference based on diffusion models. This allow us to extract and quantify the diffusion coefficients of the foci which describe the physical mechanisms of the formation of the foci.

Dansk Resumé

Metoder som næste-generation sekventering i genetik og brugen af elektroniske journaler i sundhedsvæsenet har i løbet af de seneste år drastisk øget mængden af data. Nye laboratorieprotokoller har inden for arkæogenetik nu muliggjort sekventering af DNA som er millioner år gammelt. Med indførslen af elektroniske patientjournaler blev den tilgængelige mængde data øget kraftigt, hvilket har muliggjort brugen af moderne maskinlæringsmodeller. Tilsammen har disse moderne metoder ført til et øget behov for nye værktøjer til at analysere og fortolke denne enorme mængde information – information som ser ud til at fortsætte med at vokse i størrelse i de kommende år. Denne afhandling fokuserer på udviklingen og brugen af moderne statistiske metoder på forskellig biologisk data.

Indholdet af denne afhandling er delt op i fire dele baseret på hver sin artikel. I den første artikel introducerer vi en ny statistisk metode til analyse af DNA-skade i arkæogenetik. Vi er ikke bekendt med nogen tidligere metoder der er designet til at dække dette specifikke anvendelsesområde. Vi viser i artiklen at produktet af vores forskning, metaDMG softwaren, er både hurtigt og præcist til at estimere DNA-skade – selv med kun ganske lidt data (helt ned til kun 100 DNA-sekvenser). Dette viser at metaDMG er et førende værktøj indenfor feltet til estimering af DNA-skade for både simple og komplekse arkæogenetiske datasæt

I den anden artikel præsenterer vi en ny tilgang til at forudsige medicinske komplikationer efter en knæ- eller hofteoperation ved brug af moderne maskinlæringsmodeller. Brugen af maskinlæring er stadig forholdsvis ny indenfor anæstesi og dette er et første skridt i at anvende maskinlæring indenfor dette felt. Vi viser i artiklen at moderne maskinlæringsmetoder kan anvendes til at forudsige medicinske komplikationer med højere præcision end de klassiske metoder der ofte er benyttet inden for feltet. Vi finder en 9,7% forbedring i præcision og 1,6 procentpoint forøgelse i arealet-under-ROC-kurven når man sammenligner maskinlæringsmodellen med en logistisk regression. Vi viser yderligere at metoder relateret til model-forklaring ikke blot kan bruges til at forstå modellens inderste dele, og dermed selve risikoforudsigelserne, men også kan hjælpe lægerne i deres beslutningsprocesser.

Vi beskriver i den tredje artikel hvordan rumlige uensartetheder påvirker de teoretiske forudsigelser af en epidemikurve, hvis man baserer sine forudsigelser på data fra den tidlige fase af en epidemi. Vi udviklede i samarbejde med Statens Serum Institut en agent-baseret model. Denne model var bygget på de klassiske SIR-modeller som ofte er anvendt i epidemiologien. Brugen af agent-baserede modeller

tillod os at modellere spredningen af sygdom i den danske befolkning og introducere komplekse interaktionsmønstre mellem agenterne i form af uensartetheder baseret på geografisk tæthed. Vi fandt at forudsigelser baseret på SIR-lignende modeller overestimerer det maksimale antal samtidig smittede, og det samlede antal smittede, med en faktor to, hvis man kun kigger på data fra den tidlige fase af en epidemi.

Alle levende celler deler det samme DNA, dog er der stor forskel på hvilke gener som hver enkel celle rent faktisk udtrykker. Mekanismerne bag denne genregulering og dæmpningen af specifikke gener er stadig ikke forklaret fuldstændig, men man ved at den fysiske struktur af cellekernen spiller en stor rolle. Især dæmpnings- og reoperationsfokuserne i cellekernen er særdeles vigtige i denne sammenhæng. I den fjerde artikel analyserer vi disse fokuser ved hjælp af Bayesiansk inferens baseret på diffusionsmodeller. Ud fra dette måler vi diffusionskoefficienterne af fokuserne, hvilket kan bruges til at beskrive de fysiske processer som ligger til grund for skabelsen af fokuserne.

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: A Fast and Accurate Ancient DNA Damage Toolkit for Metagenomic Data”. Submitted to *Methods in Ecology and Evolution*.
- Paper 2:** **Christian Michelsen**[†], Christoffer C. Jørgensen[†], 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”. In review at *BMJ Open*.
- Paper 3:** Mathias S. Heltberg[†], **Christian Michelsen**[†], Emil S. Martiny, Lasse E. Christensen, Mogens H. Jensen, Tariq Halasa and Troels C. Petersen (2022). “Spatial Heterogeneity Affects Predictions from Early-Curve Fitting of Pandemic Outbreaks: A Case Study Using Population Data from Denmark”. Published in: *Royal Society Open Science* 9.9. issn: 2054-5703. doi: 10.1098/rsos.220018.
- Paper 4:** Susmita Sridar[†], Mathias S. Heltberg[†], **Christian Michelsen**[†], Judith M. Hattab, Angela Taddei (2022). “Microscopic single molecule dynamics suggest underlying physical properties of the silencing foci”. Unpublished paper draft.

Shared first authorship is indicated with a dagger (†) next to the name.

The appendix contains two papers of which I am a co-author, see Appendix A and Appendix B. The appendix further contains two reports published by Statens Serums Institut that are based on my research during my Ph.D., see Appendix C and Appendix D. These appendices are further explained in Chapter 1.

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 at the time. 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 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 I met during in Trieste; thank you for making my stay in Italy so enjoyable.

I want to thank my friends for always being there for me. A special thanks to my friends from NBI and Borchon who I know I can always count on, whether or not that includes a trip in a 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 they have given me and for the sacrifices 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. And Blob, I cannot wait to meet you!

1 *Introduction*

The primary content of my thesis is the four papers included in the thesis in Chapter 2 to Chapter 5. 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 all the statistical methods and bioinformatic tools used in the papers. The original research motivation supporting the funding of this Ph.D. was multi-disciplinary and the papers included in my thesis are also highly 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 a modified geometric damage profile (exponential decay).

Sometimes the model is not known and the data generation 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 by using an agent-based model it allows for more complex and realistic behaviour of the disease and the transmission process. The model is used to simulate the spread of 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*

The similarity between family members and the degree to which siblings resemble one another has long been a mystery in human history. People have always thought about the balance between nature and nurture, as in the famous fairy tale “The Ugly Duckling” by Hans Christian Andersen from 1843. These questions were addressed two decades later, when Gregor Mendel founded genetics as a modern, scientific discipline with his studies on trait inheritance in pea plants (Mendel, Gregor, 1866).

A century later, a major breakthrough occurred when Watson and Crick discovered the double helix structure of DNA (Watson and Crick, 1953). This led to other important discoveries within genetics, such as the development of DNA sequencing allowing scientists to identify the genetic makeup for a specific cell. Until the mid 1980s, studies within archaeogenetics were limited to analysis of fossilised samples of plants, animals or other species (Parducci and Petit, 2004). Following the first successful recovery of ancient DNA from 5000 year old ancient Mummies, it was shown that it was indeed possible to extract and sequence DNA (Pääbo, 1985a; Pääbo, 1985b). This discovery, along with a dozen other, pushed the boundary for what is scientifically possible with ancient DNA, and 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 (NGS) technology which revolutionized the speed and throughput of genomic sequencing, while decimating the cost (Slatko, Gardner and Ausubel, 2018). This technological advance has led 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). In 2008, the first human genome was sequenced and since then multiple NGS methods have allowed for cheap, high-quality, in-depth sequencing of genetical samples (Genomics and Mobley, 2021). All of this shows, that the field of genetics has grown exponentially and become a central part of modern biology.

Leaving the homocentric world view, aDNA also allows for the study of archaic animals. The age limitation for when aDNA can be sequenced has in the recent years increased; 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 sequen-

cing of entire communities of organisms, so-called metagenomics. By analysing environmental DNA (eDNA) from a set of samples, one can survey the rich plant and animal assemblages of a given area and at a specific time in the past. Our new paper in Nature shows 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 can help to push the boundary of what is possible with ancient DNA.

Ancient DNA is difficult to work with since it often contains only a limited amount of biological material due to bad preservation, leading to low endogenous content with high duplication rates, making high-depth sequencing difficult¹ (Renaud et al., 2019). Here endogenous content refers to DNA from the species of interest and not e.g. ancient bacteria or modern contamination. In addition to this, ancient DNA is often highly degraded. In particular, the two prominent issues with aDNA is fragmentation and deamination (Dabney, Meyer and Pääbo, 2013; Peyrégne and Prüfer, 2020). Fragmentation refers to the fact that through time the DNA is broken into very short fragments, often with a size of less than 50 bp. A consequence of this, upon alignment, is low mapping quality, multimapping, and reference bias, which can somewhat be mitigated by the use variant graphs (Martiniano et al., 2020).

¹ Genotype likelihoods are often used to alleviate the problem of low-coverage data (Nielsen et al., 2011).

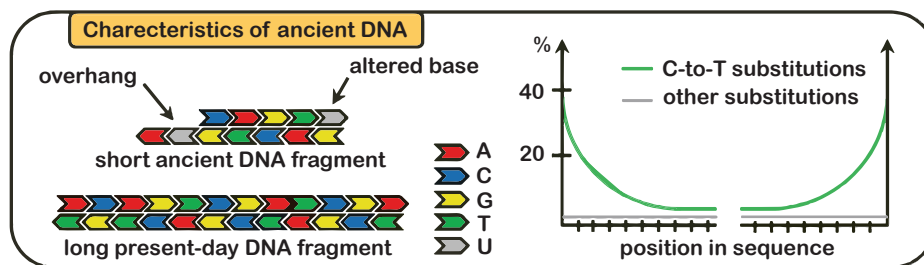


Figure 1. Illustration of DNA damage. Ancient DNA is often highly fragmented with short reads compared to modern, present-day DNA. Due to deamination, aDNA can contain uracils (U), which will be misread as thymines (T) while sequencing, leading to C-to-T nucleotide misincorporations. This is primarily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

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 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). Due to the short fragment sizes in ancient DNA, the fragments will often contain overhangs with over-expressed C-to-T frequency. In the case of single-genome analysis, previous solutions have been to either remove all transitions and only keep transversions, apply trimming at the read ends, or enzymatically remove them with USER treatment (Schubert et al., 2012; Rohland et al., 2015). For an illustration of both fragmentation and deamination of ancient DNA, see Figure 1.

Measuring DNA damage is thus a way to prove authentic aDNA. Currently, a handful of different methods for quantifying ancient DNA damage exist. In particular, the mapDamage software has been the standard for how to measure ancient DNA damage in the field (Jónsson et al., 2013). While mapDamage allows for estimating all of the four Briggs parameters, it is often the empirical deamination patterns that mapDamage computes that are used. Newer and faster methods for estimating ancient DNA damage are continuously being developed, including PyDamage (Borry et al., 2021), which tackles some of mapDamage’s limitations. However, within metagenomics, which studies the genetic material of all organisms collected from an environmental sample, faster methods suited to analyse this large-scale dataset are still lacking.

Paper I, see Chapter 2, introduces the metaDMG software which utilizes the C-to-T deamination pattern² to identify ancient DNA damage. One of the key features of this method is the beta-binomial model which allows the uncertainty of the deamination frequency to be fitted independently of the mean of the frequencies leading to improved accuracy of the damage estimation. The deamination frequencies are based on the number of C-to-T transitions, k , out of the total number of C’s, N , for a given position within the fragment. 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:

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

where p is the probability of success (a C-to-T substitution). One of the issues, however, is that the variance of the binomial distribution is proportional to the mean. The binomial distribution is thus not flexible enough to accommodate large amounts of variance in the data, so-called overdispersion (McElreath, 2020). One way to accommodate overdispersion is to instead use a beta-binomial model. The beta-binomial model is a generalization of the binomial distribution where the variance is independent of the mean. Technically, the beta-binomial model assumes that p is a random variable which follows a beta distribution $p \sim \text{Beta}(\mu, \varphi)$ where the beta distribution is parameterized³ in terms of its mean, μ , and dispersion parameter, φ , (Cepeda-Cuervo and Cifuentes-Amado, 2017). The mean and variance of this beta-binomial model is then given by:

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

² for the forward strand and the G-to-A deamination pattern for the reverse strand.

³ This can be reparameterization in term of the classical α, β parameterization by: $\mu = \alpha/(\alpha + \beta)$ and $\varphi = \alpha + \beta$.

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

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

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

Here A is the scale factor, or amplitude, q is the decay rate, and c is a constant offset. The offset can be interpreted as the baseline C-to-T background substitution rate or baseline damage rate. Since x is discrete, this is similar to a (modified) geometric sequence starting from $x = 1$. The combination of equation (2) and (3) is illustrated in Figure 2, which shows the position-dependent decreasing damage frequency. The figure also shows the increase in uncertainty in the beta-binomial model compared to the binomial-only model.

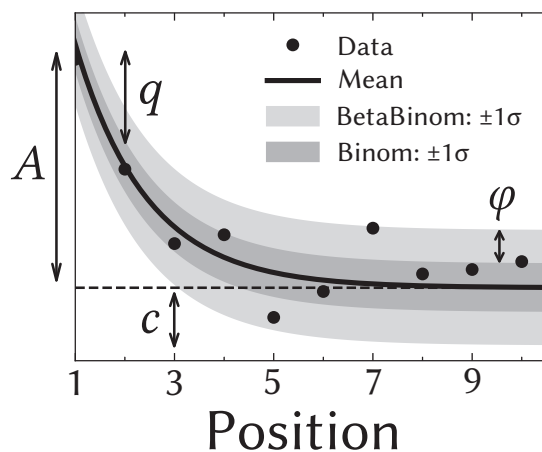
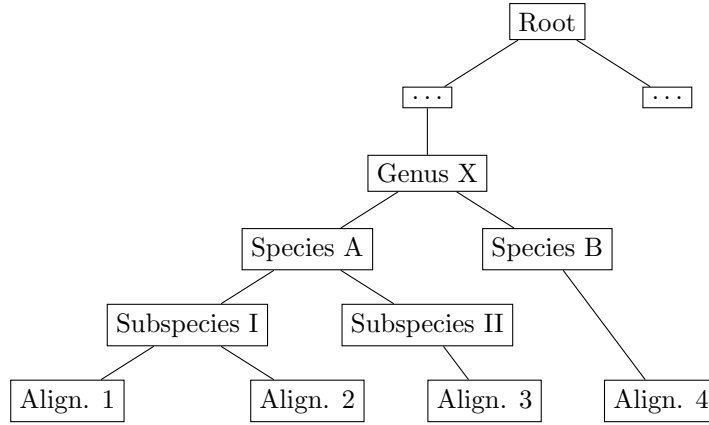


Figure 2. Illustration of the damage model. The figure shows data points as circles and the fitted damage frequency, $y(x)$, as a solid line. The amplitude of the damage is A , the offset is c , and the relative decrease in damage pr. position is given by q . The damage uncertainty for a binomial model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

The damage framework described above is based on the nucleotide misincorporations, i.e. the C-to-T transitions. The background for this data can be from either DNA sequence files mapped to a single genome or from metagenomic data consisting of multiple mapped reads. As such, the damage framework is a general tool for estimating damage based on DNA alignment files.

In the metagenomic case, metaDMG identifies the lowest common ancestor (LCA) based on the algorithm from ngsLCA (Wang et al., 2022). For each read that maps to multiple reference genomes from separate species, i.e. has multiple alignments, the taxonomic tree is traversed for each alignment until a common ancestor is found. Figure 3 illustrates the LCA for a read that maps to different (sub)species. In this example, the LCA of alignment 1 and 2 is the Subspecies I while the LCA for all four alignments is the Genus X. metaDMG works by default with the NCBI taxonomic database but can also be used with custom databases.

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



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

$$P(A \cap B) = P(B)P(A|B), \quad (4)$$

where $P(B)$ is the probability of B and $P(A|B)$ is the conditional probability of A given B . Similarly, $P(A \cap B)$ can also be expressed in terms of the probability of A :

$$P(A \cap B) = P(A)P(B|A). \quad (5)$$

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

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)}, \quad (6)$$

with a change of variables where D refers to the observed data and θ the parameter(s) of the model⁴. The first term in the numerator, $P(\theta)$, is the prior distribution and describes the probability distribution assigned to θ before observing any data. The second term is the likelihood function, $P(D|\theta)$, which is the probability of observing the data given the parameter(s). Together these two terms combine to a compromise between data and prior information.

⁴ In the case of metaDMG, D would be the observed deamination frequencies and θ the four fit parameters.

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

$$P(D) = \int_{\theta} P(D|\theta)P(\theta) d\theta. \quad (7)$$

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

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

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

et al., 2018) under the hood for automatic differentiation and just-in-time (JIT) compilation.

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

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

where the second equality is due to the evidence being independent of θ . Since this is a point estimate, $\hat{\theta}_{\text{MAP}}$ does not fully explain the full posterior, however, it is often a good approximation⁵. Comparing $\hat{\theta}_{\text{MAP}}$ to the maximum likelihood estimate (MLE):

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

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

1.2 Anesthesiology – a Machine Learning Approach

This section explains the technical background behind Paper II, see Chapter 3. This study investigates the potential advantages of using a modern machine-learning model compared to classical logistic regression to predict the risk of patients being re-hospitalized after fast-track hip and knee replacements. In particular, the patients were grouped into two groups. The first group were the so-called “risk-patients” that stayed at least 4 days in the hospital post surgery or were re-hospitalized within 90 days of surgery. The second group were the non-risk-patients. As such, this is a binary classification problem where the patient’s risk-score is predicted based on historical data. The machine learning models were trained on 33 variables, of which 7 were continuous, related to the patient’s medical record, such as age, gender, the use of walking aid, anaemia, diabetes, etc. A total of 22,017 patients were included in the study, of which 1,476 were risk-patients.

⁵ Especially when the posterior is unimodal, which is generally the case for metaDMG.

Most classification and regression problems fall under the same machine learning (ML) branch called supervised learning. In supervised learning, the goal is to find the hypothesis h^* in the hypothesis set \mathcal{H} that matches the unknown, “true” data-generating function $f : \mathcal{X} \rightarrow \mathcal{Y}$ optimally, where \mathcal{X} is the input space and \mathcal{Y} is the output space. Assuming that we have access to realizations of f , the so-called training data $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, we can use a learning algorithm \mathcal{A} combined with the training data to estimate h^* (Abu-Mostafa, Magdon-Ismail and Lin, 2012). Here N refers to the number of training samples and \mathbf{x}_i is the i th observation with the true label y_i . This process is illustrated in Figure 4.

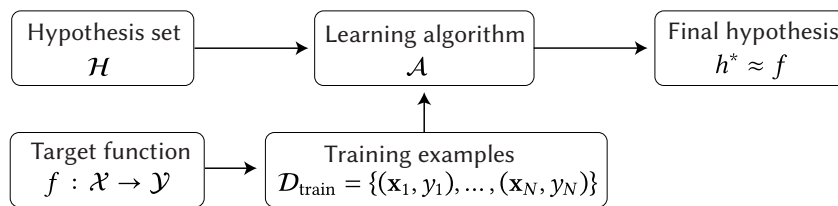


Figure 4. Illustration of how to learn from data in a supervised learning setting. Adapted from (Abu-Mostafa, Magdon-Ismail and Lin, 2012).

Both logistic regression (LR) and ML models can be viewed through the lens of Figure 4, just with $|\mathcal{H}_{\text{LR}}| \ll |\mathcal{H}_{\text{ML}}|$, i.e. the machine learning model is a lot more complex than the logistic regression model and the hypothesis space thus significantly larger. While sufficiently parameterized ML methods can in theory achieve perfect performance on the labelled training set, one is rarely interested in the predictive power of h^* on the training set, as the truth is already known. Instead, one often wish to apply the trained model to new, unseen data where the truth is unknown.

Assessing the performance of h^* on unlabelled data can be difficult. A naive estimate would be to assume that the performance on new, unseen data is the same as on the training data. However, this would likely be a poor estimate due to overfitting and thus bias the predicted performance, especially for high cardinality hypothesis sets. (Abu-Mostafa, Magdon-Ismail and Lin, 2012). The concept of overfitting is illustrated in Figure 5, which shows the training loss as a function of model complexity. The figure shows how more advanced models can achieve lower and lower training losses, however, at some point they start to overfit, leading to higher validation losses. The validation loss is the error on unseen data and is thus the quantity of interest. The goal is to find the sweet spot between underfitting and overfitting.

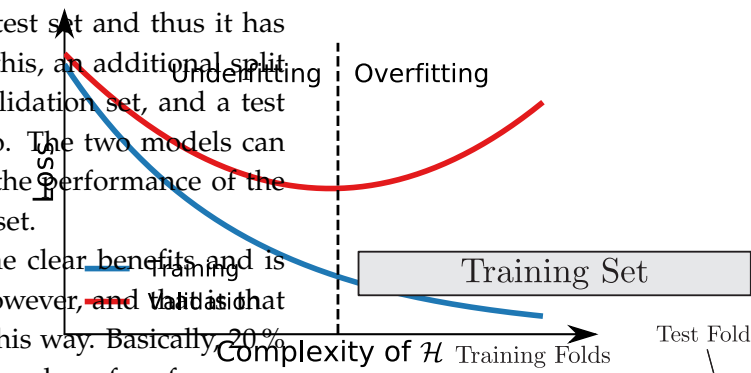
To avoid overfitting and get accurate estimates of the performance of h^* , we use a technique called cross-validation (CV). In the simplest way, this can be done by splitting the data into two sets, one called the training and one called the validation set, and then only train on the training set. Afterwards the trained

is too simple to prop-
 d *underfitting*¹¹. When
 g the inherent noise in
 as a function of model
 trated in Figure 2.2.¹⁰ As

¹¹ Here the error from $\hat{L}(h, S)$ domi-
 nates.

¹² Here the error from $\Omega(N, \mathcal{H}, S)$ domi-
 nates.

loss decreases. Initially,
 "the purity of the test set,
 me point the behavior
 and the loss increases)



This way of splitting up the data has some clear benefits, and is
 as also often used. There is a drawback, however, and that is that
 e are not fully utilizing a lot of the data in this way. Basically, 20 %
 the data are only used to provide a single number of performance

Figure 2.2: Illustration of the empirical

important issues in machine
 learning algorithms have
 erfitting and it thus has
 the issue a number of

loss as a function of n
 The training error is shown in blue and
 the test error in red

Most of them are
 ken advantage of in a
 introduced in subsec-

Figure 2.6: k-fold cross validation.
 Figure 2.6: k-fold cross validation.

and early stopping
 and the variability of the performance
 set and the performance of CV is that the performance estimate

Figure 2.6: k-fold cross validation.
 Figure 2.6: k-fold cross validation.

has some clear benefits, and is
 ggest disadvantage is the computational burden related to doing k-
 fold CV where $k \gg 1$. A compromise often used in applied machine

Figure 2.6: k-fold cross validation.
 Figure 2.6: k-fold cross validation.

single number of performance
 uncertainty, which increases and dis-
 crepancy between the training and test error is often introduced inadvertently.

Figure 2.6: k-fold cross validation.
 Figure 2.6: k-fold cross validation.

regular linear regression
 n of squares written in

Figure 2.6: k-fold cross validation.
 Figure 2.6: k-fold cross validation.

in a single performance num-
 (2.17)

Figure 2.6: k-fold cross validation.
 Figure 2.6: k-fold cross validation.

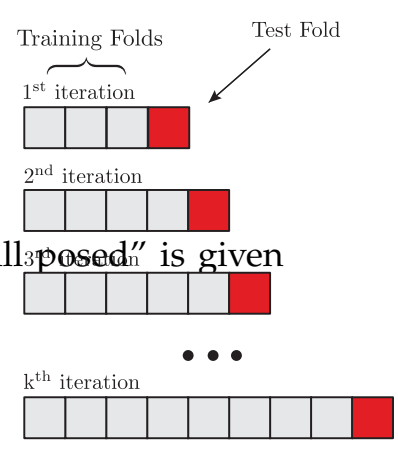
Early stopping
 estimate is
 ect is generally very small. The
 ost modern machine learning models are trained iteratively. This is

Figure 2.6: k-fold cross validation.
 Figure 2.6: k-fold cross validation.

used in this project.
 The model starts off with an
 initial guess of the parameters of the

Figure 2.6: k-fold cross validation.
 Figure 2.6: k-fold cross validation.

²² Special care has to be taken here since the k different performance values are not independent.



(b) Temporal cross validation for time series data.

The actual training of the learning model \mathcal{A} is model-dependent and will not be covered in this thesis. The term training refers of the optimization of the internal parameters in the ML model. In most cases, the training depends on the gradient of the loss function with respect to internal parameters to be computed, see Michelsen, 2020 for a more detailed description of the training process.

Training is not the only way to optimize the performance of \mathcal{A} , albeit it is the primary one. In addition to the internal parameters of the model, some parameters are external to the model in the sense that they are not optimized by the model itself, but rather by the user. These are called hyperparameters and are often optimized using a technique called hyperparameter optimization (HPO). In the case of logistic regression, the number of variables to include would be an example of a hyperparameter; in the case of a decision tree model, the depth of the tree. Hyperparameter optimization can be performed in many ways, where the common one is through grid search, see Figure 7.

In grid search, all combinations of the hyperparameters (the cartesian product) are tested and the best combination is chosen. This is a simple and intuitive approach, however, it scales exponentially with the number of hyperparameters. As such, grid search suffers from the curse of dimensionality. In addition to this, it depends on the user-defined grid, which might not be optimal. To circumvent this, a technique called random search (RS) was developed (Bergstra and Bengio, 2012). Random search is a randomized version of grid search, where the hyperparameters are sampled randomly from a distribution. This allows for a more efficient sampling of the hyperparameter space, see Figure 8. Another advantage is that RS lets the user decide on the number of iterations beforehand.

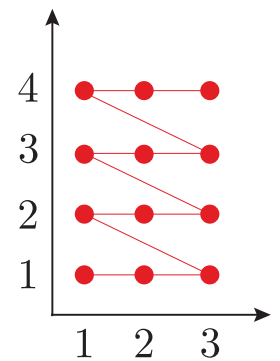


Figure 7. Illustration of grid search. Figure from Michelsen, 2020.

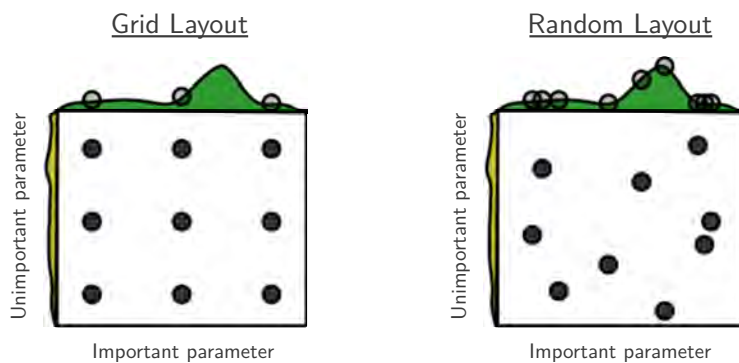


Figure 8. Illustration comparing grid search to random search. The height of green curve is the score-function which has to be optimized. Figure from Bergstra and Bengio, 2012.

Figure 1: Grid and random search of nine trials for optimizing a function $f(x, y) = g(x) + h(y) \approx g(x)$ with low effective dimensionality. Above each square $g(x)$ is shown in green, and left of each square $h(y)$ is shown in yellow. With grid search, nine trials only test $g(x)$ in three distinct places. With random search, all nine trials explore distinct values of g . This failure of grid search is the rule rather than the exception in high dimensional hyper-parameter optimization.

given learning algorithm, looking at several relatively similar data sets (from different distributions) reveals that on different data sets, different subspaces are important, and to different degrees. A grid with sufficient granularity to optimizing hyper-parameters for all data sets must consequently be inefficient for individual data sets, and a more efficient method is to sample a subset of

black with uncertainty shown in blue. This is a result of fitting GPs to the two previous points, $t = 2$. This surrogate function is supposed to fit the unknown hyperparameter-dependent evaluation function (called objective in the figure) shown as a dashed black line. Below we see the acquisition function in green. This is a function of the blue curve and the position of its maximum decides where the next guess of λ should be. With the chosen acquisition function and exploration willingness, we see that the next guess should be slightly to the left of the right-most point. This is a simple 1D toy problem, but one should imagine this happening in a high-dimensional space. After making a new guess, Bayesian optimization (Brochu, Cora and Freitas, 2010) function changes since it learnt that this gave a worse evaluation value than the right-most point. Therefore, the next proposal for λ is slightly to the right of the right-most point. The process continues in the performance of the model. This is illustrated in Figure 9. This leaves the user with the task of choosing between “exploitation” and “exploration” of the hyperparameter space in the definition of the acquisition function, yet most implementations of bayesian optimization have decent default settings.

Figure 9.

Illustration of the learning process of Bayesian optimization. The previous observations are shown as black dots and the true objective function is shown as a dashed black line. This line is fitted with Gaussian processes which is shown as the solid line with its uncertainty in purple. The acquisition function is shown in green and its maximum decides what the next iteration of the hyperparameter value(s) should be (Michelsen, 2020).

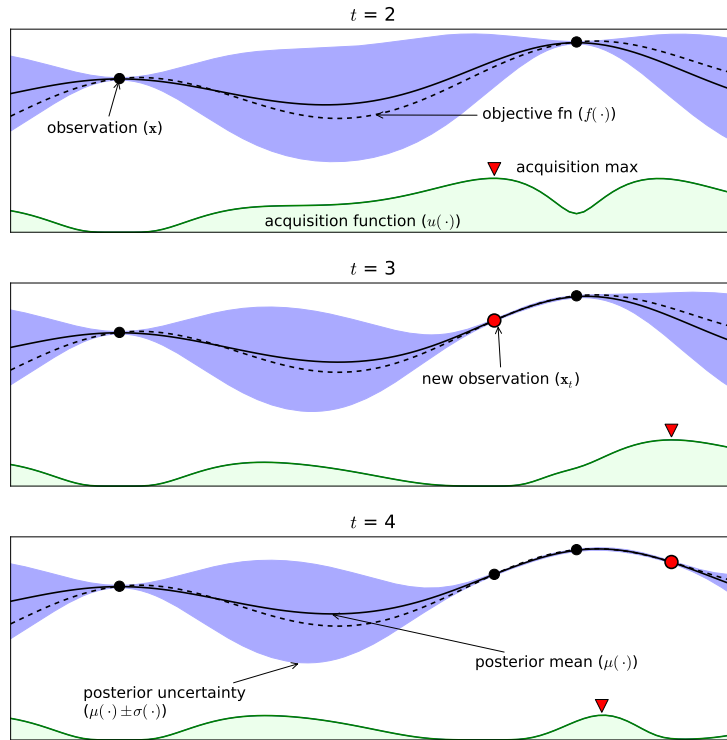


Figure 2.13: Illustration of the learning process of Bayesian optimization. The previous observations are shown as black dots and the true objective function is shown as a dashed black line. This line is fitted with Gaussian processes (GPs) which is shown as the solid line with its uncertainty in purple. The acquisition function is shown in green and its maximum decides what the next iteration of the hyperparameter value should be. Adapted from Brochu [28].

We use the Python package Optuna (Akiba et al., 2019) for HPO in Paper II due to its ease of use and its support for Bayesian optimization. In particular, we use the Tree-structured Parzen Estimator algorithm for the Bayesian optimization and a median stopping rule to minimize optimization time (Bergstra, Bardenet et al., 2011). This allowed for a good compromise between optimization time and performance.

While model performance is often paramount, in some fields – such as medicine – being able to explain the model’s predictions is almost as important. This is especially true in the case of medical decision support systems, where the model is used to make decisions about the patient’s treatment. Model explainability helps to build trust in the model, for both the patient and the medical staff alike.

In Paper II, we employ the SHapley Additive exPlanations (SHAP) values which provide estimates on which variables contribute most to the risk score predictions (Scott M Lundberg and Lee, 2017; Scott M. Lundberg, Erion et al., 2020). SHAP values allow for not only a global explanation of the model, i.e. which features are most important generally, but also a local explanation, i.e. which features led to a single patient being predicted at risk of being re-hospitalized. It has previously been shown that the interaction between SHAP values and medical doctors can improve the performance of anaesthesiologists (Scott M. Lundberg, Nair et al., 2018).

While the aim of Paper II is to show how modern machine learning techniques can be used to improve the risk prediction process, the usefulness of the SHAP values in a medical context is demonstrated in our paper in Appendix B. The paper uses the SHAP values to compare the preoperative haemoglobin level in the patient with the risk-score, stratified by sex and operation type (knee vs. hip replacement). Currently, the WHO guidelines for the haemoglobin levels are gender specific, however, our study finds no significant gender difference and a haemoglobin threshold close to the WHO suggestions for men (Anaemias and Organization, 1968).

1.3 *COVID-19 and Agent Based Models*

In early 2020, a contagious disease called COVID-19 started to spread in Europe, including Denmark. With new infections showing up faster and faster, governments started to implement different measures to limit the spread of the contagious disease, including lockdowns, travel restrictions, and social distancing, measures not previously seen in peacetime since the Spanish flu in 1918. This was the background for the work that we did in 2020 which became the basis for Paper III, see Chapter 4. This paper deals with the development of a new agent based model for COVID-19 in Denmark in collaboration with Statens Serum Institut (SSI), the Danish Center for Disease Control.

Historically, most mathematical models of infectious diseases were variations of the SIR model, which describe the evolution of a pandemic by approximating all individuals as one population (Kermack, McKendrick and Walker, 1927). As one of the simplest compartmental models, the susceptible-infectious-recovered (SIR) model is based on a system of three non-linear differential equations that describe the transition between each state, or compartment, of the model (Kröger and Schlickeiser, 2020). Initially the entire population is susceptible. At $t = 0$ an outbreak happens where some number of random agents are infected and become infectious, allowing the disease to spread. After having been infectious, the in-

dividuals recover and become immune to the disease and stop being infectious. Several variations of the SIR model exist, including the SIS model, where the recovered individuals become susceptible again (Hethcote, 1989). Another variation is the SEIR model, which includes an exposed state, where individuals are infected but not yet infectious, which is the basis for the model used in Paper III.

SIR-like models suffer from several shortcomings, including the assumptions that the population is homogeneous, and that agents are equally infectious throughout their infectious period. In reality, neither the population nor the transmission rates are homogeneous. While multistage SEIR and multicompartment models can help mitigate some of the issues none of these can handle the geographical interactions between agents, which is why we chose to develop an agent based model (ABM) (Tang et al., 2020; Wu et al., 2022). Agent based models simulate individual agents in a population in a way that allows for complex interaction patterns, e.g. based on geographical features such as agent density (Wilensky and Rand, 2015).

In particular, we implemented an event-based, stochastic, spatial ABM using the Gillespie algorithm, a stochastic simulation algorithm (Gillespie, 1977). The model is JIT compiled with Numba (Lam, Pitrou and Seibert, 2015) to speed up the simulation, allowing the simulation of the Danish population of 5.8 million people in a couple of hours instead of days. The model allows for the individual tuning of the three main effects; A) heterogeneities in the infection strength⁶, B) heterogeneities in the number of connections⁷, C) and the spatial clustering of the agents. In the absence of any of these effects, we find that the ABM's predictions matches the SEIR model's predictions within $\pm 5\%$. Once we allowed for spatial clustering, we found that the epidemic developed faster and with a higher infection peak compared to the SEIR model, but that the total number of infected in the end of the epidemic was lower.

In real-life scenarios, one does not have the opportunity to let the epidemic run loose and afterwards evaluate the strength of the epidemic; the goal is to predict the intensity in the very beginning of the epidemic and implement lockdown-related measures based on this estimate. In the second part of Paper III, we show that once spatial clustering is introduced, fitting standard SEIR-models to infection numbers from the first few days of the epidemic, predictions are overestimated by a factor of two. The result is a significant over-estimation of the impact of the epidemic, in particular the reproduction number \mathcal{R}_0 and thus also the number of infected, both the maximal number of simultaneously infected and the endemic steady state number of infected. Since the population is highly susceptible in the beginning of an epidemic, this also highlights the benefits of early lockdowns to reduce the effect of the super-connectors.

⁶ allowing *super-shedders*

⁷ allowing *super-connectors*

The developed ABM was further used by SSI to estimate the effect of contact tracing related to COVID-19 in Denmark, see Appendix C. It was further used to estimate spread of the “alpha” variant of COVID-19 (B.1.1.7) in Denmark, see Appendix D. Based on data available January 2nd 2021, the model predicted that the “alpha” variant would be the dominant variant in Denmark February 10–20, 2021. It became the dominant variant in Week 7: February 15–21, 2021 (Bager et al., 2021).

1.4 Diffusion Models and Bayesian Model Comparison

While Section 1.1 discusses the behaviour of ancient DNA, Paper IV focusses on how living cells work and, in particular, how they regulate the transcription of DNA in the cell nucleus. Despite the fact that all cells share the same DNA, the regulation and expression of the genes stored within can vary. The mechanism of the cell-specific expression and silencing of specific genomic regions are one of the most fundamental biological challenges.

Currently, different biological models try to explain the physical principles creating the heterogeneous environment in the cell nucleus of eukaryotic cells. One of these is the polymer-bridging model (PBM) that models the micro compartments called the foci. The cell nucleus contains two different types of loci; the repair foci and the silencing foci. Paper IV studies the physical mechanism of the formation of the silencing foci.

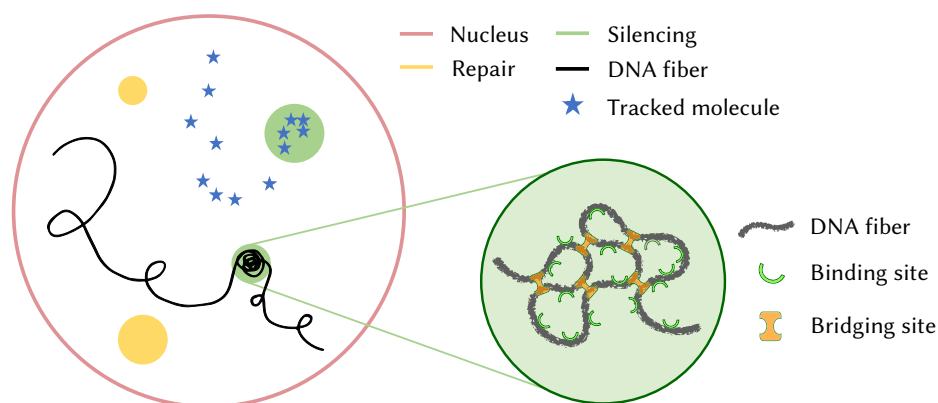


Figure 10. Illustration of the cell nucleus. The nucleus membrane is shown in red and the repair foci in yellow. The black line represents the DNA fiber which is curled up in the silencing foci in green. The right side of the figure shows a zoomed in view of the silencing foci according to the polymer-bridging model with the binding and bridging sites that interact with the SIR proteins. The tracking of the SIR proteins is shown as blue stars. Partly adapted from (Heltberg et al., 2021).

Figure 10 illustrates the parts of the cell nucleus relevant to the polymer-bridging model. Inside the nucleus, DNA fibers are curled up and some parts of the DNA locate inside the silencing foci. Inside the silencing foci, the PBM predicts binding and bridging sites that interact with the DNA fiber through the SIR proteins, which are up-regulated inside the the region of the foci (Heltberg

et al., 2021). The silent Information Regulator (SIR) proteins repress the underlying genes, and, due to the increased concentration inside the focus, the foci are termed silencing foci.

With the use of single particle tracking and photoactivated localization microscopy, it is possible to track the individual SIR protein at high temporal and spatial resolution (Manley et al., 2008; Oswald et al., 2014). As the SIR proteins are assumed to follow a diffusion process, the tracking allows for the determination of the diffusion coefficients of cell nucleus, which help quantify the heterogeneous structure in the nucleus.

Assuming classical Brownian motion in 2D, the displacement lengths, Δr_i , defined as the distances between subsequent observations \vec{x} :

$$\Delta r_i = \|\vec{x}_{i+1} - \vec{x}_i\|, \quad (10)$$

follows a Rayleigh distribution:

$$\text{Rayleigh}(r; \sigma) = \frac{r}{\sigma^2} e^{-r^2/(2\sigma^2)} \quad r > 0, \quad (11)$$

with scale parameter $\sigma = \sqrt{2d\tau}$, where d is the diffusion coefficient and τ is the time between observations (Anderson et al., 1992). Using Bayesian mixture models, the switch diffusion process is a simple model describing the system, (Baker, 2021). With $K = 2$ diffusion states, Figure 11 illustrates the model in directed factor graph notation (Dietz, 2022). It shows how the two diffusion coefficients, d_1 and d_2 , each define their own Rayleigh distribution, \mathcal{R}_k , which are then combined to a mixture distribution, $\mathcal{R}_{1,2}$, with mixing probabilities $\vec{\theta}$. The measured data, Δr , are modelled as N realisations from this mixture distribution.

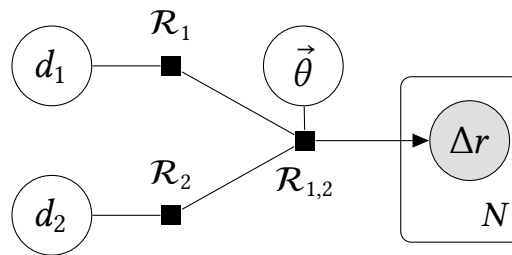


Figure 11.

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 $\vec{\theta}$ prior.

The diffusion model illustrated in Figure 11 with $K = 2$ diffusion states can be extended to K states, where data shows that both a simpler $K = 1$ model (K_1), the $K = 2$ model (K_2), and a more advanced model with $K = 3$ diffusion states (K_3), all yields appropriate results. Remembering that the formation of the foci depends on the physical properties of the cell nucleus, it is important to be able to evaluate the different models since they provide different diffusion estimates.

The models are compared using the Widely Applicable Information Criterion (WAIC) (Watanabe, 2010) which is a generalized version of the Akaike information criterion (AIC), useful for Bayesian model comparison (Gelman, Hwang and Vehtari, 2014). The WAIC is an approximation of the out-of-sample loss of the model and is defined as:

$$\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}}), \quad (12)$$

where the log-pointwise-predictive-density (lppd) is a Bayesian version of the accuracy of the model and p_{WAIC} is a penalty term that penalizes the model for the effective number of parameters (McElreath, 2020). To compare two models, the model with the lowest WAIC is preferred, however, the difference between the WAICs should also be considered. The results for the WT1 dataset from Paper IV is shown in Figure 12. This figure shows the WAIC in black for the K_1 , K_2 and K_3 models along with their uncertainties and it is easily seen that the model with only a single diffusion component does not perform well. The difference between the WAIC of the model and the best performing model (K_3) is shown in grey, $\Delta_{A,B}$, where the z -value above the error bars are the number of sigmas the difference is from zero:

$$z = \frac{\Delta_{A,B}}{\sigma_{\Delta_{A,B}}}. \quad (13)$$

Following Occam's razor, the K_2 model is chosen as the optimal model, since the difference between the K_2 model and the K_3 model, the best performing one, is statistically non-significant ($z = 0.57 < 2$).

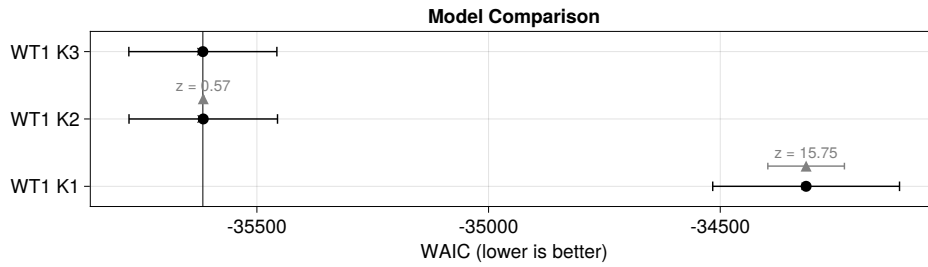


Figure 12.

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.

Bibliography

- Abu-Mostafa, Yaser S., Malik Magdon-Ismael and Hsuan-Tien Lin (2012). *Learning From Data*. S.l.: AMLBook. 213 pp. ISBN: 978-1-60049-006-4.
- Akiba, Takuya et al. (2019). 'Optuna: A Next-generation Hyperparameter Optimization Framework'. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. New York, NY, USA: Association for Computing Machinery, pp. 2623–2631. ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330701.
- Anaemias, WHO Scientific Group on Nutritional and World Health Organization (1968). *Nutritional Anaemias : Report of a WHO Scientific Group [Meeting Held in Geneva from 13 to 17 March 1967]*.
- Anderson, C.M. et al. (1992). 'Tracking of Cell Surface Receptors by Fluorescence Digital Imaging Microscopy Using a Charge-Coupled Device Camera. Low-density Lipoprotein and Influenza Virus Receptor Mobility at 4 Degrees C'. In: *Journal of Cell Science* 101.2, pp. 415–425. ISSN: 0021-9533. DOI: 10.1242/jcs.101.2.415.
- Bager, Peter et al. (2021). 'Risk of Hospitalisation Associated with Infection with SARS-CoV-2 Lineage B.1.1.7 in Denmark: An Observational Cohort Study'. In: *The Lancet Infectious Diseases* 21.11, pp. 1507–1517. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(21)00290-5.
- Baker, Lewis R. (2021). 'Inference of Diffusion Coefficients from Single Particle Trajectories'. PhD thesis. University of Colorado, Boulder. 71 pp.
- Barlow, R. J. (1993). *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*. Chichester, England ; New York: Wiley. 222 pp. ISBN: 978-0-471-92295-7.
- Bergstra, James, Rémi Bardenet et al. (2011). 'Algorithms for Hyper-Parameter Optimization'. In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc.
- Bergstra, James and Yoshua Bengio (2012). 'Random Search for Hyper-Parameter Optimization'. In: *Journal of Machine Learning Research* 13.10, pp. 281–305.
- Betancourt, Michael (2018). 'A Conceptual Introduction to Hamiltonian Monte Carlo'. arXiv: 1701.02434 [stat].
- Bingham, Eli et al. (2019). 'Pyro: Deep Universal Probabilistic Programming'. In: *Journal of Machine Learning Research* 20, 28:1–28:6.

- Borry, Maxime et al. (2021). ‘PyDamage: Automated Ancient Damage Identification and Estimation for Contigs in Ancient DNA de Novo Assembly’. In: *PeerJ* 9, e11845. ISSN: 2167-8359. DOI: 10.7717/peerj.11845.
- Bradbury, James et al. (2018). *JAX: Composable Transformations of Python NumPy Programs*. Version 0.2.5.
- 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.
- Brochu, Eric, Vlad M. Cora and Nando de Freitas (2010). *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*. DOI: 10.48550/arXiv.1012.2599. arXiv: 1012.2599 [cs].
- Carpenter, Bob et al. (2017). ‘Stan: A Probabilistic Programming Language’. In: *Journal of statistical software* 76.1.
- Cepeda-Cuervo, Edilberto and MARÍA VICTORIA Cifuentes-Amado (2017). ‘Double Generalized Beta-Binomial and Negative Binomial Regression Models’. In: *Revista Colombiana de Estadística* 40.1, pp. 141–163. ISSN: 0120-1751. DOI: 10.15446/rce.v40n1.61779.
- Dabney, Jesse, Matthias Meyer and Svante Pääbo (2013). ‘Ancient DNA Damage’. In: *Cold Spring Harbor Perspectives in Biology* 5.7, a012567. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a012567. pmid: 23729639.
- Dembinski, Hans et al. (2021). *Scikit-Hep/Iminuit: V2.8.2*. Version v2.8.2. Zenodo. DOI: 10.5281/ZENODO.3949207.
- 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’. In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1682–1690.
- Gelman, Andrew, John B. Carlin et al. (2015). *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC. 675 pp. ISBN: 978-0-429-11307-9. DOI: 10.1201/b16018.
- Gelman, Andrew, Jessica Hwang and Aki Vehtari (2014). ‘Understanding Predictive Information Criteria for Bayesian Models’. In: *Statistics and Computing* 24.6, pp. 997–1016. ISSN: 1573-1375. DOI: 10.1007/s11222-013-9416-2.
- Genomics, Front Line and Immy Mobley (2021). *A Brief History of Next Generation Sequencing (NGS)*. Front Line Genomics. URL: <https://frontlinegenomics.com/a-brief-history-of-next-generation-sequencing-ngs/> (visited on 2022).

- Gillespie, Daniel T. (1977). 'Exact Stochastic Simulation of Coupled Chemical Reactions'. In: *The Journal of Physical Chemistry* 81.25, pp. 2340–2361. ISSN: 0022-3654. DOI: 10.1021/j100540a008.
- Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2016). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7.
- Heltberg, Mathias L et al. (2021). 'Physical Observables to Determine the Nature of Membrane-Less Cellular Sub-Compartments'. In: *eLife* 10. Ed. by Agnese Seminara, José D Faraldo-Gómez and Pierre Ronceray, e69181. ISSN: 2050-084X. DOI: 10.7554/eLife.69181.
- Hethcote, Herbert W. (1989). 'Three Basic Epidemiological Models'. In: *Applied Mathematical Ecology*. Ed. by Simon A. Levin, Thomas G. Hallam and Louis J. Gross. Biomathematics. Berlin, Heidelberg: Springer, pp. 119–144. ISBN: 978-3-642-61317-3. DOI: 10.1007/978-3-642-61317-3_5.
- Homan, Matthew D. and Andrew Gelman (2014). 'The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo'. In: *The Journal of Machine Learning Research* 15.1, pp. 1593–1623. ISSN: 1532-4435.
- Jónsson, Hákon et al. (2013). 'mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters'. In: *Bioinformatics* 29.13, pp. 1682–1684. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btt193.
- Karolinska Institutet, The Nobel Assembly at (2022). *The Nobel Prize in Physiology or Medicine 2022*. NobelPrize.org. URL: <https://www.nobelprize.org/prizes/medicine/2022/press-release/> (visited on 2022).
- Kermack, William Ogilvy, A. G. McKendrick and Gilbert Thomas Walker (1927). 'A Contribution to the Mathematical Theory of Epidemics'. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 115.772, pp. 700–721. DOI: 10.1098/rspa.1927.0118.
- 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.
- Kröger, M and R Schlickeiser (2020). 'Analytical Solution of the SIR-model for the Temporal Evolution of Epidemics. Part A: Time-Independent Reproduction Factor'. In: *Journal of Physics A: Mathematical and Theoretical* 53.50, p. 505601. ISSN: 1751-8113, 1751-8121. DOI: 10.1088/1751-8121/abc65d.
- Lam, Siu Kwan, Antoine Pitrou and Stanley Seibert (2015). 'Numba: A LLVM-based Python JIT Compiler'. In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*. LLVM '15. New York, NY, USA: Association for Computing Machinery, pp. 1–6. ISBN: 978-1-4503-4005-2. DOI: 10.1145/2833157.2833162.

- Lundberg, Scott M and Su-In Lee (2017). 'A Unified Approach to Interpreting Model Predictions'. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.
- Lundberg, Scott M., Gabriel Erion et al. (2020). 'From Local Explanations to Global Understanding with Explainable AI for Trees'. In: *Nature Machine Intelligence* 2.1 (1), pp. 56–67. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0138-9.
- Lundberg, Scott M., Bala Nair et al. (2018). 'Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia during Surgery'. In: *Nature Biomedical Engineering* 2.10 (10), pp. 749–760. ISSN: 2157-846X. DOI: 10.1038/s41551-018-0304-0.
- Manley, Suliana et al. (2008). 'High-Density Mapping of Single-Molecule Trajectories with Photoactivated Localization Microscopy'. In: *Nature Methods* 5.2 (2), pp. 155–157. ISSN: 1548-7105. DOI: 10.1038/nmeth.1176.
- Martiniano, Rui et al. (2020). 'Removing Reference Bias and Improving Indel Calling in Ancient DNA Data Analysis by Mapping to a Sequence Variation Graph'. In: *Genome Biology* 21.1, p. 250. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02160-7.
- McElreath, Richard (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd ed. CRC Texts in Statistical Science. Boca Raton: Taylor and Francis, CRC Press. ISBN: 978-0-367-13991-9.
- Mendel, Gregor (1866). *Versuche Über Pflanzen-Hybriden*. Brünn, Im Verlage des Vereines, 1866, p. 464.
- Michelsen, Christian (2020). 'A Physicist's Approach to Machine Learning – Understanding the Basic Bricks'. University of Copenhagen.
- Mullis, K. et al. (1986). 'Specific Enzymatic Amplification of DNA in Vitro: The Polymerase Chain Reaction'. In: *Cold Spring Harbor Symposia on Quantitative Biology* 51 Pt 1, pp. 263–273. ISSN: 0091-7451. DOI: 10.1101/sqb.1986.051.01.032. pmid: 3472723.
- Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press. ISBN: 0-262-01802-0.
- Neal, Radford M. (2011). *MCMC Using Hamiltonian Dynamics*. Routledge Handbooks Online. ISBN: 978-1-4200-7941-8 978-1-4200-7942-5. DOI: 10.1201/b10905-7.
- Nielsen, Rasmus et al. (2011). 'Genotype and SNP Calling from Next-Generation Sequencing Data'. In: *Nature reviews. Genetics* 12.6, pp. 443–451. ISSN: 1471-0056. DOI: 10.1038/nrg2986. pmid: 21587300.
- Orlando, Ludovic et al. (2013). 'Recalibrating Equus Evolution Using the Genome Sequence of an Early Middle Pleistocene Horse'. In: *Nature* 499.7456 (7456), pp. 74–78. ISSN: 1476-4687. DOI: 10.1038/nature12323.

- Oswald, Felix et al. (2014). 'Imaging and Quantification of Trans-Membrane Protein Diffusion in Living Bacteria'. In: *Physical Chemistry Chemical Physics* 16.25, pp. 12625–12634. ISSN: 1463-9084. DOI: 10.1039/C4CP00299G.
- Pääbo, Svante (1985a). 'Molecular Cloning of Ancient Egyptian Mummy DNA'. In: *Nature* 314.6012 (6012), pp. 644–645. ISSN: 1476-4687. DOI: 10.1038/314644a0.
- (1985b). 'Preservation of DNA in Ancient Egyptian Mummies'. In: *Journal of Archaeological Science* 12.6, pp. 411–417. ISSN: 0305-4403. DOI: 10.1016/0305-4403(85)90002-0.
- Parducci, Laura and Rémy J. Petit (2004). 'Ancient DNA: Unlocking Plants' Fossil Secrets'. In: *The New Phytologist* 161.2, pp. 335–339. ISSN: 0028646X, 14698137. JSTOR: 1514319.
- Peyrègne, Stéphane and Kay Prüfer (2020). 'Present-Day DNA Contamination in Ancient DNA Datasets'. In: *BioEssays* 42.9, p. 2000081. ISSN: 1521-1878. DOI: 10.1002/bies.202000081.
- Phan, Du, Neeraj Pradhan and Martin Jankowiak (2019). 'Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro'. arXiv: 1912.11554 [cs, stat].
- Renaud, Gabriel et al. (2019). 'Authentication and Assessment of Contamination in Ancient DNA'. In: *Ancient DNA: Methods and Protocols*. Ed. by Beth Shapiro et al. Methods in Molecular Biology. New York, NY: Springer, pp. 163–194. ISBN: 978-1-4939-9176-1. DOI: 10.1007/978-1-4939-9176-1_17.
- Rohland, Nadin et al. (2015). 'Partial Uracil-DNA-glycosylase Treatment for Screening of Ancient DNA'. In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370.1660, p. 20130624. ISSN: 1471-2970. DOI: 10.1098/rstb.2013.0624. pmid: 25487342.
- Schubert, Mikkel et al. (2012). 'Improving Ancient DNA Read Mapping against Modern Reference Genomes'. In: *BMC Genomics* 13, p. 178. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-178. pmid: 22574660.
- Slatko, Barton E., Andrew F. Gardner and Frederick M. Ausubel (2018). 'Overview of Next Generation Sequencing Technologies'. In: *Current protocols in molecular biology* 122.1, e59. ISSN: 1934-3639. DOI: 10.1002/cpmb.59. pmid: 29851291.
- Tang, Lu et al. (2020). 'A Review of Multi-Compartment Infectious Disease Models'. In: *International Statistical Review* 88.2, pp. 462–513. ISSN: 1751-5823. DOI: 10.1111/insr.12402.
- Tashman, Leonard J. (2000). 'Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review'. In: *International Journal of Forecasting* 16.4, pp. 437–450. ISSN: 0169-2070.

- Van der Valk, Tom et al. (2021). 'Million-Year-Old DNA Sheds Light on the Genomic History of Mammoths'. In: *Nature* 591.7849 (7849), pp. 265–269. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03224-9.
- Wang, Yucheng et al. (2022). 'ngsLCA—A Toolkit for Fast and Flexible Lowest Common Ancestor Inference and Taxonomic Profiling of Metagenomic Data'. In: *Methods in Ecology and Evolution* n/a.n/a. ISSN: 2041-210X. DOI: 10.1111/2041-210X.14006.
- 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.
- Watson, J. D. and F. H. C. Crick (1953). 'Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid'. In: *Nature* 171.4356 (4356), pp. 737–738. ISSN: 1476-4687. DOI: 10.1038/171737a0.
- Wilensky, Uri and William Rand (2015). *An Introduction to Agent-Based Modeling*. The MIT Press. ISBN: 978-0-262-73189-8. JSTOR: j.ctt17kk851.
- Wu, Zhuang et al. (2022). 'A Multistage Time-Delay Control Model for COVID-19 Transmission'. In: *Sustainability* 14.21 (21), p. 14657. ISSN: 2071-1050. DOI: 10.3390/su142114657.

2 *Paper I*

The following 68 pages contain the paper:

Christian Michelsen, Mikkel W. Pedersen, Antonio Fernandez-Guerra, Lei Zhao, Troels C. Petersen, Thorfinn S. Korneliussen (2022). “metaDMG: A Fast and Accurate Ancient DNA Damage Toolkit for Metagenomic Data”. Submitted to *Methods in Ecology and Evolution*.

metaDMG – A Fast and Accurate Ancient DNA Damage Toolkit for Metagenomic Data

✉ For correspondence:

christianmichelsen@gmail.com

(CM); mwpedersen@sund.ku.dk

(MW);

tskorneliussen@sund.ku.dk

(TSK).

[†]Authors contributed equally.

Data availability: The source code for metaDMG is available on [Zenodo](#) or at the [Github](#) repository. All code used in the statistical analysis can be found at the following DOI:

[10.5281/zenodo.7368194](https://doi.org/10.5281/zenodo.7368194).

Sequencing data and supporting material used in simulations can be found at [ERDA](#).

Funding: CM and TP is funded by the Lundbeck Foundation.

MWP is funded by the ERC project LASTJOURNEY (ERC_Adv_834514). TSK is funded by Carlsberg grant CF19-0712.LZ. was funded by Lundbeck Foundation Centre for Disease Evolution: R302-2018-2155

Competing interests: The author declare no competing interests.

⁴ Christian Michelsen ^{1,2 †} ✉, Mikkel Winther Pedersen ^{2 †} ✉, Antonio Fernandez-Guerra ² ✉, Lei Zhao², Troels C. Petersen ¹ ✉, Thorfinn Sand Korneliussen ² ✉ ✉

¹Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark; ²GLOBE, Section for Geogenetics, Øster Voldgade 5-7, 1350, Copenhagen, Denmark

Abstract

- Motivation** Under favourable conditions DNA molecules can persist for hundreds of thousands of years. Such genetic remains make up invaluable resources to study past assemblages, populations, and even the evolution of species. However, DNA is subject to degradation, and hence over time decrease to ultra low concentrations which makes it highly prone to contamination by modern sources. Strict precautions are therefore necessary to ensure that DNA from modern sources does not appear in the final data is authenticated as ancient. The most generally accepted and widely applied authenticity for ancient DNA studies is to test for elevated deaminated cytosine residues towards the termini of the molecules: DNA damage. To date, this has primarily been used for single organisms and recently for read assemblies, however, these methods are not applicable for estimating DNA damage for ancient metagenomes with tens and even hundreds of thousands of species.
- Methods** We present metaDMG, a novel framework and toolkit that allows for the estimation,

quantification and visualization of postmortem damage for single reads, single genomes
26 and even metagenomic environmental DNA by utilizing the taxonomic branching structure.
It bypasses any need for initial classification, splitting reads by individual organisms, and
28 realignment. We have implemented a Bayesian approach that combines a modified
geometric damage profile with a beta-binomial model to fit the entire model to the
30 individual misincorporations at all taxonomic levels.

3. **Results** We evaluated the performance using both simulated and published environmental
32 DNA datasets and compared to existing methods when relevant. We find *metaDMG* to be an
order of magnitude faster than previous methods and more accurate – even for complex
34 metagenomes. Our simulations show that *metaDMG* can estimate DNA damage at taxonomic
levels down to 100 reads, that the estimated uncertainties decrease with increased number
36 of reads and that the estimates are more significant with increased number of C to T
misincorporations.

38 4. **Conclusion** *metaDMG* is a state-of-the-art program for aDNA damage estimation and allows
for the computation of nucleotide misincorporation, GC-content, and DNA fragmentation
40 for both simple and complex ancient genomic datasets, making it a complete package for
ancient DNA damage authentication.

42 **keywords:** [ancient DNA](#), [DNA damage estimation](#), [DNA damage](#), [metaDMG](#), [metagenomics](#).

44 1 | INTRODUCTION

Throughout the life of an organism it contaminates its environment with DNA, cells, or tissue, thus
46 leaving genetic traces behind. As the cell leaves its host, DNA repair mechanisms stops and the DNA
is subjected to intra and extra cellular enzymatic, chemical, and mechanical degradation, resulting
48 in fragmentation and molecular alterations that over time lead to the characteristics of ancient
DNA (Briggs et al., 2007; Dabney, Meyer, and Pääbo, 2013). Ancient DNA (aDNA) has been shown
50 to persist in a diverse variety of environmental contexts, e.g. within fossils such as bones, soft-
tissue, and hair, as well as in geological sediments, archaeological layers, ice-cores, permafrost soil
52 for hundreds of thousands of years (Valk et al., 2021; Zavala et al., 2021). Common for all is that
they have an accumulated amount of deaminated cytosines towards the termini of the DNA strand,

54 which, when amplified, results in misincorporations of thymines on the cytosines (Ginolhac et al.,
2011; Dabney, Meyer, and Pääbo, 2013).

56 Even though postmortem DNA damage (PMD) is characterized by the four Briggs parameters
(Briggs et al., 2007), they are rarely used directly for asserting “ancientness”. Researchers work-
58 ing with ancient DNA tend to simply use the empirical C→T on the first position of the fragment
together with other supporting summary statistic of the experiment (Jónsson et al., 2013). Quanti-
60 fying PMD have become standard for single individual sources like hair, bones, teeth and also ap-
plied to smaller subsets of species in ancient environmental metagenomes (Pedersen et al., 2016;
62 Murchie et al., 2021; Wang, Pedersen, et al., 2021; Zavala et al., 2021). While this is a relatively fast
process for single individuals it becomes increasingly demanding, iterative, and time consuming as
64 the samples and the diversity within increases, as in the case for metagenomes from ancient soil,
sediments, dental calculus, coprolites, and other ancient environmental sources. It has therefore
66 been practice to estimate damage for only the key taxa of interest in a metagenome, as metage-
nomic samples easily include tens of thousands of different taxonomic entities, which makes a
68 complete estimate across the metagenomes computationally intractable, if not an impossible task
(Pedersen et al., 2016). To overcome these limitations, we designed a toolkit called *metaDMG* (pro-
70 nounced *metadamage*) which allows for the rapid computation of various statistics relevant for the
quantification of PMD at read level, single genome level, and even metagenomic level by taking into
72 account the intricate branching structure of the taxonomy of the possible multiple alignments for
the single reads.

74 Our thorough analysis with both simulated and real data shows that *metaDMG* is both faster at
ancient DNA damage estimation and provides more accurate damage estimates. Furthermore, as
76 *metaDMG* is designed with the increasingly large datasets that are currently generated in the field
of ancient environmental DNA in mind, *metaDMG* is able to process complex metagenomes within
78 hours instead of days. At the same time, it outperforms standard tools that estimate DNA damage
for single genomes and samples with low complexity. Furthermore, it can compute a global dam-
80 age estimate for a metagenome as a whole. Lastly, *metaDMG* is compatible with the NCBI taxonomy
and use *ngsLCA* (Wang, T. S. Korneliussen, et al., 2022) to perform a lowest common ancestor (LCA)
82 classification of the aligned reads to get precise damage estimates at all taxonomic levels. It also
allows for custom taxonomies and thus also the use of metagenomic assembled genomes (MAGs)
84 as references.

This paper is organized as follows. In *section 2* we present our statistical models including two
86 novel test statistics, D_{fit} and Z_{fit} . We quantify the performance of our test statistics using various
simulation approaches in *section 3*. The results of these simulations is shown in *section 4* and
88 finally, the method and results are discussed in *section 5*.

2 | METHODS & MATERIALS

90 To quantify ancient damage, one can either compute it on a per read level or across an entire
taxa. A priori, the actual biochemical changes which characterizes post mortem damage in a
92 single read cannot be directly observed, but by aligning each fragment and considering the ob-
served difference between the reference and read, the possible PMD can be computed. We have
94 (re)implemented the approach used in PMDtools (Skoglund et al., 2014) which allows for the ex-
traction of single DNA reads which are estimated to contain PMD, see *Appendix 1*. This approach,
96 will preferentially choose reads that has excess of C→T in the first positions and can not be used
directly for asserting or quantifying to what degree a given library might contain damaged frag-
98 ments. We have therefore developed a novel statistical method that aims to mitigate this caveat
by using all reads or reads that aligns to specific taxa. First we will define the mismatch matrices
100 in *subsection 2.1* followed by the lowest common ancestor method in *subsection 2.2*. The mis-
match matrices can further be improved by multinomial regression, see *subsection 2.3*, however,
102 this requires more data than than what is usually available in metagenomic studies. As such, we
present the beta-binomial damage model in *subsection 2.4* which aims to work even on extremely
104 low-coverage data.

2.1 | Mismatch matrices/nucleotide misincorporation patterns

106 We seek to obtain the pattern or signal of damage across multiple reads by generating what is
called the mismatch matrix or the nucleotide misincorporation matrix. This matrix represents
108 the nucleotide substitution counts across reads and provides us with the position dependent mis-
match matrices, $M(x)$, with x denoting the position in the read, starting from 1. At a specific position
110 x , $M_{\text{ref} \rightarrow \text{obs}}(x)$ describes the number of nucleotides that was mapped to a reference base B_{ref} but
was observed to be B_{obs} , where B is one of the four bases: A, C, G, T. The number of C→T transitions
112 at the first position, e.g., is denoted as $M_{C \rightarrow T}(x = 1)$.

Alignments for a read can be discarded based on their mapping quality, and we also give the

114 user the possibility of filtering out specific nucleotides of the read if the base quality score fall below
some threshold. The quality scores could also be used as probabilistic weights, however, due to
116 the four-bin discretization of quality scores on modern day sequencing machines, we limit the use
of these to filtering.

118 **2.2 | Lowest Common Ancestor and Mismatch matrices**

For environmental DNA (eDNA) studies a competitive alignment approach is routinely applied.
120 Here all possible alignments for a given read are considered. Each read is mapped against a multi
species reference databases (e.g. nucleotide or RefSeq from NCBI or custom downloaded). A single
122 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,
124 whereas a read that maps solely to a single species, e.g. would be indicative of the read being well
classified. We limit the tabulation and construction of the mismatch matrices to the subset of reads
126 that are well classified.

For each read, we compute the lowest common ancestor using all alignments contained within
128 the user defined taxonomic threshold (species, genus or family) and tabulate the mismatches ma-
trices for each cycle (Wang, T. S. Korneliussen, et al., 2022). If none of the alignments pass the
130 filtering thresholds (excess similarity, mapping quality, etc.), the read is discarded. Depending on
the run mode, we allow for the construction of these mismatch matrices on three different levels.
132 Firstly, we can 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 mapDamage (Jónsson et al., 2013).
134 Secondly, we can obtain the per reference counts, or, finally, if a taxonomy database has been
supplied, we can build mismatch matrices at the species level and aggregate from leaf nodes to
136 the internal taxonomic ranks (genus, kingdom etc) towards the root. We will use the term “taxa”
to refer to either of these levels; i.e. a specific taxa can either refer to a specific LCA, a specific
138 reference, or all reads in a global estimate, depending on the run-mode.

When aggregating the mismatch matrices for the internal nodes in our taxonomic tree, two
140 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 normalized by the number of alignments of each
142 read; weight-type 1, which is the default.

2.3 | Regression Framework

144 The nucleotide misincorporation frequencies are routinely used as the basis for assessing whether
 or not a given library is ancient by looking at the expected drop of C→T (or its complementary G→A)
 146 frequencies as a function of the position of the reads. This signal is caused by a higher deamination
 rate in the single-strand part of the damaged fragment than that in the double strand part. The
 148 mismatch matrix is constructed based on the empirical observations and are subject to stochastic
 noise. The effect of noise in the mismatch matrix can be limited by the use of the multinomial
 150 regression model. We continue the work of Cabanski et al., 2012 to provide four different regres-
 sion methods to stabilize the raw mismatch matrix across all combinations of reference bases,
 152 observed bases, strands and positions, see [Appendix 2](#) for details, derivation and results. Given
 enough sequencing data, this approach will provide an improved, noise-reduced mismatch ma-
 154 trix which would be relevant for single genome ancient DNA studies. However, for extremely low
 coverage studies, such as environmental DNA, the method is likely to overfit and would not be as
 156 suitable as the simplified model described in the [subsection 2.4](#).

2.4 | Damage Estimation

158 In standard ancient DNA context it is generally not possible to obtain vast amounts of data and
 thus we propose two novel tests statistics, D_{fit} and Z_{fit} , that are especially suited for this common
 160 scenario. 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
 162 patterns seen in ancient DNA by looking exclusively at the C→T transitions in the forward direction
 (5') and the G→A transitions in the reverse direction (3'). For each taxa, we denote the number of
 164 transitions, $k(x)$, as:

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

166 and the number of reference counts $N(x)$:

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

170 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 likelihood lacks the flexibility needed to deal with the large amount of variance (overdispersion) we found in the data due to poorly curated references and possible misalignments.

To accommodate overdispersion, we instead apply a beta-binomial distribution, $\mathcal{P}_{\text{BetaBinomial}}$, which treats the probability of deamination, p , as a random variable following a beta distribution¹ with mean μ and concentration ϕ : $p \sim \text{Beta}(\mu, \phi)$. The beta-binomial distribution has the following probability density function:

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

where B is defined as the beta function:

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

with $\Gamma(\cdot)$ being the gamma function (Cepeda-Cuervo and Cifuentes-Amado, 2017).

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, $k \sim \mathcal{P}_{\text{BetaBinomial}}(N, \mu, \phi)$:

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

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.

Note that both equation (3) and (5) relates to the damage at a specific base position (cycle), i.e. for a single k and N . To estimate the overall damage in the entire read using the position dependent counts, $k(x)$ and $N(x)$, we model μ as being position dependent, $\mu(x)$, and assume a position-independent concentration, ϕ . We model the damage frequency with a modified geometric sequence, i.e. exponentially decreasing for discrete values of x :

$$y(x; A, q, c) = A(1 - q)^{|x|-1} + c. \quad (6)$$

Here A is the amplitude of the damage and q is the relative decrease of damage pr. position. A background, c , was added to reflect the fact that the mismatch between the read and reference might be due to other factors than just ancient damage. As such, we allow for a non-zero amount of damage, even as $x \rightarrow \infty$. This is visualized in [Figure 1](#) along with a comparison between the classical binomial model and the beta-binomial model.

¹ Note that we do not parameterize the beta distribution in terms of the common (α, β) parameterization, but instead using the more intuitive (μ, ϕ) parameterization. One can reparameterize $(\alpha, \beta) \rightarrow (\mu, \phi)$ using the following two equations: $\mu = \frac{\alpha}{\alpha + \beta}$ and $\phi = \alpha + \beta$ (Cepeda-Cuervo and Cifuentes-Amado, 2017).

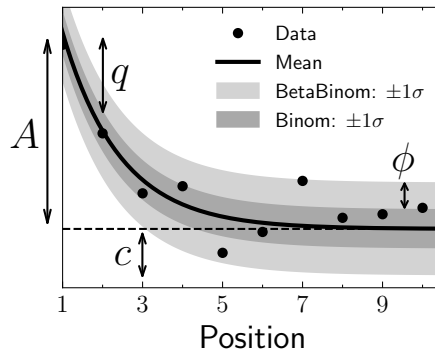


Figure 1. Illustration of the damage model. The figure shows data points as circles and the damage, $f(x)$, as a solid line. The amplitude of the damage is A , the offset is c , and the relative decrease in damage per position is given by q . The damage uncertainty for a binomial model is shown in dark grey and the uncertainty for a beta-binomial model in light grey.

² Parameterized as (μ, ϕ)

To estimate the four fit parameters, A , q , c , and ϕ , we apply Bayesian inference to utilize domain specific knowledge in the form of priors. We assume weakly informative beta-priors² for both A , q , and c . In addition to this, we assume an exponential prior on ϕ with the requirement of $\phi > 2$ to avoid too much focus on 0-or-1 probabilities (McElreath, 2020). The final model is thus:

$$\begin{aligned}
 [A \text{ prior}] \quad & A \sim \text{Beta}(0.1, 10) \\
 [q \text{ prior}] \quad & q \sim \text{Beta}(0.2, 5) \\
 [c \text{ prior}] \quad & c \sim \text{Beta}(0.1, 10) \\
 [\phi \text{ prior}] \quad & \phi \sim 2 + \text{Exponential}(1/1000) \\
 [\text{likelihood}] \quad & k_i \sim \mathcal{P}_{\text{BetaBinomial}}(N_i, y(x_i; A, q, c), \phi),
 \end{aligned} \tag{7}$$

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: $D \equiv y(x = \pm 1) - c$. 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 D :

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

Since D estimates the overexpression of damage due to ancientness, not only is the mean of D , D_{fit} , relevant but also the certainty of it being non-zero (and positive). We quantify this through the

significance $Z_{\text{fit}} = D_{\text{fit}}/\sigma_D$ which is thus the number of standard deviations (“sigmas”) away from
220 zero. Assuming a Gaussian distribution of D , $Z_{\text{fit}} > 2$ would indicate a probability of D being larger
than zero, i.e. containing ancient damage, with more than 97.7% probability. This assumption
222 works well in the case of many reads or a high amount of damage due to central limit theorem.
When the assumption breaks down, the significance is still a relevant test statistic, it is only the
224 conversion to a probability that will become biased.

These two values allows us to not only quantify the amount of ancient damage (D_{fit}) but also the
226 certainty of this damage (Z_{fit}) without having to run multiple models and comparing these. An intu-
itive interpretation of our D_{fit} statistic is, that this is the excess deamination in the beginning of the
228 read, taking all cycle positions into account and excluding the constant deamination background
(c). This is visually similar to the A parameter in *Figure 1*.

230 We perform the Bayesian inference of the parameters models using Hamiltonian Monte Carlo
(HMC) sampling which is a particular of Monte Carlo Markov Chain (MCMC) algorithm (Betancourt,
232 2018). Specifically, we use the NUTS implementation in NumPyro (Phan, Pradhan, and Jankowiak,
2019), a Python package which uses JAX (Bradbury et al., 2018) under the hood for automatic differ-
234 entiation and JIT compilation. We treat each taxa as being independent and generate 1000 MCMC
samples after an initial 500 samples as warm up.

236 Since running the full Bayesian model is computationally expensive, we also allow for a faster,
approximate method by fitting the maximum a posteriori probability (MAP) estimate. We use iMi-
238 nuit (Dembinski et al., 2021) for the MAP optimization with Numba acceleration (Lam, Pitrou, and
Seibert, 2015) for even faster run times. On a Macbook M1 Pro model from 2021, the timings for
240 running the full Bayesian model is 1.41 ± 0.04 s pr. fit and for the MAP it is 4.34 ± 0.07 ms pr. fit,
showing more than a 2 order increase in performance (around 300x) for the approximate model.
242 Both models allow for easy parallelisation to decrease the computation time.

2.5 | Visualisation

244 We provide an interactive graphical user interface (dashboard) to visualise, explore, and manip-
ulate the results from the modelling phase. An interactive example of this can be found online
246 (<https://metadm.onrender.com/>). The structure of the dashboard is explained in *Figure 2*. The dash-
board allows for filtering, styling and variable selection, visualizing the mismatch matrix related to
248 a specific taxa, and exporting of both fit results and plots. By filtering, we include both filtering by

sample, by the summary statistics of the data (e.g. requiring D_{fit} to be above a certain threshold),
 250 and even by taxonomic level (e.g. only looking at taxa that are part of the Mammalia class). We
 greatly believe that a visual overview of the fit results increase understanding of the data at hand.
 252 The dashboard is implemented with Plotly plots and incorporated into a Dash dashboard (Plotly,
 2015).

254 3 | SIMULATION STUDY

To determine metaDMG's performance, we performed a set of rigorous in-silica simulations to identify
 256 and quantify any possible biases as well the accuracy of our test statistics. Overall, the simulations
 can be split two groups. The first is based on a genome from a single species and is used to mea-
 258 sure the performance of the actual damage estimation and damage model. The second is based
 on syntethic ancient metagenomic datasets using the statistics and nature of a set of published
 260 ancient metagenomes.

3.1 | Single-genome simulations

262 The first simulations follow a simple setup in which we extract reads from a set of representa-
 tive genomes having variable length and GC-content. We next added post-mortem damage mis-
 264 incorporations using NGSNGS (Henriksen, Zhao, and T. Korneliussen, 2022) a recent implemen-
 tation of the original Briggs model similar to Gargammel (Neukamm, Peltzer, and Nieselt, 2021)
 266 and lastly added sequencing errors (Renaud et al., 2017). All reads are hereafter mapped using
 Bowtie2 against each of the respective reference genomes and ancient DNA damage estimated
 268 the DNA damage using metaDMG. The simulations were computed with varying amount of damage
 added by changing the single-stranded DNA deamination, δ_{ss} in the original Briggs model (Briggs
 270 et al., 2007).

³ NCBI: NC_012920.1

⁴ NCBI: KX703002.1

⁵ NCBI: NZ_CP024731.1

⁶ NCBI: NZ_LS483369.1

⁷ NCBI: GCA_001929375.1

In detail, we focused on the following genomes; Homo Sapiens mitochondrial³, a *Betula nana*
 272 chloroplast⁴, and three microbial genomes (*Fusobacterium pseudoperiodonticum*⁵, *Neisseria cinerea*⁶,
 and *Actinomyces oris* strain S64C⁷) with the varying GC-content, low (28%), medium (37%), and high
 274 (50%) respectively. For each simulation, we performed 100 independent replications to measure
 the variability of the parameter estimation and quantify the robustness of the estimates. We fur-
 276 ther simulated eight different sets of damage (0%, 1%, 2%, 5%, 10%, 15%, 20%, and 30% damage
 on position 1), all with 13 sets of different number of reads (10, 25, 50, 100, 250, 500, 1.000, 2.500,

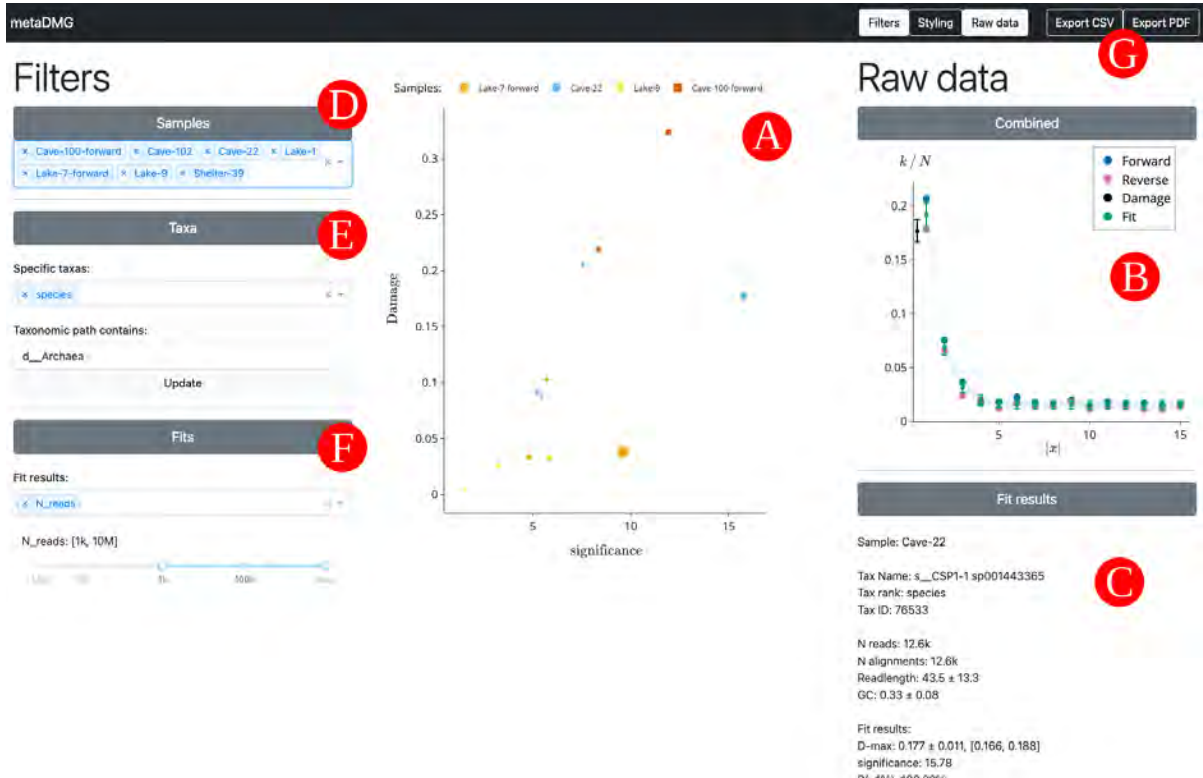


Figure 2. Overview of the interactive metaDMG dashboard. A) The main damage plot shows the damage (D_{fit}) on the y-axis and the significance (Z_{fit}) on the x-axis. Each point is a single taxa from one of the metagenomic samples, see [Table 1](#). Once clicked on a specific taxa, the right-hand window shows information about the selected taxa and related fit. B) The top window shows a plot of the damage frequency for both the forward and reverse direction along with the estimated fit and damage. C) Below, the results of the fit are shown, including taxonomic information, read-specific information, the fit results, and the full taxonomic path. D) In the left filtering window, the samples to include can be selected. E) This windows allows for selection based on taxa-specific criteria. Here we show a selection of only taxa with “species” as their LCA and taxa that are part of the archaea domain. F) The final filtering window allows for setting fit related thresholds such as the minimum damage or significance. Here it is shown discarding taxa with fewer than 1000 reads. G) In the top right, after the selection and filtering process is finished, the final taxa can be exported to a CSV file along with all of the fit information, or the damage plots can be generated and saved.

278 5.000, 10.000, 25.000, 50.000, and 100.000 reads). We also sought to measure the effect of the
 fragment lengths using three sets of different fragment length distributions sampled from a *log-*
 280 *normal* distribution with mean 35, 60, and 90, each with a standard deviation of 10). Furthermore,
 to investigate whether the damage estimation by metaDMG is independent of contig size, we artifi-
 282 cially created three different genomes by sampling 1.000, 10.000 or 100.000 different basepairs
 from a uniform categorical distribution of A, C, G, and T. Based on these three genomes, we added
 284 artificial deamination for a different number of reads, as for the other simulations. Lastly, we also
 created 1000 repetitions of non-damaged simulations for Homo Sapiens to measure the rate of
 286 false positives. The exact commands used can be found in **Appendix 3**.

To compare the damage estimates to known values, for each of the genomes mentioned above
 288 and for each amount of artificial damage, we generated 1.000.000 reads using NGSNGS without
 any added sequencing noise. The values we compare is the difference in damage frequency at
 290 position 1 and 15:

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

292 which is the average of the C→T damage frequency difference and the G→A damage frequency
 difference.

294 3.2 | Metagenomic Simulations

A metagenome contains a complex mixture of organisms, all with highly different characteristics
 296 in GC content, read length, abundance, or degree of DNA damage, and there are large differences
 between different environments. It is therefore far from simple to obtain DNA damage estimates
 298 for such multitude of organisms. In order to test the accuracy and sensitivity of metaDMG, we simu-
 lated six of the nine ancient metagenomes (with more than 1 million reads) covering a wide span
 300 of environments and ages (**Table 1**).

First, we mapped all reads of each metagenome with bowtie2 against a database consisting of
 302 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
 304 al., 2021). We then used bam-filter v1.0.11 (Fernandez-Guerra, 2022a) with the flag `--read-length-freqs`
 to get read length distributions for each genome reads aligned to and their respective abundance.
 306 Next, we filtered genomes with an observed-to-expected coverage ratio greater than 0.75 using

Table 1. Metagenomic samples. “Name” is the name of the sample used throughout this paper. “Site” is the type of metagenomic site. “Type” is the type of environment. “Age” is the approximate age of the sample in kyr Bp. “Sediment” is the name type of sediment. “Instrument” is the Illumina model. “Library” is the library type where D. means double stranded and S. means single stranded. “Reads” is the raw number of reads (in millions). “Source” is the source of the data. The dagger (†) indicates samples that were not a part of the metagenomic simulation pipeline.

Name	Site	Type	Age (kyr)	Sediment	Instrument	Library	Reads (M)	Source
Library-0 [†]	Control	Control	0	Reagents	HiSeq4000	D.	19.7	(Ardelean et al., 2020)
Pitch-6	Syltholmen pitch	Chewed organic material	5.7	Organic material	HiSeq2500	D.	150.3	(Jensen et al., 2019)
Lake-1 [†]	Spring Lake	Lake gyttja/sediment	1.4	Organic material	HiSeq 100	D.	49.8	(Pedersen et al., 2016)
Lake-7	Lake CH12	Lake gyttja/sediment	6.7	Organic material	HiSeq2500	S.	291.9	(Schulte et al., 2021)
Lake-9	Spring Lake	Lake gyttja/sediment	9.2	Organic material	HiSeq 100	D.	128.4	(Pedersen et al., 2016)
Shelter-39 [†]	Abri Pataud	Rock shelter	39.4	Sediment	MiSeq	S.	0.4	(Braadbaart et al., 2020)
Cave-22	Chiquihuite cave	Cave sediment	22.2	Carbonate rock	HiSeq4000	D.	5.7	(Ardelean et al., 2020)
Cave-100	Eustatus Cave	Cave sediment	100	Carbonate rock	HiSeq2500	S.	21.8	(Vernot et al., 2021)
Cave-102	Pesturina Cave	Neanderthal tooth	102	Dental calculus	HiSeq4000	D.	12.3	(Fellows Yates et al., 2021)

bamfilter. The filtered BAM files were then processed by metaDMG to obtain misincorporation matrices for each genome. The abundance tables, fragment length distribution, and misincorporation matrices were then used in aMGSIM-smk v0.0.1 (Fernandez-Guerra, 2022b), a Snakemake workflow (Mölder et al., 2021) that facilitates the generation of multiple synthetic ancient metagenomes. The underlying tools in this workflow is the gargammel toolkit (Renaud et al., 2017), that based on input read length distribution extract a subset of sequences (FragSim) with similar length. This is then followed by the addition of $C \rightarrow T$ substitutions (DeamSim) which mimics the postmortem damage process. Finally the deaminated sequences are passed to the ART (Huang et al., 2012) for sequence simulation. The data used and generated by the workflow can be obtained from ERDA. We then performed taxonomic profiling and damage estimation using identical parameters as for the synthetic reads generated by aMGSIM-smk.

4 | RESULTS

We tested the accuracy and performance of the metaDMG damage estimates, D_{fit} , using a set of different simulation scenarios and subsequently tested on 9 real-life ancient metagenomic dataset.

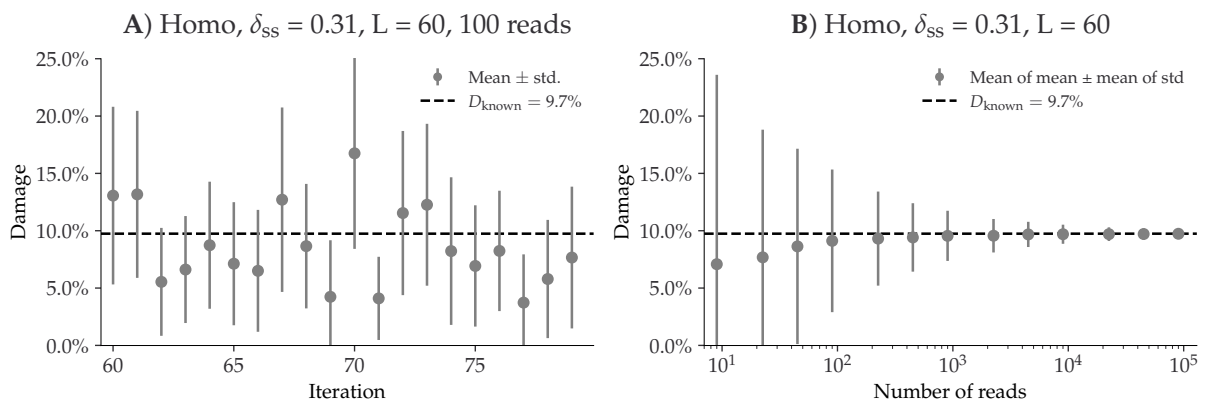


Figure 3. Overview of the single-genome simulations based on the Homo Sapiens genome with a fragment length distribution with mean 60 and the Briggs parameter $\delta_{\text{SS}} = 0.31$ (approximately 10% damage). **A)** This plot shows the estimated damage (D_{fit}) of 20 replicates, each with 100 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).

4.1 | Single-genome simulation results

322 To illustrate the results the performance on single-genomes, we first focus on a single, specific set
of simulation parameters. This simulation is based on the Homo Sapiens genome with the Briggs
324 parameter $\delta_{\text{SS}} = 0.31$ (approximately 10% damage) and a mean fragment length of 60. In general,
we use $\delta = 0.0097$, $\nu = 0.024$, and $\lambda = 0.36$ as Briggs parameters, while varying δ_{SS} (Briggs et al.,
326 2007). We show the metaDMG damage results for the 100 independent replications in **Figure 3**. The
left part of the figure shows the individual metaDMG damage estimates for an arbitrary choice of 20
328 replications (iteration 60 to 79). When the damage estimates are very low, the distribution of D_{fit} is
skewed (restricted to positive values), sometimes leading to errorbars going into negative damage,
330 which represents unrealistic estimates. The right hand side of the figure visualizes the average
amount of damage based on all 100 replications across a varying number of reads. This shows
332 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) for
334 this specific set of simulation parameters.

Across multiple simulations, each with 8 different damage levels, 13 different numbers of reads,
336 and 100 replications, we find no significant difference in test statistic across different species (**Fig-**

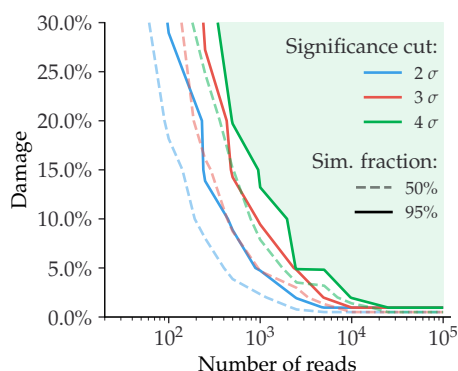


Figure 4. Relationship between the damage and the number of reads for simulated data (single-genome). Given a specific significance cut, the solid contour line shows the relationship between the amount of damage and the number of reads required to be able to correctly infer damage in 95% of the taxa. 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 4σ confidence.

ure S5 and Figure S6), across different GC-levels (Figure S7–Figure S9), different fragment length
 338 distributions (Figure S10–Figure S12), or even different contig lengths (Figure S13–Figure S15), see
 Appendix 4. Based on the single-genome simulations, we compute the relationship between the
 340 amount of damage in a taxa and the number of reads required to correctly infer that the reads
 from that taxa are damaged, see Figure 4. If we want to assert damage with a significance of more
 342 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, whereas we only need 100 reads if our target
 344 organism has 30 % damage.

Finally, to quantify the risk of incorrectly classifying a non-ancient taxa as damaged, we created
 346 1000 independent replications for a varying number of reads, where none of them had any artificial
 ancient damage applied, only sequencing noise. Figure 5 shows the damage (D_{fit}) as a function of
 348 the significance (Z_{fit}) for the case of 1000 reads. Even though the estimated damage is larger than
 zero, the damage is non-significant since the significance is less than one. When looking at all the
 350 figures across the different number of reads, see Appendix 5, we note that a relaxed significance
 threshold requiring that $D_{fit} > 1\%$ and $Z_{fit} > 2$ would filter out all of non-damaged points. Overall
 352 the conclusion being that our novel test statistic is conservative and has low false positive rate.

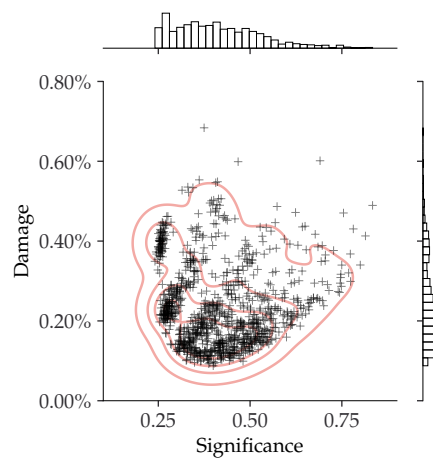


Figure 5. Inferred damage of modern, simulated data (single-genome). The plot shows the inferred damage estimates of 1000 replicates, each with 1000 reads and no artificial ancient damage applied. Each single cross corresponds to a simulation and the red lines outlines the kernel density estimate (KDE) of the damage estimates. The marginal distributions are shown as histograms next to the scatter plot.

4.2 | Metagenomic simulation results

354 With the full metagenomic simulation pipeline we can further probe the performance of metaDMG.
 By considering the different metagenomic scenarios, see [Table 1](#), at different steps in the pipeline,
 356 we are able to show that metaDMG provides relevant and accurate damage estimates.

To verify that the risk of getting false positives is non-significant, we run metaDMG on the metage-
 358 nomic assemblages after fragmentation with FragSim, but before any no deamination with Deam-
 Sim has yet been added. We find that the previously established relaxed significance threshold
 360 ($D_{\text{fit}} > 1\%$ and $Z_{\text{fit}} > 2$) correctly filters out all of the taxa, see [Figure 6](#). This is as expected, as there
 has not yet been added any artificial post mortem damage in the form of deamination.

362 We see a clear difference in the damage estimates between the ancient and the non-ancient
 taxa once we add deamination with DeamSim and sequencing errors with ART, see [Figure 7](#). The
 364 non-ancient taxa would still not pass the relaxed threshold, in contrast to the taxa in the ancient
 samples.

366 The results of [Figure 7](#) are summarized in [Table 2](#). We find that Cave-100-forward, Cave-102,
 Pitch-6 all have more than 60% of their ancient taxa correctly labelled as damaged according to the
 368 relaxed threshold, while it for Cave-22 and Lake-7-forward is a bit lower and Lake-9 does not show

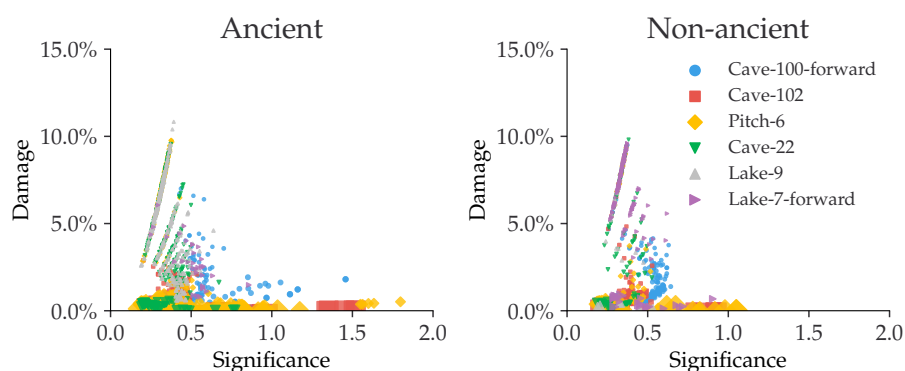


Figure 6. Estimated amount of damage as a function of significance for metagenomic simulations. This figure shows the metagenomic simulations after FragSim has been applied, but before including any deamination or sequencing errors. We generate both non-ancient and ancient taxa in the simulation pipeline. The left subfigure shows the damage of the ancient taxa and similarly for the non-ancient taxa in the right subfigure.

any clear support of damage. However, once we condition on the requirement of having more than
 370 100 reads, the fraction of ancient taxa correctly identified as ancient increases to more than 90%
 for most of the samples. A small investigation of one of the ancient taxa (*Stenotrophomonas Mal-*
 372 *tophilia*) in the simulation that did not meet the criteria to be ancient metaDMG, i.e. a false negative,
 can be found in [Appendix 6](#).

374 4.3 | Real Data

The results from running the real metagenomic data through the metaDMG pipeline show clear ev-
 376 idence of taxa with significant DNA damage present in the metagenome and a layered pattern
 similar to what was observed in the simulated ancient metagenomes, see [Figure 8](#).

378 As DNA damage is not a function of time, we cannot expect that there is a direct relation be-
 tween damage and time, however, we do see that the oldest samples, Cave-100 and Cave-102, see
 380 [Table 1](#), which are 100 and 102 thousand years BP, show the highest amount of damage of all the
 metagenomes. Both the Pitch-6 and Cave-22 samples, which are 6 and 22 thousand year old and
 382 thus younger than two above mentioned cave samples, have almost similar levels of damage. This
 is not unexpected as the micro environment surrounding the layer in which the metagenome was
 384 found plays a significant role in the state of DNA. In our case, the younger Pitch-6 derives from a
 water logged but open air site, while the Cave-22 sample was obtained in dry but cool (~11 degree
 386 Celsius year around) cave layers.

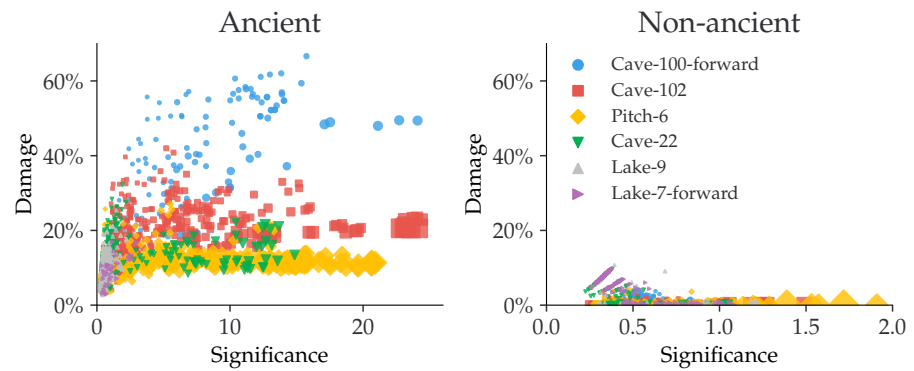


Figure 7. Estimated amount of damage as a function of significance for metagenomic simulations. This figure shows the metagenomic simulations after fragmentation, deamination, and sequencing errors have been applied. The left subfigure shows the damage of the ancient taxa and similarly for the non-ancient taxa in the right subfigure.

Table 2. metaDMG damage results for the six different metagenomic simulations. The first column is the total number of taxa, the second column is the total number of taxa that would pass the threshold of $D_{fit} > 1\%$ and $Z_{fit} > 2$, the third column is the number of taxa with more than 100 reads, and the final column is the number of taxa 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%

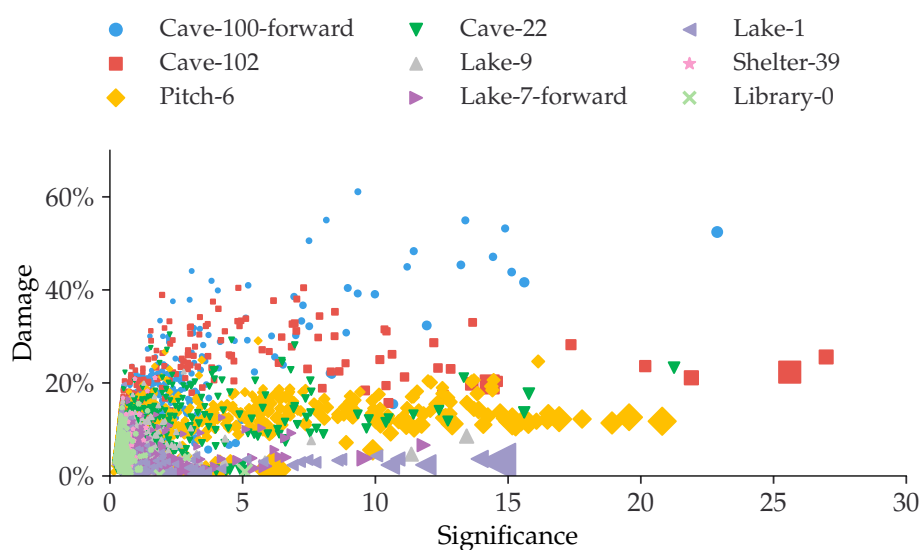


Figure 8. Estimated amount of damage as a function of significance using the real data, see [Table 1](#).

The metagenomes with the least DNA damage are the ones from the lake sediments (Lake-1, Lake-7 and Lake-9). These samples do show some taxa with significant DNA damage, although they do not have a strong damage signal.

Importantly, we find that in the true metagenomes, `metaDMG` is able to assign low significance to the taxa that likely are not damaged or that have too little data, see e.g. the upper right hand corner of [Figure 9](#). This subfigure shows the damage plot for the *Gallus Gallus* species (red junglefowl) from the Lake-1 sample. This particular species only has $D_{\text{fit}} = 2.2\%$ and $Z_{\text{fit}} = 1.0$, which does not satisfy the relaxed DNA damage threshold ($D_{\text{fit}} > 1\%$, $Z_{\text{fit}} > 2$). In addition to the *Gallus Gallus* species, [Figure 9](#) further shows examples of species with large amounts of data (*Homo Sapiens* in the Pitch-6 sample and *Crocota Crocuta* in the Cave-100 sample, based only on forward data), and an example of medium damage (*Equisetum Arvense* in Lake-7, based only on forward data).

Interestingly, and of high importance for downstream interpretation, is that for certain samples, some taxa were found to have a high significance although with lower DNA damage than what is observed across the given metagenome as a whole. This underlines the need to evaluate the DNA damage variation within each metagenome, perform a proper outlier test and the basic setting of logical thresholds.

We find that when using the relaxed DNA damage threshold, `metaDMG` falsely classifies a single

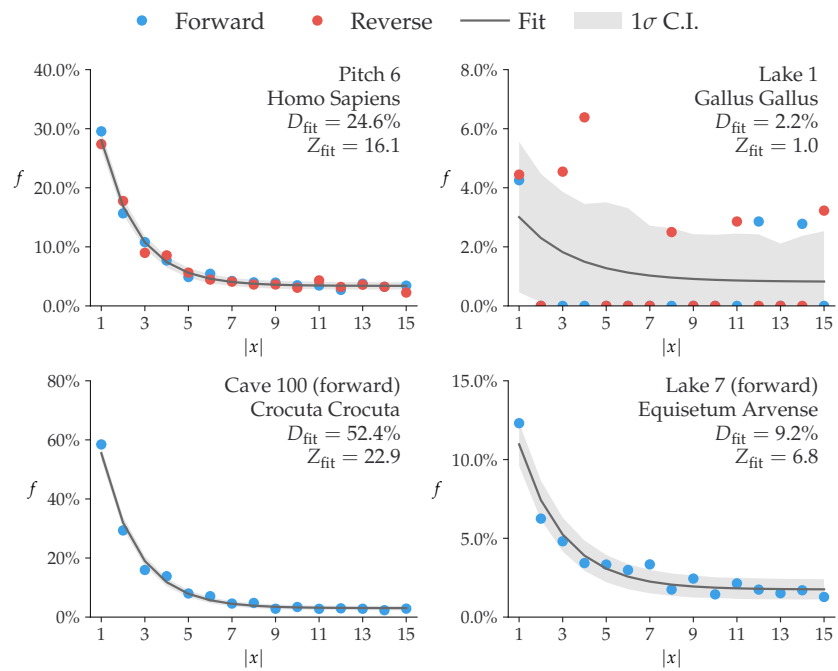


Figure 9. Damage plots of four representative species from the real-data metagenomic samples, see [Table 1](#). Each subfigure shows the damage rate $f(x) = k(x)/N(x)$ as a function of position x for both forward (C→T) and reverse (G→A). The metaDMG fit is shown in grey with the 68% credible intervals as shaded regions. In the upper right corner of each subfigure, the information about the sample and the species together with the metaDMG damage estimates is shown.

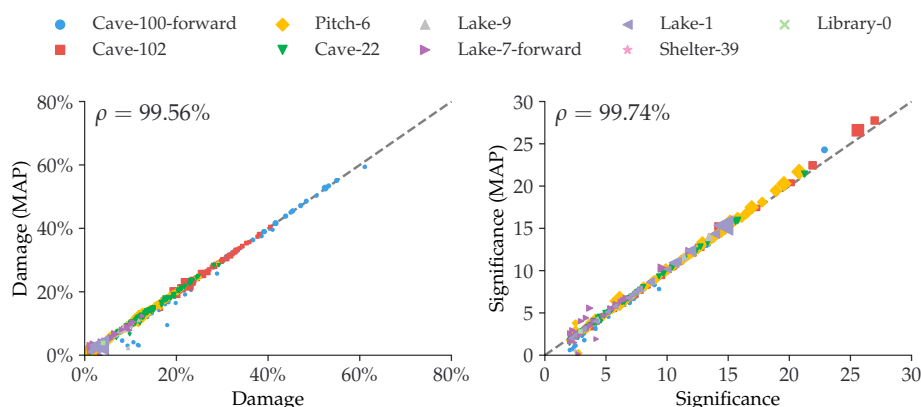


Figure 10. Comparison between the full Bayesian model and the fast, approximate, MAP model for the estimated damage and significance. The figure shows data after a loose cut of $D_{\text{fit}} > 1\%$, $Z_{\text{fit}} > 2$ and more than 100 reads. The dashed, grey line shows the 1:1 ratio and the correlation, ρ , is shown in the upper left corner.

404 of the taxa from the control test Library-0 as being ancient. However, with a more conservative
 405 damage threshold ($D_{\text{fit}} > 2\%$, $Z_{\text{fit}} > 3$, more than 100 reads), none of the taxa from the library
 406 control are classified as ancient.

4.4 | Bayesian vs. MAP

408 Due to the higher computational burden of computing the full Bayesian model compared to the
 409 faster, approximate MAP model in samples with several thousand taxa, the MAP model is in prac-
 410 tice the model of choice due to lower computational complexity. We compared the performance
 411 of D_{fit} and Z_{fit} on the real datasets in [Table 1](#), see [Figure 10](#). This figure compares the estimated
 412 damage between the Bayesian model and the MAP model (left subfigure) and the estimated sig-
 413 nificances (right subfigure) for taxa passing a threshold of $D_{\text{fit}} > 1\%$, $Z_{\text{fit}} > 2$, and more than 100
 414 reads. The figure shows that the vast majority of taxa map 1:1 between the Bayesian and the MAP
 415 model. It should be noticed that the taxa with the worst correspondence in damage estimates
 416 are all based on forward-only fits, i.e. with no information from the reverse strand, which leads
 417 to less data to base the fits on. For the comparison with no thresholds applied, see [Figure S23](#) in
 418 [Appendix 7](#). We recommend to use the full, Bayesian model in the case of extremely low-coverage
 data or when used on only a small number of taxa (e.g. when using metaDMG in global-mode).

420 4.5 | Existing Methods

To our knowledge there are not currently available methods for assessing and quantifying post-mortem DNA damage in a metagenomic context. We compare the performance of the D_{fit} statistic in metaDMG to existing methods such as those found in PyDamage (Borry et al., 2021). Since PyDamage is based solely on single genome analysis we use the non-LCA mode of metaDMG. This mode iterates through the different referenceIDs for all mapped reads and estimates the damage for each. In general, we find that metaDMG is more conservative, accurate and precise in its damage estimates.

One example of this can be found in *Figure 11*, which shows both the metaDMG and PyDamage results of the simulations described in *subsection 3.1*, in particular the 100 replications of the Homo Sapiens single-genome with 100 reads and 15% added artificial damage (and a fragment length distribution with mean 60). *Figure 11* shows that the metaDMG estimates are between 5% and 25% damage, while PyDamage estimates up to more than 50% damage, in a sample with 15% artificially added damage. The comparisons between metaDMG and PyDamage for the other sets of simulation parameters can be found in *Figure S24–Figure S31* in *Appendix 8*.

To compare the computational performance, we use the real-life Pitch-6 sample (i.e. non-simulated), see *Table 1*. This alignment file (in BAM-format) takes up 857 MB of space and has 3.7 millions reads with a total of 19 million alignments to 11.433 unique taxa. When using only a single core, PyDamage took 1105s to compute all fits, while metaDMG took 88s, a factor of 12.6x faster. The rest of the timings are shown in *Table 3*. PyDamage requires the alignment files 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 computational load before running the actual damage estimation. metaDMG on the other hand requires the reads to be sorted by name to minimize the time it takes to run the LCA.

444 5 | DISCUSSION

To our knowledge there are no currently available methods other than metaDMG that is geared towards damage analysis in a metagenomic setting. It is the first general framework designed specifically for the quantification of ancient damage in all contexts. The toolkit contains various inter-linked and independent modules including a state-of-the-art graphical user interface that allow

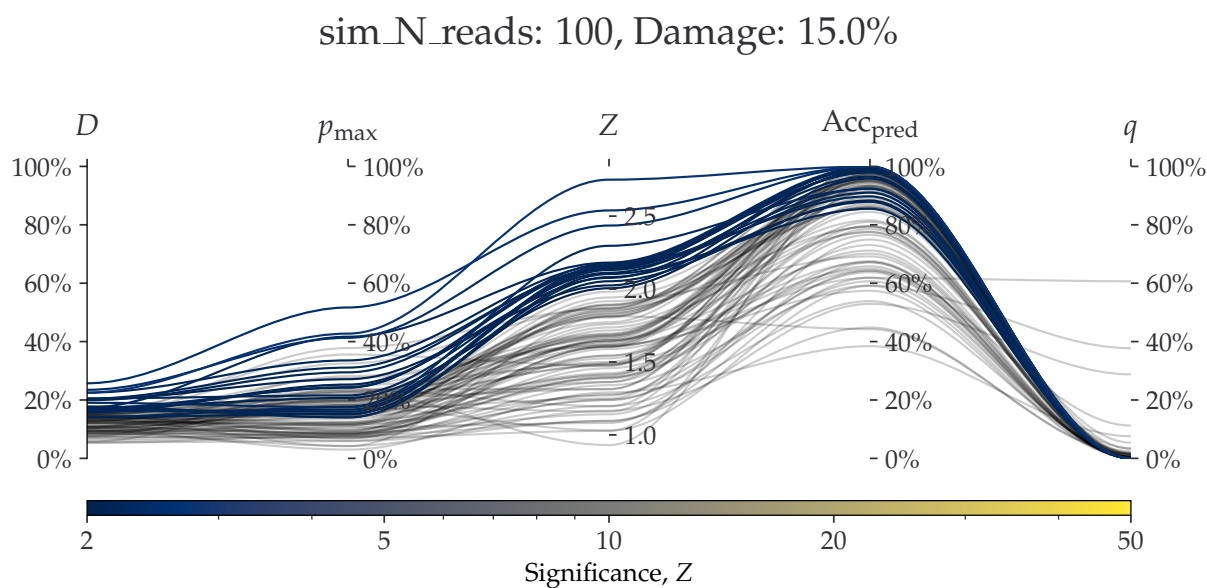


Figure 11. Parallel coordinates plot comparing metaDMG and PyDamage for the Homo Sapiens single-genome simulation with 100 reads and 15% added artificial damage. The two first axes show the estimated damage: D_{fit} by metaDMG and p_{max} by PyDamage. The following two axes show the fit quality: significance (Z_{fit}) by metaDMG and the predicted accuracy (Acc_{pred}) by PyDamage. The final axis shows the q -value by PyDamage. Each of the 100 replications are plotted as single lines. Replications passing the relaxed metaDMG damage threshold ($D_{fit} > 1\%$ and $Z_{fit} > 2$) are shown in color proportional to their significance. Replications that did not pass are shown in semi-transparent black lines.

Table 3. 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 taxa. 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	Pydamage (s)	metaDMG (s)	Improvement (x)
1	1105	88 s	12.6
2	592	66 s	9.0
4	398	54 s	7.4

researchers to explore their data.

450 Multiple areas of future improvements exists. Currently, our novel test statistic for the damage
estimation D_{fit} is based on a statistical model where we only consider the C→T and G→A transitions
452 and where each taxa is modelled as being fully independent, even for closely related species when
provided a taxonomic tree. This could be improved upon with the use of a hierarchical model
454 where information across taxonomic leaf nodes is shared. The current implementation, however,
allows for easy parallelization of the individual fits which reduces the time spent on the inference.
456 In addition to the mismatch matrices, another improvement would be to include the read length
distribution as a covariate in the damage model, as, in addition to deamination, the fragment length
458 distribution is also an indicator of ancient damage (Dabney, Meyer, and Pääbo, 2013; Peyrégne and
Prüfer, 2020).

460 We show that the D_{fit} statistic that metaDMG provides is accurate across different damage levels
and different number of reads. In the single-genome reference case, we further show that the
462 estimates are stable across different species and fragment length distributions. In addition to this,
we find that the results are independent of the contig size, in contrast to PyDamage (Borry et al.,
464 2021).

The basis for the D_{fit} statistic is the leaf node mismatch matrices which contains the raw ob-
466 served substitution frequencies. The computation of these could also take into account the com-
puted mapping uncertainty and the uncertainty of the assigned called nucleotide. We include a
468 regression approach for stabilizing the mismatch matrices across all covariates but this requires
much more data than our current approach. Rather than regressing on all covariates, it might also
470 be more biological meaningful to regress on the four Briggs parameters.

In our toolkit we have included the PMDtools approach (Skoglund et al., 2014) that allows for
472 the separation of highly damaged reads from undamaged reads. The method offers a reasonable
way to distinguish the endogenous ancient DNA from possible modern contamination. But this
474 method may suffer from the fact that some fixed empirical parameters are applied. A possible
extension can be using several statistics estimated from the specific sample (e.g., taxa specific D_{fit}
476 and the ancient fragment lengths) as priors in an empirical Bayes inference framework to learn the
categories of reads unsupervisedly.

478 Our research indicate that the metaDMG results are conservative with very low false positive rates.
This is particularly important with metagenomic samples as the number of taxa, and thus the num-

480 ber of damage estimates, tend to be large. As the number of fits increases, we strongly believe that
a graphical user interface is important. This helps to select and filter the fit results, and to better un-
482 derstand the data at hand. We have tested `metaDMG` using a state of the art metagenomic simulation
pipeline based on multiple metagenomic real-life sample from a variety of different environments.
484 We hope that `metaDMG` can improve the knowledge about DNA damage degradation in different
environments and be the foundation of a more general, metagenomic ancient damage study.

486 **6 | AUTHOR CONTRIBUTIONS**

CM developed and implemented the damage model and all aspect of the python code including
488 the CLI, all fits, and the dashboard. TP helped develop the model and with statistical discussions.
TSK implemented the C/C++ code relating to the lowest common ancestor and mismatch matrices.
490 LZ implemented the PMDtools and full multinomial regression subfunctionality. AFG and MWP
designed the metagenomic simulation study and the application of `metaDMG` to real data. CM and
492 MWP ran all analyses. CM, MWP and TSK initiated and designed the project. All authors contributed
to writing the manuscript.

494 REFERENCES

- Ardelean, Ciprian F. et al. (2020). "Evidence of human occupation in Mexico around the Last Glacial
 496 Maximum". en. In: *Nature* 584.7819. Number: 7819 Publisher: Nature Publishing Group, pp. 87–
 92. ISSN: 1476-4687. DOI: [10.1038/s41586-020-2509-0](https://doi.org/10.1038/s41586-020-2509-0).
- 498 Betancourt, Michael (2018). "A Conceptual Introduction to Hamiltonian Monte Carlo". In: *arXiv:1701.02434*
[stat]. arXiv: 1701.02434.
- 500 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:
 502 2167-8359. DOI: [10.7717/peerj.11845](https://doi.org/10.7717/peerj.11845).
- Braadbaart, F. et al. (2020). "Heating histories and taphonomy of ancient fireplaces: A multi-proxy
 504 case study from the Upper Palaeolithic sequence of Abri Pataud (Les Eyzies-de-Tayac, France)".
 en. In: *Journal of Archaeological Science: Reports* 33, p. 102468. ISSN: 2352-409X. DOI: [10.1016/j.
 506 jasrep.2020.102468](https://doi.org/10.1016/j.jasrep.2020.102468).
- Bradbury, James et al. (2018). *JAX: composable transformations of Python NumPy programs*.
- 508 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
 510 Sciences Section: Biological Sciences, pp. 14616–14621. ISSN: 0027-8424, 1091-6490. DOI: [10.
 1073/pnas.0704665104](https://doi.org/10.1073/pnas.0704665104).
- 512 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-
 514 2105. DOI: [10.1186/1471-2105-13-221](https://doi.org/10.1186/1471-2105-13-221).
- Cepeda-Cuervo, Edilberto and MARÍA VICTORIA Cifuentes-Amado (2017). "Double Generalized Beta-
 516 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.
 518 15446/rce.v40n1.61779](https://doi.org/10.15446/rce.v40n1.61779).
- Dabney, Jesse, Matthias Meyer, and Svante Pääbo (2013). "Ancient DNA Damage". In: *Cold Spring*
 520 *Harbor Perspectives in Biology* 5.7, a012567. ISSN: 1943-0264. DOI: [10.1101/cshperspect.a012567](https://doi.org/10.1101/cshperspect.a012567).
- Dembinski, Hans et al. (2021). *scikit-hep/iminuit: v2.8.2*. DOI: [10.5281/ZENODO.3949207](https://doi.org/10.5281/ZENODO.3949207).

- 522 Fellows Yates, James A. et al. (2021). "The evolution and changing ecology of the African hominid
oral microbiome". en. In: *Proceedings of the National Academy of Sciences* 118.20, e2021655118.
524 ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.2021655118](https://doi.org/10.1073/pnas.2021655118).
- Fernandez-Guerra, Antonio (2022a). *BAM-filter*. original-date: 2021-10-19T09:14:18Z.
- 526 — (2022b). *genomewalker/aMGSIM-smk: v0.0.1*. DOI: [10.5281/zenodo.7298422](https://doi.org/10.5281/zenodo.7298422).
- Ginolhac, Aurelien et al. (2011). "mapDamage: testing for damage patterns in ancient DNA se-
528 quences". In: *Bioinformatics* 27.15, pp. 2153–2155. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/
btr347](https://doi.org/10.1093/bioinformatics/btr347).
- 530 Henriksen, Ramus, Lei Zhao, and Thorfinn Korneliussen (2022). *NGSNGS: v0.5.0*. DOI: [10.5281/zenodo.
7326212](https://doi.org/10.5281/zenodo.7326212).
- 532 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).
- 534 Jensen, Theis Z. T. et al. (2019). "A 5700 year-old human genome and oral microbiome from chewed
birch pitch". en. In: *Nature Communications* 10.1. Number: 1 Publisher: Nature Publishing Group,
536 p. 5520. ISSN: 2041-1723. DOI: [10.1038/s41467-019-13549-9](https://doi.org/10.1038/s41467-019-13549-9).
- Jónsson, Hákon et al. (2013). "mapDamage2.0: fast approximate Bayesian estimates of ancient DNA
538 damage parameters". en. In: *Bioinformatics* 29.13, pp. 1682–1684. ISSN: 1367-4803, 1460-2059.
DOI: [10.1093/bioinformatics/btt193](https://doi.org/10.1093/bioinformatics/btt193).
- 540 Lam, Siu Kwan, Antoine Pitrou, and Stanley Seibert (2015). "Numba: a LLVM-based Python JIT com-
piler". In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*. LLVM
542 '15. New York, NY, USA: Association for Computing Machinery, pp. 1–6. ISBN: 978-1-4503-4005-2.
DOI: [10.1145/2833157.2833162](https://doi.org/10.1145/2833157.2833162).
- 544 Langmead, Ben and Steven L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2". en. In:
Nature Methods 9.4. Number: 4 Publisher: Nature Publishing Group, pp. 357–359. ISSN: 1548-
546 7105. DOI: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923).
- McElreath, Richard (2020). *Statistical rethinking: a Bayesian course with examples in R and Stan*. 2nd ed.
548 CRC texts in statistical science. Boca Raton: Taylor and Francis, CRC Press. ISBN: 978-0-367-
13991-9.
- 550 Mölder, Felix et al. (2021). *Sustainable data analysis with Snakemake*. en. Tech. rep. 10:33. Type: arti-
cle. F1000Research. DOI: [10.12688/f1000research.29032.2](https://doi.org/10.12688/f1000research.29032.2).

- 552 Murchie, Tyler J. et al. (2021). "Collapse of the mammoth-steppe in central Yukon as revealed by
ancient environmental DNA". en. In: *Nature Communications* 12.1. Number: 1 Publisher: Nature
554 Publishing Group, p. 7120. ISSN: 2041-1723. DOI: [10.1038/s41467-021-27439-6](https://doi.org/10.1038/s41467-021-27439-6).
- Nayfach, Stephen et al. (2021). "CheckV assesses the quality and completeness of metagenome-
556 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).
- 558 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/
560 gkx1095](https://doi.org/10.1093/nar/gkx1095).
- Neukamm, Judith, Alexander Peltzer, and Kay Nieselt (2021). "DamageProfiler: fast damage pattern
562 calculation for ancient DNA". In: *Bioinformatics* 37.20, pp. 3652–3653. ISSN: 1367-4803. DOI: [10.
1093/bioinformatics/btab190](https://doi.org/10.1093/bioinformatics/btab190).
- 564 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:
566 Nature Publishing Group, pp. 996–1004. ISSN: 1546-1696. DOI: [10.1038/nbt.4229](https://doi.org/10.1038/nbt.4229).
- Pedersen, Mikkel et al. (2016). "Postglacial viability and colonization in North America's ice-free
568 corridor". en. In: *Nature* 537.7618. Number: 7618 Publisher: Nature Publishing Group, pp. 45–
49. ISSN: 1476-4687. DOI: [10.1038/nature19085](https://doi.org/10.1038/nature19085).
- 570 Peyrégne, Stéphane and Kay Prüfer (2020). "Present-Day DNA Contamination in Ancient DNA Datasets".
en. In: *BioEssays* 42.9. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.202000081>,
572 p. 2000081. ISSN: 1521-1878. DOI: [10.1002/bies.202000081](https://doi.org/10.1002/bies.202000081).
- Phan, Du, Neeraj Pradhan, and Martin Jankowiak (2019). "Composable Effects for Flexible and Accel-
574 erated Probabilistic Programming in NumPyro". In: *arXiv:1912.11554 [cs, stat]*. arXiv: 1912.11554.
- Plotly, Technologies (2015). *Collaborative data science*. Place: Montreal, QC Publisher: Plotly Tech-
576 nologies Inc.
- Renaud, Gabriel et al. (2017). "gargammel: a sequence simulator for ancient DNA". eng. In: *Bioinfor-
578 matics (Oxford, England)* 33.4, pp. 577–579. ISSN: 1367-4811. DOI: [10.1093/bioinformatics/btw670](https://doi.org/10.1093/bioinformatics/btw670).
- Schulte, Luise et al. (2021). "Hybridization capture of larch (*Larix Mill.*) chloroplast genomes from
580 sedimentary ancient DNA reveals past changes of Siberian forest". en. In: *Molecular Ecology Re-
sources* 21.3. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13311>, pp. 801–
582 815. ISSN: 1755-0998. DOI: [10.1111/1755-0998.13311](https://doi.org/10.1111/1755-0998.13311).

- Skoglund, Pontus et al. (2014). "Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal". In: *Proceedings of the National Academy of Sciences* 111.6. Publisher: Proceedings of the National Academy of Sciences, pp. 2229–2234. DOI: [10.1073/pnas.1318934111](https://doi.org/10.1073/pnas.1318934111).
- Valk, Tom van der et al. (2021). "Million-year-old DNA sheds light on the genomic history of mammoths". en. In: *Nature* 591.7849. Number: 7849 Publisher: Nature Publishing Group, pp. 265–269. ISSN: 1476-4687. DOI: [10.1038/s41586-021-03224-9](https://doi.org/10.1038/s41586-021-03224-9).
- Vernot, Benjamin et al. (2021). "Unearthing Neanderthal population history using nuclear and mitochondrial DNA from cave sediments". In: *Science* 372.6542. Publisher: American Association for the Advancement of Science, eabf1667. DOI: [10.1126/science.abf1667](https://doi.org/10.1126/science.abf1667).
- Wang, Yucheng, Thorfinn Sand Korneliussen, et al. (2022). "ngsLCA—A toolkit for fast and flexible lowest common ancestor inference and taxonomic profiling of metagenomic data". In: *Methods in Ecology and Evolution* n/a.n/a. Publisher: John Wiley & Sons, Ltd. ISSN: 2041-210X. DOI: [10.1111/2041-210X.14006](https://doi.org/10.1111/2041-210X.14006).
- Wang, Yucheng, Mikkel Pedersen, et al. (2021). "Late Quaternary dynamics of Arctic biota from ancient environmental genomics". en. In: *Nature* 600.7887. Number: 7887 Publisher: Nature Publishing Group, pp. 86–92. ISSN: 1476-4687. DOI: [10.1038/s41586-021-04016-x](https://doi.org/10.1038/s41586-021-04016-x).
- Zavala, Elena I. et al. (2021). "Pleistocene sediment DNA reveals hominin and faunal turnovers at Denisova Cave". en. In: *Nature* 595.7867. Number: 7867 Publisher: Nature Publishing Group, pp. 399–403. ISSN: 1476-4687. DOI: [10.1038/s41586-021-03675-0](https://doi.org/10.1038/s41586-021-03675-0).

Appendix 1

606
604

PMDTOOLS

608

610

612

614

$$D_x = C + p(1 - p)^{|x|}, \quad (10)$$

616

618

620

622

Three non-mutually exclusive events can lead to an observation of C→T or G→A (Skoglund et al., 2014), namely (i) a true biological polymorphism (occurring at rate π), (ii) a sequencing errors (rate ϵ , can be extracted from the base quality scores of the site on the sampled strand), and (iii) in the case of damaged DNA, the damaged nucleotide frequencies are assumed to be only related to its position from either termini of the ancient fragment (C→T from 5' end, and G→A from 3' end). The error probability of the postmortem nucleotide misincorporation is under the pmdtools model given by:

here $C = 0.01$ and $p = 0.3$ are both suitable constants. Skoglund et al., 2014 defines the likelihood ratio of a strand between the PMD model and the NULL model as its postmortem damage score (PMDS),

$$\text{PMDS} = \log \left\{ \frac{\prod_x L(\text{PMD} | S_x)}{\prod_x L(\text{NULL} | S_x)} \right\}, \quad (11)$$

The reads with the PMDS exceeding an empirical p-value threshold can then be used for filtering intensively damaged fragments.

Appendix 2

624

MULTINOMIAL LOGISTIC REGRESSIONS

626

Full Multinomial Logistic Regression

628

630

632

634

Postmortem damages have impacts on the next generation sequencing reads. A common phenomenon is the increasing of the calling error rates from nucleotide C→T due to the cytosine deamination process. Unawareness of this will lead to inaccurate inferences. Evidences show that the magnitude of such changes are related to the positions the site is within a read (the fraction of the ancient DNA). Here we present four slightly different ways (i.e., full unconditional regression, full conditional regression, folded unconditional regression and folded conditional regression) 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.

636

Data Description

638

640

We perform the regressions based on the summary statistic of the mismatch matrix, i.e., $M(x)$, which 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 S1* and *Table S2* give an example of the data format we use for the inference.

Read Counts								
Ref.	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	55598
-15	12660634	4746	11753	6732	14815	9561359	5184	59633

642 **Appendix 2—table S1.** The read counts per position given the reference nucleotides are A or C of an
644 ancient human data. The negative position indices are the position on the reversed strand. In the
646 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 $M_{A \rightarrow i}(x)$ or $M_{C \rightarrow i}(x)$.

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	4595	9337266	9889	5986	13928	4403	12490990
-15	305014	4541	9310617	10065	5919	13442	4232	12529684

648 **Appendix 2—table S2.** The read counts per position given the reference nucleotides are G or T of the
650 same human data as in Table S1. The negative position indices are the position on the reversed
652 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 $M_{G \rightarrow i}(x)$ or $M_{T \rightarrow i}(x)$.

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 x , i.e., $P_{i \rightarrow j}(x)$ while the conditional regression's target is to estimate the probability of observing a read of nucleotide j given its reference is i at position x , i.e., $P_{j|i}(x)$. Their

relationship is as follows:

$$P_{j|i}(x) = \frac{P_{i \rightarrow j}(x)}{\sum_{j \in B} P_{i \rightarrow j}(x)}.$$

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

The unconditional regression's log-likelihood function is defined as follows,

$$\begin{aligned} l_{\text{uncond}} &= \sum_x \sum_{i,j \in B} M_{i \rightarrow j}(x) \log P_{i \rightarrow j}(x) \\ &= \sum_x \left[M(x) \log P_{T \rightarrow T}(x) + \sum_{(i,j) \neq (T,T)} M_{i \rightarrow j}(x) \log \frac{P_{i \rightarrow j}(x)}{P_{T \rightarrow T}(x)} \right], \end{aligned} \quad (12)$$

where $M(x) = \sum_{i,j \in B} M_{i \rightarrow j}(x)$. According to the multinomial logistic regression, we assume,

$$\log \frac{P_{i \rightarrow j}(x)}{P_{T \rightarrow T}(x)} = \sum_{n=0}^{\text{order}} \alpha_{i,j,x,n} x^n \quad (13)$$

Applying Equation 13 to Equation 12, we have

$$l_{\text{uncond}} = \sum_x \left\{ -M(x) \log \left[1 + \sum_{(i,j) \neq (T,T)} \exp \left(\sum_{n=0}^{\text{order}} \alpha_{i,j,x,n} x^n \right) \right] + \sum_{(i,j) \neq (T,T)} M_{i \rightarrow j}(x) \sum_{n=0}^{\text{order}} \alpha_{i,j,x,n} x^n \right\} \quad (14)$$

The number of inferred parameters ($\alpha_{i,j,x,n}$), for the full conditional regression is $30 \times (\text{order} + 1)$.

And the relevant derivatives of the unconditional regression likelihood are as follows,

$$\frac{\partial l_{\text{uncond}}}{\partial \alpha_{i,j,x,n}} = -M(x) \frac{x^n \exp \left(\sum_{n=0}^{\text{order}} \alpha_{i,j,x,n} x^n \right)}{1 + \sum_{(i,j) \neq (T,T)} \exp \left(\sum_{n=0}^{\text{order}} \alpha_{i,j,x,n} x^n \right)} + M_{i \rightarrow j}(x) x^n. \quad (15)$$

Conditional Regression Likelihood

Viewed as the sum of log-likelihoods given the reference nucleotide $i \in B$, the conditional regression's log-likelihood function is,

$$\begin{aligned} l_{\text{cond}} &= \sum_{i \in B} \sum_x \sum_{j \in B} M_{i \rightarrow j}(x) \log P_{j|i}(x) \\ &= \sum_{i \in B} \sum_x \left[M_i(x) \log P_{T|i}(x) + \sum_{j \neq T} M_{i \rightarrow j}(x) \log \frac{P_{j|i}(x)}{P_{T|i}(x)} \right], \end{aligned} \quad (16)$$

where $M_i(x) = \sum_{j \in B} M_{i \rightarrow j}(x)$. Furthermore, if we assume,

$$\log \frac{P_{j|i}(x)}{P_{T|i}(x)} = \sum_{n=0}^{\text{order}} \beta_{i,j,x,n} x^n \quad (17)$$

By applying Equation 17 to Equation 16, we can obtain,

$$l_{\text{cond}} = \sum_{i \in B} \sum_x \left\{ -M_i(x) \log \left[1 + \sum_{j \neq T} \exp \left(\sum_{n=0}^{\text{order}} \beta_{i,j,x,n} x^n \right) \right] + \sum_{j \neq T} M_{i \rightarrow j}(x) \sum_{n=0}^{\text{order}} \beta_{i,j,x,n} x^n \right\} \quad (18)$$

The number of inferred parameters ($\beta_{i,j,x,n}$) for the full unconditional regression is $24 \times (\text{order} + 1)$. And the relevant derivatives of the conditional likelihood are as follows,

$$\frac{\partial l_{\text{cond}}}{\partial \beta_{i,j,x,n}} = -M_i(x) \frac{x^n \exp \left(\sum_{n=0}^{\text{order}} \beta_{i,j,x,n} x^n \right)}{1 + \sum_{j \neq T} \exp \left(\sum_{n=0}^{\text{order}} \beta_{i,j,x,n} x^n \right)} + M_{i \rightarrow j}(x) x^n. \quad (19)$$

Folded Multinomial Logistic Regression

The folded regressions use the same log-likelihood functions as the full regression (i.e., Equation 14 and 18) but are conducted based on a presumable symmetric PMD pattern, i.e., the probability of $C \rightarrow T$ at the position x of an random chosen ancient DNA strand is assumed to equal to the probability of $G \rightarrow A$ at the position $-x$. Such an theoretical assumption go match the current ancient library preparation process (Dabney, Meyer, and Pääbo, 2013; Henriksen, Zhao, and T. Korneliussen, 2022).

$$\alpha_{i,j,x,n} = \alpha_{c(i),c(j),-x,n}, \quad (20)$$

$$\beta_{i,j,x,n} = \beta_{c(i),c(j),-x,n}, \quad (21)$$

where $c(i)$ means the complimentary nucleotide of the nucleotide i , e.g., $c(A) = T$ and $c(G) = C$.

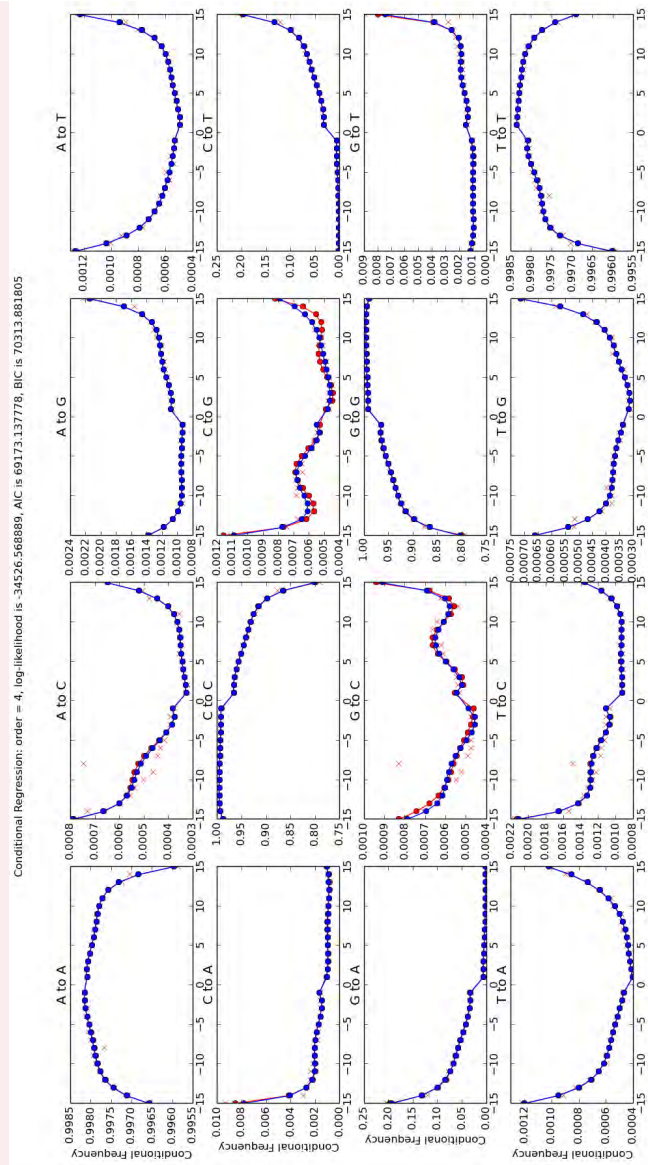
By doing the folded regression, we halve the number of inferred parameters ($\alpha_{i,j,x,n}$ or $\beta_{i,j,x,n}$). Hence The number of inferred parameters for the folded unconditional regression is $15 \times (\text{order} + 1)$, and that of folded conditional regression is $12 \times (\text{order} + 1)$.

Results for multinomial logistic regression

The optimization of the likelihood functions are based on the C++ library of gsl and use the function `gsl_multimin_fminimizer_nmsimplex2` with the initial searching point set to be the results of logistic regression. We here present here 4 figures pertaining to showcase the

performance of our model. The regression methods are based on the summary statistic of the counts of mismatches and the optimization is therefore in the scale of milliseconds. *Figure S1* and *Figure S2* are the conditional regression results of the ancient and control human data correspondingly. And *Figure S3* and *Figure S4* are the folded conditional regression results of the same data as above.

728



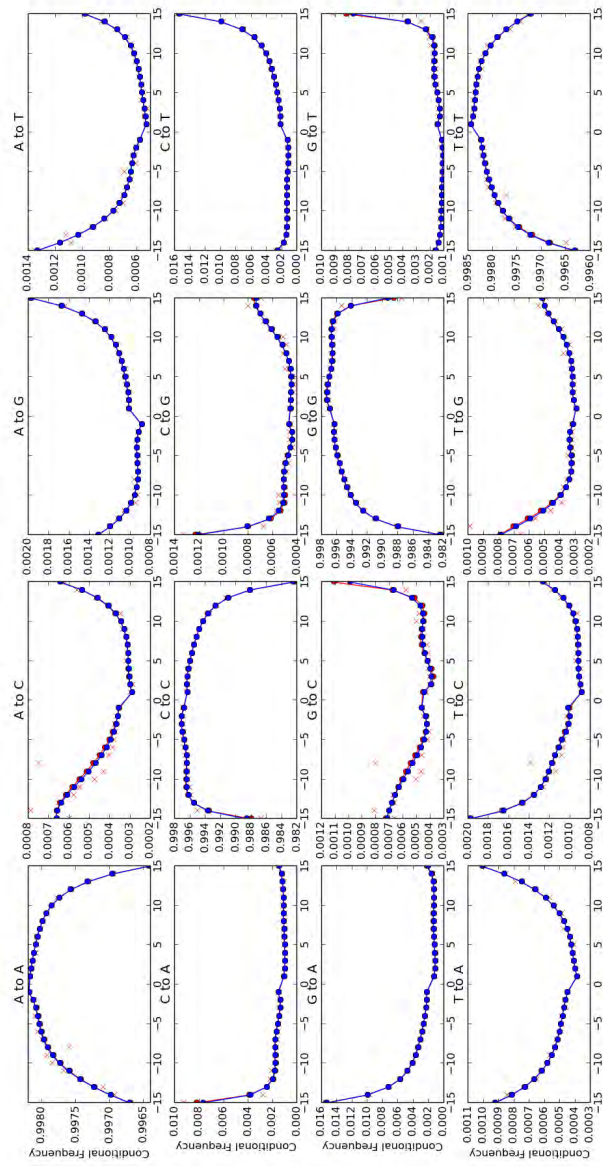
730

Appendix 2—figure S1. Conditional regression results with the order 4 of the ancient human data.

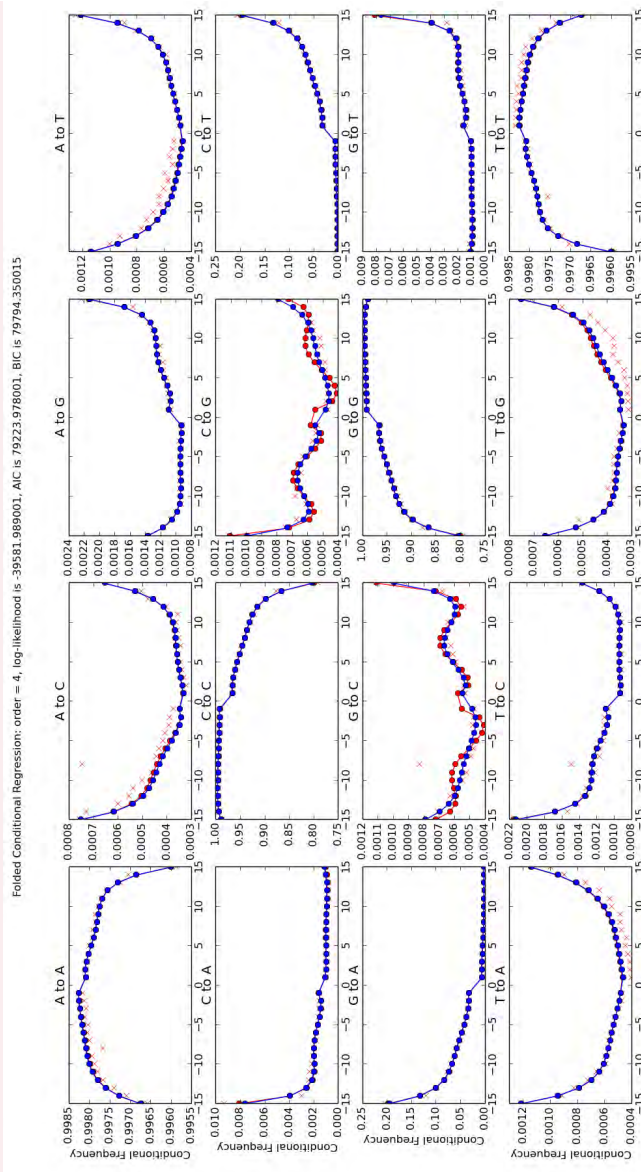
732

Each panel of figure represents a specific reference/read pair and plots its frequency across different positions. The positions from left to right are -1 to -15 and 15 to 1 .

Conditional Regression: order = 4, log-likelihood is 9508.304647, AIC is 19136.609294, BIC is 20252.301722



734 **Appendix 2—figure S2.** Conditional regression results with the order 4 of the control human data.
 736 Each panel of figure represents a specific reference/read pair and plots its frequency across different
 positions. The positions from left to right are -1 to -15 and 15 to 1 .



738

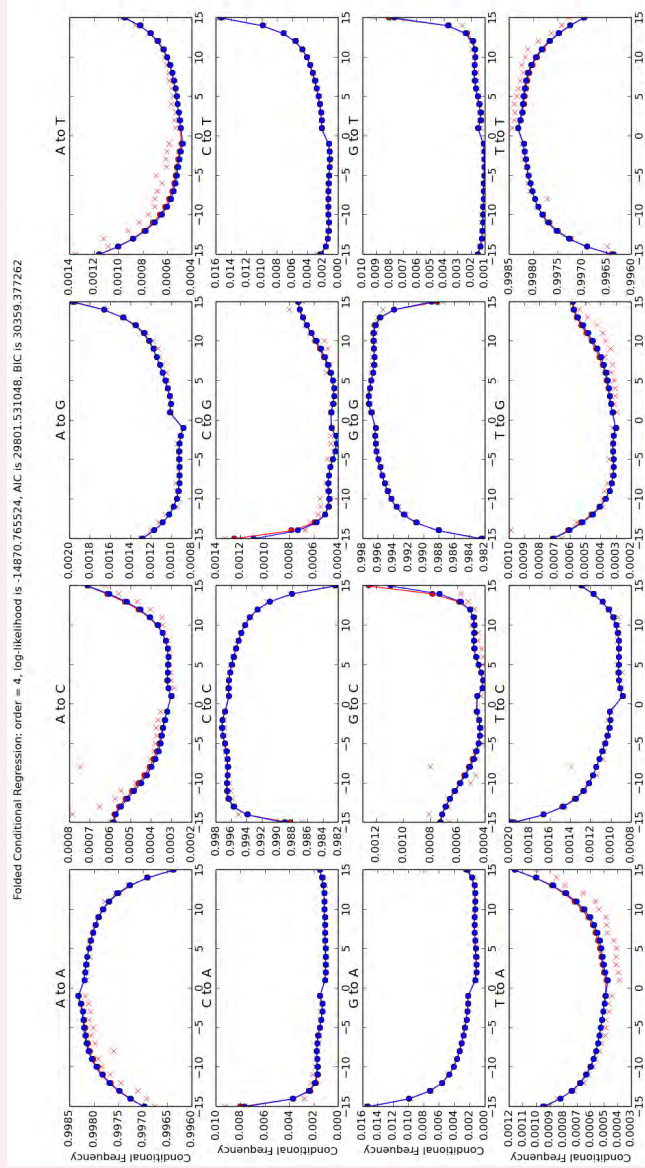
Appendix 2—figure S3. Folded conditional regression results with the order 4 of the ancient human

740

data. Each panel of figure represents a specific reference/read pair and plots its frequency across

742

different positions. The positions from left to right are -1 to -15 and 15 to 1 .



744 **Appendix 2—figure S4.** Folded conditional regression results with the order 4 of the control human
 746 data. Each panel of figure represents a specific reference/read pair and plots its frequency across
 different positions. The positions from left to right are -1 to -15 and 15 to 1 .

748 As shown in the figures, the regression models stabilize the coarse mismatch matrices
and describe a much more detailed PMD pattern (not only C→T and G→A, but also all other
750 reference and read combinations), but they might suffer from an overfitting issue espe-
cially when the data is limited, while the simpler regression model in the main text (*sub-*
752 *section 2.4*) shows an acceptable statistic power even with extremely small amount of data,
we thus recommend the readers to use the simpler regression model unless used with ex-
754 tremely high-coverage data.

Our code can also perform the unconditional regression, but as the unconditional regres-
sion needs to estimate more parameters based on the same dataset, it is more vulnerable
756 to a possible overfitting issue. We thus only present the figures of the conditional results.

758 Appendix 3**NGSNGS COMMANDS**

760 The resulting read data files (fastq files) were simulated with NGSNGS using the above
mentioned simulation parameters, all with the same quality scores profiles as used in ART
762 (Huang et al., 2012), based on the Illumina HiSeq 2500 (150 bp). The mapping was performed
using Bowtie-2 (Langmead and Salzberg, 2012):

```
764 ./ngsngs -i $genome -r $Nread -ld LogNorm,$lognorm_mean,$lognorm_std -seq SE \  
-f fq -q1 $quality_scores -m b,0.024,0.36,$damage,0.0097 -o $fastq  
766 bowtie2 -x $genome -q $fastq.fq --no-unal
```

Appendix 4

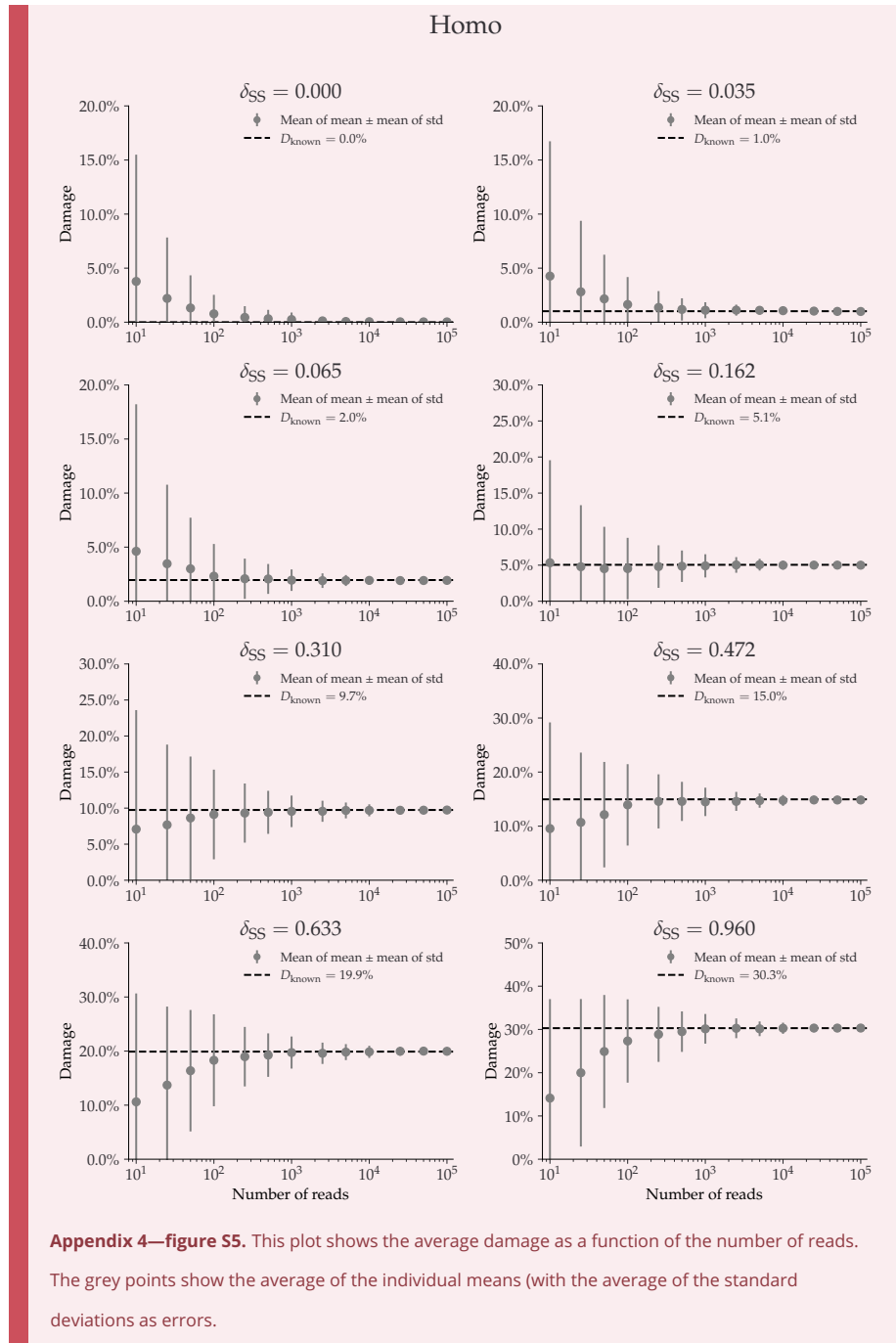
768

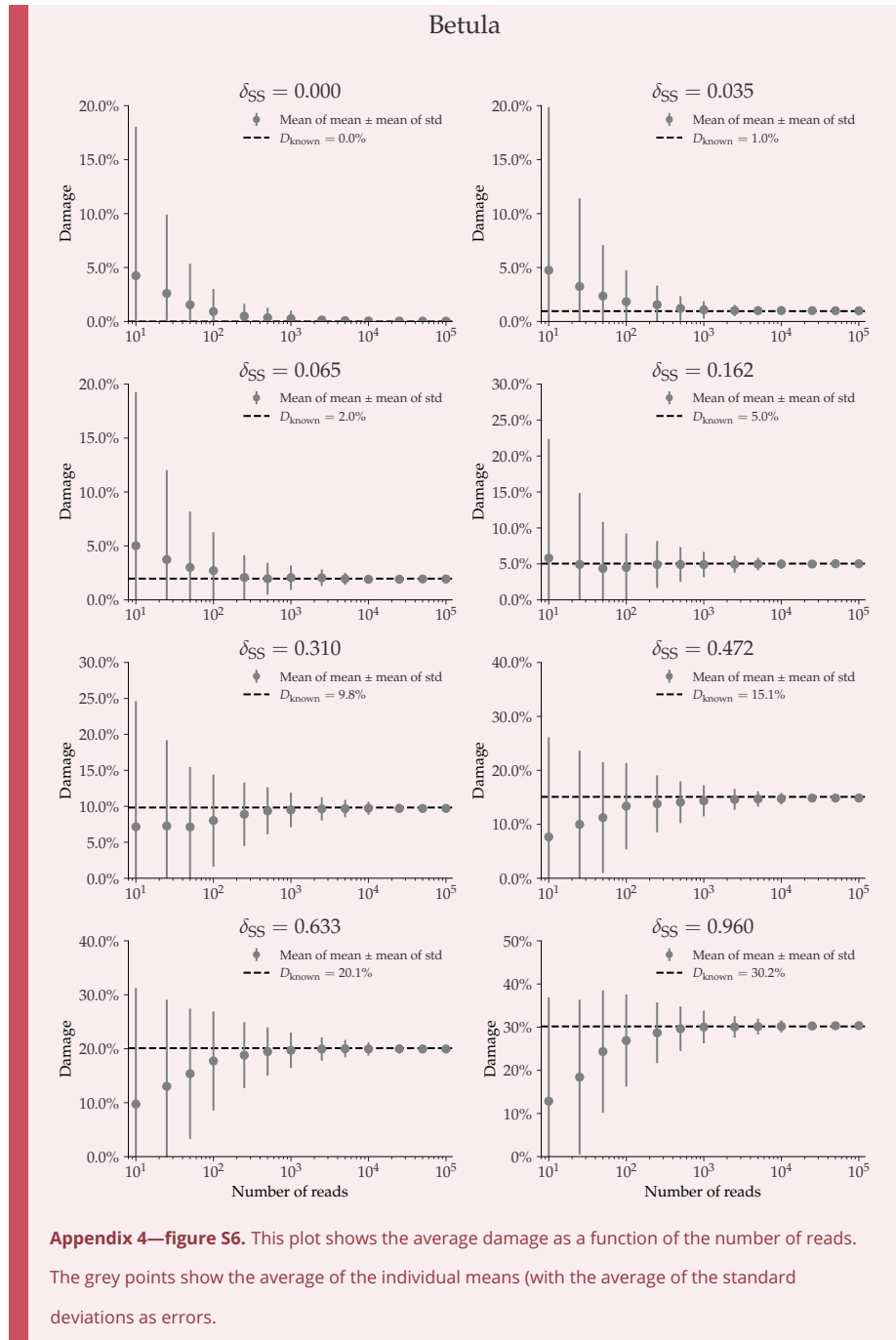
NGSNGS SIMULATIONS

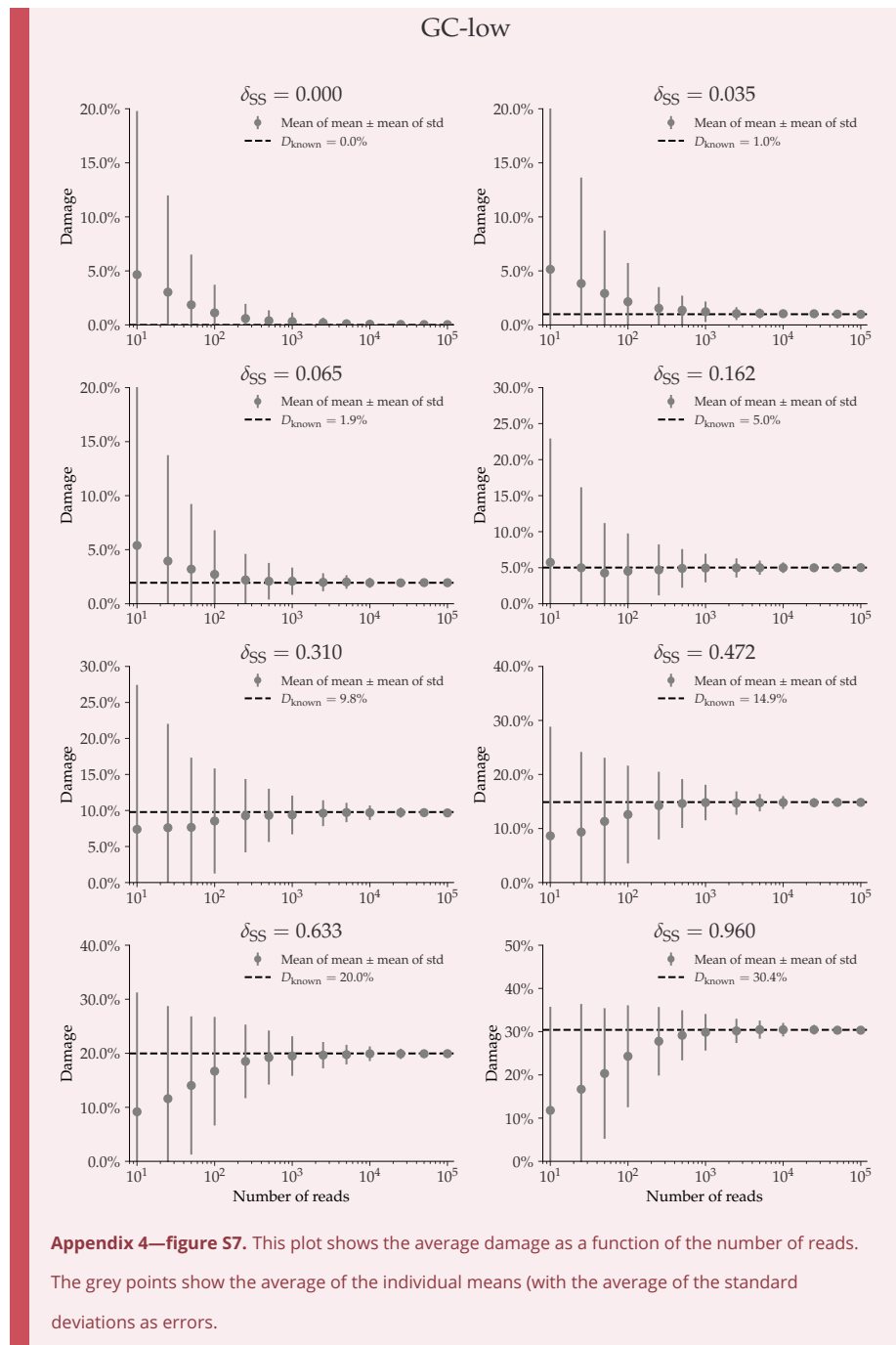
770

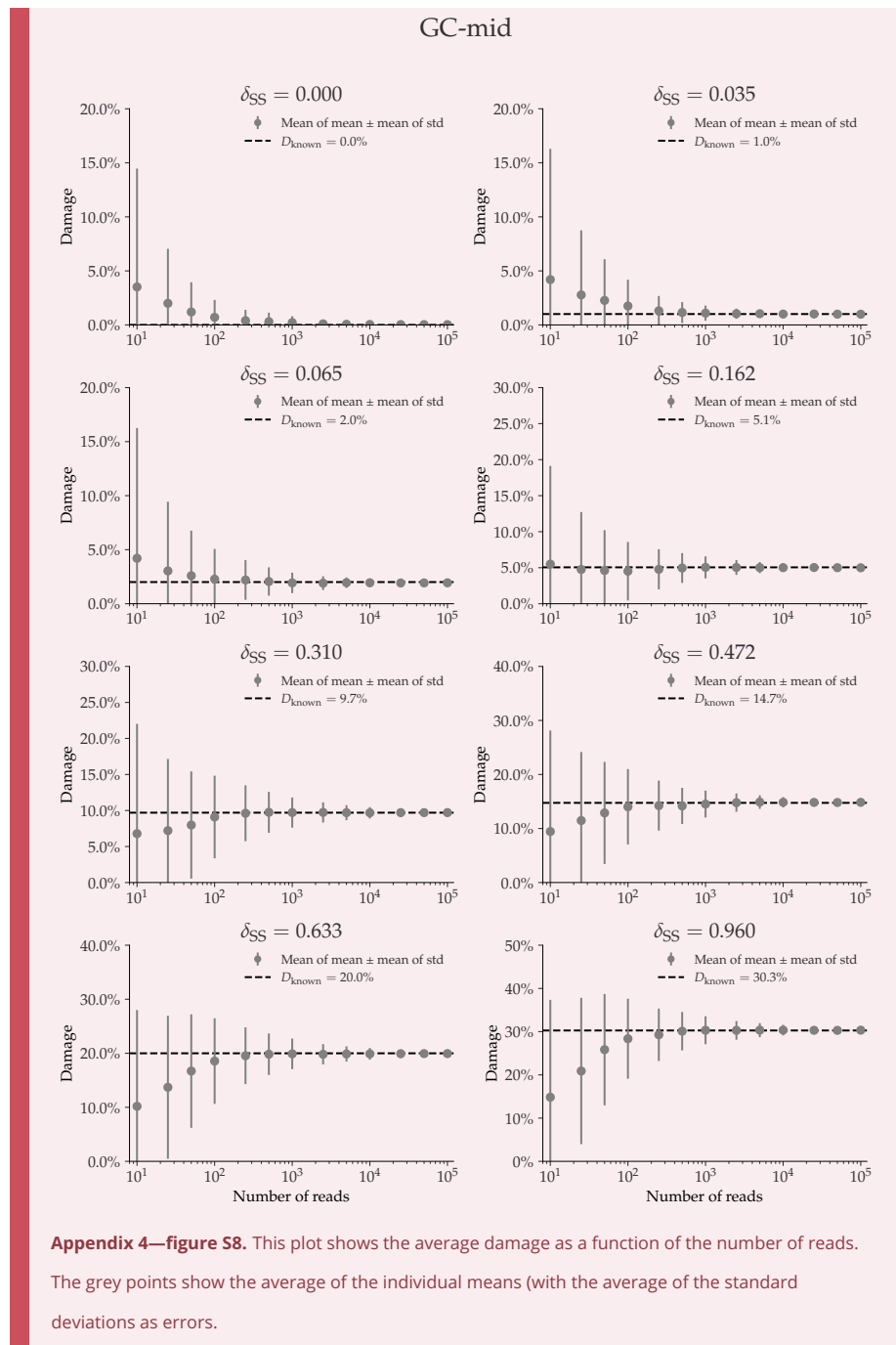
772

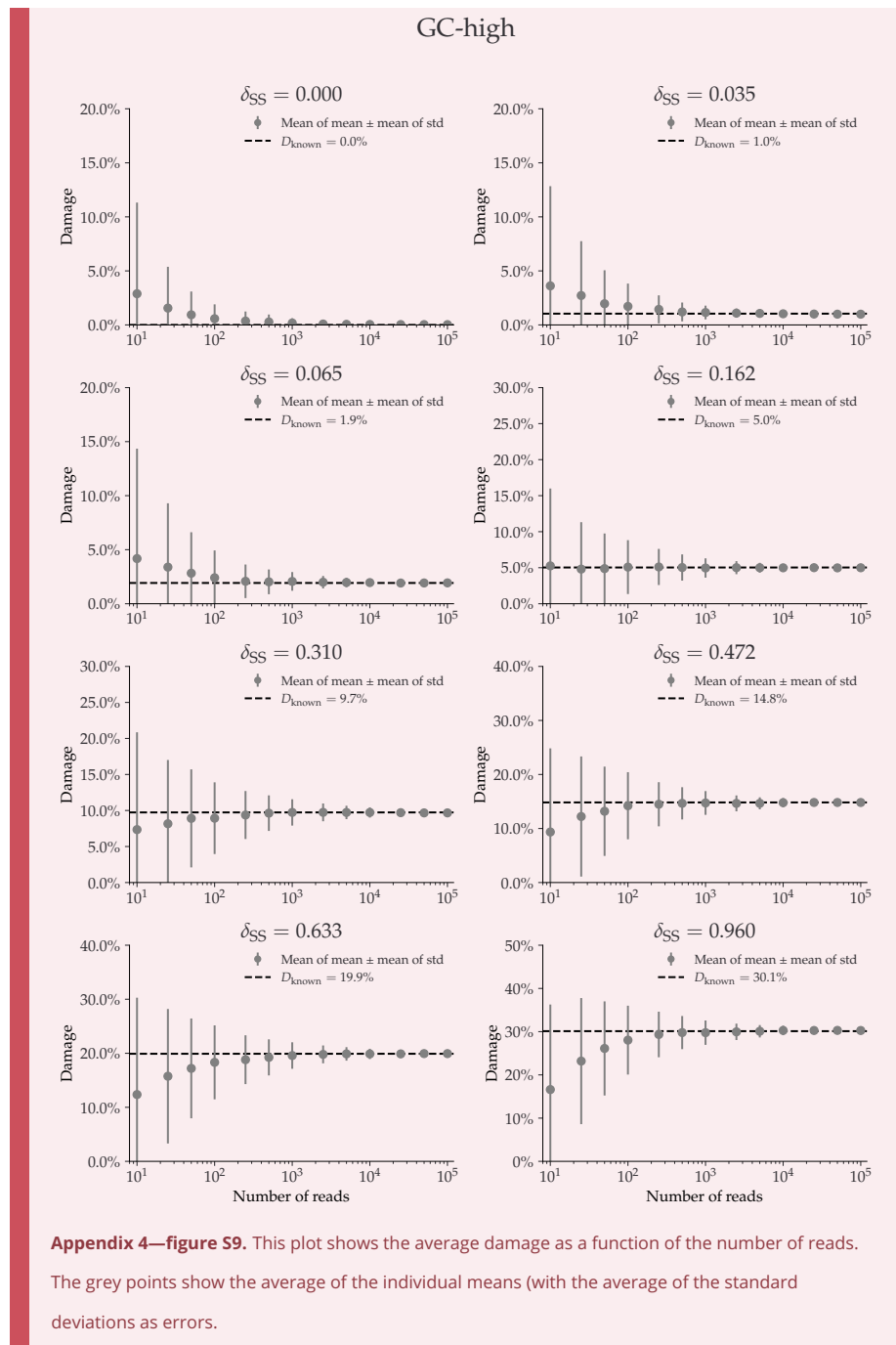
The following figures show the metaDMG damage estimates for the different NGSNGS simulations (Henriksen, Zhao, and T. Korneliussen, 2022). These simulations include different species (Homo Sapiens and Betula), different GC-levels (low, middle, high), different fragment length distributions (with mean 35, 60, and 90), and different contig lengths (length 1.000, 10.000, 100.000), see **subsection 3.1** for more information.

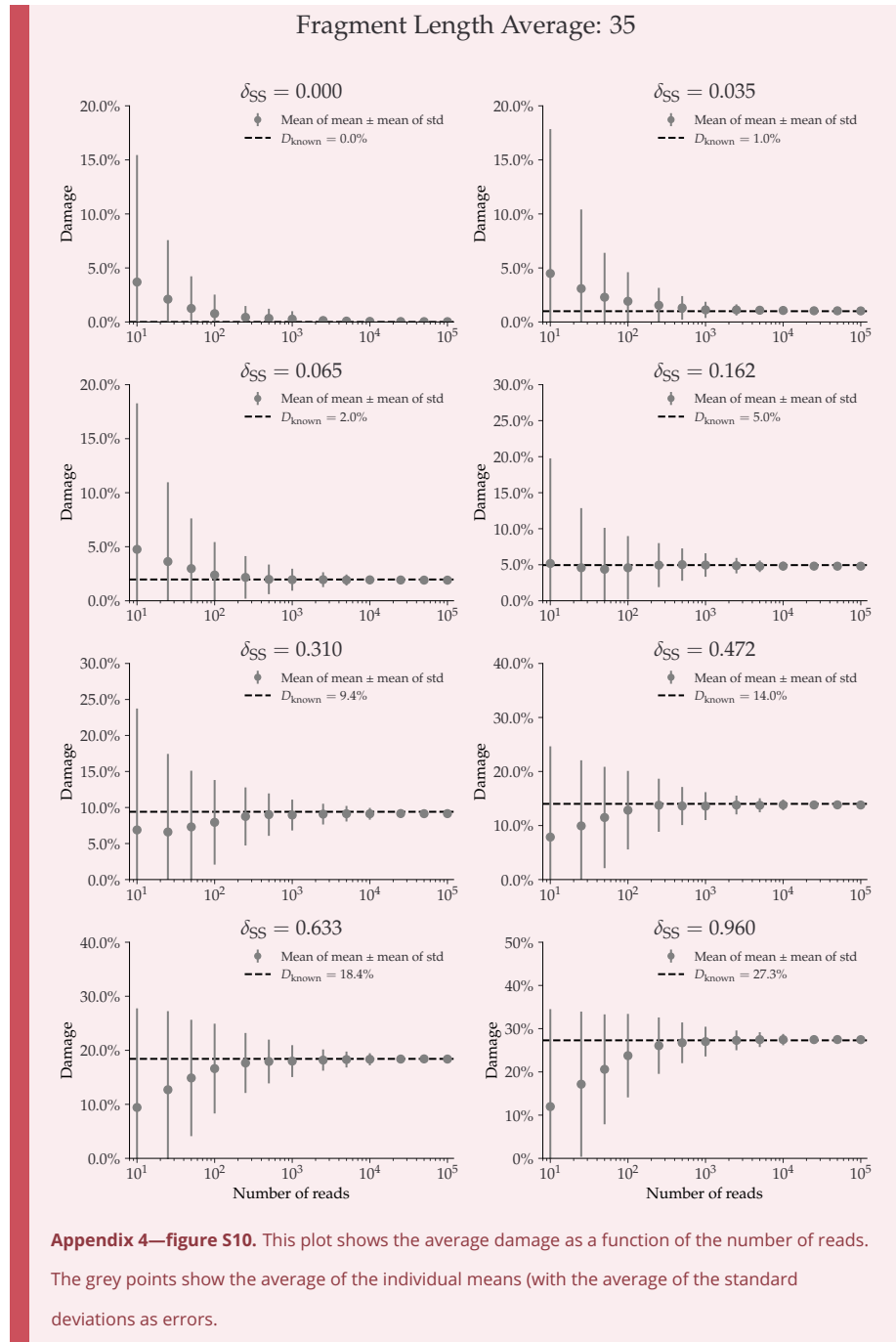


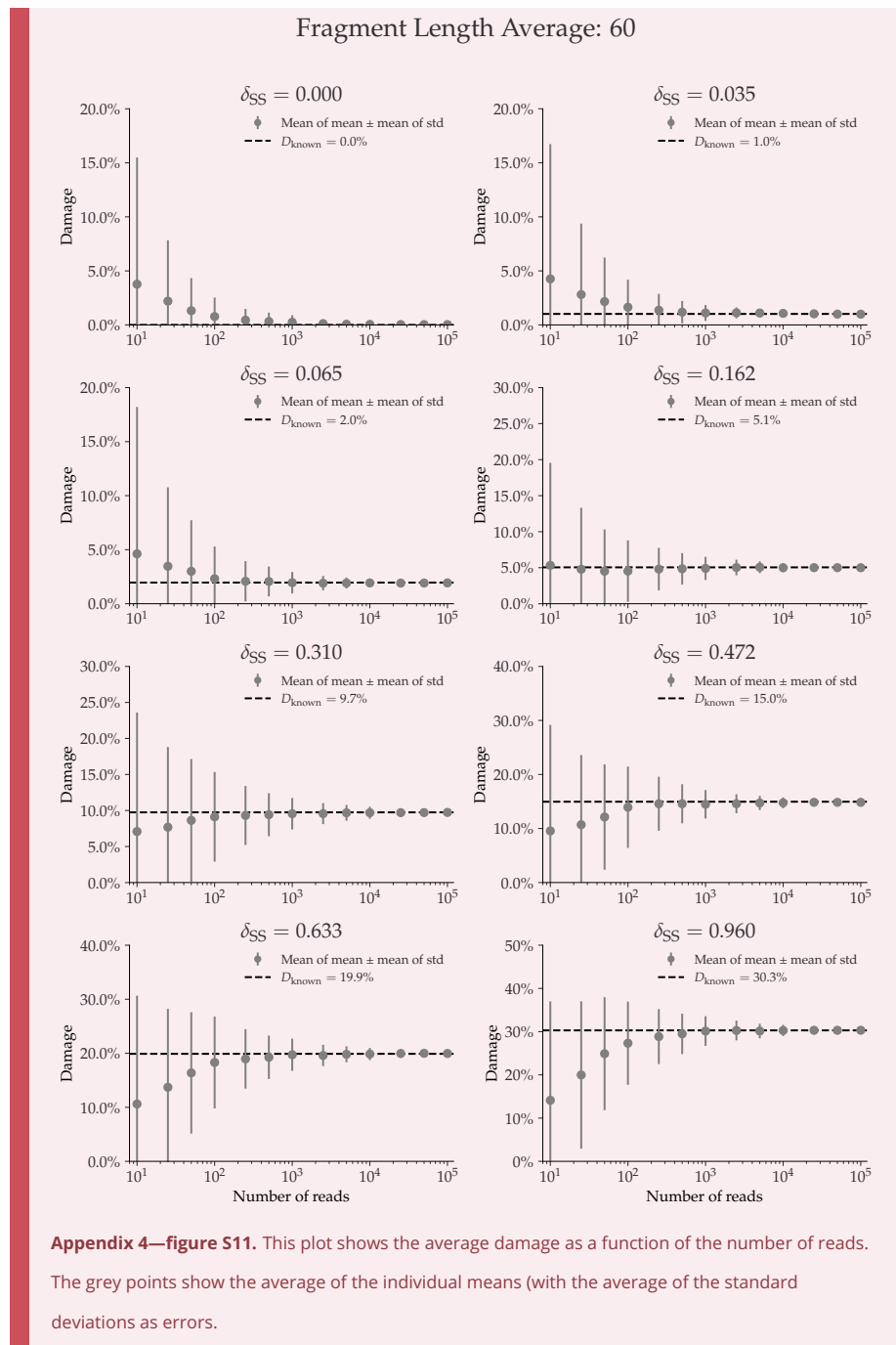


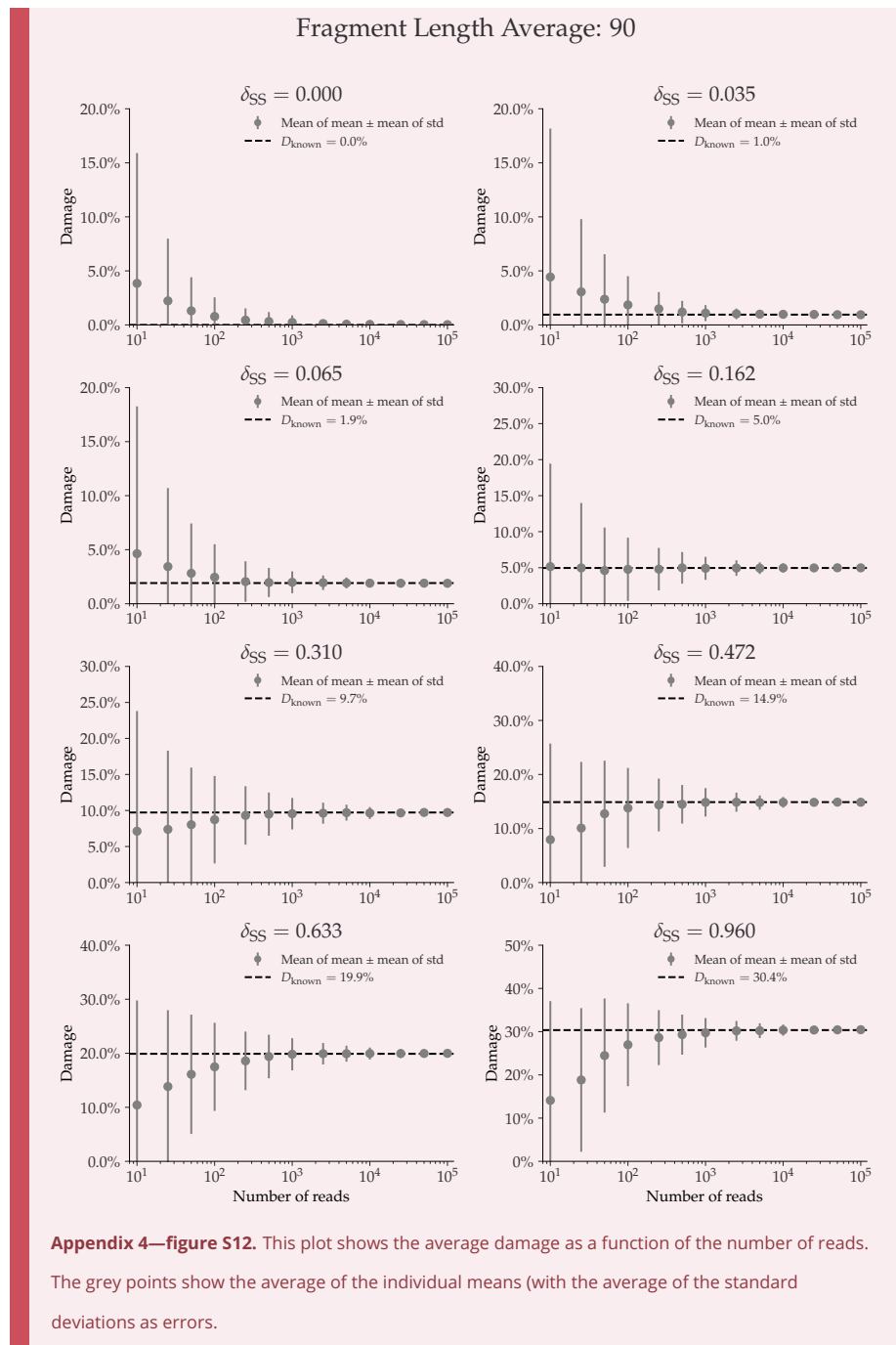


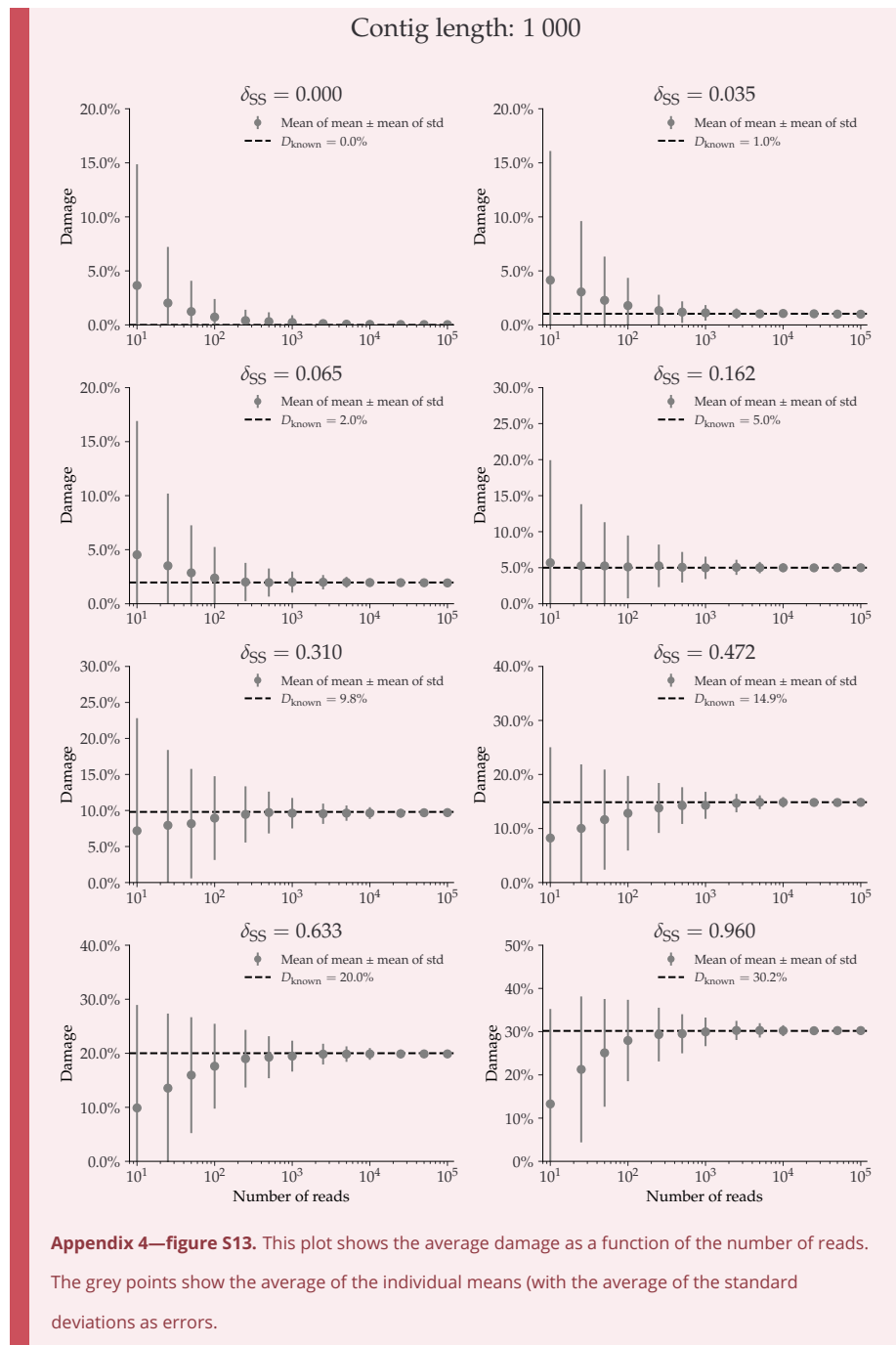


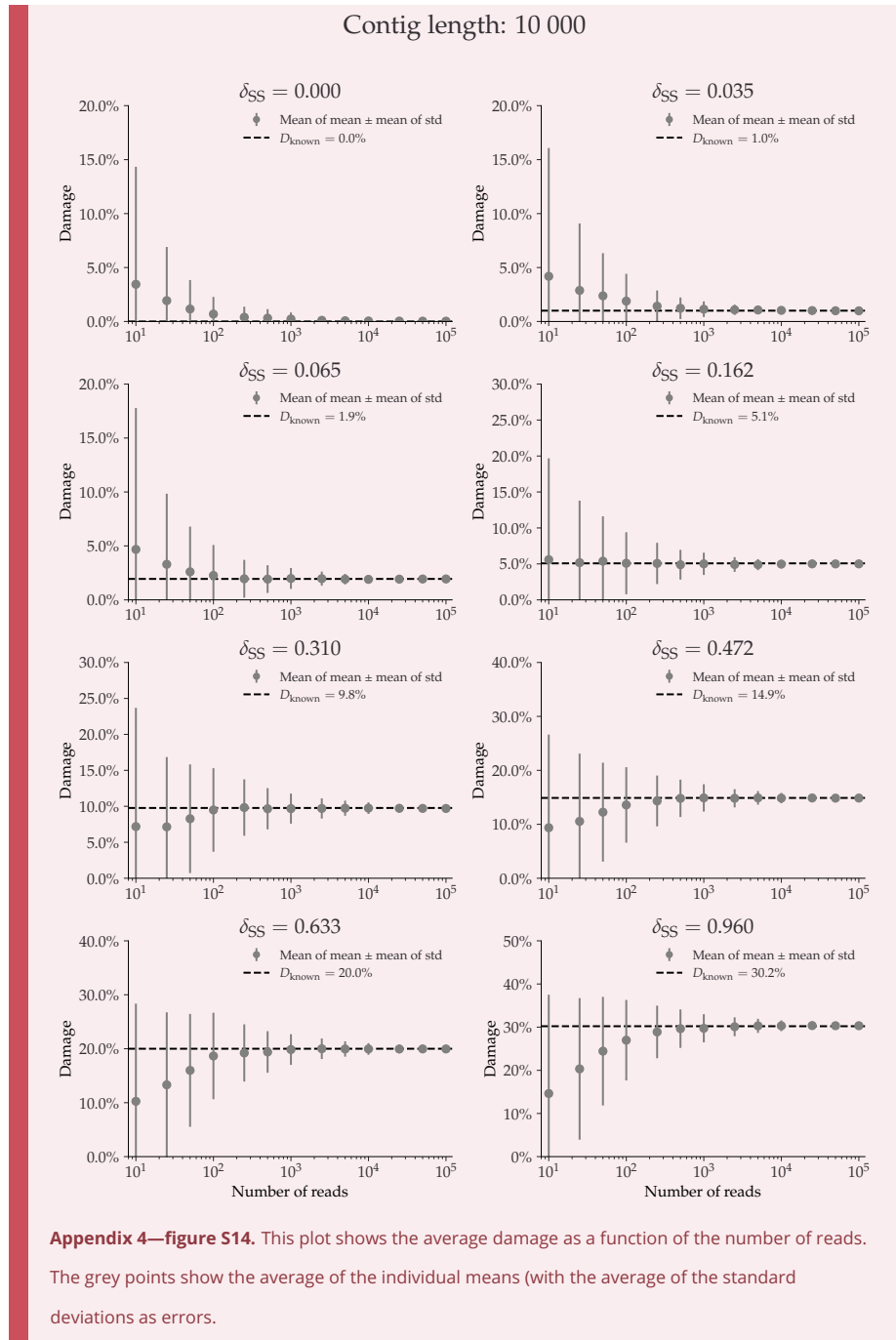


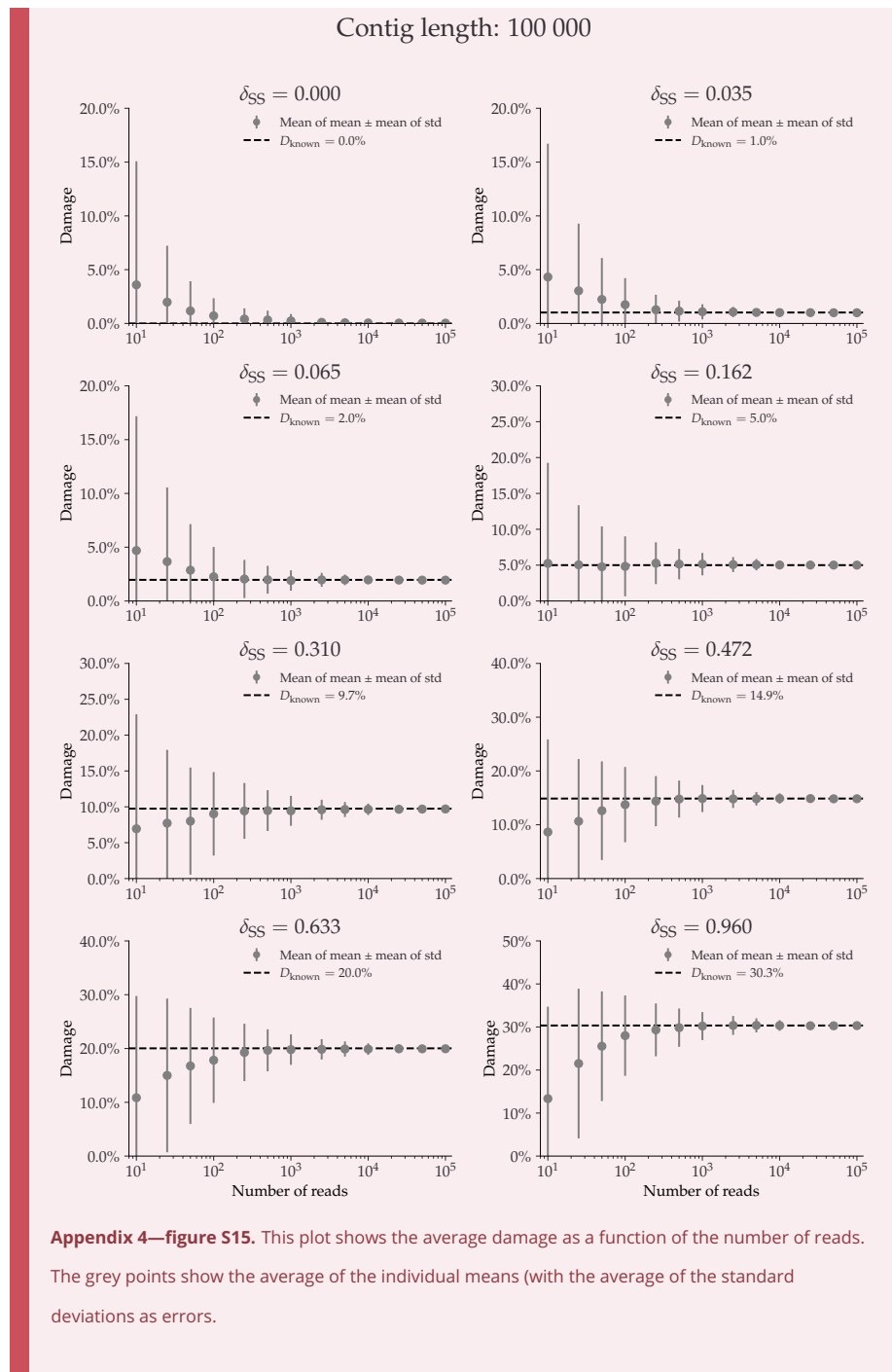












Appendix 5

830

NGSNGS SIMULATIONS – ZERO DAMAGE

832

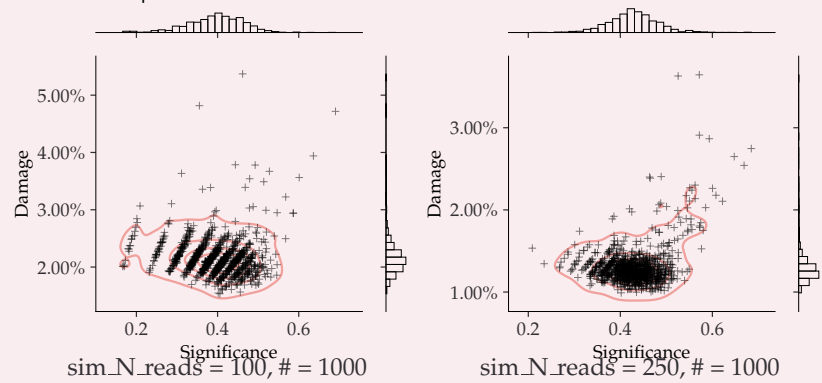
Damage estimates for non-damaged simulated data, each with 1000 replications, see *sub-*
section 3.1. The inferred damage is shown on the y-axis and the significance on the x-axis.

834

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
 $\text{sim_N_reads} = 25, \# = 1000$ $\text{sim_N_reads} = 50, \# = 1000$
 next to the scatter plot.

836

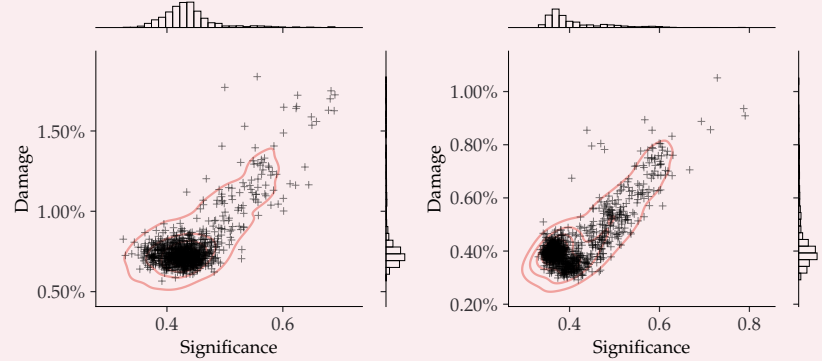
838



Appendix 5—figure S16. Left) 25 simulated reads. Right) 50 simulated reads.

840

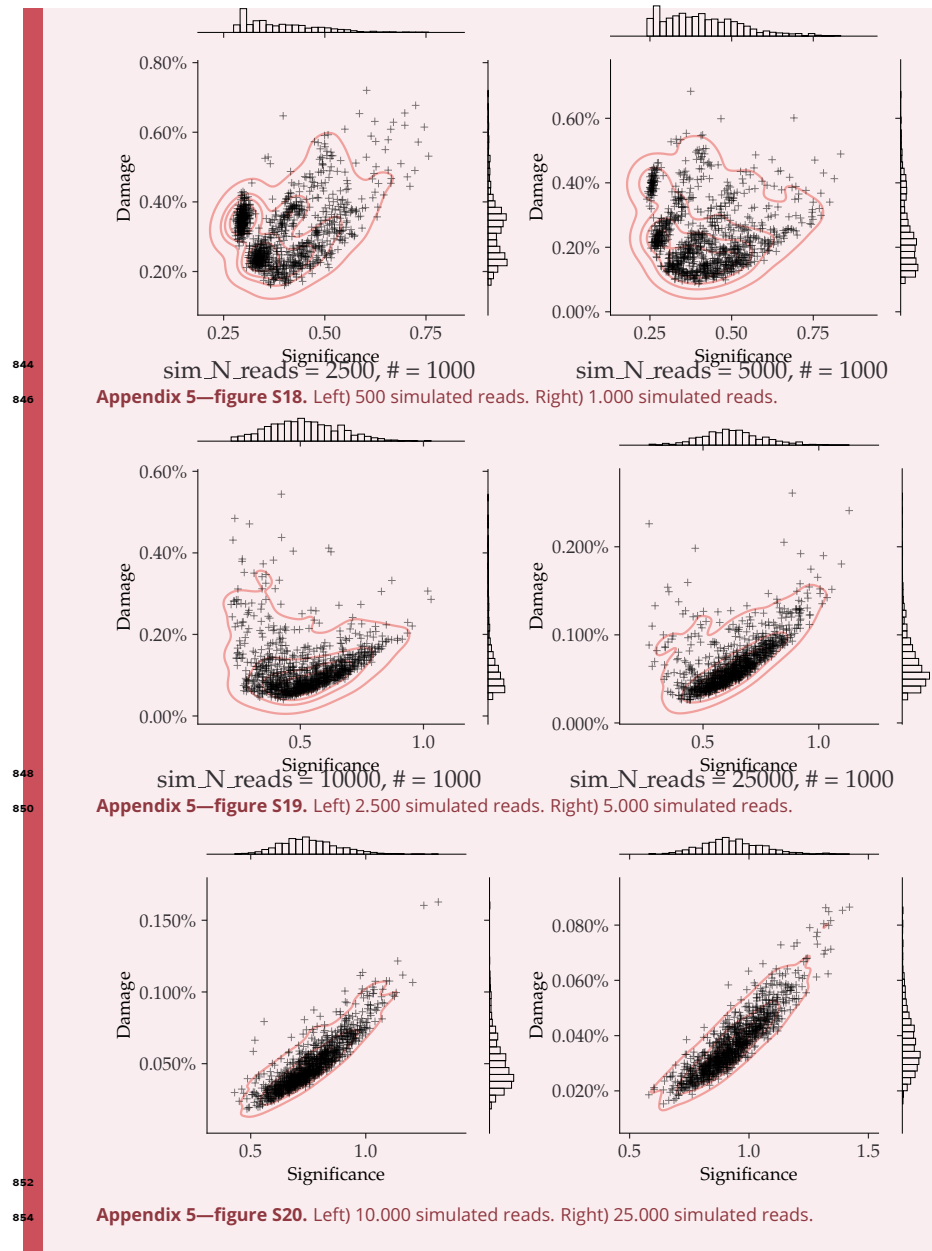
842



Appendix 5—figure S17. Left) 100 simulated reads. Right) 250 simulated reads.

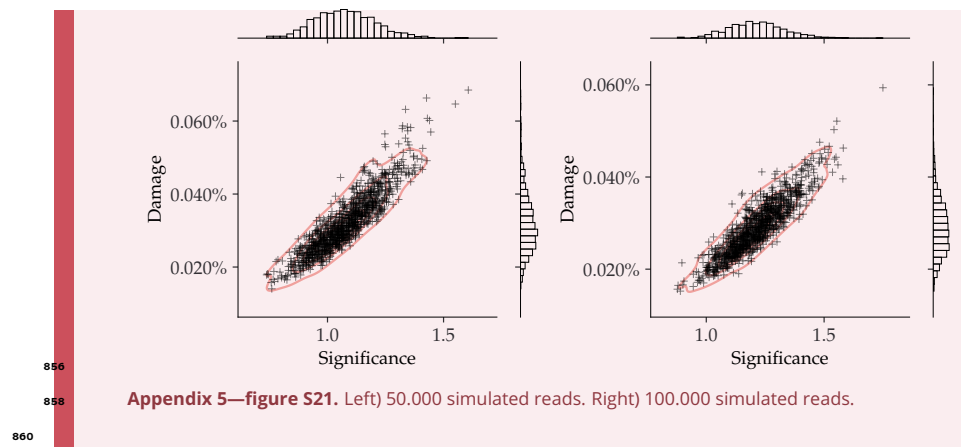
sim_N_reads = 500, # = 1000

sim_N_reads = 1000, # = 1000



sim_N_reads = 50000, # = 1000

sim_N_reads = 100000, # = 1000



Appendix 6

862

FALSE NEGATIVES

864

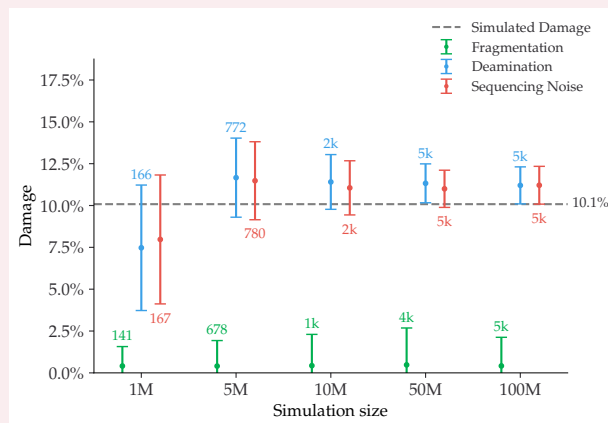
866

868

870

872

Even though the simple requirement of having more than 100 reads drastically improves the performance of the damage estimates, see [subsection 4.2](#), it does not identify all of the species that were simulated to be ancient. One of these non-identified taxa is the *Stenotrophomonas Maltophilia* species in the Pitch-6 sample. We show the damage estimates for different simulations for this particular taxa in [Figure S22](#) to quantify the behaviour of the damage estimate at the different stages of the simulation pipeline. For the final stage in the gargammel pipeline, ie. including fragmentation, deamination, and sequencing noise (red in the figure), only 167 reads are assigned to this specific taxa after mapping, when a total of 1 million reads were simulated. The significance is $Z_{\text{fit}} = 1.9$, just below the damage threshold.



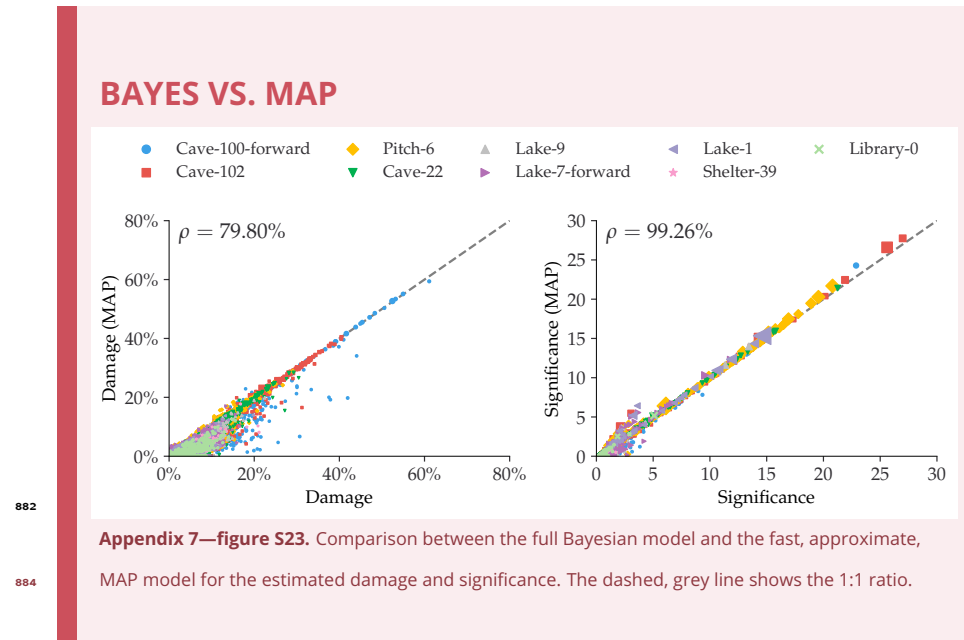
874

876

878

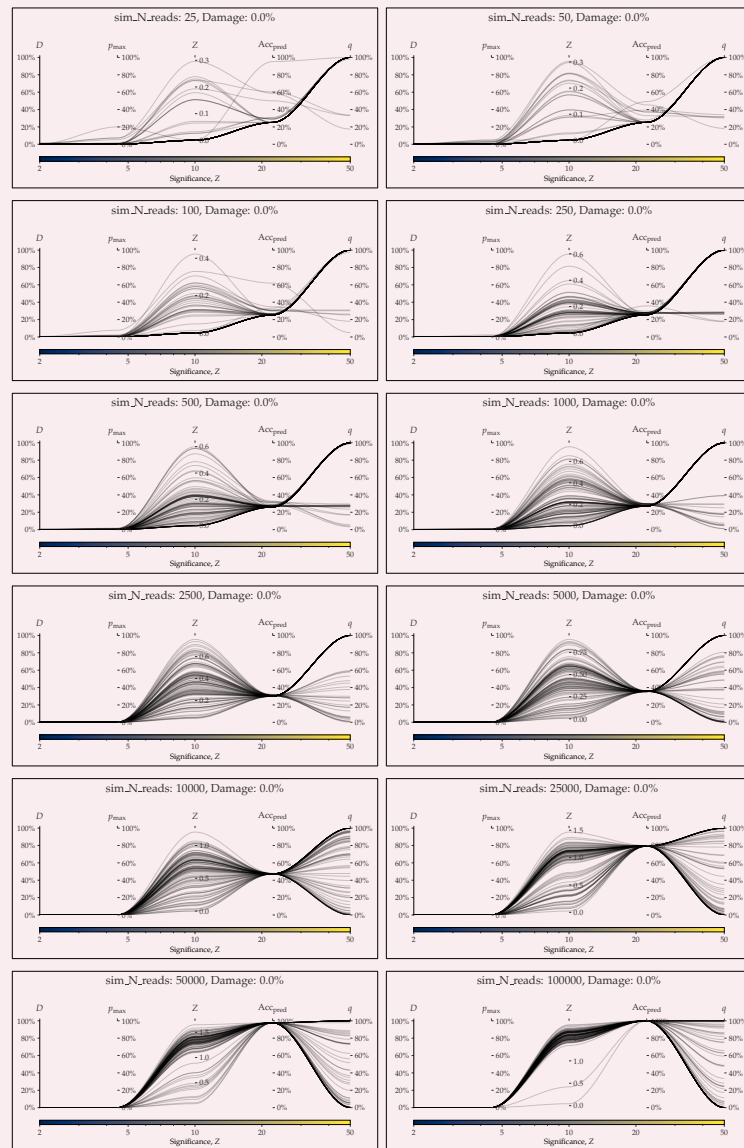
Appendix 6—figure S22. Damage estimates of the *Stenotrophomonas maltophilia* species from the Pitch-6 sample. Damage is shown as a function of the total simulation size, with the fragmentation files in green, the deamination files in blue and the final files including sequencing errors in red. All errors are 1σ error bars (standard deviation). The number of reads for each fit is shown as text the simulated amount of damage is shown as a dashed grey line.

880 Appendix 7



886 **Appendix 8****PYDAMAGE COMPARISON**

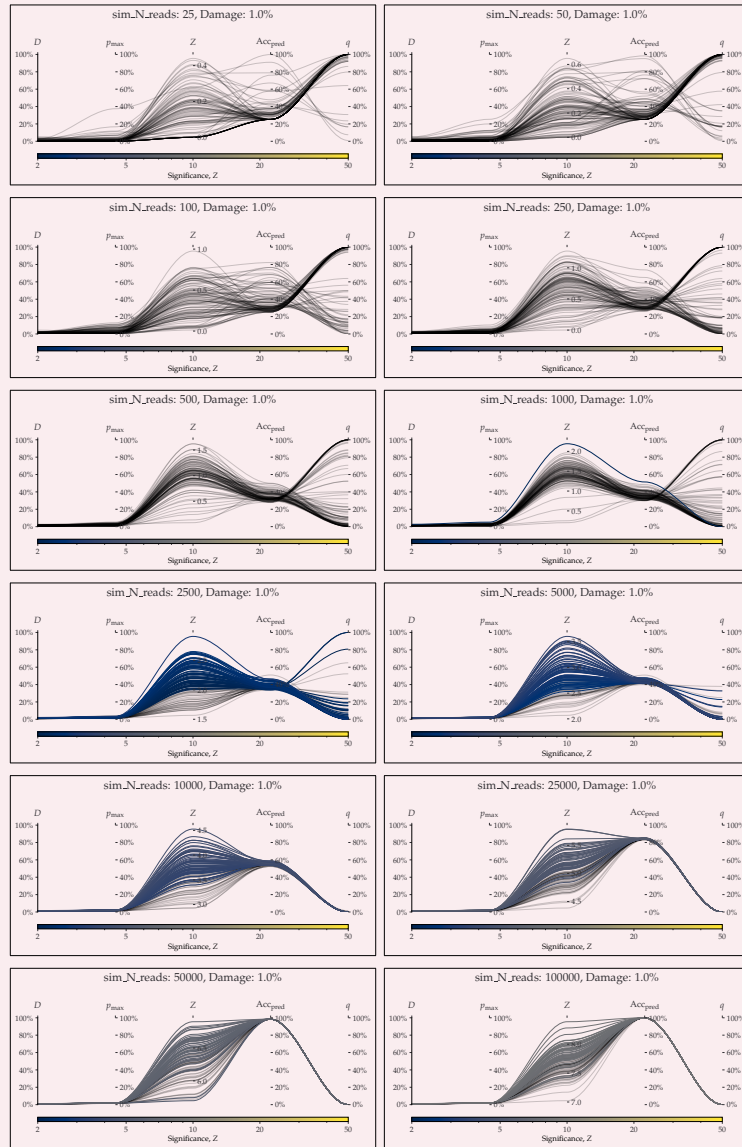
888 The following figures show the parallel coordinates plot comparing metaDMG and PyDamage
for the Homo Sapiens single-genome simulation with 100 reads for different amount of ar-
890 tificially added damage, see **subsection 4.5**. The two first axes show the estimated damage:
 D_{fit} by metaDMG and p_{max} by PyDamage. The following two axes show the fit quality: signif-
892 icance (Z_{fit}) by metaDMG and the predicted accuracy (Acc_{pred}) by PyDamage. The final axis
shows the q -value by PyDamage. Each of the 100 replications are plotted as single lines.
894 Replications passing the relaxed metaDMG damage threshold ($D_{\text{fit}} > 1\%$ and $Z_{\text{fit}} > 2$) are
shown in color proportional to their significance. Replications that did not pass are shown
896 in semi-transparent black lines.



898

Appendix 8—figure S24. parallel coordinates plot comparing metaDMG and PyDamage for 0% artificial damage.

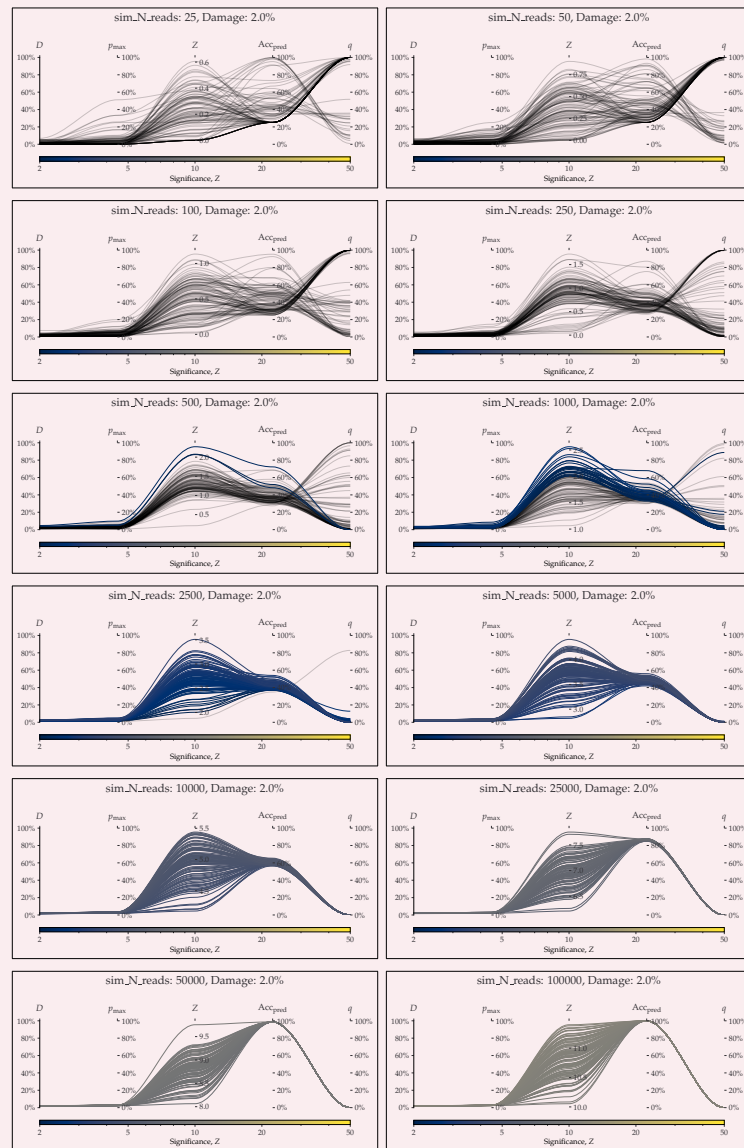
900



902

Appendix 8—figure S25. parallel coordinates plot comparing metaDMG and PyDamage for 1% artificial damage.

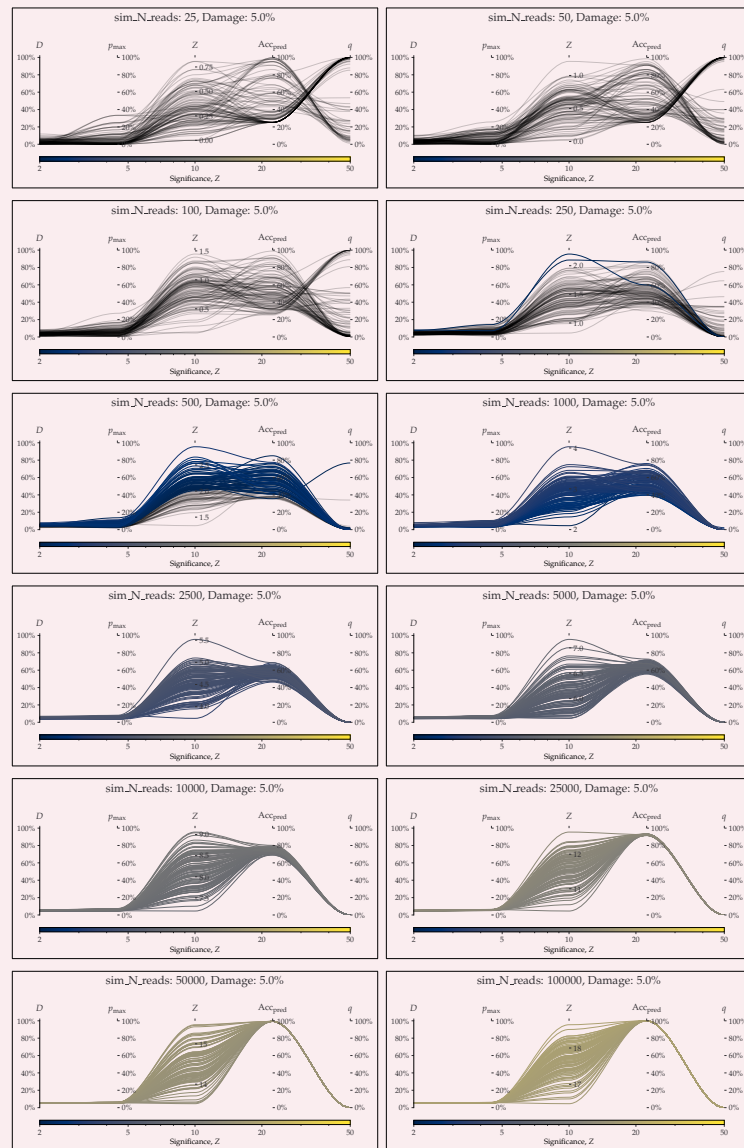
904



906

Appendix 8—figure S26. parallel coordinates plot comparing metaDMG and PyDamage for 2% artificial damage.

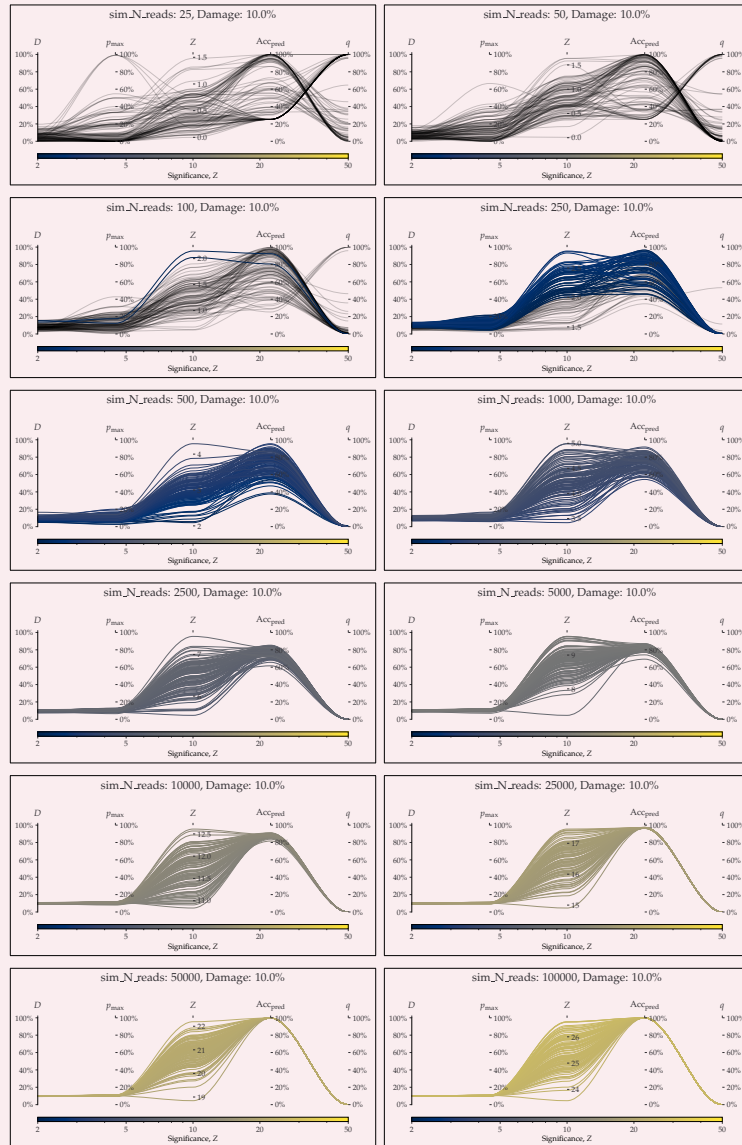
908



910

Appendix 8—figure S27. parallel coordinates plot comparing metaDMG and PyDamage for 5% artificial damage.

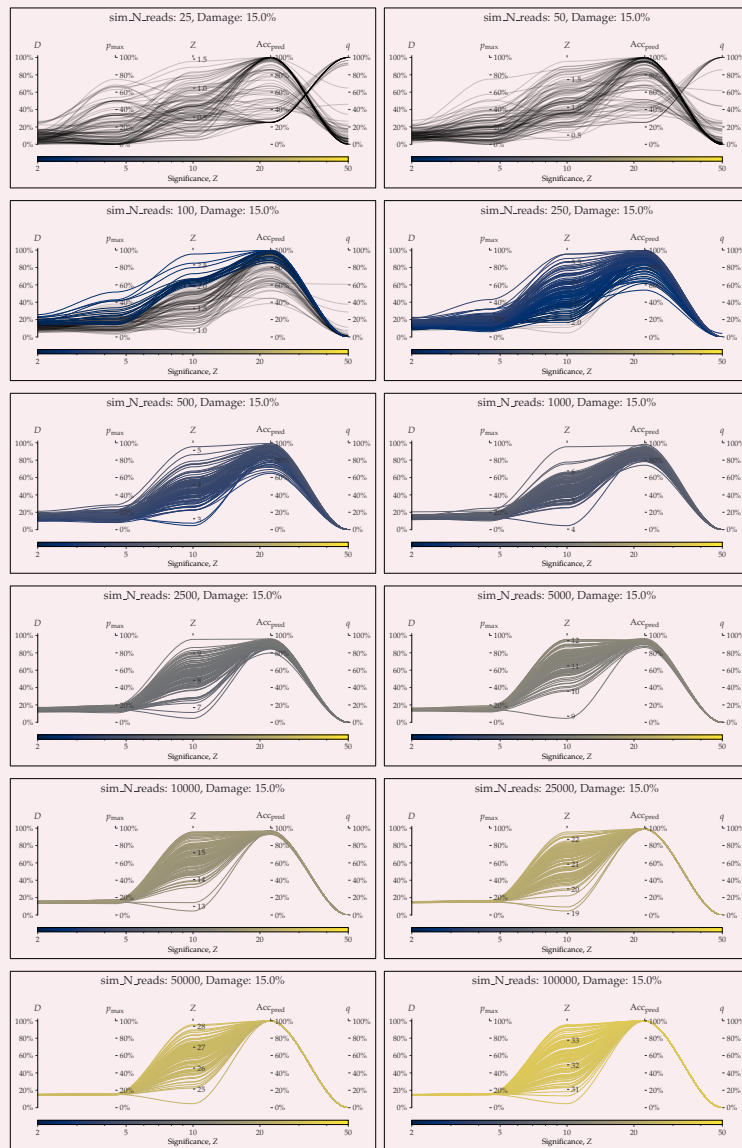
912



914

Appendix 8—figure S28. parallel coordinates plot comparing metaDMG and PyDamage for 10% artificial damage.

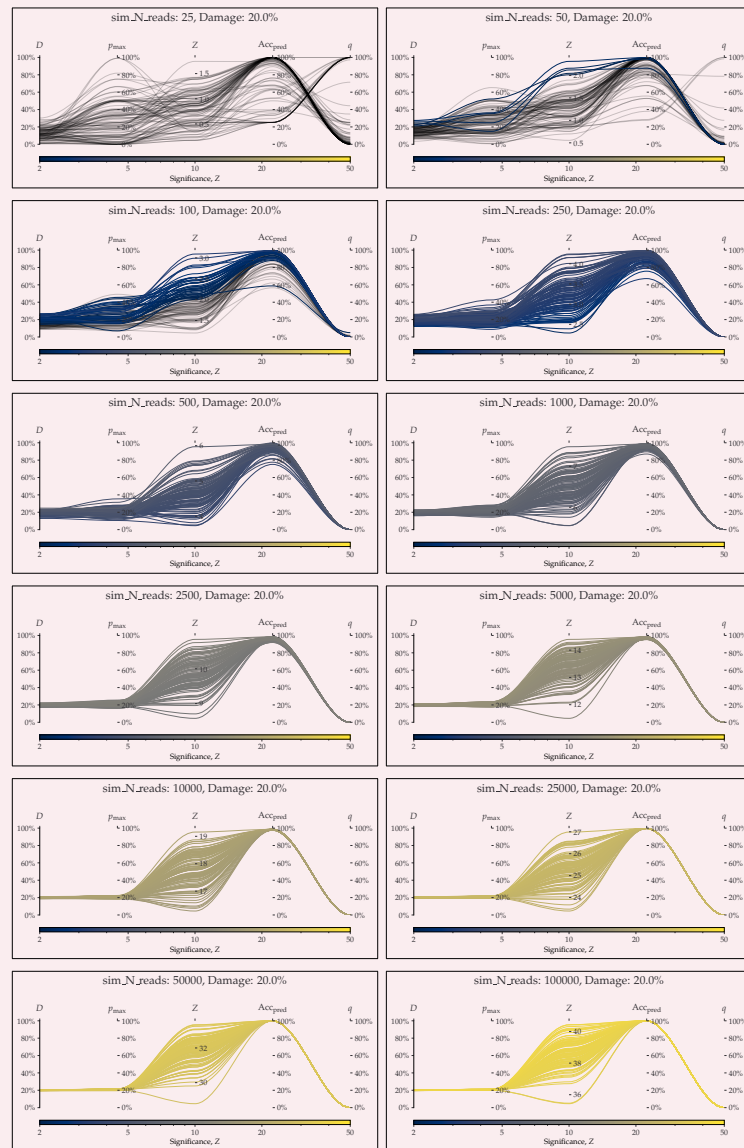
916



918

Appendix 8—figure S29. parallel coordinates plot comparing metaDMG and PyDamage for 15% artificial damage.

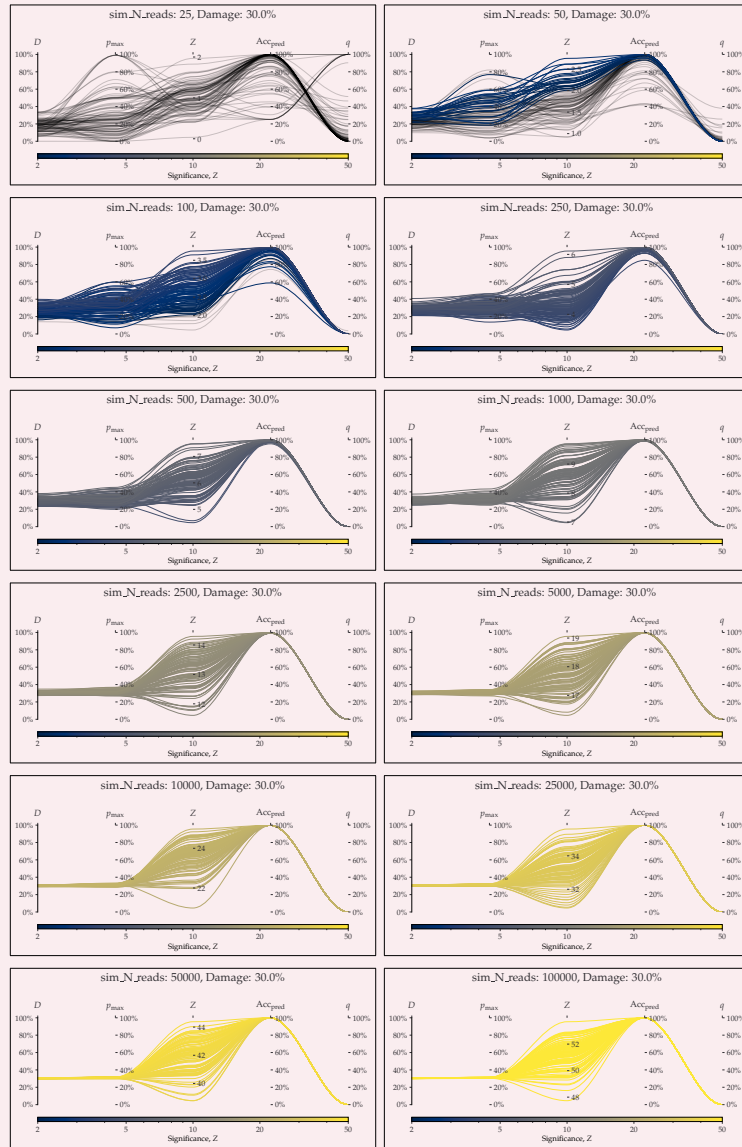
920



922

Appendix 8—figure S30. parallel coordinates plot comparing metaDMG and PyDamage for 20% artificial damage.

924



926

Appendix 8—figure S31. parallel coordinates plot comparing metaDMG and PyDamage for 30% artificial damage.

928

3 *Paper II*

The following 38 pages contain the paper:

Christian Michelsen, Christoffer C. Jørgensen, 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”. In review at BMJ Open.

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

ABSTRACT

Objectives: Machine-learning models may improve prediction of length of stay (LOS) and morbidity after surgery. However, few studies include fast-track programs, and most rely on administrative coding with limited follow-up and information on perioperative care. This study investigates benefits of machine-learning models for prediction of postoperative morbidity in fast-track total hip (THA) and knee arthroplasty (TKA).

Design: Cohort study with prospective recording of comorbidity and prescribed medication. Information on length of stay and readmissions through the Danish National Patient Registry and medical records.

Participants: Consecutive unselected primary THA or TKAs between 2014-2017 from seven Danish centers with established fast-track protocols. Data from 2014-2016 (n:18013) was used for training and data from 2017 (n:3913) was used for testing.

Outcomes: Ability of a machine-learning model based on boosted decision trees with 33 preoperative variables for predicting "medical" morbidity leading to LOS >4 days or 90-days readmissions vs. a logistic regression model. We also evaluated a parsimonious machine-learning and logistic regression model using the ten most important variables. Model performances were analyzed using precision, area under receiver operating (AUROC) and precision recall curves (AUPRC) among other performance measures. Variable importance was analyzed using Shapley Additive Explanations values.

Results: Using a threshold of 20% "risk-patients" (n:782), precision, AUROC and AUPRC were 13.6%, 76.3% and 15.5% vs. 12.4%, 74.7% and 15.6% for the machine-learning and logistic regression model, respectively. The parsimonious machine-learning model performed better than the full logistic regression model. Of the top ten variables, eight were shared between the machine-learning and logistic regression models, but with a considerable age-related variation in importance of specific types of medication.

Conclusion: Machine-learning algorithms using preoperative characteristics and prescriptions slightly improved identification of patients in high-risk of "medical" complications after fast-track THA and TKA. Such algorithms could help identify patients who benefit from intensified perioperative care.

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

STRENGTHS AND LIMITATIONS

Strengths

- Fully implemented fast-track protocols with complete follow-up through nationwide registries and medical records.
- State of the art machine-learning techniques
- Novel analysis on the importance of preoperative prescriptions in predicting postoperative morbidity.

Limitations

- Limited amount of multilevel continuous data, potentially limiting full realization of the machine-learning model.
- Registration of preoperative prescriptions dependent on reimbursement and lack of information on actual use on day of surgery

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

INTRODUCTION

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 (THA) and knee arthroplasty (TKA), 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.¹⁻³ However, despite such progress, a fraction of patients still have postoperative complications leading to prolonged length of stay or readmissions.^{1,3,4}

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.⁴⁻⁸ 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.⁴⁻⁸

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.^{9,10} This is also the case in THA, TKA and uni-compartmental knee replacement, where several publications on machine-learning algorithms for prediction of length of stay,^{11,12} complications,¹³ disability,¹⁴ potential outpatient setup,¹⁵ readmissions¹⁶ or payment models,^{17,18} have shown promising predictive value compared to conventional statistical methods.¹⁹

However, few papers have included fast-track 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 THA and TKAs within a fully implemented fast-track 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.²⁰ 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.²⁰ Thus, machine-learning models remain promising and could provide an improved basis for identifying a potential "high-risk" surgical population who may benefit from more extensive preoperative evaluation and postoperative medical care.

Consequently, we used a large consecutive cohort of patients undergoing fast-track total hip and knee replacement within a national public health-care system¹ to develop an improved

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

machine-learning model for preoperative prediction of “medical” complications resulting in prolonged length of stay and readmissions. Model performances were subsequently compared to a traditional logistic regression model. In addition to well-defined patient-reported preoperative risk-factors, we also included information on dispensed reimbursed prescriptions 6 months prior to surgery using a nationwide registry.²¹

METHODS

Reporting of the study is done in accordance with the Transparent reporting of multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement²² and the Clinical AI Research (CAIR) checklist proposal.²³

The study is based on the Centre for Fast-track Hip and Knee Replacement database which is a prospective database on preoperative patient characteristics and enrolling consecutive patients from 7 departments between 2010 and 2017. The database is registered on ClinicalTrials.gov as a study registry (NCT01515670). Patients completed a preoperative questionnaire with nurse assistance if needed. Additional information on reimbursed prescriptions 6 months prior to surgery was acquired using the Danish National Database of Reimbursed Prescriptions (DNDRP) which records all dispensed prescriptions with reimbursement in Denmark.²¹ Finally, data were combined with the Danish National Patient Registry (DNPR) for information on length of stay (counted as postoperative nights spent in hospital), 90-days readmissions with overnight stay and mortality. In case of length of stay >4 days or readmission, patient discharge summaries were reviewed for information on postoperative morbidity and in case of insufficient information, the entire medical records were reviewed. Readmissions were only included if considered related to the surgical procedure, thus excluding planned procedures like cancer workouts, cataract surgery, etc. Readmissions due to urinary tract infection or dizziness after day 30 were also considered unrelated to the surgical procedure. In case of postoperative mortality the entire medical record, including potential readmissions, was reviewed to identify cause of death. Evaluation of discharge and medical records was performed by PP supervised by CJ. In case of disagreement, records were conferred with HK. Subsequently, causes of length of stay >4, readmissions or mortality were classified as “medical” when related to perioperative care (renal failure, falls, pain, thrombosis, anemia, venous thromboembolism or infection etc.) and “surgical” if related to surgical technique (prosthetic infection, revision surgery, periprosthetic fracture, hip dislocation, etc.).¹ In case of a length of stay 4-6 days with a standard discharge summary describing a successful postoperative course, it was assumed that no clinically relevant postoperative complications had occurred. If length of stay was >6 days

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

but with standard discharge summary, the entire medical record was evaluated to confirm that no relevant complications had occurred.

For the present study, only cases between 2014 and 2017 were used to provide the most up-to-date data. All patients had elective unilateral total hip and knee replacement in dedicated arthroplasty departments with similar fast-track protocols, including multimodal opioid sparing analgesia with high-dose (125mg) methylprednisolone, preference for spinal anesthesia, only in-hospital thromboprophylaxis when length of stay ≤ 5 days, early mobilization, functional discharge criteria and discharge to own home.¹ There were no selection criteria for the fast-track protocol as it is considered standard of care, but we excluded patients with previous major hip or knee surgery within 90-days of THA or TKA and THA due to severe congenital joint disorder or cancer (Supplemental Material 1).

Patient and Public Involvement

There was no involvement of patients or the public in the planning or conduction of the study.

Outcomes

The primary outcome was to develop a machine-learning model to predict the occurrence of “medical” complications resulting in a length of stay >4 days or readmission and compare model performance with a traditional logistic regression model (primary outcome). Secondly, we investigated how inclusion of cases with a length of stay >4 days but no reported “medical” complication as a positive outcome influenced the model (secondary outcome). For both outcomes, we also investigated whether a parsimonious model including only the top ten variables would perform equally well as the full model. All figures and tables in the main text are based on the primary outcome; the corresponding figures for the secondary outcome are reported in the Supplemental Material.

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

Statistical Analysis

Data consisted of 33 input variables, of which 7 were continuous. All variables were collected prospectively, either through the patient completed questionnaire, through the DNDRP or a combination of both (table 1). Initially we trimmed the dataset 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 and allow for unbiased evaluation of model performance, data was subsequently split into a training set consisting of 18013 (82.2%) procedures from 2014-2016 and a test set of 3913 (17.8%) procedures from 2017, as is standard in modelling of data with a temporal component.²⁴ (Supplemental Material 1). These sample sizes are larger than the proposed minima of 3656, when assuming the model will explain 20% of the variability.²⁵ The data analysis was performed in Python and is available online at <https://zenodo.org/record/7330268>.

As reference model, we used logistic regression with missing values being handled by multiple imputations. All variables were then normalized to have zero mean and unit standard deviation by subtracting the original mean and dividing by the original standard deviation. In addition, we used boosted decision trees (LightGBM)²⁶ for the machine-learning models, as such methods work well with categorical data and missing values. We used cross entropy as the objective function for the machine-learning model.

The full machine-learning model was trained and hyperparameter optimized using the state of the art framework [Optuna](#)²⁷ with the [Tree-structured Parzen Estimator](#) algorithm²⁸ to efficiently sample hyperparameters and with a median stopping rule to minimize optimization time. The models were trained on the training data and then used for making predictions on the unseen test data (Supplemental Material 1). The classification threshold was calibrated such that 20% of the total number of patients were predicted as positive by the model (positive predictive fraction of 20%). We also included results for values of 25% and 30%. Furthermore, we trained two parsimonious models using machine-learning and logistic regression with only the 10 most important features. All mentioned models were calibrated using Platt's method (Supplemental Material 2).²⁹ Finally, we constructed a model based on age alone (Age) to explore the added value of multiple variable prediction.

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.^{30 31} Finally, we investigated a potential relation between reimbursed prescribed cardiac drugs, anticoagulants, psychotropics and pulmonary drugs and age, as the

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

relation between polypharmacy and postoperative outcomes have mainly been found in older patients.³²

For evaluating model performance, we computed the number of true positives (TP), false positives (FP), false negatives (FN), true negatives (TN), sensitivity (true positive rate = TP / (TP+FN)), precision (positive predictive value = TP / (TP+FP)). 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 imbalance.^{33 34} The MCC ranges between -1 (the 100% wrong classifier), 0 (the random classifier), and +1 (the perfect classifier). Finally, we computed the area under the receiver operating characteristic curve (AUROC) and the area under the precision recall curve (AUPRC). To evaluate the statistical difference between the classifiers, we applied a Bayesian metric comparison P(sensitivity),³⁵ which is the probability that a model will perform better than the machine-learning model relative to the sensitivity. Thus, for two equally performing models P(sensitivity) is \approx 50%.

RESULTS

Median age in the 3913 patients was 70 years (IQR 62-76), 59% were female and 58% had THA (table 1).

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

Preoperative characteristics n (%) unless otherwise specified	training set (n:18013)	test set (n:3913)
mean age (SD)	69.0 (62.0-75.0)	70.0 (62.0-76.0)
mean number of reimbursed prescriptions ¹ (SD)	2.0 (0.0-3.0)	2.0 (0.0-3.0)
female gender	755 (64.0)	12133 (58.2)
hip arthroplasty	9918 (54.8)	2260 (57.8)
mean weight in kg (SD)	80.5 (70.0-93.0)	81.0 (70.0-92.0)
mean height in cm (SD)	170.0 (164.0-177.0)	170.0 (164.0-177.0)
mean body mass index (SD)	27.5 (24.6-31.2)	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	5914 (32.9)	1381 (35.7)
with others	11971 (66.5)	2469 (63.8)
institution	116 (0.6)	21 (0.5)
missing	12 (0.6)	42 (1.1)
Hemoglobin (SD)	8.6 (8.1-9.1)	8.6 (8.1-9.2)
Missing	291 (1.5)	55 (1.4)
>2 units of alcohol/day	1382 (7.7)	286 (7.4)
Missing	57 (0.8)	36 (0.9)

1			
2			
3	active smoker	130 (11.0)	2751 (13.2)
4	missing	11 (0.9)	141 (0.7)
5	cardiac disease	2527 (14.0)	529 (13.7)
6	missing	17 (0.6)	53 (1.4)
7	hypercholesterolemia	5396 (29.9%)	1133 (29.3%)
8	missing	83 (0.5)	44 (1.2)
9	hypertension	9030 (51.4)	1849 (49.5)
10	missing	546 (3.0)	179 (4.6)
11	pulmonary disease	1668 (9.2)	355 (9.2)
12	missing	63 (0.4)	38 (1.0)
13	previous cerebral attack	1038 (5.8)	213 (5.6)
14	missing	157 (1.3)	77 (2.0)
15	previous VTE	1331 (7.5)	283 (7.4)
16	missing	283 (1.6)	66 (1.7)
17	malignancy (undefined)	1469 (8.1)	134 (3.4)
18	previous radically treated malignancy	1752 (9.7)	440 (11.2)
19	missing	136 (0.8)	40 (1.0)
20	chronic kidney disease	266 (1.5)	57 (1.5)
21	missing	276 (1.5)	50 (1.3)
22	family member with VTE	2235 (14.1)	430 (12.5)
23	missing	2189 (12.6)	479 (12.2)
24	regular snoring	266 (22.5)	5522 (26.5)
25	uncertain about snoring	208 (17.6)	3781 (18.1)
26	missing	259 (21.9)	3309 (15.9)
27	not feeling rested	7272 (42.4)	9340 (44.8)
28	uncertain about being rested	48 (4.1)	809 (3.9)
29	missing	105 (8.9)	1230 (5.9)
30	psychiatric disorder	1464 (8.4)	282 (7.6)
31	missing	580 (3.2)	182 (4.7)
32	<hr/> Characteristic based on combination of questionnaire and DNDRP <hr/>		
33	<u>Diabetes</u>		
34	diet treated diabetes ²	251 (1.4)	52 (1.3)
35	oral antidiabetics	1294 (7.2)	291 (7.5)
36	insulin treated diabetes ³	405 (2.2)	68 (1.8)
37	missing	68 (0.4)	36 (0.9)

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

¹Antirheumatica, steroids, anticoagulants, cardiac, cholesterol lowering, respiratory and psychotropic drugs.

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

Details on prescribed drug types are shown in Supplemental Material 3. Median length of stay was 2 (IQR: 1-2) days with 7.6% 90-days readmissions and the primary outcome occurring in 182 (4.7%) patients. When applying any model with a positive prediction fraction of 20% to the 3913 patients, 782 qualified as "risk-patients". The results are summarized in figure 1 and table 2.

57
58
59
60

Table 2: Performance of the models with a predefined positive prediction fraction of 20% for primary outcome.

Positive prediction fraction 20%	TP	FP	FN	TN	Sensitivity %	Precision %	MCC %	AUROC %	AUPRC %	Brier %	P (sensitivity) %
Full machine-learning model	106	676	76	3055	58.2	13.6	21.1	76.3	15.5	4.19	-
Full logistic regression model	97	685	85	3046	53.3	12.4	18.4	74.7	15.6	4.32	17.2
Parsimonious machine-learning model	100	682	82	3049	54.9	12.8	19.3	75.9	17.3	4.34	26.4
Parsimonious logistic regression model	90	692	92	3039	49.5	11.5	16.3	73.8	15.8	4.33	4.86
Age-only model	87	676	95	3055	47.8	11.4	15.8	69.7	12.1	38.8	3.55

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

When considering risk scores from the full machine-learning (figure 1a) and full logistic regression model leading to this risk-patient selection, 106 and 98 had the primary outcome, respectively. Correspondingly, the sensitivity and precision were 58.2% and 13.6% for the full machine-learning and 53.3% and 12.4% for the full logistic regression model, respectively. The full machine-learning model was superior (figure 1b) on all parameters (except AUPRC) compared to any of the other models, although the differences were minor (table 2).

The results were similar when using positive prediction fractions of 25% and 30%, but with the sensitivity for the full machine-learning model increasing to 64.3% and 69.2% and precision decreasing to 12.0% and 10.7%, respectively (Supplemental Material 4). Despite age being the single most important variable, age alone had a significantly lower sensitivity at 47.8%.

When evaluating feature importance, we found a strong correlation between the full machine-learning and full logistic regression model, with age and use of walking aids being the most important variables in both (figure 2a). From the combined importance of variables outside the top ten, the machine-learning approach extracted more information with fewer variables than logistic regression (figure 1b).

For the full machine-learning model, there was a clear signal that increasing age, number of reimbursed prescriptions, and presence of comorbidity, all contributed to an increased risk score. In contrast, a recent date of surgery and an increased hemoglobin level seemed to

1
2
3 reduce the calculated risk (figure 2b). Individual analysis of the SHAP interaction values for
4 types of anticoagulant prescriptions revealed that prescriptions on vitamin-K antagonists (VKA)
5 or adenosine diphosphate (ADP) antagonists increased, while acetylic salicylic acid and direct
6 oral anticoagulants (DOAC) reduced the risk score of the full machine-learning model,
7 regardless of age (figure 3a). The SHAP analysis of prescribed cardiac drugs revealed that
8 prescriptions on Ca²⁺-antagonists and betablockers in combination with one or two other
9 antihypertensives increased the risk-score, as did prescriptions on nitrates, other
10 antihypertensives and antiarrhythmics. For the remaining cardiac drugs, prescriptions either
11 reduced or had minor influence, and with limited relation with age (figure 3b). Preoperative
12 psychotropic prescriptions increased the risk-score except for antipsychotics (0.6%). For users
13 of selective serotonin inhibitors there was a clear age-related distinction with the risk score
14 being increased in elderly patients but decreased in those < 60 years (figure 3c). Finally, the risk
15 score increased with prescriptions on inhalation steroid and β -blockers, and more accentuated
16 in the younger patients (figure 3d).

17 The results including patients with a length of stay >4 days, but no reported postoperative
18 complications (secondary outcome) were similar as for the primary outcome. In general, we
19 found that the full machine-learning model was superior to the others, although the differences
20 were smaller than for the primary outcome. (Supplemental Material 5 listing outcome
21 parameters and Supplemental material 6 figure S1a-b showing distributions and ROC curves for
22 the secondary outcome). While the ten most important variables for the full machine-learning
23 model remained unchanged, familiar disposition for venous thromboembolism replaced gender
24 as one of the top ten important variables in the full logistic regression model (Supplemental
25 material 7 figure S2a-b showing SHAP values for the secondary outcome). Furthermore, the
26 SHAP analysis on specific prescribed drugs demonstrated that the machine-learning model
27 found no benefits from information on prescriptions on respiratory drugs, why all SHAP values
28 were zero. In addition, the reduced risk with acetylsalicylic acid and DOAC prescriptions, as well
29 as the influence of practically all cardiac drugs except for nitrates, other antihypertensives and
30 antiarrhythmics, was attenuated (Supplemental material 8 figure S3a-d showing SHAP-values of
31 prescriptions of specific drugs for the secondary outcome).

32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 **DISCUSSION**

52
53 We found that using a machine-learning algorithm including all 33 available variables and a
54 parsimonious machine-learning-algorithm encompassing only the 10 most important predictors
55
56
57
58
59
60

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

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 the primary and the secondary outcome, it was less suited for prediction of postoperative medical complications after fast-track THA and TKA 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.

A previous systematic review also found that machine-learning algorithms may provide better prediction of postoperative outcomes in THA and TKA.³⁶ 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.³⁶ 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.¹³ 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,³⁷ 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 uni-compartmental knee replacement¹² and spine surgery,⁹ they are not standard in fast-track protocols and not easy to interpret from a pathophysiological point of view.

Most decisions on which patients may benefit from more extensive postoperative care will likely need to be conducted preoperatively, as there is an increasing need to prioritize limited health-care resources. Thus, although postoperative information such as duration of surgery, perioperative blood length of stays or postoperative hemoglobin have been included in other

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

studies³⁷, we decided against the use of peri- and postoperative data. The same approach has been used by Ramkumar *et al.* who used U.S. National Inpatient Sample data including 15 preoperative variables, to predict length of stay, patient charges and disposition after both TKA³⁸ and THA.¹⁸ However, these studies were not conducted in a socialized health care system, and the main focus was on the need for differentiated payment bundles and without specific information on the reason for increased length of stay or non-home discharge.³⁸ Wei *et al.* used an artificial neural network model to predict same-day discharge after TKA, based on the NSQUIP database from 2018 and found that six of the ten most important variables were the same compared with logistic regression, similar to our findings.³⁹ However, patients with one-day length of stay were intentionally excluded due to variations in in-patient vs. out-patient registration.³⁹

Age has traditionally been a major factor when predicting surgical outcomes and remained the single most important predictor in our study.. However, although elderly patients had increased risk of postoperative complications, likely related to decline of physical reserves,⁴⁰ the use of chronological age alone was inferior compared to both machine-learning and logistic regression models incorporating comorbidity and functional status. Thus, using age by itself for identifying the high-risk population resulted in missing 18% of the “true risk-patients” (87 compared to 106 in the full ML model).

We used the SHAP values for estimation of feature importance, thus providing a better understanding of the otherwise “black-box” machine-learning model. The SHAP values showed which variables contribute most to the risk-score predictions. In this context, inclusion of specific data on reimbursed prescriptions 6 months prior to surgery based upon the unique Danish registries, unsurprisingly found increased risk-scores with increased number of prescriptions and with the majority being in elderly patients. Similarly, a Canadian study in elective non-cardiac surgery found decreased survival and increased length of stay and readmissions and costs in patients >65 years with polypharmacy.³² However, this is a complex relationship where some patients benefit from their treatments, while other may suffer from undesirable side-effects. Consequently, the authors cautioned against altering perioperative practices based on current evidence.³² However, the information from the included prescriptions with SHAP analysis may provide inspiration for new hypothesis-generating studies such as investigation of the potential differences in risk-profile between having preoperative prescribed VKA and DOAKs. Also, the age-related differences in risk from SSRI's in our study could guide further studies on “deprescription”.

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

Another requirement for machine-learning-algorithms to be clinically useful is user friendliness and not depending on excessive additional data collection by the attending clinicians. In this context, it was a bit disappointing that the parsimonious machine-learning algorithm with only the ten most important variables was slightly worse at predicting the secondary outcome than the full logistic regression model. A reason for this could be that when including a length of stay >4 days but without described medical complications, the combination of all variables provides information not available by merely including the ten most important ones. This highlights the need for as much detailed, and preferably non-binary, data as possible to fulfill the true potential of machine-learning algorithms.

Our study has some limitations. First, one of the strengths of machine learning compared to logistic regression is the analysis of multilevel continuous data, whereas we included only a limited number of, often binary, preoperative variables. This could have limited the full realization of our machine-learning model. As previously mentioned, we excluded intraoperative information, including type of anesthesia, surgical approach etc. all of which may influence postoperative outcomes. The observational design of this study means that we cannot exclude unmeasured confounding or confounding by indication. Also, despite that the DNDRP has a near complete registration of dispensed medicine in Denmark, some types or drugs, especially benzodiazepines, are exempt from general reimbursement and thus not sufficiently captured.²¹ Furthermore, it is doubtful whether the patients used all types of drugs at the time of surgery (e.g. heparin which is rarely for long-term use). Finally, classification of a complication being "medical" depended on review of the discharge records which can also introduce bias. However, we believe our approach to be superior to depending only on diagnostic codes which often are inaccurate⁴¹ and provide limited details on whether the complication may be attributed to a medical or surgical adverse event. The strengths of our study include the use of national registries with high degree of completion (>99% of all somatic admissions in case of the DNDRP),⁴² prospective recording of comorbidity, extensive information on prescription patterns 6 months prior to surgery and similar established enhanced recovery protocols in all departments.

In summary, our results suggest that machine-learning-algorithms likely provide clinically relevant improved predictions for defining patients in high-risk of medical complications after fast-track THA and TKA compared to a logistic regression model. Future studies could benefit from using such algorithms to find a manageable population of patients who benefit from intensified perioperative care.

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

For peer review only

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

Competing interests: Prof. Kehlet is a board member of “Rapid Recovery”, by Zimmer Biomet. Mr. Heltberg is sponsored by a grant from the Lundbeck Foundation, independently of the present study. Dr. Petersen is an advisory member of Sanofi outside of the present study. The remaining authors declare no conflicts of interest.

Funding: The study received funding from the Lundbeck Foundation, Copenhagen, Denmark grant number R25-A2702.

Author Contributions: CJ and HK contributed to the original idea of the study. CJ, PP and HK contributed to data collection and review of medical records. CM, MH, MJ, AL and TP contributed to the statistical methods, designed the prediction models and conducted the statistical analysis. CM, CJ, HK and TP wrote the original draft. All authors contributed to revision of the initial draft and agreed on the final version of the manuscript. The members of the Centre for Fast-track Hip and Knee Replacement Database collaborative group all contributed by implementing the fast-track protocol at their respective departments and reviewing the final manuscript.

Collaborators:

Frank Madsen M.D. Consultant, Department of Orthopedics, Aarhus University Hospital, Aarhus, Denmark
Torben B. Hansen M.D., Ph.D., Prof. Department of Orthopedics, Holstebro Hospital, Holstebro, Denmark
Kirill Gromov, M.D., Ph.D., Ass.Prof. Department of Orthopedics Hvidovre Hospital, Hvidovre Denmark
Thomas Jakobsen, M.D., Ph.D., DM.Sci., Ass. Prof. Department of Orthopedics, Aalborg University Hospital, Farsø, Denmark
Claus Varnum, M.D., Ph.D., Ass. Prof. Department of Orthopedic Surgery, Lillebaelt Hospital - Vejle, University Hospital of Southern Denmark, Denmark
Soren Overgaard, M.D., DM.Sci., Prof, Department of Orthopedics, Bispebjerg Hospital, Copenhagen, Denmark
Mikkel Rathsach, M.D., Ph.D., Ass. Prof. Department of Orthopedics, Gentofte Hospital, Gentofte, Denmark
Lars Hansen, M.D., Consultant, Department of Orthopedics, Sydvestjysk Hospital, Grindsted, Denmark

Data sharing: The original dataset is not publicly available due to Danish data-protection law but can be acquired from the corresponding author by request. All statistical code can be freely accessed from <https://zenodo.org/record/7330268>

Ethics and Permissions

No Ethics Committee approval was necessary as the National Danish Ethics committee exempt non-interventional observational studies. Permission to review and store information from medical records without informed consent was acquired from Center for Regional Development (R-20073405) and the Danish Data Protection Agency (RH-2007-30-0623).

References

1. Petersen PB, Kehlet H, Jorgensen CC, et al. Improvement in fast-track hip and knee arthroplasty: a prospective multicentre study of 36,935 procedures from 2010 to 2017. *Scientific reports* 2020;**10**(1):21233.
2. Khan SK, Malviya A, Muller SD, et al. Reduced short-term complications and mortality following Enhanced Recovery primary hip and knee arthroplasty: results from 6,000 consecutive procedures. *Acta Orthop* 2014;**85**(1):26-31.
3. Partridge T, Jameson S, Baker P, et al. Ten-Year Trends in Medical Complications Following 540,623 Primary Total Hip Replacements from a National Database. *The Journal of bone and joint surgery American volume* 2018;**100**(5):360-67.
4. Jorgensen CC, Gromov K, Petersen PB, et al. Influence of day of surgery and prediction of LOS > 2 days after fast-track hip and knee replacement. *Acta orthopaedica* 2021;**92**(2):170-75.
5. 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**(1):e009813.
6. Johns WL, Layon D, Golladay GJ, et al. Preoperative Risk Factor Screening Protocols in Total Joint Arthroplasty: A Systematic Review. *J Arthroplasty* 2020;**35**(11):3353-63.
7. 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**(6):1466-73 e1.
8. Shah A, Memon M, Kay J, et al. Preoperative Patient Factors Affecting Length of Stay following Total Knee Arthroplasty: A Systematic Review and Meta-Analysis. *J Arthroplasty* 2019;**34**(9):2124-65 e1.
9. Li Q, Zhong H, Girardi FP, et al. Machine Learning Approaches to Define Candidates for Ambulatory Single Level Laminectomy Surgery. *Global spine journal* 2021:2192568220979835.
10. Chiew CJ, Liu N, Wong TH, et al. Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission. *Annals of surgery* 2020;**272**(6):1133-39.
11. Li H, Jiao J, Zhang S, et al. 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. *The journal of knee surgery* 2022;**35**(1):7-14.
12. Lu Y, Khazi ZM, Agarwalla A, et al. Development of a Machine Learning Algorithm to Predict Nonroutine Discharge Following Unicompartamental Knee Arthroplasty. *J Arthroplasty* 2021;**36**(5):1568-76.
13. Shah AA, Devana SK, Lee C, et al. Development of a Novel, Potentially Universal Machine Learning Algorithm for Prediction of Complications After Total Hip Arthroplasty. *J Arthroplasty* 2021;**36**(5):1655-62 e1.
14. Sniderman J, Stark RB, Schwartz CE, et al. Patient Factors That Matter in Predicting Hip Arthroplasty Outcomes: A Machine-Learning Approach. *J Arthroplasty* 2021;**36**(6):2024-32.
15. Kugelman DN, Teo G, Huang S, et al. A Novel Machine Learning Predictive Tool Assessing Outpatient or Inpatient Designation for Medicare Patients Undergoing Total Hip Arthroplasty. *Arthroplasty today* 2021;**8**:194-99.
16. Mohammadi R, Jain S, Namin AT, et al. Predicting Unplanned Readmissions Following a Hip or Knee Arthroplasty: Retrospective Observational Study. *JMIR medical informatics* 2020;**8**(11):e19761.

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

17. Ramkumar PN, Navarro SM, Haeberle HS, et al. Development and Validation of a Machine Learning Algorithm After Primary Total Hip Arthroplasty: Applications to Length of Stay and Payment Models. *J Arthroplasty* 2019;**34**(4):632-37.
18. Ramkumar PN, Karnuta JM, Navarro SM, et al. Preoperative Prediction of Value Metrics and a Patient-Specific Payment Model for Primary Total Hip Arthroplasty: Development and Validation of a Deep Learning Model. *J Arthroplasty* 2019;**34**(10):2228-34 e1.
19. Haeberle HS, Helm JM, Navarro SM, et al. Artificial Intelligence and Machine Learning in Lower Extremity Arthroplasty: A Review. *J Arthroplasty* 2019;**34**(10):2201-03.
20. Johannesdottir KB, Kehlet H, Petersen PB, et al. Machine learning classifiers do not improve prediction of hospitalization > 2 days after fast-track hip and knee arthroplasty compared with a classical statistical risk model. *Acta orthopaedica* 2022;**93**:117-23.
21. Johannesdottir SA, Horvath-Puho E, Ehrenstein V, et al. Existing data sources for clinical epidemiology: The Danish National Database of Reimbursed Prescriptions. *ClinEpidemiol* 2012;**4**:303-13.
22. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine* 2015;**162**(1):W1-73.
23. Olczak J, Pavlopoulos J, Prijs J, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. *Acta orthopaedica* 2021;**92**(5):513-25.
24. Tashman L. Out-of-Sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting* 2000;**16**(4):437-50.
25. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *Bmj* 2020;**368**:m441.
26. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T. LightGBM: a highly efficient gradient boosting decision tree. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc, 2017:3149-57.
27. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. Secondary Optuna: A Next-generation Hyperparameter Optimization Framework 2019. <http://arxiv.org/abs/1907.10902>.
28. Bergstra J, Bardenet R, Bengio Y, et al. Algorithms for Hyper-Parameter Optimization. 2011; 24. <https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf>.
29. Platt J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. CiteSeer, 2000.
30. Lundberg SM, Erion G, Chen H, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature machine intelligence* 2020;**2**(1):56-67.
31. Lundberg SMLS. A Unified Approach to Interpreting Model Predictions. In: Guyon I, ed. Adv Neural Inf Process Syst [Internet]: Curran Associates, Inc., 2017.
32. McIsaac DI, Wong CA, Bryson GL, et al. Association of Polypharmacy with Survival, Complications, and Healthcare Resource Use after Elective Noncardiac Surgery: A Population-based Cohort Study. *Anesthesiology* 2018;**128**(6):1140-50.
33. Chicco D. Ten quick tips for machine learning in computational biology. *BioData mining* 2017;**10**(1):35 (2017).
34. Chicco D, Totsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining* 2021;**14**(1):13 (2021).

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

35. Totsch N, Hoffmann D. Classifier uncertainty: evidence, potential impact, and probabilistic treatment. *PeerJ Computer science* 2021;**7**:e398.
36. Lopez CD, Gazgalis A, Boddapati V, et al. Artificial Learning and Machine Learning Decision Guidance Applications in Total Hip and Knee Arthroplasty: A Systematic Review. *Arthroplasty today* 2021;**11**:103-12.
37. Han C, Liu J, Wu Y, et al. To Predict the Length of Hospital Stay After Total Knee Arthroplasty in an Orthopedic Center in China: The Use of Machine Learning Algorithms. *Frontiers in surgery* 2021;**8**:606038.
38. Ramkumar PN, Karnuta JM, Navarro SM, et al. Deep Learning Preoperatively Predicts Value Metrics for Primary Total Knee Arthroplasty: Development and Validation of an Artificial Neural Network Model. *J Arthroplasty* 2019;**34**(10):2220-27 e1.
39. Wei C, Quan T, Wang KY, et al. Artificial neural network prediction of same-day discharge following primary total knee arthroplasty based on preoperative and intraoperative variables. *Bone Joint J* 2021;**103-B**(8):1358-66.
40. Griffiths R, Beech F, Brown A, et al. Peri-operative care of the elderly. *Anaesthesia* 2014;**69** Suppl 1:81-98.
41. Bedard NA, Pugely AJ, McHugh MA, et al. Big Data and Total Hip Arthroplasty: How Do Large Databases Compare? *J Arthroplasty* 2018;**33**(1):41-45.e3.
42. Schmidt M, Schmidt SA, Sandegaard JL, et al. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clinical epidemiology* 2015;**7**:449-90.

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

FIGURE LEGENDS

Figure 1a-b

1a) Distribution of full machine learning model risk scores for patients +/- the primary outcome. The dashed line marks the classification threshold of 20% positive prediction fraction.

1b) 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) and the age-only model (AM).

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 the primary outcome (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 the primary outcome, where positive increase and negative values decrease the risk score. Each dot represents a patient and the color is related to the value of the variable with blue being lowest and red highest..

Figure 3a-d

SHAP scatter-plot on the contributions to the full machine-learning model on the primary outcome (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

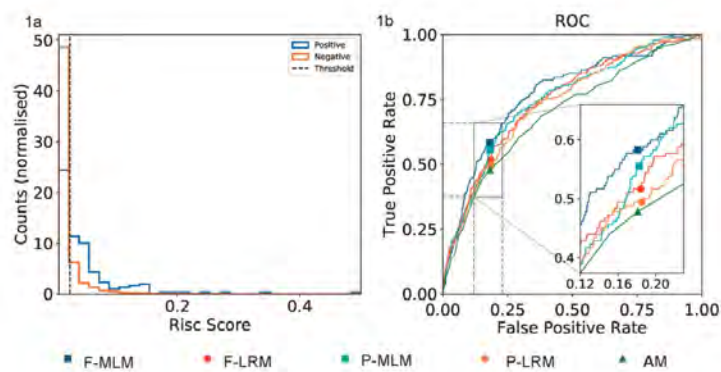
ACE: angiotensin converting enzyme AHT: antihypertensive. Other AHT were defined as AHT different from diuretics ANG-II/ACE inhibitors or Ca²⁺antagonists. IHD: Ischemic heart disease

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.

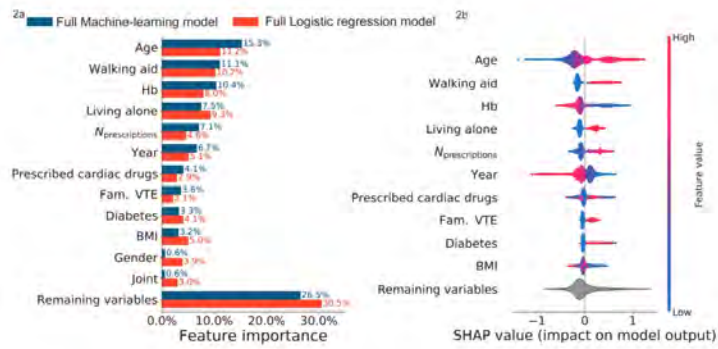


1a) Distribution of full machine learning model risk scores for patients +/- the primary outcome. The dashed line marks the classification threshold of 20% positive prediction fraction.

1b) 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) and the age-only model (AM).

297x209mm (300 x 300 DPI)

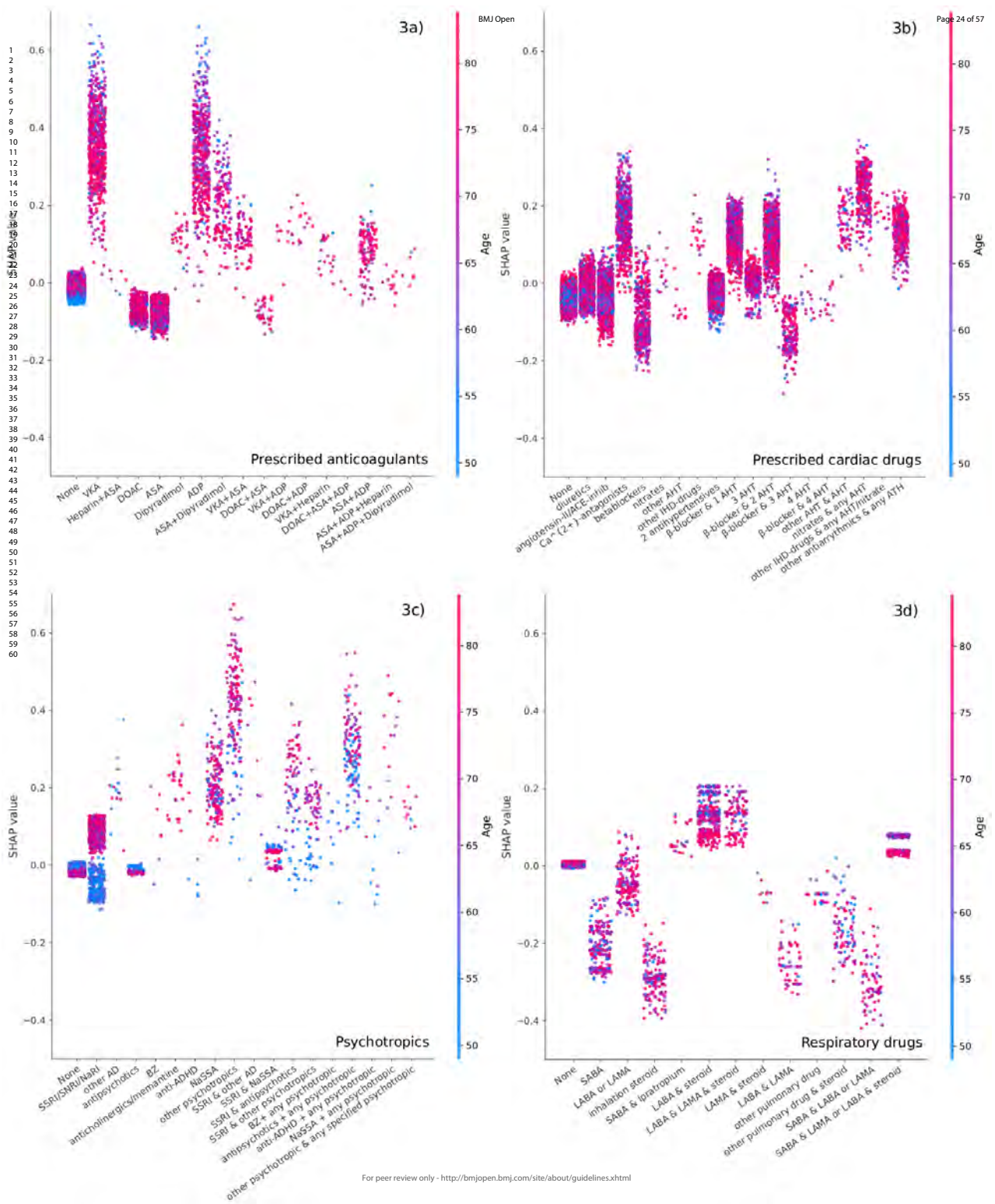
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



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 the primary outcome (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.

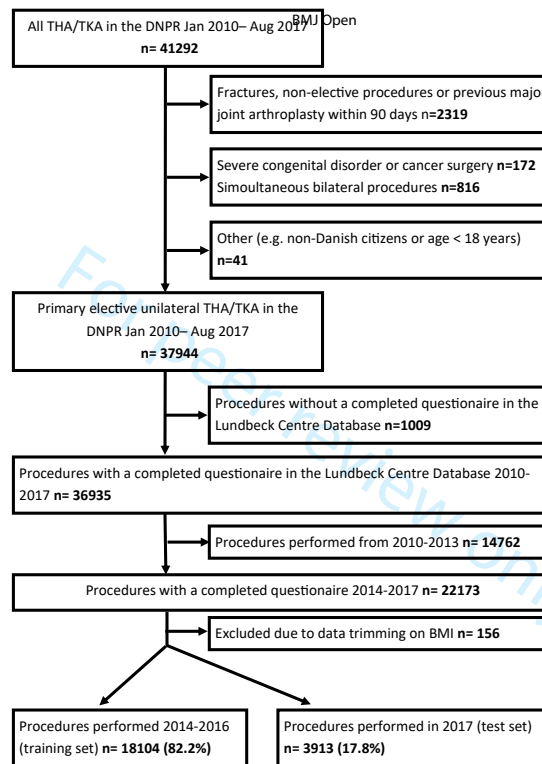
2b) The SHAP-values for the full machine-learning model on the primary outcome, where positive increase and negative values decrease the risk score. Each dot represents a patient and the color is related to the value of the variable with blue being lowest and red highest.

297x209mm (300 x 300 DPI)



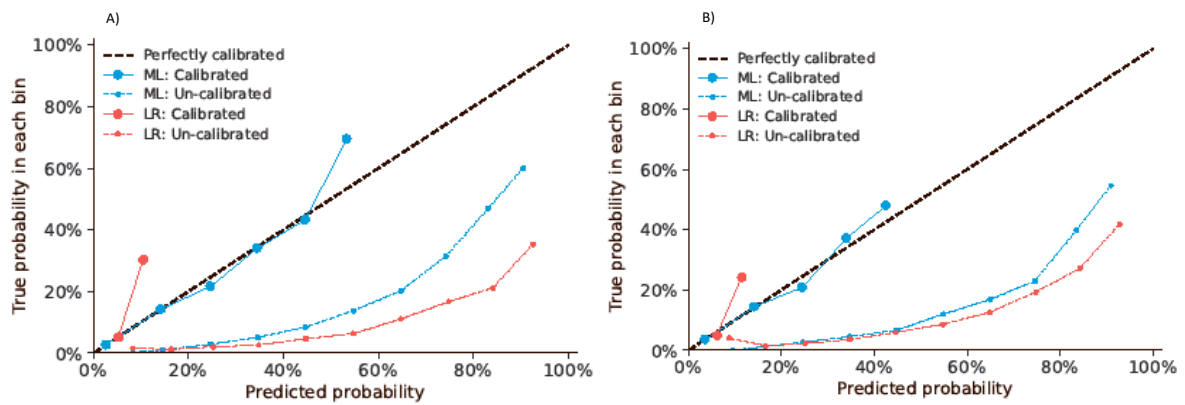
Page 25 of 57
Supplemental Material 1

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



Flowchart of the study population and final sample size. THA: total hip arthroplasty TKA: total knee arthroplasty DNPR: the Danish National Patient Register <http://www.dnpr.dk/site/about/guidelines.xhtml>

Supplemental Material 2



Calibration plot of the machine learning model (ML) and the logistic regression (LR) for both the calibrated and un-calibrated models for the primary (A) and secondary (B) outcome.

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

Supplemental material table 3

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

Reimbursed prescriptions within 3 months preoperatively	training set (n:18104)	test set (n:3913)
<u>Anticoagulants</u>		
none	13570 (75.0)	2953 (75.5)
VKA	729 (4.0)	127 (3.2)
Heparin & Acetylsalicylic acid	6 (0.0)	1 (0.0)
DOAC	526 (2.9)	181 (4.6)
Acetylsalicylic acid	2235 (12.3)	462 (11.8)
Dipyradimol	31 (0.2)	3 (0.1)
ADP-antagonist	522 (2.9)	122 (3.1)
Acetylsalicylic acid & Dipyradimol	169 (0.9)	16 (0.4)
VKA & Acetylsalicylic acid	81 (0.4)	7 (0.2)
DOAC & Acetylsalicylic acid	42 (0.2)	5 (0.1)
VKA & ADP-antagonist	13 (0.1)	2 (0.1)
DOAC & ADP-antagonist	12 (0.1)	5 (0.1)
VKA & Heparin	22 (0.1)	0 (0.0)
DOAC & Acetylsalicylic acid & ADP-antagonist	3 (0.0)	1 (0.0)
Acetylsalicylic acid & ADP-antagonist	124 (0.7)	26 (0.7)
Acetylsalicylic acid & ADP-antagonist & Heparin	12 (0.1)	1 (0.0)
Acetylsalicylic acid & ADP-antagonist & Dipyradimol	7 (0.0)	1 (0.0)
<u>Cardiac prescriptions</u>		
none	7741 (42.8)	1780 (45.5)
diuretics	1070 (5.9)	191 (4.9)
angiotensin-II/ACE-inhibitors	2287 (12.6)	528 (13.5)
Ca ²⁺ antagonists	688 (3.8)	140 (3.6)
β-blocker	492 (2.7)	96 (2.5)
nitrates	14 (0.1)	5 (0.1)
other antihypertensives	12 (0.1)	0 (0.0)
other types of medication for IHD	22 (0.1)	1 (0.0)
2 antihypertensives	2360 (13.0)	513 (13.1)
β-blocker & 1 antihypertensive ¹	966 (5.3)	195 (5.0)
3 antihypertensives	515 (2.8)	83 (2.1)
β-blocker & 2 antihypertensives ¹	902 (5.0)	168 (4.3)
β-blocker & 3 antihypertensives ¹	235 (1.3)	55 (1.4)
4 antihypertensives	16 (0.1)	4 (0.1)
β-blocker & 4 antihypertensives	15 (0.1)	6 (0.2)
other antihypertensive & antihypertensives ¹	78 (0.4)	18 (0.5)
nitrates & any hypertensive	323 (1.8)	57 (1.5)
other drugs for IHD & any antihypertensive and/or nitrate	16 (0.1)	4 (0.1)
other antiarrhythmics & any antihypertensives	352 (1.9)	69 (1.8)
<u>Anticholesterols</u>		
none	12665 (70.0)	2762 (70.6)
statins	5218 (28.8)	1105 (28.2)
other anti-lipids	118 (0.7)	24 (0.6)
statins +other anti-lipids	103 (0.6)	22 (0.6)
<u>Systemic steroids</u>		
	1038 (5.7)	234 (6.0)

1			
2			
3			
4	<u>Antirheumatics</u>		
5	none	17709 (97.8)	3822 (97.7)
6	disease-modifying antirheumatic drugs	392 (2.2)	91 (2.3)
7	other antirheumatics	3 (0.0)	0 (0.0)
8			
9	<u>Respiratory prescriptions</u>		
10	none	16256 (89.8)	3498 (89.4)
11	SABA	235 (1.3)	54 (1.4)
12	LABA or LAMA	194 (1.1)	42 (1.1)
13	inhalation steroid only	176 (1.0)	43 (1.1)
14	SABA & Ipratropium (+/- others)	24 (0.1)	0 (0.0)
15	LABA & steroid	432 (2.4)	87 (2.2)
16	LABA & LAMA & steroid	115 (0.6)	26 (0.7)
17	LAMA & steroid	10 (0.1)	1 (0.0)
18	LABA & LAMA	56 (0.3)	31 (0.8)
19	other pulmonary drugs	26 (0.1)	9 (0.2)
20	other pulmonary drugs & steroid	95 (0.5)	12 (0.3)
21	SABA & LABA or LAMA	76 (0.4)	26 (0.7)
22	SABA & LABA or LAMA & steroid	409 (2.3)	84 (2.1)
23			
24	<u>Psychotropic prescriptions</u>		
25	none	16113 (89.0)	3496 (89.3)
26	SSRI/SNRI/NaRI	1055 (5.8)	209 (5.3)
27	other antidepressants	16 (0.1)	2 (0.1)
28	antipsychotics	104 (0.6)	20 (0.5)
29	benzodiazepines ²	7 (0.0)	0 (0.0)
30	anti-cholinergics or memantine	27 (0.1)	6 (0.2)
31	anti-ADHD drugs	7 (0.0)	4 (0.1)
32	NaSSA	177 (1.0)	32 (0.8)
33	other psychotropics	166 (0.9)	44 (1.1)
34	SSRI + other antidepressants	9 (0.0)	1 (0.0)
35	SSRI + NaSSA	86 (0.5)	16 (0.4)
36	SRRRI + antipsychotics	80 (0.4)	18 (0.5)
37	SRRRI + other psychotropics	72 (0.4)	19 (0.5)
38	benzodiazepines + any psychotropic	11 (0.1)	4 (0.1)
39	antipsychotics + any psychotropic	137 (0.8)	32 (0.8)
40	anti-ADHD + any psychotropic	11 (0.1)	3 (0.1)
41	NaSSA + any psychotropic	16 (0.1)	6 (0.2)
42	other psychotropics + any specified psychotropic	10 (0.1)	1 (0.0)
43			
44			
45			

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

¹either diuretics, ACE/ANG-II inhibitors or Ca²⁺antagonists ²likely underreported due to limited general reimbursement for benzodiazepines in Denmark

53
54
55
56
57
58
59
60

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

Supplemental material 4

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

Positive prediction fraction 25%	TP	FP	FN	TN	Sensitivity %	Precision %	MCC %	AUROC %	AUPRC %	Brier %	P(sensitivity) %
Full machine-learning model	120	858	62	2873	65.9	12.3	20.9	77.0	15.3	4.32	-
Full logistic regression model	108	870	74	2861	59.3	11.0	17.5	74.6	15.6	4.32	10.4
Parsimonious machine-learning model	114	864	68	2867	62.6	11.7	19.2	74.9	14.1	4.35	26.2
Parsimonious logistic regression model	103	875	79	2856	56.6	10.5	16.1	73.6	15.2	4.33	3.9
Age-model	94	824	88	2907	51.6	10.2	14.7	69.7	12.2	38.8	1.2
Positive prediction fraction 30%	TP	FP	FN	TN	Sensitivity %	Precision %	MCC %	AUROC %	AUPRC %	Brier %	P(sensitivity) %
Full machine-learning model	130	1043	52	2688	71.4	11.1	20.0	77.0	15.3	4.32	-
Full logistic regression model	117	1056	65	2675	64.2	10.0	16.5	74.6	15.6	4.32	8.4
Parsimonious machine-learning model	124	1049	58	2682	68.1	10.6	18.4	74.9	14.1	4.35	30.0
Parsimonious logistic regression model	118	1055	64	2676	64.8	10.1	16.8	73.6	15.2	4.33	14.0
Age-model	100	955	82	2776	54.9	9.5	13.9	69.7	12.2	38.8	0.9

TP: true positives FP: false positives FN: false negatives TN: true negatives MCC: Matthews correlation coefficient AUROC: area under the receiver operating 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 material 5

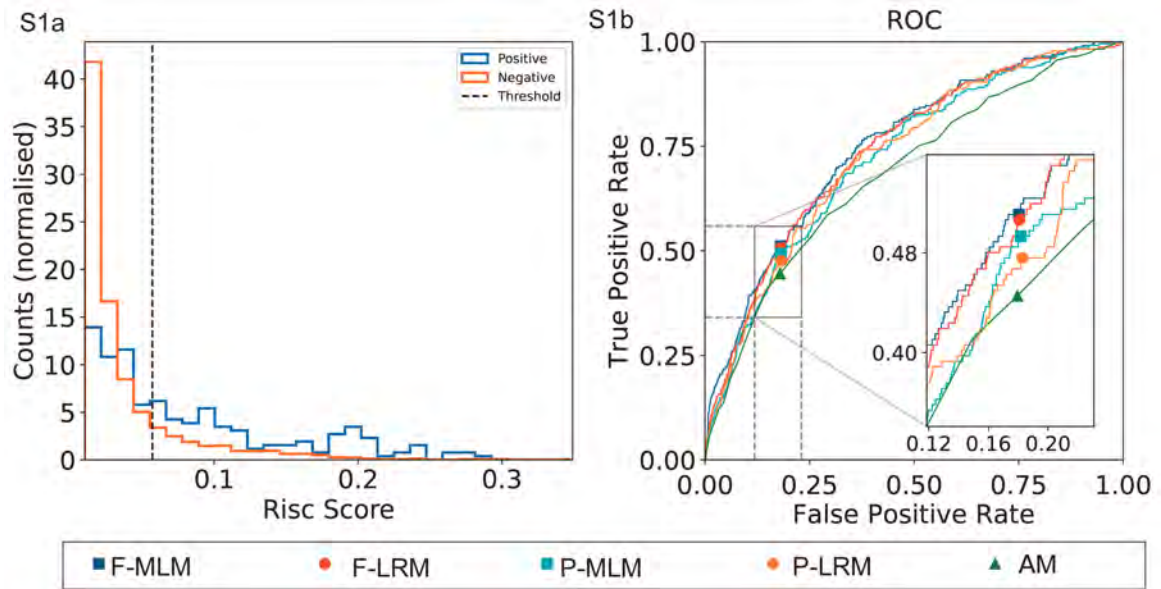
Table S1: Performance of different models for the secondary outcome (LOS >4 days or readmissions due to "medical" morbidity or LOS >4 days but without recorded morbidity)

Positive prediction fraction 20%	TP	FP	FN	TN	Sensitivity %	Precision %	MCC %	AUROC %	AUPRC %	Brier %	P(sensitivity) %
Full machine-learning model	117	665	112	3019	51.1	15.0	19.4	75.0	18.1	5.23	-
Full logistic regression model	115	667	114	3017	50.2	14.7	18.9	74.1	16.7	5.35	46.4
Parsimonious machine-learning model	109	673	120	3011	47.6	13.9	17.2	72.1	15.8	5.33	35.2
Parsimonious logistic regression model	109	673	120	3011	47.6	13.9	17.2	72.9	16.7	5.37	22.6
Age-model	102	661	127	3023	44.5	13.4	15.8	68.7	13.4	38.3	10.3
Positive prediction fraction 25%	TP	FP	FN	TN	Sensitivity %	Precision %	MCC %	AUROC %	AUPRC %	Brier %	P(sensitivity) %
Full machine-learning model	128	850	101	2834	55.9	13.1	17.8	75.0	18.1	5.23	-
Full logistic regression model	133	845	96	2839	58.1	13.6	19.1	74.1	16.7	5.35	68.0
Parsimonious machine-learning model	121	857	108	2827	52.8	12.3	16.3	72.1	15.8	5.33	25.5
Parsimonious logistic regression model	127	851	102	2833	55.5	13.0	17.5	72.9	16.7	5.37	46.6
Age-model	113	805	116	2879	49.3	12.3	15.2	68.7	13.4	38.3	17.2
Positive prediction fraction 30%	TP	FP	FN	TN	Sensitivity %	Precision %	MCC %	AUROC %	AUPRC %	Brier %	P(sensitivity) %
Full machine-learning model	146	1027	83	2657	63.4	12.4	18.4	75.0	18.1	5.23	-
Full logistic regression model	144	1029	85	2655	62.9	12.3	17.9	74.1	16.7	5.35	42.4
Parsimonious machine-learning model	135	1038	94	2651	59.0	11.5	15.8	72.1	15.8	5.33	14.9
Parsimonious logistic regression model	140	1033	89	2651	61.1	11.9	17.0	72.9	16.7	5.37	28.3
Age-model	122	933	107	2751	53.3	11.6	14.8	68.7	13.4	38.3	7.9

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

Supplemental material 6 Figure S1a-b

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

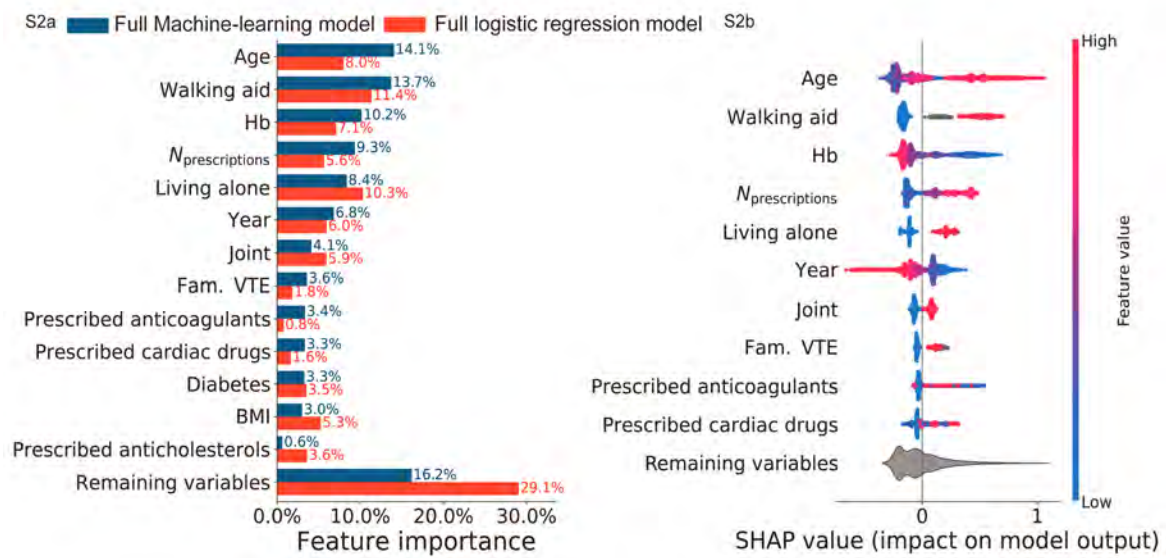


S1a) Distribution of full machine-learning model risk-scores for patients +/- the secondary outcome (LOS >4 days or readmissions due to "medical" morbidity or LOS >4 days with no recorded morbidity). The dashed line marks the classification threshold of a 20% positive prediction fraction.

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) and the age-only model (AM).

Supplemental Material 7

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 the secondary outcome (LOS >4 days or readmissions due to "medical" morbidity or LOS >4 days with no recorded morbidity).

Only the importance of prescribed anti-cholesterols 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. Positive SHAP-values increase the risk score while negative values decrease the risk score. Each dot repre-

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

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

Supplemental material 8

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 Ca²⁺-antagonists. IHD: Ischemic heart disease

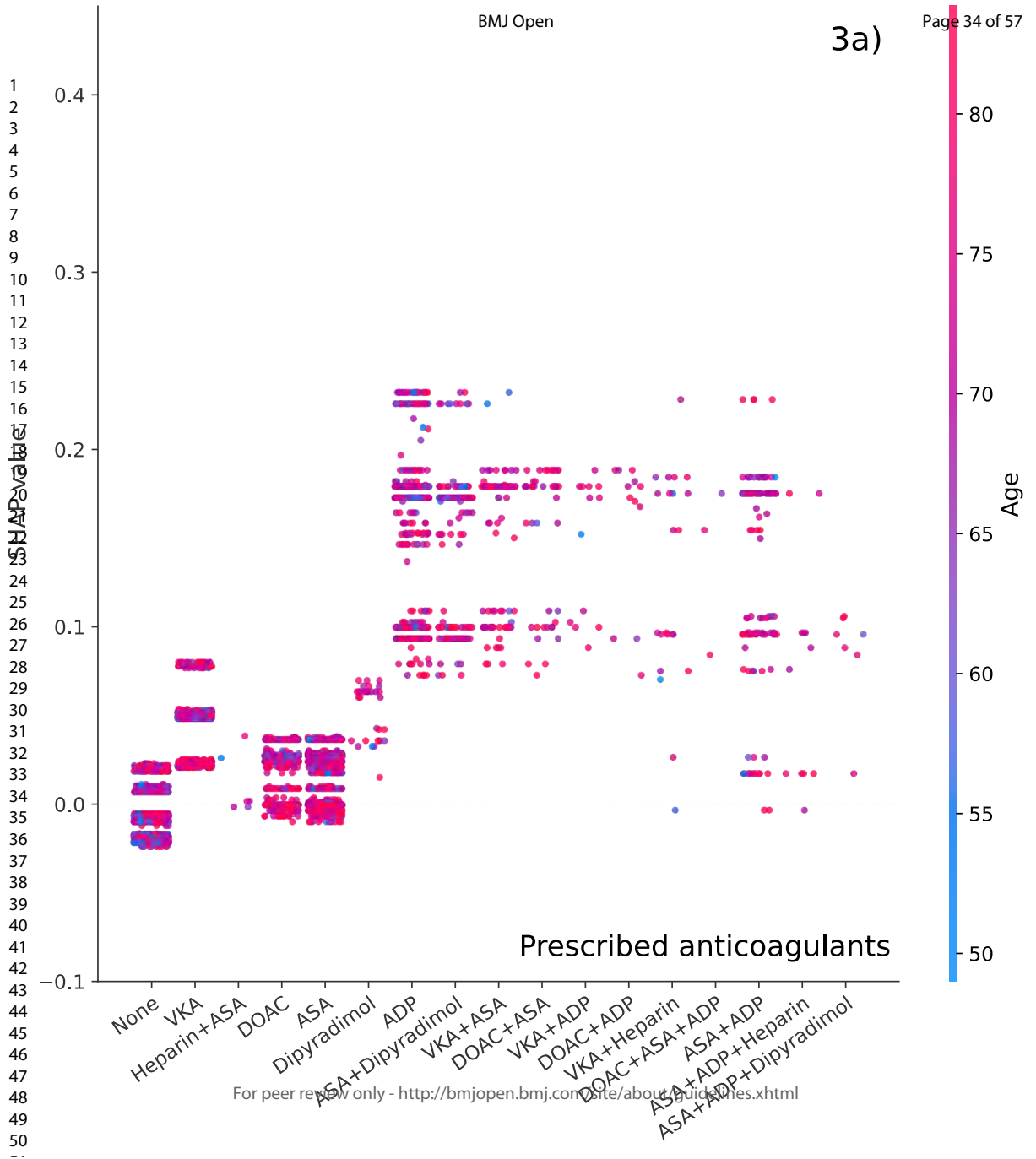
3c) Prescribed psychotropics

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.

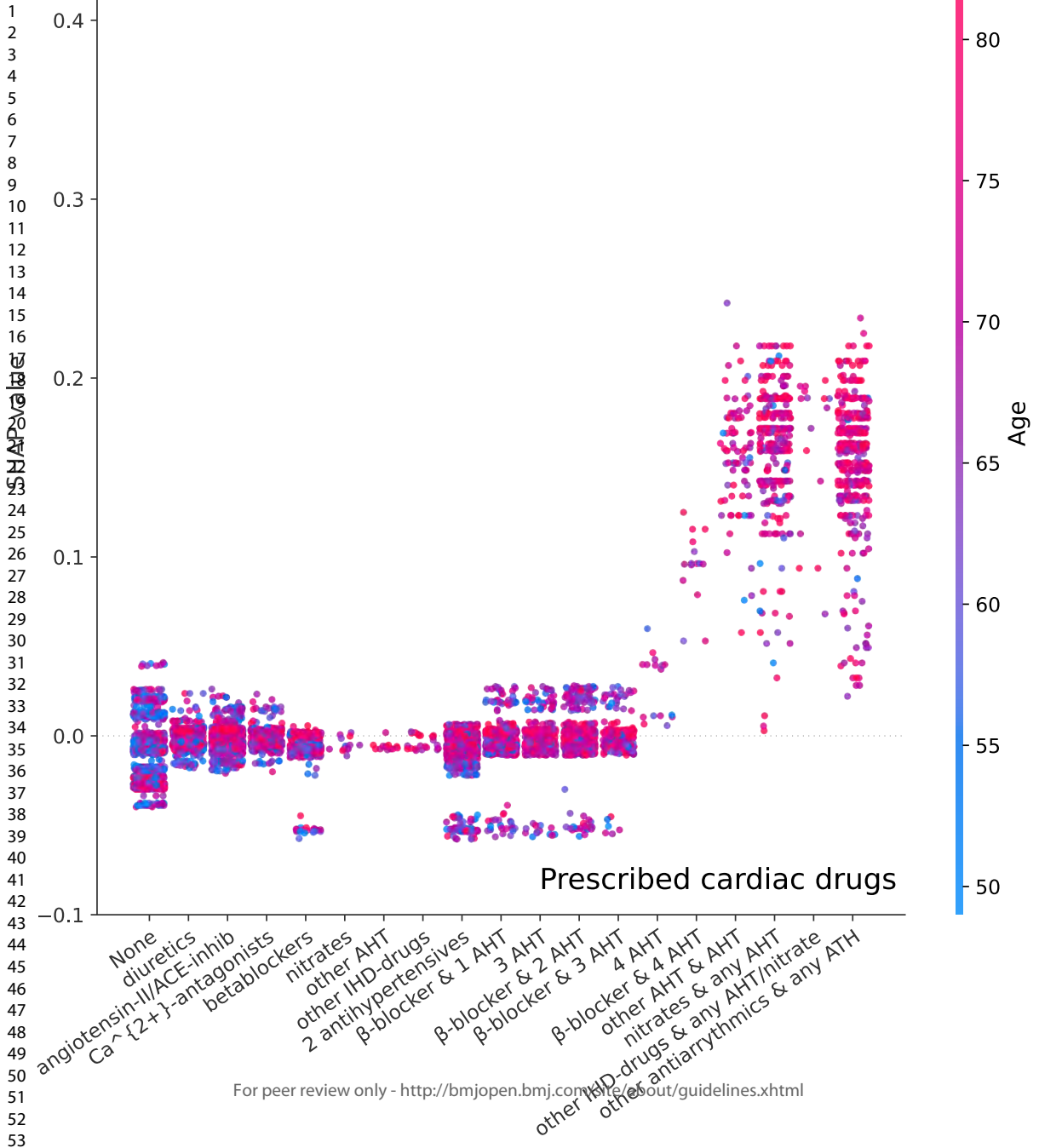
SABA: Short-acting beta agonist LABA: long-acting beta agonist LAMA: Long-acting muscarinic antagonist.

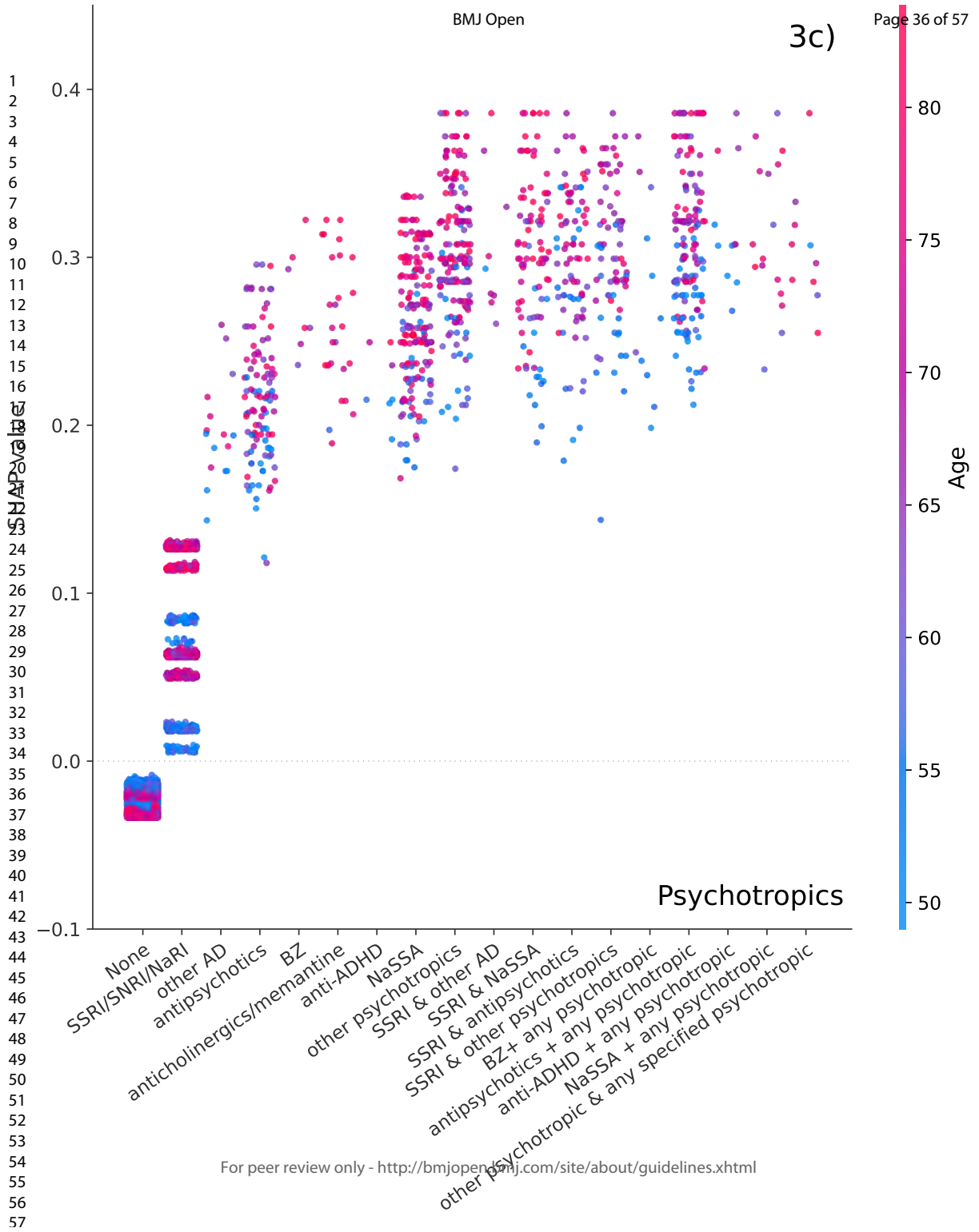


Page 35 of 57

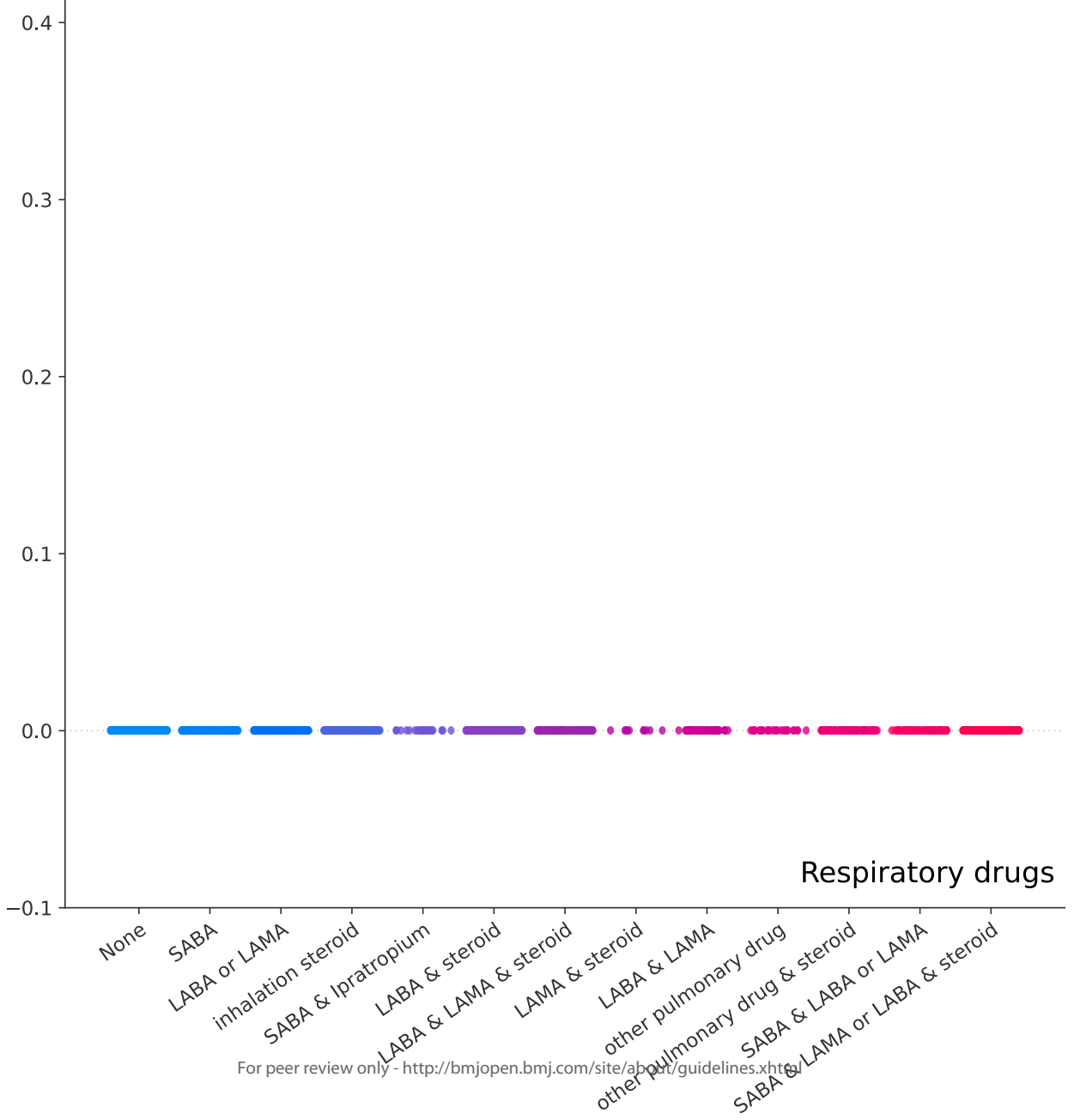
BMJ Open

3b)





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



TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	4-5
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	4
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	5
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	5
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5
	5b	D;V	Describe eligibility criteria for participants.	5
	5c	D;V	Give details of treatments received, if relevant.	5
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	6
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	N/A
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	7
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	N/A
Sample size	8	D;V	Explain how the study size was arrived at.	N/A
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	7
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	7
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	7
	10c	V	For validation, describe how the predictions were calculated.	7
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	8
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	7 and Sup mat 2
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	Tbl1
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	Sup mat 1
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	Tbl 1
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	Tbl 1
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	7-9
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	N/A
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	7
	15b	D	Explain how to use the prediction model.	Fig 2a-b
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	Tbl2
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	N/A
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	14
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	10-14
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	14
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	16
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	16

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

For peer review only - <http://bmjopen.bmj.com/site/about/guidelines.xhtml>

4 *Paper III*

The following 9 pages contain the paper:

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

ROYAL SOCIETY
OPEN SCIENCE

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

[†]These authors contributed equally.

THE ROYAL SOCIETY
PUBLISHING

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

Mathias L. Heltberg^{1,2,3,†}, Christian Michelsen^{1,†},
Emil S. Martiny^{1,†}, Lasse Engbo Christensen⁴, Mogens
H. Jensen¹, Tariq Halasa⁵ and Troels C. Petersen¹

¹Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, Copenhagen E 2100, Denmark

²Laboratoire de Physique, Ecole Normale Supérieure, Rue Lhomond 15, Paris 07505, France

³Infektionsberedskab, Statens Serum Institute, Artillerivej, Copenhagen S 2300, Denmark

⁴DTU Compute, Section for Dynamical Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Anker Engélunds Vej 101A, Kongens Lyngby 2800, Denmark

⁵Animal Welfare and Disease Control, University of Copenhagen, Grønnegårdsvej 8, Frederiksberg 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 is 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.

© 2022 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License <http://creativecommons.org/licenses/by/4.0/>, which permits unrestricted use, provided the original author and source are credited.

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 1*a*. 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

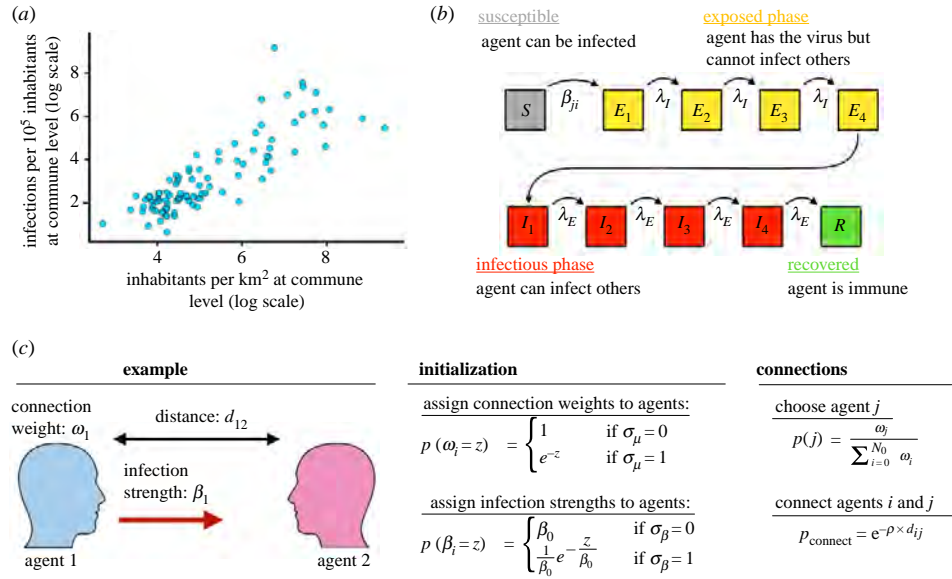


Figure 1. (a) Population density (x -axis) and the number of infections per 10^5 inhabitants (y -axis) for each commune in Denmark. (b) Illustration of the modified susceptible-exposed-infected-removed (SEIR) model used. It consists of 10 consecutive states (S, E_{1-4}, I_{1-4} and R), with transition rates governed by β, λ_E and λ_r , 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 σ_β , 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 $\tilde{\beta} = \beta(\mu/2N_0)$. 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 σ_μ 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 ρ , 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 $\varepsilon_\rho = 4\%$ (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	5.8×10^6	10^5 – 10^7	—
N_{init}	number of individuals initially infected	100	1 – 10^4	—
μ	average number of network contacts	40	10–100	—
β	typical infection strength	0.01	0.001–0.1	d^{-1}
$\lambda_{\mathcal{E}}$	rate to move through $\frac{1}{4}$ of latency period	1	0.5–4	d^{-1}
$\lambda_{\mathcal{I}}$	rate to move through $\frac{1}{4}$ of infectious period	1	0.5–4	d^{-1}
σ_{μ}	population clustering spread	0	0–1	—
σ_{β}	interaction strength spread	0	0–1	—
ρ	typical acceptance distance	0.1	0–0.5	km^{-1}
ϵ_{ρ}	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}_{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_{peak} , but also subsided quicker, leading to a lower number of infected once reaching endemic steady state, R_{∞} (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 ρ) 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 β (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

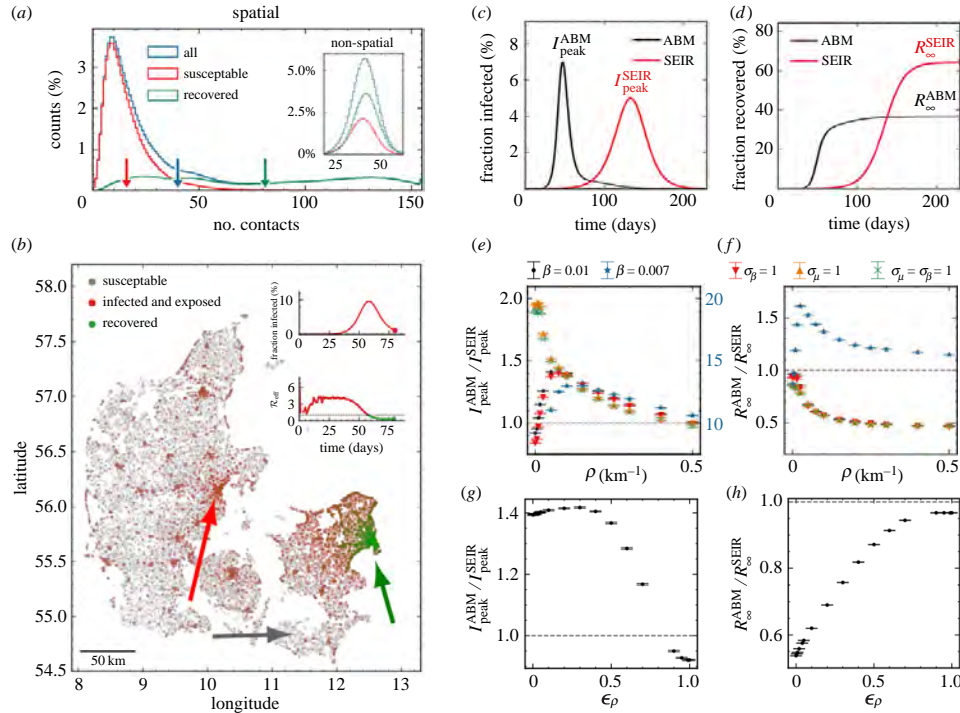


Figure 2. (a) Histograms showing the number of susceptible (red) and recovered (green) individuals at the end of an epidemic with $\rho = 0.1 \text{ km}^{-1}$. The distribution before the epidemic is shown in blue. The arrows show the mean of each distribution. The inset shows the same for $\rho = 0 \text{ km}^{-1}$. (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_{peak} , between deterministic (SEIR) and ABM as a function of ρ , 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_{∞} , between deterministic (SEIR) and ABM as a function of ρ . Colours similar to (e). (g) Same as (e), but as a function of ϵ_{ρ} . (h) Same as (f), but as a function of ϵ_{ρ} .

(i.e. including *super connectors*). 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 connectors) and green (super connectors 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_{∞} , and this converged towards half of the SEIR model prediction as a function of ρ 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_{∞} 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, ϵ_{ρ} , we found that $I_{\text{peak}}^{\text{ABM}}$ is almost unaffected for $\epsilon_{\rho} < 0.5$ (figure 2g), while R_{∞}^{ABM} increases linearly towards the SEIR model R_{∞}^{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 ρ , we found that the SEIR model predictions overestimated the infection peak height I_{peak} and the total number of infected R_{∞} 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 ϵ_p are below 10%, the bias in the estimated infection peak height, I_{peak} , was constant within statistical uncertainty (figure 3e). For the total number of infected, R_{∞} , we observed an almost linear regression to the SEIR model as ϵ_p approaches one. However, even when $\epsilon_p = 0.25$, 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, asymptomatic 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

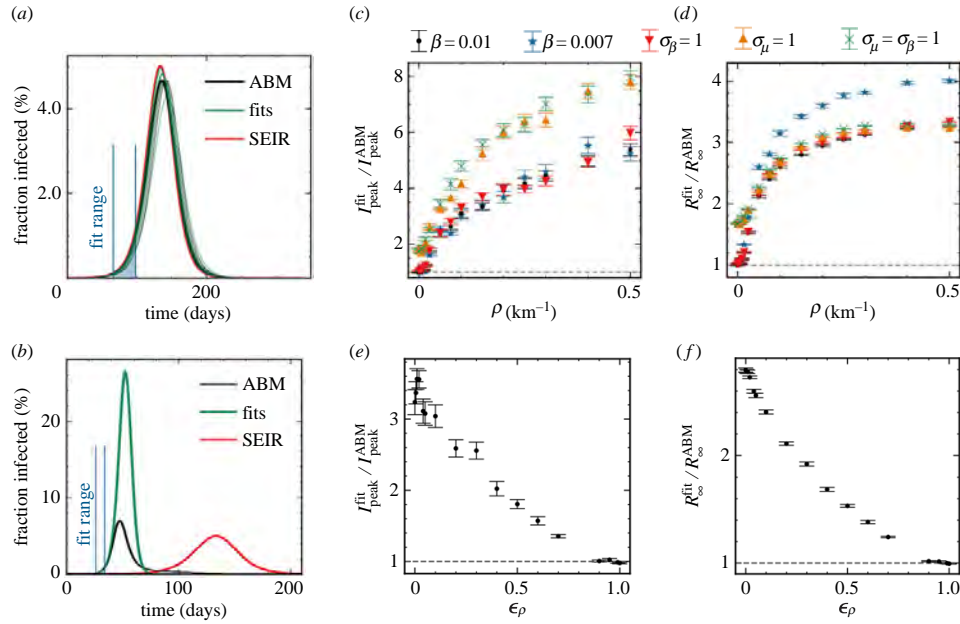


Figure 3. (a) Number of infected individuals for the ABM in black, the SEIR model in red and the SEIR fits to the ABM data in green. Blue lines show the interval where parameters are fitted (also shown below the curves). Here, $\rho = 0 \text{ km}^{-1}$. (b) Same as (a) but with population clustering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

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 β_{ij} for each link. The average contact number, μ , 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

ρ 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 we 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 β and a time delay, τ , to the SEIR model with a χ^2 -fit (assuming Poissonian statistics) and kept λ_E and λ_I fixed to the true numbers (used in the simulation). The initial number of infected, N_{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_{∞}^{fit} were extracted from the fitted model and compared to the $I_{\text{peak}}^{\text{ABM}}$ and R_{∞}^{ABM} from the ABM simulation.

Data accessibility. Data and relevant code for this research work are stored in GitHub: 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 Græsbo, 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.aba9757)
- WHO: see 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)
- 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)
- 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)
- Kuniya T. 2020 Prediction of the epidemic peak of coronavirus disease in Japan, 2020. *J. Clin. Med.* **9**, 789. (doi:10.3390/jcm9030789)
- Ghafari-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. *Heliyon* **7**, e08143. (doi:10.1016/j.heliyon.2021.e08143)
- Fokas AS, Dikaios N, Katsis 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)
- 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)
- 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)
- 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)
- van Bunnik BA, Morgan AL, Bessell P, Calder-Gerver 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).
- 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).
- 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)
- Sneppen K, Nielsen BF, Taylor RJ, Simonsen L. 2021 Overdispersion in COVID-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control. *Proc. Natl Acad. Sci. USA* **118**, e2016623118. (doi: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 (R0). *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. Ganasegeran K, Jamil MFA, Ch'ng 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/ijerph17155477)
 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, Hoogesteijn AL, Piccinini R, Maciorowski D, Kempaiah 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, Dai 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/NEJMp2016822)
 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, Hoogesteijn AL, Konah SN, Febles JL, Perkins DJ, Smith SD. 2012 Connecting network properties of rapidly disseminating epizoonotics. *PLoS ONE* **7**, e39778. (doi:10.1371/journal.pone.0039778)

5 *Paper IV*

The following 17 pages contain the paper:

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”. Unpublished paper draft.

Microscopic single molecule dynamics suggest underlying physical properties of the silencing foci

Susmita Sridar^{1 †}✉, Mathias Spliid Heltberg^{2 †}✉, Christian Michelsen^{2 †}✉,
Angela Taddei¹, Judith Mine Hattab¹

¹Institut Curie, PSL University, Sorbonne Universite, CNRS, Nuclear Dynamics, Paris,
France; ²Niels Bohr Institute, University of Copenhagen

✉ For correspondence:

mathias.heltberg@nbi.ku.dk
(MH)

[†]Authors contributed equally.

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.

Competing interests: The
author declare no competing
interests.

Abstract

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 Understanding the physical principles of how cells can express and silence specific regions of the genome presents one of the most fundamental challenges in biology. As a model to study this, 28 budding yeast chromosomes is a strong candidate, since it has very few repetitive sequences outside of the rDNA compared to other eucaryotes that contain centromeric hetero-chromatin. When 30 haploid cells grow at their maximal rate, one characteristic aspect is that 32 telomeres accumulate at the nuclear envelope allowing them to form ≈ 3 –5 foci. The sizes of these are in the order of a few hundreds of nanometer and therefore below the diffraction limit of conventional epifluorescence microscopes.

34 Inside such foci, the silent regulatory factors Sir2, Sir3 and Sir4 concentrate into the form of the SIR complex (Palladino et al., 1993). These are therefore termed silencing foci, since they can 36 press the expression of the underlying genes through interaction with the telomeric protein Rap1, and thereby spread on chromatin and potentially forming a compact chromatin structure. Studies 38 in vitro has revealed that this complex associates with nucleosome in a 1:2:1 stoichiometry and can significantly compact chromatin (Swygert et al., 2018).

40 The sequestration of SIR proteins from silent chromatin favor the subtelomeric repression and the position of telomeres inside these foci favors faithful recombination events upon double strand 42 break (Batté et al., 2017). Furthermore, it also prevents the binding of the SIRs at specific groups of promoters in the genome (Maillet et al., 1996; Marcand et al., 1996; Taddei et al., 2009).

44 In the foci, the telomere composition is not fixed, however telomeres show preferential attachment to other telomeres coupled to chromosome arms of approximately equal length (Therizols 46 et al., 2010; Schober et al., 2008; Duan et al., 2010). This process of telomeres grouping in a limited number of foci requires Sir3 association to telomeres but is independent of heterochromatin 48 formation (Ruault et al., 2011) and these foci has been revealed to fuse into bigger foci or hyper-clusters when SIR3 is overexpressed, suggesting a regulatory role on telomere clustering for SIR3 50 (Ruault et al., 2011).

In this work we investigate the physical mechanism of the formation of silencing foci. In particular 52 we use using Single Particle Tracking (SPT) and Photo Activable Localization Microscopy (PALM) in

Saccharomyces cerevisiae cells in order to obtain precise information about the dynamics of single particles in the heterogenous environment. In this, SPT is a powerful technique that makes the microscopic steps taken by the molecules observable, by taking “live” recordings of individual molecules in a cell at high temporal and spatial resolution (50 Hz, 30 nm) (Dolgin, 2019; Manley et al., 2008; Oswald et al., 2014). Based on this in vivo movement, SPT allows for grouping specific proteins into subpopulations defined by the measured diffusion coefficients. From this it is possible to quantify the motion of each subpopulation and thereby estimating the residence times in different parts of the nucleus, allowing us to estimate the free-energy of the system. To assist the SPT measurements, PALM can establish a density maps of the molecules of interest by their position at 30 nm resolution.

Using these methods we have assessed the dynamics of SIR3 cells with silencing foci. We find that inside the silencing foci, SIR3 moves significantly slower and we relate this to the motion of the whole focus itself. This allow us to identify the diffusion properties of both free telomeres, and telomeres inside a focus. Next we apply, Sir4 deprived mutants and observe that the foci has disappeared, allowing us to extract the free diffusion coefficient of SIR3. Finally we use this to extract the free energy of the molecules inside the repair foci, and we compare this to the theoretical prediction, assuming that the repair foci belongs to the Polymer-Bridging model. Here we find a good agreement, thus suggesting that the physical nature of these foci is really a dense collection of multiple binding sites that suppress the movement of molecules while enhancing their concentration in the formed region.

2 | METHODS & MATERIALS

2.1 | Diffusion model

For each of the different types of data, we load in the cells and group them by cell number and ID. For each group we compute the distance Δr between the subsequent observations \vec{x}_i :

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

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

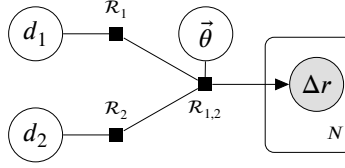


Figure 1. A graphical representation of the Bayesian model case of two diffusion components using the directed factor graph notation (Dietz, 2022). Here d_1 is the diffusion coefficient, \mathcal{R}_1 is the d -parameterized Rayleigh distribution and $\mathcal{R}_{1,2}$ is the mixture model of the Rayleigh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

80 bution is given by:

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

82 In this study, we parameterize the Rayleigh distribution in terms of the diffusion coefficient d , which is related to the scale parameter σ in eq. (2), through the time resolution parameter, τ :

$$84 \quad \sigma = \sqrt{2d\tau}, \quad (3)$$

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:

$$\begin{aligned} & [d \text{ prior}] && d \sim \text{Exponential}(0.1) \\ & 88 \quad [\text{transformation}] && \sigma = \sqrt{2d\tau} \\ & && \\ & 90 \quad [\text{likelihood}] && \Delta r_i \sim \text{Rayleigh}(\sigma). \end{aligned} \quad (4)$$

A more realistic diffusion model include more than a single diffusion coefficient. Figure 1 shows 92 this for the two-component case in directed factor graph notation (Dietz, 2022). In particular, the 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 ν -prior. We model each of the distances as independent, indicated by the N -replications plate. In equations, the figure is similar to:

$$\begin{aligned} & 96 \quad [d_1 \text{ prior}] && d_1 \sim \text{Exponential}(0.1) \\ & && [d_2 \text{ prior (ordered)}] && d_2 \sim \text{Exponential}(0.1), \quad d_1 < d_2 \\ & 98 \quad [\vec{\theta} \text{ prior}] && \theta_1 \sim \text{Uniform}(0, 1), \quad \vec{\theta} = [\theta_1, 1 - \theta_1] \\ & && [\text{mixture model}] && \mathcal{R}_{1,2}(d_1, d_2, \vec{\theta}) = \text{MixtureModel} \left([\mathcal{R}(d_1), \mathcal{R}(d_2)], \vec{\theta} \right) \\ & 100 \quad [\text{likelihood}] && \Delta r_i \sim \mathcal{R}_{1,2}(d_1, d_2, \vec{\theta}). \end{aligned} \quad (5)$$

¹ ordered such that $d_1 < d_k < d_K$ to prevent the classical label-switching problem in the case of mixture models (McLachlan and Peel, 2004)

102 2.2 | Model comparison

We can generalize the $K = 2$ diffusion model to higher values of K by having d_1, \dots, d_K ordered¹ diffusion coefficients and letting the mixture model's $\bar{\theta}$ -prior be a random variable from a flat Dirichlet distribution (such that $\sum_k \theta_k = 1$). We find that including up to three diffusion coefficients yields appropriate results. To compare the three models of different complexity, we compute the Widely Applicable Information Criterion (WAIC) (Watanabe, 2010) which is a generalized version of the Akaike information criterion (AIC) useful for Bayesian model comparison (Gelman, Hwang, and Vehtari, 2014). In short, the WAIC is an approximation of the out-of-sample performance of the model and consists of two terms, the log-pointwise-predictive-density, lppd, and the effective number of parameters p_{WAIC} :

$$112 \quad \text{WAIC} = -2 (\text{lppd} - p_{\text{WAIC}}). \quad (6)$$

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

2.3 | Implementation

120 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

126 3.1 | Two diffusive populations identified at for SIR3 mobility in WT

We started out by using SPT to investigate the mobility of individual SIR3 proteins in vivo in WT cells. To obtain this imaging of SIR3 without altering its normal expression level, we constructed a line of haploid cells that express the endogenous SIR3 fused to Halo (Figure 2A and Materials

130 and methods). Before we visualized this on a PALM microscope (see Materials and methods), we
incubated the exponentially growing cells with fluorescent and fluorogenic JF647. This is a dye that
132 emits light when it is bound to Halo. We then used a low concentration of JF647 in order to obtain
visible individual molecules (Ranjan et al., 2020; Figure 1B). With this setup, the SIR3-Halo bound
134 to JF647 (SIR3-Halo/JF647) were visualized at 20 ms time intervals (50 Hz) in 2-dimensions during
1000 frames until all signal had decayed. A typical individual cell is shown in Figure 2B and the
136 tracking of the individual molecules is visualised in Figure 2C and based on these we moved on
to calculate the density and displacement maps of the SIR3 molecules. Here it should be noted
138 that the tracking of SIR3 is performed in 2-dimensions and the molecules are observable as long
as are inside the focal plan which is the z-section of about 400 nm (Figure 2D). After measuring
140 all the traces, we computed the Probability Density Function for the trace lengths, and here we
found that while the shortest traces seemed to follow an exponential decay, there was a tail with
142 some very long traces (Figure 2E). Here it is important to note that the half-life time of JF6467 is
approximately 2 seconds, meaning that the short traces are due to molecules moving out of the
144 observable z-section and not the photo bleaching of the JF647 dyes.

We aimed to estimate the effective diffusion coefficient of SIR3 in the WT environment, and
146 therefore we computed the displacement for all points in each trace separately, and grouped these
into the displacement histogram (Hansen et al., 2018; Klein et al., 2019; Stracy and Kapanidis, 2017).
148 In this way we could test the naive hypothesis that SIR3 molecules simply exhibit a single diffusive
motion. We therefore fit the displacement histograms to a Rayleigh distribution (a one parameter
150 fit), and use the resulting fit quality to determine if this hypothesis is sufficient to describe the
obtained data. By using Maximum-likelihood minimisation, we extract the most likely value for the
152 diffusion coefficient and based on this we use the Kolmogorov-Smirnoff (KS) test, obtaining a p-
value of $p = 0.0001$, indicating that more complex motion takes place. To take into account that the
154 molecules can diffuse inside the silencing foci and outside these, we introduce two subpopulations
characterized by distinct diffusion coefficients (see Materials and methods). By analysing individual
156 cells, we observe that single traces can be very long in a small region of space, indicating a lower
diffusion coefficient (Figure 2F). By fitting the displacement histogram with the two-population fit,
158 we reveal that this leads to a good agreement. We further introduce a third population of diffusion
coefficients, but obtain similar quality of the fit arguing that two diffusion coefficients is sufficient
160 to explain the motion of the data (Figure 2G).

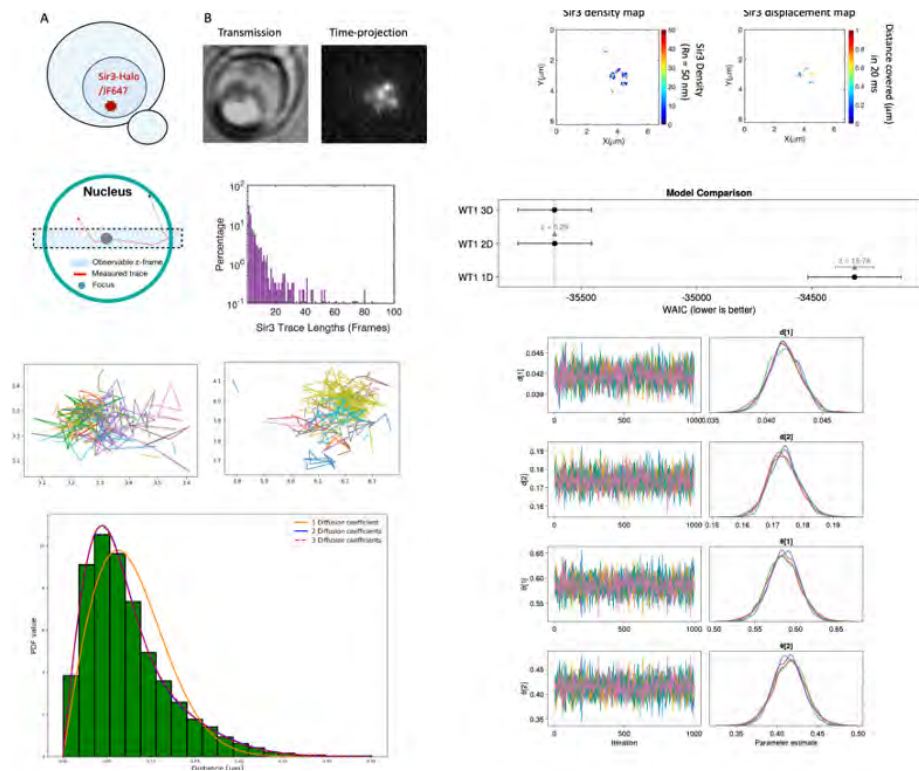


Figure 2. Mobility of Sir3 inside foci in WT cells.

Based on this we conclude that SIR3 in WT seem to have a motion defined by two distinct pop-
 162 ulations, significantly different from each other. While one of these populations has a very small
 diffusion coefficient, representing the motion inside the focus, the other seem to be slow compared
 164 to free molecules (compare to the free RAD52 for instance in Miné-Hattab et al., 2022). Therefore
 we hypothesise that this could be related to motion of SIR3 molecules attached to single telomeres,
 166 that are not part of the foci and therefore has higher mobility. In order to test this hypothesis, we
 tried to remove the existing foci and obtain the motion in this environment.

168 3.2 | Increased mobility of SIR3 in SIR2D-4D mutants

From the theory of silencing foci, it is well established that the proteins SIR2D and SIR4D should be
 170 present in order for the foci to assemble. Therefore, we hypothesised that by deleting these two
 related genes and thereby removing the availability of SIR2D and SIR4D, the silencing foci should
 172 not be able to form (Figure 3A). We succeeded in doing this, and observed that the motion of SIR3

seemed less dense at specific locations compared to the motion of WT (Figure 3B, compare to
174 Figure 2B). Furthermore, we also noted that the traces seemed to be shorter and no traces were
as long as the ones we observed in the WT conditions (Figure 3B). Therefore again computed the
176 displacement histogram and used similar methods as described in Figure 2G to extract the dif-
fusion coefficients. Again, we found that one diffusion coefficient could not explain the motion
178 of SIR3, but two- and three subpopulations did indeed fit the data sufficiently well. Even though
the three-diffusion coefficient fit did lead to a slight improvement in the fit, the two-population
180 fitted the data very well (Figure 3C). This was further confirmed by turning to the Bayesian analysis,
where we obtain a well-defined and unimodal distributions for each of the fitting parameters in
182 the two-population fit (Figure 3D). By comparing the related WAIC scores, we also found that the
three-population fit leads to a 1.28σ increase in the fit quality, but since this is within statistical
184 uncertainty, we conclude that the two-population fit has the most explanatory power of the ob-
served data. By inspecting the diffusion coefficients here we note a very interesting aspect: While
186 the slow diffusion coefficient, found in the WT motion, has disappeared in the SIR2D-4D mutant,
the high diffusion coefficient for the WT is also identified in the motion of SIR3 in the SIR2D-4D
188 mutant, but that a new faster population also has emerged. This supports our hypothesis that
the slow observed diffusion coefficient in the WT is a result of the motion inside the foci, but that
190 the fast diffusion coefficient does not represent freely diffusing molecules, but rather molecules
attached to the semi-mobile telomeres. This also means that effectively all SIR3 molecules are
192 bound in the WT suggesting a high number of binding sites and a high binding rate of these sites.

To further support these claims, we constructed a SIR2D mutant, that was deprived of SIR2 but
194 still had SIR4 (Figure 3E). Here we again found a well distributed map of SIR3 (Figure 3F), and by com-
puting the displacement histogram we revealed that approximately the same diffusion coefficients
196 existed in this mutant (Figure 3G – compare to Figure 3C). Here it seemed that the two-population
fit differed slightly more from the three-population fit than the double mutant. To compare the
198 quality of all hypotheses we again turned to the Bayesian analysis, where we could again find
that all parameters in the two-population fit were smoothly, unimodally distributed, and while the
200 three population fit had slightly better predictive power it was still only 1.78σ and therefore within
statistical uncertainty (Figure 3H).

202 Based on this, we conclude that by depriving the cell of SIR2 (and SIR4), the foci disappears and
the mobility of SIR3 is increased and is described by two populations: A slow population repre-

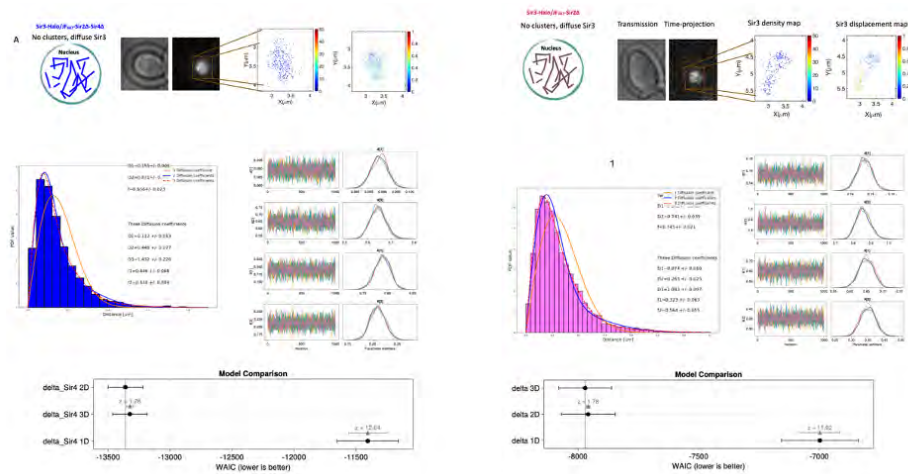


Figure 3. Mobility of Sir3 inside foci in mutant strains.

204 sending the bound molecules to single telomeres and a fast population representing the free SIR3
 molecules. To understand the nature of the actual foci, we needed to understand the mobility of
 206 SIR3 inside the foci better. Therefore, we investigated the movement of the foci itself.

3.3 | Mobility of silencing foci is comparable to the motion of SIR3 inside 208 the foci

Our aim was now to extract the motion of the foci as a single structure and compare this to the
 210 motion of the single molecules. In order to obtain this, we used high photo-activation illumination
 to simultaneously activate all SIR3-mMaple and image the silencing foci as a single entities. Here
 212 we are aware that the observed movement should now be dominated by the focus, but since the
 binding of single SIR3 molecules to the single telomeres, we should be aware that this could also be
 214 observed in the data (Figure 4A). We extracted the traces of these whole-mobility structures, and
 we obtained some confined slowly diffusing traces (blue part in Figure 4B) but also many faster
 216 moving traces (multiple colours in Figure 4B). By eye, this does suggest that some movement takes
 place as a well-defined structure (a silencing-focus) while other motion might be due to the more
 218 mobile single telomeres. To test this, we now generated the displacement histograms for the en-
 tities, and extracted the diffusion coefficients (Figure 4C). In order to compare the hypotheses of
 220 the subpopulations, we directly applied the bayesian analysis and while the two-parameter fit did
 again lead to well-defined parameters, the three-population fit did lead to a better fit (6.08 σ). We

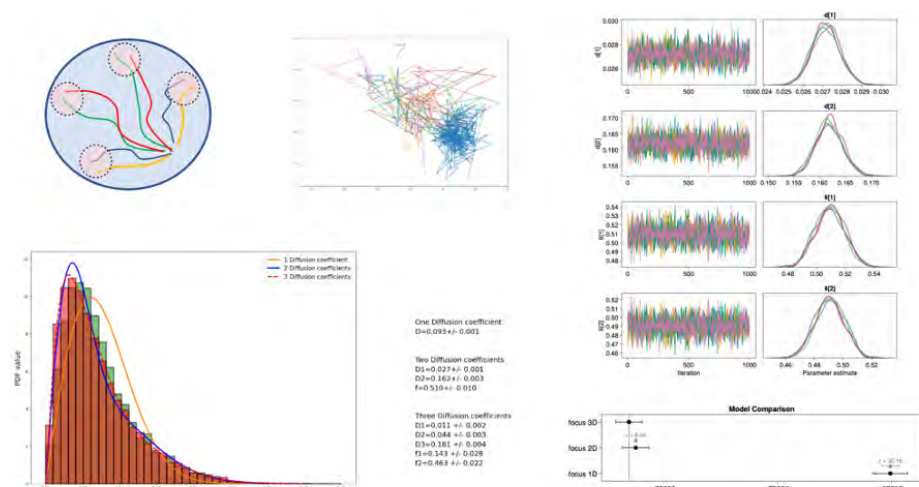


Figure 4. Individual Sir3 vs. whole focus.

222 note however that in both fits, the fast part of the population does match the diffusion coefficient
of the single telomeres observed for single molecule SIR3 in both WT and SIR2D-SIR4D mutants.
224 Now focusing on the slow part of the diffusion coefficients, we note that these match the extracted
diffusion coefficients we found for the single molecule movement of SIR3. However to obtain a bet-
226 ter understanding for the similarities of these, and in particular in order to extract the experimental
noise levels in the measurements since these might differ significantly for the measurements of
228 the entire focus and and measurements of the single SIR3, we moved on to measure the mean
squared distance (MSD) and use these to extract the actual diffusion coefficients.

230 3.4 | Diffusion of SIR3 inside the silencing focus match the predicted move- ment of a Polymer Bridging Model

Our aim was now to extract the motion of the foci as a single structure and compare this to the
motion of the single molecules. In order to obtain this, we used the previously derived theoretical
result that connects the diffusion coefficient inside the foci structures to the free energy of these
(Heltberg et al., 2021). Here the exact diffusion coefficient is extremely important and the result we
obtained in section two is affected by the experimental noise level and this has a significant impact
since the diffusion coefficient is so low. In order to separate these we used the method of Mean-
Square Distances (MSD). Here we take the slow part of the population in the WT data, and for traces
belonging to this family of diffusion coefficients we generate the mean square distances. Finally,

we fit the three first datapoints to a straight line, and use the slope as the diffusion coefficient whereas the intersection is a parameter determined by the experimental noise level. In order to obtain the free energy, we compare the fraction of traces belonging to the slow population, relative to the fast part of the population (See Miné-Hattab et al., 2022 for similar application). In this we take the size of the observable frame compared to the overall size of the cell nucleus into account, as well as we estimate an average of four foci on average. With this we obtain a relation between the free energy and the diffusion coefficient. We know that in a polymer bridging model this should scale as:

$$U = k_{\text{BT}} \ln \left(\frac{D_{\text{inside}} - D_{\text{focus}}}{D_{\text{outside}} - D_{\text{focus}}} \right). \quad (7)$$

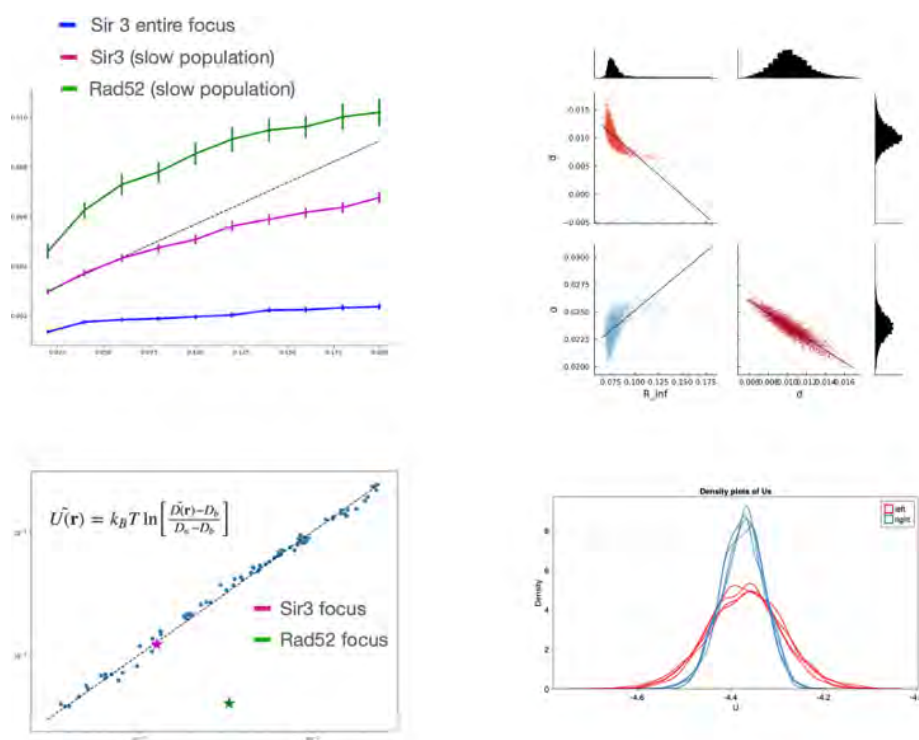


Figure 5. Free energy and diffusion relation. Relation to Rad52.

232 Here we assume that the diffusion coefficient of the focus is similar to the diffusion coefficient
of the binding sites that would diffuse in a bridging model. We then used the simulation results of
234 Heltberg et al., 2021, to show that in the simulations this type of structure always yield this relation
and we showed the result of the relation in the repair foci that is markedly off this line (Figure 5B).

236 Finally we plotted the result for the silencing foci for the values obtained in this study and here
we obtained a remarkable agreement. To further validate these results we used the Bayesian
238 approach, where we tested the mutual correlation of the parameters investigated in the MSD curve
(Figure 5C). Next we used this method to extract the free energy from the populations and compare
240 this to the free energy estimate based on the extracted diffusion coefficients (Figure 5D). These
were completely comparable, which further strengthens the conclusion that the motion inside the
242 silencing foci is really comparable to what would be theoretically expected in the polymer bridging
model.

244 4 | DISCUSSION

The two leading hypotheses for describing the nature of nuclear foci is the polymer bridging model
246 and the liquid droplet model. In this work we have used the data obtained from SPT experiments
to investigate the underlying nature of the silencing foci, experienced by the motion of SIR3. We
248 find that the behaviour is comparable with the theoretical expectations of the polymer bridging
model and this work therefore strengthens the hypothesis that these structures are indeed a dense
250 collection of binding sites.

From a theoretical perspective, it is noteworthy that the method we apply here cannot directly
252 falsify the hypothesis of a liquid structure, but rather it fails to disprove the hypothesis of a polymer
bridging structure. We use a statistical mechanics formulation, derived a mean field, for the PBM.
254 This shares the same functional form as the LPM in the sense that the diffusion coefficient follow a
step function with one value inside the focus and another value outside. However for the PBM we
256 have an additional constraint that precisely links the concentration of proteins and their relative
diffusion inside the focus: the more time spent inside the focus, the slower the effective diffusion.
258 In this sense, if the diffusion coefficient is higher than some value (To the right of the diagonal in
Figure 4A), then this would typically represent a liquid droplet where diffusion can be faster.

260 From a functional perspective, it is also interesting to consider the role of a polymer bridging
model, compared to a liquid model. The simplest form of a silencing foci, would simply keep away
262 the transcription factors and activators. We have previously shown (Heltberg et al., 2021) that the
existence of a polymer bridging model, would typically increase the first passage time to find a
264 target, whereas a liquid model could greatly enhance this. Therefore it is a tempting hypothesis,

that foci formed with the aim of slowing down rates would be of a polymer bridging model, whereas
266 the foci with the aim of increasing rates (for instance in the repair foci) could be liquid droplets. On a
more general point, foci are formed inside the nucleus for various reasons with different roles, and
268 it is clear that they can remain stable very different timescales. Here it is interesting that repair foci
are maintained for relative short periods (timescale of hours) and they have the ability to quickly
270 dissolve as long-term stability is not so important. On the other hand, gene expression foci can be
very stable (Hnisz et al., 2017; Bing et al., 2020), and this could be explained by the hypothesis that
272 these would typically be polymer bridging structures.

4.1 | Acknowledgment

274 Acknowledgements here

4.2 | Data availability

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

REFERENCES

- 278 Batté, Amandine et al. (2017). "Recombination at subtelomeres is regulated by physical distance,
double-strand break resection and chromatin status". eng. In: *The EMBO journal* 36.17, pp. 2609–
280 2625. ISSN: 1460-2075. DOI: [10.15252/embj.201796631](https://doi.org/10.15252/embj.201796631).
- 282 Betancourt, Michael (2018). "A Conceptual Introduction to Hamiltonian Monte Carlo". In: *arXiv:1701.02434*
[stat]. arXiv: 1701.02434.
- 284 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/>.
- 286 Bing, X. Y. et al. (2020). "Snapshot: The Regulatory Genome". en. In: *Cell* 182.6, 1674–1674.e1. ISSN:
0092-8674. DOI: [10.1016/j.cell.2020.07.041](https://doi.org/10.1016/j.cell.2020.07.041). URL: <https://www.sciencedirect.com/science/article/pii/S0092867420309491> (visited on 2022).
- 288 Dietz, Laura (2022). "Directed factor graph notation for generative models". In.
- 290 Dolgin, Elie (2019). "The sounds of science: biochemistry and the cosmos inspire new music". en. In:
Nature 569.7755. Bandiera_abtest: a Cg_type: Books And Arts Number: 7755 Publisher: Nature
Publishing Group Subject_term: Arts, Culture, pp. 190–191. DOI: [10.1038/d41586-019-01422-0](https://doi.org/10.1038/d41586-019-01422-0).
292 URL: <https://www.nature.com/articles/d41586-019-01422-0> (visited on 2022).
- 294 Duan, Zhijun et al. (2010). "A three-dimensional model of the yeast genome". eng. In: *Nature* 465.7296,
pp. 363–367. ISSN: 1476-4687. DOI: [10.1038/nature08973](https://doi.org/10.1038/nature08973).
- 296 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/
298 ge18b.html](https://proceedings.mlr.press/v84/ge18b.html) (visited on 2022).
- 300 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-
1375. 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
302 2022).
- 304 Hansen, Anders S et al. (2018). "Robust model-based analysis of single-particle tracking experi-
ments with Spot-On". In: *eLife* 7. Ed. by David Sherratt. Publisher: eLife Sciences Publications,
Ltd, e33125. ISSN: 2050-084X. DOI: [10.7554/eLife.33125](https://doi.org/10.7554/eLife.33125). URL: <https://doi.org/10.7554/eLife.33125>
306 (visited on 2022).

- Heltberg, Mathias L et al. (2021). "Physical observables to determine the nature of membrane-less cellular sub-compartments". In: *eLife* 10. Ed. by Agnese Seminara, José D Faraldo-Gómez, and Pierre Ronceray. Publisher: eLife Sciences Publications, Ltd, e69181. ISSN: 2050-084X. DOI: 10.7554/eLife.69181. URL: <https://doi.org/10.7554/eLife.69181> (visited on 2022).
- Hnisz, Denes et al. (2017). "A Phase Separation Model for Transcriptional Control". en. In: *Cell* 169.1, pp. 13–23. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.02.007. URL: <https://www.sciencedirect.com/science/article/pii/S009286741730185X> (visited on 2022).
- 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.
- Klein, Hannah L. et al. (2019). "Guidelines for DNA recombination and repair studies: Cellular assays of DNA repair pathways". en. In: *Microbial Cell* 6.1. Publisher: Shared Science Publishers, pp. 1–64. ISSN: 2311-2638. DOI: 10.15698/mic2019.01.664. URL: <http://microbialcell.com/researcharticles/2019a-klein-microbial-cell/>, %20http://microbialcell.com/researcharticles/2019a-klein-microbial-cell/ (visited on 2022).
- Maillet, L. et al. (1996). "Evidence for silencing compartments within the yeast nucleus: a role for telomere proximity and Sir protein concentration in silencer-mediated repression." en. In: *Genes & Development* 10.14. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1796–1811. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.10.14.1796. URL: <http://genesdev.cshlp.org/content/10/14/1796> (visited on 2022).
- Manley, Suliana et al. (2008). "High-density mapping of single-molecule trajectories with photoactivated localization microscopy". en. In: *Nature Methods* 5.2. Number: 2 Publisher: Nature Publishing Group, pp. 155–157. ISSN: 1548-7105. DOI: 10.1038/nmeth.1176. URL: <https://www.nature.com/articles/nmeth.1176> (visited on 2022).
- Marcand, S. et al. (1996). "Silencing of genes at nontelomeric sites in yeast is controlled by sequestration of silencing factors at telomeres by Rap 1 protein". eng. In: *Genes & Development* 10.11, pp. 1297–1309. ISSN: 0890-9369. DOI: 10.1101/gad.10.11.1297.
- 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.

- 338 McLachlan, Geoffrey J. and David Peel (2004). *Finite Mixture Models*. en. Google-Books-ID: c2_fAox0DQoC.
John Wiley & Sons. ISBN: 978-0-471-65406-3.
- 340 Miné-Hattab, Judith et al. (2022). "Single molecule microscopy reveals key physical features of repair
foci in living cells". In: *eLife* 10 (), e60577. ISSN: 2050-084X. DOI: [10.7554/eLife.60577](https://doi.org/10.7554/eLife.60577). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7924958/> (visited on 2022).
- 342 Oswald, Felix et al. (2014). "Imaging and quantification of trans-membrane protein diffusion in liv-
ing bacteria". en. In: *Physical Chemistry Chemical Physics* 16.25. Publisher: The Royal Society of
344 Chemistry, pp. 12625–12634. ISSN: 1463-9084. DOI: [10.1039/C4CP00299G](https://doi.org/10.1039/C4CP00299G). URL: <https://pubs.rsc.org/en/content/articlelanding/2014/cp/c4cp00299g> (visited on 2022).
- 346 Palladino, F. et al. (1993). "SIR3 and SIR4 proteins are required for the positioning and integrity of
348 yeast telomeres". eng. In: *Cell* 75.3, pp. 543–555. ISSN: 0092-8674. DOI: [10.1016/0092-8674\(93\)90388-7](https://doi.org/10.1016/0092-8674(93)90388-7).
- 350 Ranjan, Anand et al. (2020). "Live-cell single particle imaging reveals the role of RNA polymerase
II in histone H2A.Z eviction". In: *eLife* 9. Ed. by Geeta J Narlikar et al. Publisher: eLife Sciences
352 Publications, Ltd, e55667. ISSN: 2050-084X. DOI: [10.7554/eLife.55667](https://doi.org/10.7554/eLife.55667). URL: <https://doi.org/10.7554/eLife.55667> (visited on 2022).
- 354 Ruault, Myriam et al. (2011). "Clustering heterochromatin: Sir3 promotes telomere clustering inde-
pendently of silencing in yeast". In: *The Journal of Cell Biology* 192.3, pp. 417–431. ISSN: 0021-
356 9525. DOI: [10.1083/jcb.201008007](https://doi.org/10.1083/jcb.201008007). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3101097/>
(visited on 2022).
- 358 Schober, Heiko et al. (2008). "Controlled exchange of chromosomal arms reveals principles driving
telomere interactions in yeast". In: *Genome Research* 18.2, pp. 261–271. ISSN: 1088-9051. DOI:
360 [10.1101/gr.6687808](https://doi.org/10.1101/gr.6687808). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2203624/> (visited on
2022).
- 362 Stracy, Mathew and Achillefs N. Kapanidis (2017). "Single-molecule and super-resolution imaging
of transcription in living bacteria". en. In: *Methods*. Transcriptional dynamics 120, pp. 103–114.
364 ISSN: 1046-2023. DOI: [10.1016/j.ymeth.2017.04.001](https://doi.org/10.1016/j.ymeth.2017.04.001). URL: <https://www.sciencedirect.com/science/article/pii/S1046202316305011> (visited on 2022).
- 366 Swygert, Sarah G. et al. (2018). "SIR proteins create compact heterochromatin fibers". In: *Proceed-
ings of the National Academy of Sciences* 115.49. Publisher: Proceedings of the National Academy

- 368 of Sciences, pp. 12447–12452. DOI: [10.1073/pnas.1810647115](https://doi.org/10.1073/pnas.1810647115). URL: <https://www.pnas.org/doi/10.1073/pnas.1810647115> (visited on 2022).
- 370 Taddei, Angela et al. (2009). “The functional importance of telomere clustering: Global changes in gene expression result from SIR factor dispersion”. In: *Genome Research* 19.4, pp. 611–625. ISSN: 1088-9051. DOI: [10.1101/gr.083881.108](https://doi.org/10.1101/gr.083881.108). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2665780/> (visited on 2022).
- 374 Therizols, Pierre et al. (2010). “Chromosome arm length and nuclear constraints determine the dynamic relationship of yeast subtelomeres”. In: *Proceedings of the National Academy of Sciences* 107.5. Publisher: Proceedings of the National Academy of Sciences, pp. 2025–2030. DOI: [10.1073/pnas.0914187107](https://doi.org/10.1073/pnas.0914187107). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.0914187107> (visited on 378 2022).
- 380 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.

APPENDIX

A *Kap København*

The following 32 pages contain the paper published in Nature 2022:

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

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

Article


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

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

Received: 30 September 2021

Accepted: 18 October 2022

Open access

 Check for updates

Kurt H. Kjær^{1,2,3}, Mikkel W. Pedersen^{1,2,3}, Bianca De Sanctis^{2,3}, Binia De Cahsan⁴, Thorfinn S. Korneliussen¹, Christian S. Michelsen^{1,5}, Karina K. Sand¹, Stanislav Jelavić^{1,6}, Anthony H. Ruter¹, Astrid M. Z. Bonde⁷, Kristian K. Kjeldsen⁸, Alexey S. Tesakov⁹, Ian Snowball¹⁰, John C. Gosse¹¹, Inger G. Alsos¹², Yucheng Wang^{1,2}, Christoph Dockter¹³, Magnus Rasmussen¹³, Morten E. Jørgensen¹³, Birgitte Skadhauge¹³, Ana Prohaska¹, Jeppe Å. Kristensen^{9,14}, Morten Bjerager⁹, Morten E. Allentoft^{1,15}, Eric Coissac^{12,16}, PhyloNorway Consortium^{1,7}, Alexandra Rouillard^{1,17}, Alexandra Simakova⁹, Antonio Fernandez-Guerra¹, Chris Bowler¹⁸, Marc Macias-Fauria¹⁹, Lasse Vinner¹, John J. Welch², Alan J. Hidy²⁰, Martin Sikora¹, Matthew J. Collins²¹, Richard Durbin⁹, Nicolaj K. Larsen¹ & Eske Willerslev^{1,2,22}

Late Pliocene and Early Pleistocene epochs 3.6 to 0.8 million years ago¹ had climates resembling those forecasted under future warming². Palaeoclimatic records show strong polar amplification with mean annual temperatures of 11–19 °C above contemporary values^{3,4}. The biological communities inhabiting the Arctic during this time remain poorly known because fossils are rare⁵. Here we report an ancient environmental DNA (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 cycle⁷ (Fig. 1). Nearly 40 years of palaeoenvironmental and climate research at the site provide a unique perspective into a period when the site was situated at the boreal Arctic ecotone with reconstructed summer and winter average minimum temperatures of 10 °C and –17 °C respectively—more than 10 °C warmer than the present^{7–11}.

These conditions must have driven substantial ablation of the Greenland Ice Sheet, possibly producing one of the last ice-free intervals⁷ in the last 2.4 million years (Myr). Although the Kap København Formation is known to yield well-preserved macrofossils from a coniferous boreal forest and a rich insect fauna, few traces of vertebrates have been found. To date, these comprise remains from lagomorph genera, their coprolites and *Aphodius* beetles, which live in and on mammalian dung^{10,11}. However, the approximately 3.4 Myr old Fyles Leaf bed and Beaver Pond on Ellesmere Island in Arctic Canada preserve fossils of

Q5

Q6

¹Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark. ²Department of Zoology, University of Cambridge, Cambridge, UK. ³Department of Genetics, University of Cambridge, Cambridge, UK. ⁴Section for Evolutionary Genomics, Faculty of Health and Medical Sciences, The Globe Institute, Copenhagen K, Denmark. ⁵Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark. ⁶Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, IRD, Université Gustave Eiffel, ISTerre, Grenoble, France. ⁷Halsnaes Kommune, Frederiksværk, Denmark. ⁸GEUS, Geological Survey of Denmark and Greenland, Copenhagen K, Denmark. ⁹Geological Institute, Russian Academy of Sciences, Moscow, Russia. ¹⁰Department of Earth Sciences, Uppsala University, Uppsala, Sweden. ¹¹Department of Earth and Environmental Sciences, Dalhousie University, Halifax, Canada. ¹²The Arctic University Museum of Norway, UiT—The Arctic University of Norway, Tromsø, Norway. ¹³Carlsberg Research Laboratory, Copenhagen V, Denmark. ¹⁴Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, UK. ¹⁵Trace and Environmental DNA (TrEnd) Laboratory, School of Molecular and Life Sciences, Curtin University, Perth, Western Australia, Australia. ¹⁶University of Grenoble-Alpes, Université Savoie Mont Blanc, CNRS, LECA, Grenoble, France. ¹⁷Department of Geosciences, UiT—The Arctic University of Norway, Tromsø, Norway. ¹⁸Institut de Biologie de l'École Normale Supérieure (IBENS), École Normale Supérieure, CNRS, INSERM Université PSL, Paris, France. ¹⁹School of Geography and the Environment, University of Oxford, Oxford, UK. ²⁰Center for Accelerator Mass Spectrometry, Lawrence Livermore National Laboratory, Livermore, CA, USA. ²¹Department of Archaeology, University of Cambridge, Cambridge, UK. ²²MARUM, University of Bremen, Bremen, Germany. ²³These authors contributed equally: Kurt H. Kjær, Mikkel W. Pedersen. *A list of authors and their affiliations appears at the end of the paper. A full list of members and their affiliations appears in the Supplementary Information. ²⁴e-mail: kurtk@sund.ku.dk; ew482@cam.ac.uk

Article

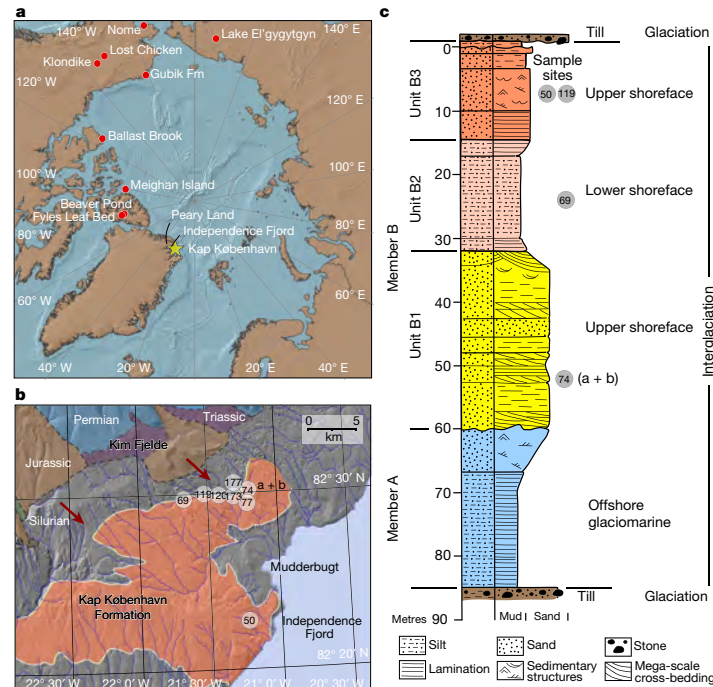


Fig. 1 | Geographic location and depositional sequence. **a.** Location of Kap København Formation in North Greenland at the entrance to the Independence Fjord (82° 24' N 22° 12' W) and locations of other Arctic Plio-Pleistocene fossil-bearing sites (red dots). **b.** Spatial distribution of the erosional remnants of the 100-m thick succession of shallow marine near-shore sediments between Mudderbugt and the low mountains towards the north. **c.** Glacial–interglacial

division of the depositional succession of clay Member A and units B1, B2 and B3 constituting sandy Member B. Sampling intervals for all sites are projected onto the sedimentary succession of locality 50. Sedimentological log modified after ref. ⁷. Circled numbers on the map mark sample sites for environmental DNA analyses, absolute burial dating and palaeomagnetism. Numbered sites refer to previous publications ^{7,10,11,14,98}.

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

The Kap København Formation is formally subdivided into two members⁷ (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 environment¹⁴. 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 facies⁷.

The specific depositional environments are also reflected in the mineralogy of the units, where the proximal B3 locality has the lowest clay and highest quartz contents (Sample compositions in Supplementary Tables 4.2.1 and 4.2.2 and unit averages in Supplementary Tables 4.2.3 and 4.2.4). The architecture of the basin infill suggests that Member B units thicken towards the present coast—that is, distal to the sediment source in the low mountains in the north (Fig. 1). Abundant organic detritus horizons are recorded in units B1 and B3, which also contain beds rich in arctic and boreal plant and invertebrate macrofossils, as well as terrestrial mosses^{10,15}. Therefore, the taphonomy of the DNA

most probably reflects the biological communities eroded from a range of habitats, fluviually 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 mineralogical composition between Kap København Formation sediments and Kim Fjelde sediments (Supplementary Tables 4.2.1 and 4.2.5).

Geological age

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

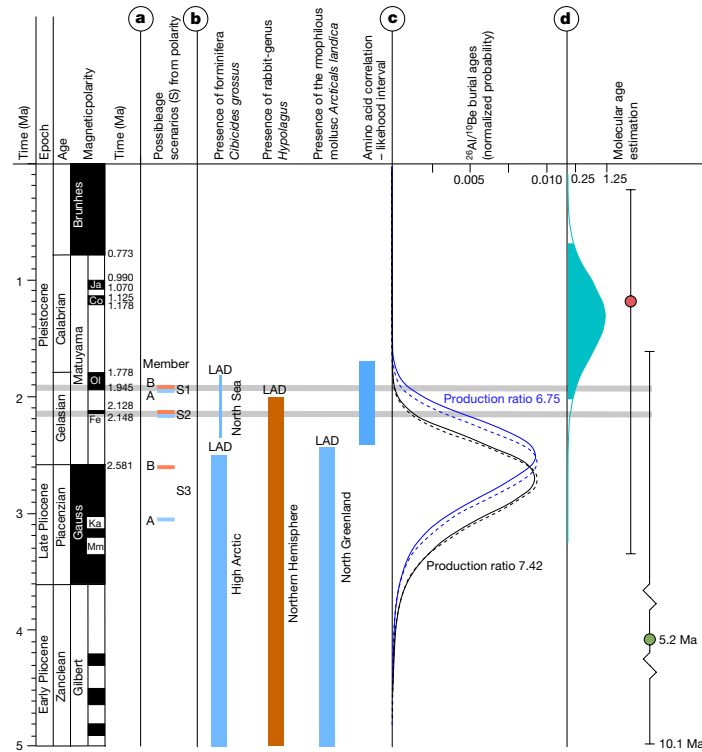


Fig. 2 | Age proxies for the Kap København Formation. **a**, Revised palaeomagnetic analysis shows unit B2 to have normal polarity and unlocks three possible ages scenarios (S1–S3) including Members A (blue) and B (brown). Normal polarity is coloured black and reverse polarity is shown in white. **b**, 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 shells⁷. **c**, Convolved probability distribution functions for cosmogenic burial ages calculated for two different production ratios

(7.42 (black) and 6.75 (blue)). The dashed line and the solid line show the distributions for steady erosion and zero erosion, respectively. These distributions are all maximum ages. **d**, Molecular dating of *Betula* sp. yielding a median age of the DNA in the sediment of 1.323 Myr, with whiskers confining the 95% height posterior density (HPD) of 0.68 to 2.02 Myr (blue density plot), running Markov chain Monte Carlo estimation over for 100 million iterations. The red dot is the median molecular age estimate found using the Mastodon mitochondrial genome restricting to radiocarbon-dated specimens, whereas the green area includes molecular clock estimated specimens in BEAST, running Markov chain Monte Carlo estimation for 400 million iterations. Whiskers confine the 95% HPD.

Q14

Q16

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

DNA preservation

DNA degrades with time owing to microbial enzymatic activity, mechanical shearing and spontaneous chemical reactions such as hydrolysis and oxidation²⁰. 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 dating²¹. 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 temperature²² (MAT) of -17°C , we found a thermal age of $2.7 \text{ kyT}_{\text{DNA}@10^\circ\text{C}}$ —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 fossils²³, we found it plausible that DNA with an average size of 50 base pairs (bp) could survive at the Kap København Formation, assuming that the site remained frozen (Supplementary Information, section 4 and Supplementary Table 4.4.2). Mechanisms that preserve DNA in sediments are likely to be different from that of bone. Adsorption at mineral surfaces modifies the DNA conformation, probably impeding molecular recognition by enzymes, which effectively hinders enzymatic degradation^{24–27}. To investigate whether the minerals found in Kap København Formation could have retained DNA during the deposition and preserved it, we determined the mineralogical composition of the sediments using X-ray diffraction and measured their adsorption capacities. Our findings

Q15

Q17

Article

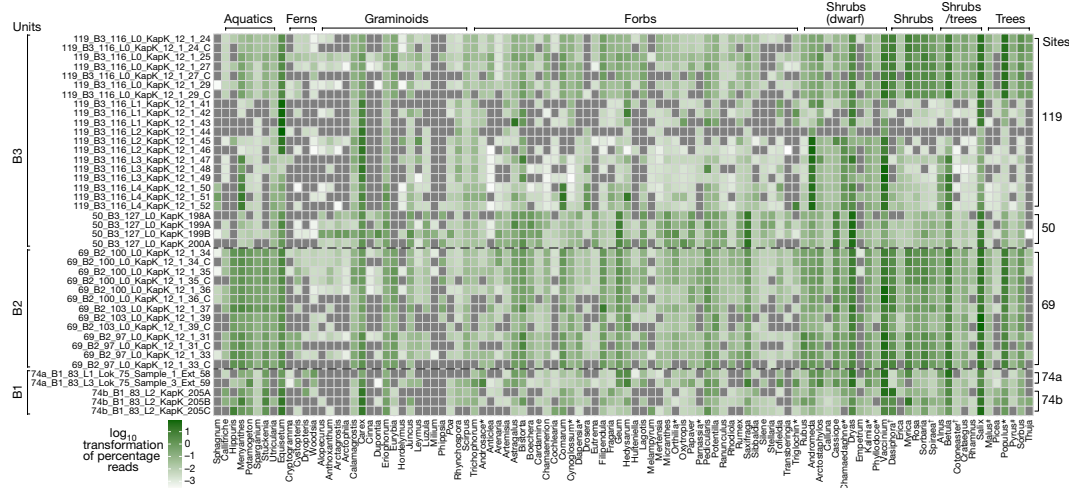


Fig. 3 | Early Pleistocene plants of Northern Greenland. Metagenomic taxonomic profiles of the plant assemblage. Taxa in bold are genera only found as DNA and not as macrofossil or pollen. Asterisks indicate those that are found at other Pliocene arctic sites. Extinct species as identified by either

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 4.3.1.1). 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 environments²⁸ (4.9 ng ml⁻¹ DNA), the DNA adsorption capacity of smectite is 200 times greater than for quartz. We applied a sedimentary eDNA extraction protocol²⁹ 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 DNA²⁹ 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 libraries³⁰. 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 1.0³¹ 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 *kmer* comparisons using *simka*³² (Supplementary Information, section 6) and then parsed for taxonomic classification using competitive mapping with *HOLI* (<https://github.com/miwiipe/KapCopenhagen.git>), which includes a recently published dataset of more than 1,500 genome skims of Arctic and boreal plant taxa^{33,34} (Methods and Supplementary Information, section 6). Considering the age of the samples and thus the potential genetic distance to recent reference genomes, we allowed each read to have a similarity between 95–100% for it to be taxonomically classified using *ngsLCA*³⁵. The *metaDMG* (v.0.14.0) program (<https://metadmg-dev.github.io/metaDMG-core/index.html>) was subsequently used to quantify and filter each taxonomic node for postmortem DNA damage for all the metagenomic samples (Methods). This method estimates the average damage at the termini position (D-max) and a likelihood ratio (λ -LR) that quantifies how much better the damage model (that is, more damage at the beginning of the read) fits the data compared with a null model (that is, a constant amount of damage; see Supplementary Information, section 6). We found the DNA damage to be highly increased, especially for eukaryotes (mean D-max = 40.7%). From this we set D-max $\geq 25\%$ as a filtering threshold for a taxonomic node to be parsed for further downstream analysis as well as a λ -LR higher or equal to 1.5. We furthermore set a threshold requiring that the minimum number of reads per taxon exceeded the median of reads assigned across all taxa divided by two to filter for taxa in low abundance. Similarly, for a sample to be considered, the total number of reads for a sample had to exceed the median number of reads per sample divided by two, to filter for samples with fewest reads. Lastly, we filtered out taxa with fewer than three replicates and subsequently reads were normalized by conversion to proportions (Figs. 3 and 4a).

DNA, pollen and macrofossils comparison

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

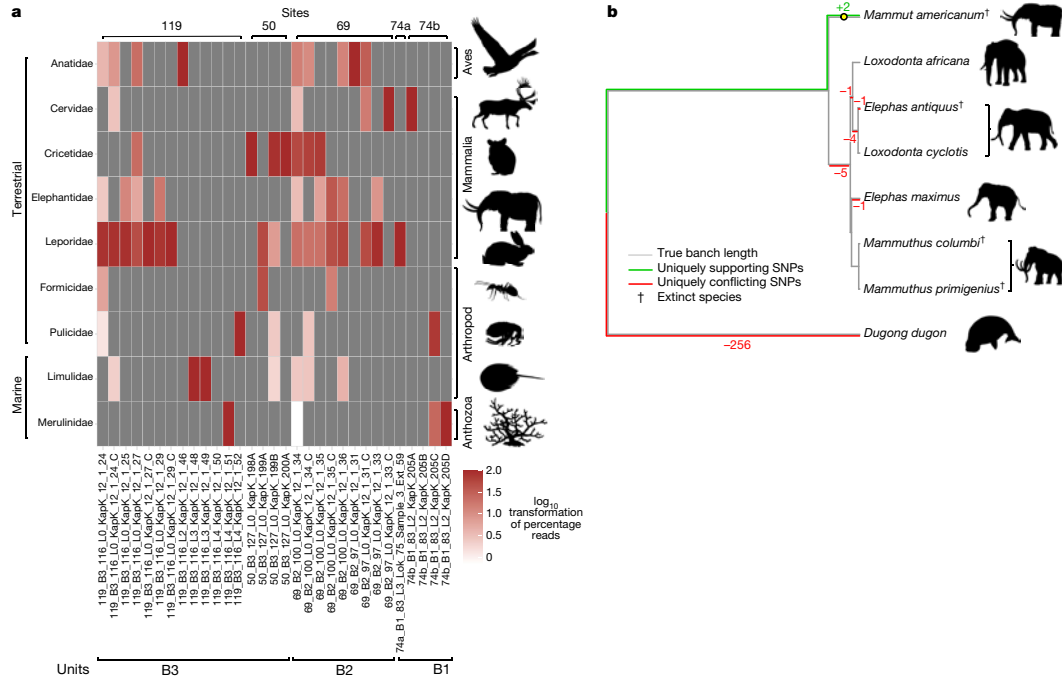


Fig. 4 | Early Pleistocene animals of Northern Greenland. **a**, Metagenomic taxonomic profiles of the animal assemblage from units B1, B2 and B3. Taxa in bold are genera only found as DNA. **b**, phylogenetic placement and pathPhynder⁴⁸ results of mitochondrial reads uniquely classified to Elephantidae or lower (Source Data 1).

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

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

A total of 39 plant genera out of the 102 identified by DNA also occurred as macrofossils or pollen at the genus level. A further 39 taxa were potentially identified as macrofossil or pollen but not to the same

taxonomic level^{10,15} (Source Data 1, sheets 1 and 2). For example, 12 genera of Poaceae were identified by DNA (*Alopecurus*, *Anthoxanthum*, *Arctagrostis*, *Arctophila*, *Calamagrostis*, *Cinna*, *Dupontia*, *Hordelymus*, *Leymus*, *Milium*, *Phippsia* and *Poa*), of these only *Hordelymus* is not found in the Arctic today (<http://panarcticflora.org/>), but these were only distinguished to family level in the pollen analysis and only one Poaceae macrofossil was found. There were 24 taxa that were recorded only as DNA. These included the boreal tree *Populus* and a few shrubs and dwarf shrubs, but mainly herbaceous plants. Of the 73 plant genera recovered as macrofossils^{10,15}, only 24 were not detected in the DNA analysis. Because macrofossils and DNA have similar taphonomies—as both are deposited locally—more overlap is expected between them than between DNA and pollen, which is typically dispersed regionally⁴³. 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 records⁴⁴, 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 Bennike⁴⁵ 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 grow¹⁰. Bennike⁴⁵ 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×, 57.06× and 27.04×, respectively) and sufficient present-day whole chloroplast reference sequences (Methods). Owing to their age and hence potential genetic distance from the modern reference genomes, we lowered the similarity threshold of uniquely classified reads to 90% and merged these by unit to increase coverage. Both *Betula* and *Salix* placed basally to most of the represented species in the respective genera, and the *Populus* placement results showed support for a mixture of different species related to *P. trichocarpa* and *P. balsamifera* (Extended Data Figs. 7–9).

We used the *Betula* chloroplast reads for a molecular dating analysis, because they were placed confidently on a single edge of the phylogenetic tree (that is, not a mixture as in *Populus*), had a large number of reference sequences, and had high coverage in the ancient sample. We used BEAST⁴⁶ 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 time⁴⁷ of 61.1 Myr for calibration of the root node. Our BEAST analysis was robust to both different priors on the age of the ancient sample, and to different nucleotide substitution models (Supplementary Fig. 10). This yielded a median age estimate of 1.323 Myr, with a 95% HPD of (0.6786, 2.0172) Myr (Fig. 2).

Animal DNA results

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

Q19

Most notably, we found reads in unit B2 and B3 assigned to the family Elephantidae, which includes elephants and mammoths, but taxonomically not mastodon (*Mammot* sp.)—which are, however, in the NCBI taxonomy, and therefore our analysis reads classified to Elephantidae or below therefore include *Mammot* sp. A consensus genome of our Elephantidae mitochondrial reads falls on the *Mammot* 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 BEAST⁴⁸. 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 record⁷. *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 Greenland⁴⁹. By aligning our reads against the Tara Oceans eukaryotic metagenomic assembled genomes (SMAGs) data (Methods), we further reveal the presence of 24 marine planktonic taxa in 14 samples, covering both zooplankton and phytoplankton (Fig. 5). These detected SMAGs belong to the supergroups Opisthokonta (6), Stramenopila (15) and Archaeplastida (3). The majority of these signals are from SMAGs associated with cold regions in the modern ocean (that is, the Arctic Ocean and Southern Ocean), such as diatoms (Bacillariophyta), Chrysophyceae and the MAST-4 group (Supplementary

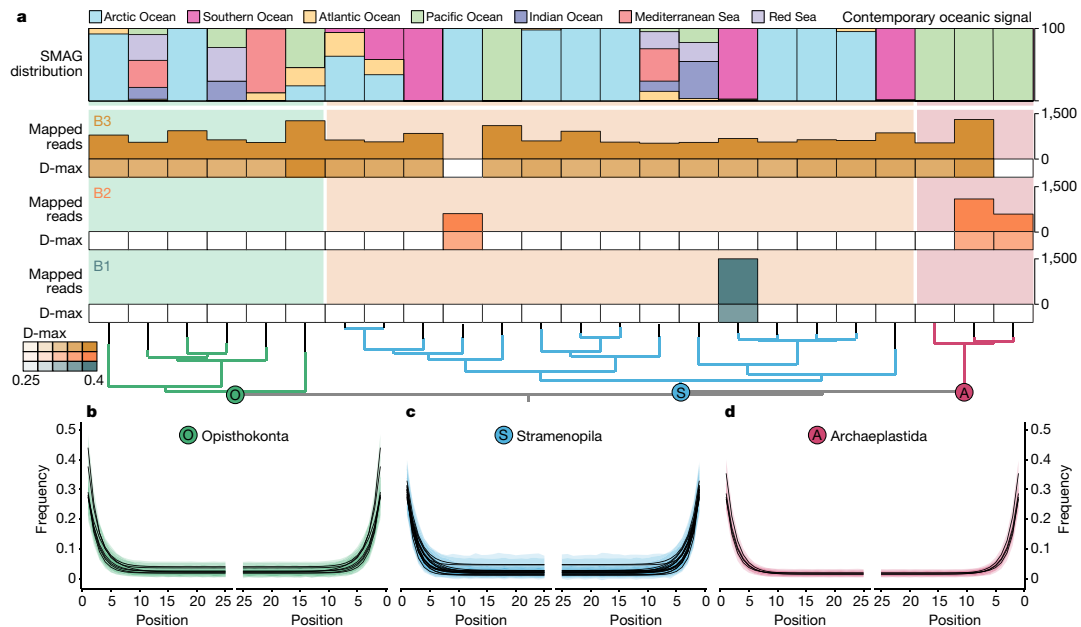


Fig. 5 | Marine planktonic eukaryotes identified at the Kap København Formation. a, Detection of SMAGs and average damage (D-max) of a SMAG within a member unit. Top, the SMAG distribution in contemporary oceans based on the data of Delmont et al.²³. The SMAGs are ordered on the basis of

phylogenomic inference from Delmont et al.⁷³. b–d, Distribution of DNA damage among the taxonomic supergroup Opisthokonta (b), Stramenopila (c) and Archaeplastida (d) (Source Data 1).

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

Discussion

The Kap København ancient eDNA record is extraordinary for several reasons; the upper limit of the 95% highest posterior density of the estimated molecular age is 2.0 Myr and independently supports a geological age of approximately 2 Myr (Fig. 2). This implies that the DNA is considerably older than any previously sequenced DNA²¹. Our DNA results detected five times as many plant genera as previous studies using shotgun sequencing of ancient sediments^{29,34,50,51}, which is well within the range of the richest northern boreal metabarcoding records⁵². 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 *Mammot* (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 sediments³⁴. 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 macrofossils⁷. 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 samaras⁵³. The Kap København units with mastodon DNA yielded macrofossils and DNA from *Betula* as well as more thermophilic arboreal taxa including *Thuja*, *Taxus*, *Cornus* and *Viburnum*, none of which range into Greenland's hydric Arctic tundra or polar deserts today. The co-occurrence of these taxa in multiple units compels a revision of previous temperature estimates as well as the presence of permafrost.

Article

No single modern plant community or habitat includes the range of taxa represented in many of the macrofossil and DNA samples from Kap København. The community assemblage represents a mixture of modern boreal and Arctic taxa, which has no analogue in modern vegetation^{10,15}. To some degree, this is expected, as the ecological amplitudes of modern members of these genera have been modified by evolution⁵⁴. Furthermore, the combination of the High Arctic photoperiod with warmer conditions and lower atmospheric CO₂ concentrations⁵⁵ 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 Ophisthokonta and Stramenophila are found together with horseshoe crabs, corals and green microalgae (Archaeplastida), which today inhabit warmer waters at more southern latitudes.

Megaherbivores, particularly mastodons, could have had a significant impact on an interglacial taiga environment, even providing a top-down trophic control on vegetation structure and composition at this high latitude. The presence of mastodons^{56,57} coupled with the absence of anthropogenic fire, which has had a role in some Holocene boreal habitats⁵⁸, are important differences. Another important factor is the proximity and biotic richness of the refugia from which pioneer species were able to disperse into North Greenland when conditions became favourable at the beginning of interglacials. The shorter duration of Early Pleistocene glaciations produced less extensive ice sheets allowing colonization from relatively species-rich coniferous-deciduous woodlands in northeastern Canada^{12,59}. More extensive glaciation later in the Pleistocene increasingly isolated North Greenland and later re-colonizations were from increasingly distant and/or less diverse refugia.

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

Online content

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

- 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).
- IPCC Climate Change 2013: *The Physical Science Basis* (eds Stocker, T. F. et al.) (Cambridge Univ. Press, 2013).
- Brigham-Grette, J. et al. Pliocene warmth, polar amplification, and stepped Pleistocene cooling recorded in NE Arctic Russia. *Science* **340**, 1421–1427 (2013).
- Gosse, J. C. et al. PoLAR-FIT: Pliocene Landscapes and Arctic Remains—Frozen in Time. *Geosci. Can.* **44**, 47–54 (2017).
- Matthews, J. V., Telka, A. Jr & Kuzmina, S. A. Late Neogene insect and other invertebrate fossils from Alaska and Arctic/Subarctic Canada. *Zool. Bespozovnoy* **16**, 126–153 (2019).
- Willerstein, E. et al. Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* **300**, 791–795 (2003).
- Funder, S. et al. Late Pliocene Greenland—the Kap København formation in North Greenland. *Bull. Geol. Soc. Den.* **48**, 117–134 (2001).
- Funder, S. & Hjort, C. A reconnaissance of the Quaternary geology of eastern North Greenland. *Rapp. Grønlands Geol. Unders.* **99**, 99–105 (1980).
- Fredskild, B. & Røen, U. Macrofossils in an interglacial peat deposit at Kap København, North Greenland. *Boreas* **11**, 181–185 (2008).
- 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).
- Böcher, J. *Palaeontology of the Kap København Formation, a Plio-Pleistocene sequence in Peary Land, North Greenland* (Museum Tusulanum Press, 1995).
- Rybczynski, N. et al. Mid-Pliocene warm-period deposits in the High Arctic yield insight into camel evolution. *Nat. Commun.* **4**, 1550 (2013).

- Wang, X., Rybczynski, N., Harington, C. R., White, S. C. & Tedford, R. H. A basal ursine bear (*Protarctos abstrusus*) from the Pliocene High Arctic reveals Eurasian affinities and a diet rich in fermentable sugars. *Sci. Rep.* **7**, 17722 (2017).
- Simonarson, L. A., Petersen, K. S. & Funder, S. *Molluscan palaeontology of the Pliocene-Pleistocene Kap København Formation, North Greenland*. *Arct. Antarct. Alp. Res.* **32**, (1998).
- Mogensen, G. S. Pliocene or Early Pleistocene mosses from Kap København, North Greenland. *Lindbergia* **10**, 19–26 (1984).
- Funder, S., Abrahamsen, N., Bennike, O. & Feyling-Hanssen, R. W. Forested Arctic: evidence from North Greenland. *Geology* **13**, 542–546 (1985).
- Abrahamsen, N. & Marcussen, C. Magnetostratigraphy of the Plio-Pleistocene Kap København Formation, eastern North Greenland. *Phys. Earth Planet. Inter.* **44**, 53–61 (1986).
- 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).
- Feyling-Hanssen, R. W. *Foraminiferal Stratigraphy in the Plio-Pleistocene Kap København Formation, North Greenland* (Museum Tusulanum Press, 1990).
- Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
- van der Valk, T. et al. Million-year-old DNA sheds light on the genomic history of mammoths. *Nature* **591**, 265–269 (2021).
- Klimaet i Grønland. <https://www.dmi.dk/klima/temaforside-klimaet-frem-til-i-dag/klimaet-i-gronland/> (DMI, 2021).
- 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).
- Nguyen, T. H. & Elimelech, M. Plasmid DNA adsorption on silica: kinetics and conformational changes in monovalent and divalent salts. *Biomacromolecules* **8**, 24–32 (2007).
- 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).
- 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).
- 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).
- Karl, D. M. & Bailiff, M. D. The measurement and distribution of dissolved nucleic acids in aquatic environments. *Limnol. Oceanogr.* **34**, 543–558 (1989).
- Pedersen, M. W. et al. Postglacial viability and colonization in North America's ice-free corridor. *Nature* **537**, 45–49 (2016).
- Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, db.prot5448 (2010).
- Murchie, T. J. et al. Optimizing extraction and targeted capture of ancient environmental DNA for reconstructing past environments using the PalaeoChip Arctic-1.0 bait-set. *Quat. Res.* **99**, 305–328 (2021).
- Benoit, G. et al. Multiple comparative metagenomics using multitaxite k-mer counting. Preprint at <https://arxiv.org/abs/1604.02412> (2016).
- Pedersen, M. W. et al. Environmental genomics of Late Pleistocene black bears and giant short-faced bears. *Curr. Biol.* **31**, 2728–2736.e8 (2021).
- Wang, Y. et al. Late Quaternary dynamics of Arctic Biota from ancient environmental genomics. *Nature* **600**, 86–92 (2021).
- 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).
- Raynolds, M. K. et al. A raster version of the Circumpolar Arctic Vegetation Map (CAVM). *Remote Sens. Environ.* **232**, 111297 (2019).
- Bay, C. Floristical and ecological characterization of the polar desert zone of Greenland. *J. Veg. Sci.* **8**, 685–696 (1997).
- Boertmann, D. & Bay, C. *Grønlands Rødliste 2018: Fortegnelse over Grønlandske Dyr og Planter Trusselformer* (Grønlands Naturinstitut, Aarhus Universitet, 2018).
- Böcher, T. W., Holman, K. & Jakobson, K. *Grønlands Flora*, 3rd Edn (Forlaget Haase & Sen, 1978).
- Elven, R., Murray, D. F., Razzhivin, V. Y. & Yurtsev, B. A. *Annotated Checklist of the Panarctic Flora (PAF)* (2011).
- Bay, C. Four decades of new vascular plant records for Greenland. *PhytoKeys* **145**, 63–92 (2020).
- Bay, C. *A Phytogeographical Study of the Vascular Plants of Northern Greenland—North of 74 Northern Latitude*, Vol. 36 (Kommissionen for Videnskabelige Undersøgelser i Grønland, 1992).
- Parducci, L. et al. Ancient plant DNA in lake sediments. *New Phytol.* **214**, 924–942 (2017).
- Alsos, I. G. et al. Plant DNA metabarcoding of lake sediments: How does it represent the contemporary vegetation. *PLoS ONE* **13**, e0195403 (2018).
- 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).
- Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
- Yang, X.-Y. et al. Plastomes of Betulaceae and phylogenetic implications. *J. Syst. Evol.* **57**, 508–518 (2019).
- Bouckaert, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
- 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).
- Graham, R. W. et al. Timing and causes of mid-Holocene mammoth extinction on St Paul Island, Alaska. *Proc. Natl Acad. Sci. USA* **113**, 9310–9314 (2016).
- Parducci, L. et al. Shotgun environmental DNA, pollen, and macrofossil analysis of lateglacial lake sediments from southern Sweden. *Front. Ecol. Evol.* **7**, 189 (2019).

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

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 Alsos¹² & Eric Coissac^{12,18}

Article

Methods

Sampling

Q20

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

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

Absolute age dating

In 2014, Be and Al oxide targets from 8 × 1 kg quartz-rich sand samples collected at modern depths ranging from 3 to 21 m below stream cut terraces were analysed by accelerator mass spectrometry and the cosmogenic isotope concentrations interpreted as maximum ages using a simple burial dating approach¹ (²⁶Al:¹⁰Be versus normalized ¹⁰Be). The ²⁶Al and ¹⁰Be 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 ²⁶Al and ¹⁰Be 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 ²⁶Al:¹⁰Be 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 ²⁶Al 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 ²⁶Al:¹⁰Be is lower, and therefore a calculated burial age assuming the initial production ratio would overestimate the final burial duration. We also consider that once buried, the sand grains may have been exposed to secondary cosmogenic muons (their depth would be too great for submarine nucleonic production). As sedimentation rates in these glaciated near-shore environments are relatively rapid, we show that even the muonic production would be negligible (see Supplemental Information). However, once the marine sediments emerged above sea level, in-situ production by both nucleogenic and muogenic production could alter the ²⁶Al:¹⁰Be. The ²⁶Al versus ¹⁰Be isochron plot reveals this complex burial history (Supplementary Information, section 3) and the concentration versus depth composite profiles for both ²⁶Al and ¹⁰Be reveal that the shallowest samples may have been exposed during a period of time (~15,000 years ago) that is consistent with deglaciation in the area (Supplemental Information). While we interpret the

individual simple burial age of all samples as a maximum limiting age of deposition of the Kap København Formation Member B, we recommend using the three most deeply shielded samples in a single depth profile to minimize the effect of post-depositional production. We then calculate a convolved probability distribution age for these three samples (KK06A, B and C). However, this calculation depends on the ²⁶Al:¹⁰Be 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 σ confidence (Supplementary Information, section 3). Taking the midpoint between the negative and positive 3 σ confidence limits, we obtain a maximum burial age of 2.70 ± 0.46 Myr. This age is also supported by the position of those three samples on the isochron plot, which suggests the true age may not be significantly different that this maximum limiting age.

Thermal age

The extent of thermal degradation of the Kap København DNA was compared to the DNA from the Krestovka Mammoth molar. Published kinetic parameters for DNA degradation⁶⁰ were used to calculate the relative rate difference over a given interval of the long-term temperature record and to quantify the offset from the reference temperature of 10 °C, thus estimating the thermal age in years at 10 °C for each sample (Supplementary Information, section 4). The mean annual air temperature (MAT) for the the Kap København sediment was taken from Funder et al. (2001)⁶ and for the Krestovka Mammoth the MAT was calculated using temperature data from the Cerskij Weather Station (WMO no. 251230) 68.80° N 161.28° E, 32 m from the IRI Data Library (<https://iri.columbia.edu/>) (Supplementary Table 4.4.1).

Q21

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

Q22

Mineralogic composition

The minerals in each of the Kap København sediment samples were identified using X-ray diffraction and their proportions were quantified using Rietveld refinement. The samples were homogenized by grinding ~1 g of sediment with ethanol for 10 min in a McCrone Mill. The samples were dried at 60 °C and added corundum (CR-1, Baikowski) as the internal standard to a final concentration of 20.0 wt%. Diffractograms were collected using a Bruker D8 Advance (θ–θ geometry) and the LynxEye detector (opening 2.71°), with Cu K_{α1,2} radiation (1.54 Å; 40 kV, 40 mA) using a Ni-filter with thickness of 0.2 mm on the diffracted beam and a beam knife set at 3 mm. We scanned from 5–90° 2θ with a step size of 0.1° and a step time of 4 s while the sample was spun at 20 rpm. The opening of the divergence slit was 0.3° and of the antiscatter slit 3°. Primary and secondary Soller slits had an opening of 2.5° and the opening of the detector window was 2.71°. For the Rietveld analysis, we used the Profex interface for the BGMN software^{62,63}. The instrumental parameters and peak broadening were determined by the fundamental parameters ray-tracing procedure⁶⁴. A detailed description of identification of clay minerals can be found in the supporting information.

Adsorption

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

Table 4.2.6. The purity of minerals was checked using X-ray diffraction with the same instrumental parameters and procedures as listed in the above section i.e., mineralogical composition. Notes on the origin, purification and impurities can be found in the supplementary information section 4. We used artificial seawater⁶⁵ and salmon sperm DNA (low molecular weight, lyophilized powder, Sigma Aldrich) as a model for eDNA adsorption. A known amount of mineral powder was mixed with seawater and sonicated in an ultrasonic bath for 15 min. The DNA stock was then added to the suspension to reach a final concentration between 20–800 µg ml⁻¹. The suspensions were equilibrated on a rotary shaker for 4 h. The samples were then centrifuged and the DNA concentration in the supernatant determined with UV spectrometry (Biophotometer, Eppendorf), with both positive and negative controls. All measurements were done in triplicates, and we made five to eight DNA concentrations per mineral. We used Langmuir and Freundlich equations to fit the model to the experimental isotherm and to obtain adsorption capacity of a mineral at a given equilibrium concentration.

Pollen

The pollen samples were extracted using the modified Grischuk protocol adopted in the Geological Institute of the Russian Academy of Science which utilizes sodium pyrophosphate and hydrofluoric acid⁶⁶. Slides prepared from 6 samples were scanned at 400× magnification with a Motic BA 400 compound microscope and photographed using a Moticam 2300 camera. Pollen percentages were calculated as a proportion of the total palynomorphs including the unidentified grains. Only 4 of the 6 samples yielded terrestrial pollen counts ≥50. In these, the total palynomorphs identified ranged from 225 to 71 (mean = 170.25; median = 192.5). Identifications were made using several published keys^{67,68}. The pollen diagram was initially compiled using Tilia version 1.5.12⁶⁹ but replotted for this study using Psimpoll 4.10⁷⁰.

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 DNA⁷¹ and converted to 65 dual-indexed Illumina sequencing libraries (including 13 negative extraction- and library controls)³⁰. 34 libraries were thereafter subjected to ddPCR using a QX200 AutoDG Droplet Digital PCR System (Bio-Rad) following manufacturer's protocol. Assays for ddPCR include a P7 index primer (5'-AGCAGAAGACGGCATAC-3') (900nM), gene-targeting primer (900 nM), and a gene-targeting probe (250nM). We screened for Viridiplantae psbD (primer: 5'-TCATAATTGGACGTTGAACC-3', probe: 5'-(FAM)ACTCCCATCATATGAAA(BHQ1)-3') and Poaceae psbA (primer: 5'-CTCACAACTTCCTCTAGAC-3', probe 5'-(HEX)AGCTGCTGTTGAAGTTC(BHQ1)-3'). Additionally, 34 of the 65 libraries were enriched using targeted capture enrichment, for mammalian mitochondrial DNA using the PaleoChip Arctic1.0 bait-set³¹ 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 simka³² (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 HOLL³³ 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 database³⁴ as well as searched the NCBI SRA for 139 genomes of boreal animal taxa (March 2020) of which 16 partial-full genomes were found and added (Source Data 1, sheet 4) and used the GTDB microbial database version 95 as decoy. All alignments were hereafter merged using samtools and sorted using gz-sort (v. 1). Cytosine deamination frequencies were then estimated using the newly developed metaDMG, by first finding the lowest common ancestor across all possible alignments for each read and then calculating damage patterns for each taxonomic level (<https://metadmg-dev.github.io/metaDMG-core/index.html>) (Supplementary Information, section 6). In parallel, we computed the mean read length as well as number of reads per taxonomic node (Supplementary Information, section 6). Our analysis of the DNA damage across all taxonomic levels pointed to a minimum filter for all samples at all taxonomic levels with a D-max ≥ 25% and a likelihood ratio (λ-LR) ≥ 1.5. This ensured that only taxa showing ancient DNA characteristics were parsed for downstream profiling and analysis and resulted in no taxa within any controls being found (Supplementary Information, section 6).

Marine eukaryotic metagenome

We sought to identify marine eukaryotes by first taxonomically labelling all quality-controlled reads as Eukaryota, Archaea, Bacteria or Virus using Kraken 2⁷² 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 eukaryotes⁷³ that contain 683 metagenome-assembled genomes (MAGs) and 30 single-cell genomes (SAGs) from *Tara Oceans*⁷⁴, following the naming convention in Delmont et al.⁷³, we will refer to them as SMAGs. Reads labelled as root, unclassified, archaea, bacteria and virus were refined through a second Kraken 2 labelling step using a high-resolution database containing archaea, bacteria and virus created by the taxdb-integration workflow. We used the same Kraken 2 parameters and filtering thresholds as the initial search. Both Kraken 2 databases were built with parameters optimized for the study read length (--kmer-len 25 --minimizer-len 23 --minimizer-spaces 4).

Reads labelled as eukaryota, root and unclassified were hereafter mapped with Bowtie2⁷⁵ 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 93% and a breadth of coverage ratio and coverage evenness of less than 0.75 were removed. We followed a data-driven approach to select the mean read ANI threshold, where we explored the variation of mapped reads as a function of the mean read ANI values from 90% to 100% and identified the elbow point in the curve (Supplementary Fig. 6.11.1). We used anvio⁷⁶ 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 checklist⁴² and NCBI. For example, since Bennike (1990)¹⁸, *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*³⁹. Plant families have changed names—for instance, Gramineae is now called Poaceae and Scrophulariaceae has been re-circumscribed to exclude Plantaginaceae and Orobanchaceae⁷⁷. We then classified the taxa into the following: category 1 all identical genera recorded by DNA and macrofossils or pollen, category 2 genera recorded by DNA also found by macrofossils or pollen including genera 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 mitochondria⁷⁸. 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 Genbank⁷⁹, including a single representative sequence from a recently diverged outgroup. For the *Betula* genus, we also included three chloroplast genomes from the PhyloNorway database^{34,80}. We changed all ambiguous bases in the fasta files to N. We used MAFFT⁸¹ to align each of these sets of reference sequences, and inspected multiple sequence alignments in NCBI MSViewer to confirm quality⁸². We trimmed mitochondrial alignments with insufficient quality due to highly variable control regions for Leporidae, Cricetidae and Anserinae by removing the d-loop in MegaX⁸³.

The BEAST suite⁴⁸ 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 BioPython⁸⁴ to create a reference consensus fasta sequence for each set of taxa. Furthermore, we used SNPSites⁸⁵ 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 bwa⁸⁶ 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 samtools⁸⁷. This produced 103,042, 39,306, 91,272, 182 and 129 reads for *Salix*, *Populus*, *Betula*, Elephantidae and Capreolinae, respectively.

We next used pathPhynder⁸⁸, 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 *Mammot*, 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 MAFFT⁸¹ 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. ⁴⁸) and repeated the previously described pathPhynder steps, with the exception that the pathPhynder tree path analysis for the *Mammot* SNPs was based on transitions and transversions, not restricting to only transversions due to low coverage.

Mammot americanum. We confirmed the phylogenetic placement of our sequence using a selection of Elephantidae mitochondrial reference sequences, GTR+G, strict clock, a birth-death substitution model, and ran the MCMC chain for 20,000,000 runs, sampling every 20,000 steps. Convergence was assessed using Tracer⁸⁹ 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 BEAST⁴⁸ v1.10.4. We used two separate approaches when dating our mastodon mitogenome, as demonstrated in a recent publication⁹⁰. 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 determined⁹⁰. We utilized the same BEAST parameters as Karpinski et al.⁹⁰ and set the age of our sample with a gamma distribution (5% quantile: 8.72×10^4 , Median: 1.178×10^6 ; 95% quantile: 5.093×10^6 ; initial value: 74,900; shape: 1; scale: 1,700,000). In short, we specified a substitution model of GTR+G4, a strict clock, constant population size, and ran the Markov Chain Monte Carlo chain for 50,000,000 runs, sampling every 50,000 steps. Convergence of the run was again determined using Tracer.

Molecular dating methods

In this section, we describe molecular dating of the ancient birch (*Betula*) chloroplast genome using BEAST v1.10.4 (ref. ⁴⁸). In principle, the genera *Betula*, *Populus* and *Salix* had both sufficiently high chloroplast genome coverage (with mean depth $24.16 \times$, $57.06 \times$ and

Q24

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

First, we downloaded 27 complete reference *Betula* chloroplast genome sequences and a single *Alnus* chloroplast genome sequence to use as an outgroup from the NCBI Genbank repository, and supplemented this with three *Betula* chloroplast sequences from the Phylo-Norway database generated in a recent study²⁹, 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/miwiipe/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 Mafft⁹¹ and checked the quality of our alignment by eye in Seqotron⁹² and NCBI MsaViewer. Next, we called a consensus sequence from this MSA using the BioAlign consensus function⁹³ 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 metaDMG⁹³, 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 Seqtk⁹⁴. We mapped these read sets against the consensus *Betula* chloroplast genome using BWA⁸⁶ 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 samtools⁸⁶. 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 bcftools⁸⁶ to make an mpileup and call a vcf file, using options for haploidy and disabling the default calling algorithm, which can slightly bias 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 qualities⁹⁵, 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 code⁹⁶,

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 Seqotron⁹² 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 BEAST⁴⁶. 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 BEAST⁴⁶ 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 height⁴⁷; standard deviation was obtained by conservatively converting the 95% HPD to z-scores. For the overall tree prior, we selected the coalescent model. The age of the ancient sequence was estimated following the overall procedures of Shapiro et al. (2011)⁹⁷. To assess sensitivity to prior choice for this unknown date, we used two different priors, namely a gamma distribution metric towards a younger age (shape = 1, scale = 1.7); and a uniform prior on the range (0, 10 Myr). We also compared two different models of rate variation among sites and substitution types within each partition, namely a GTR+G with four rate categories, and base frequencies estimated from the data, and the much simpler Jukes Cantor model, which assumed no variation between substitution types nor sites within each partition. All other priors were set at their defaults. Neither rate model nor prior choice had a qualitative effect on results (Extended Data Fig. 10). We also ran the coding regions alone, since they translated correctly and are therefore highly reliable sites and found that they gave the same median and a much larger confidence interval, as expected when using fewer sites (Extended Data Fig. 10). We ran each Markov chain Monte Carlo for a total of 100 million iterations. After removing a burn-in of the first 10%, we verified convergence in Tracer⁸⁹ 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 TreeAnnotator⁴⁶ had placed the ancient sequence phylogenetically identically to pathPhynder⁸⁸ placement, which is shown in Extended Data Fig. 9. For our major results, we report the uniform ancient age prior, and the GTR+G₄ 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.

Q30

Data availability

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

Code availability

All code used is available at <https://github.com/miwiPe/KapCopenhagen.git>.

60. Lindahl T, & Nyberg B. Rate of depurination of native deoxyribonucleic acid. *Biochemistry* **11**, 3610–3618 (1972).
61. Hansen, J., Sato, M., Russell, G. & Kharecha, P. Climate sensitivity, sea level and atmospheric carbon dioxide. *Philos. Trans. A* **371**, 20120294 (2013).
62. Taut, T., Kleeberg, R. & Bergmann, J. The new Seifert Rietveld program BGMN and its application to quantitative phase analysis. *Mater. Struct.* **5**, 57–66 (1998).
63. Doebelin, N. & Kleeberg, R. Profex: a graphical user interface for the Rietveld refinement program BGMN. *J. Appl. Crystallogr.* **48**, 1573–1580 (2015).
64. Cheary, R. W. & Coelho, A. A fundamental parameters approach to X-ray line-profile fitting. *J. Appl. Crystallogr.* **25**, 109–121 (1992).
65. Kester, D. R., Duedall, I. W., Connors, D. N. & Pytkowicz, R. M. Preparation of artificial seawater. *Limnol. Oceanogr.* **12**, 176–179 (1967).
66. Grichuk, E. D. & Zaklinskaya, V. P. *The Analysis of Fossil Pollen and Spore and Using these Data in Paleogeography* (GeographGIZ Press, 1948).
67. Kupriyanova, L. A. & Aleshina, L. A. *Pollen and Spores of the European USSR Flora* (Nauka, 1972).
68. Moore, P. D., Webb, J. A. & Collinson, M. E. *Pollen Analysis*. (Blackwell Scientific, 1991).
69. Grimm, E. C. *Tilia and Tiliagraph* (Illinois State Museum, 1991).
70. Bennett, K. D. Manual for psimpoll and pscomb. <http://www.chrono.qub.ac.uk/psimpoll/psimpoll.html> (2002).
71. Ardelean, C. F. et al. Evidence of human occupation in Mexico around the Last Glacial Maximum. *Nature* **584**, 87–92 (2020).
72. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
73. Delmont, T. O. et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.10.15.341214> (2021).
74. Karsenti, E. et al. A holistic approach to marine eco-systems biology. *PLoS Biol.* **9**, e1001177 (2011).
75. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
76. Eren, A. M. et al. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2021).
77. The Angiosperm Phylogeny Group. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot. J. Linn. Soc.* **141**, 399–436 (2003).
78. Chevigny, N., Schatz-Daas, D., Lotfi, F. & Gualberto, J. M. DNA repair and the stability of the plant mitochondrial genome. *Int. J. Mol. Sci.* **21**, 328 (2020).
79. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44**, D67–D72 (2016).
80. Alsos, I. G. et al. Last Glacial Maximum environmental conditions at Andøya, northern Norway: evidence for a northern ice-edge ecological 'hotspot'. *Quat. Sci. Rev.* **239**, 106364 (2020).
81. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
82. Yachdav, G. et al. MSASViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics* **32**, 3501–3503 (2016).
83. Kumar, S., Stecher, G., Li, M., Nuryasa, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
84. Cook, P. J. A. et al. BioPython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
85. Page, A. J. et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* **2**, e000056 (2016).
86. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
87. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
88. Martiniano, R., De Sanctis, B., Hallast, P. & Durbin, R. Placing ancient DNA sequences into reference phylogenies. *Mol. Biol. Evol.* **39**, msac017 (2022).
89. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
90. Karpinski, E. et al. American mastodon mitochondrial genomes suggest multiple dispersal events in response to Pleistocene climate oscillations. *Nat. Commun.* **11**, 4048 (2020).

91. Huang, Y., Wang, J., Yang, Y., Fan, C. & Chen, J. Phylogenomic analysis and dynamic evolution of chloroplast genomes in Salicaceae. *Front. Plant Sci.* **8**, 1050 (2017).
92. Fourment, M. & Holmes, E. C. Seqotron: a user-friendly sequence editor for Mac OS X. *BMC Res. Notes* **9**, 106 (2016).
93. Michelsen, C. et al. metaDMG: a fast and accurate ancient DNA damage toolkit for metagenomic data. *Nat. Methods*.
94. Li, H. et al. Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences (2013).
95. Li, H. Improving SNP discovery by base alignment quality. *Bioinformatics* **27**, 1157–1158 (2011).
96. R Core Team. R: A language and environment for statistical computing (R Foundation for Statistical Computing, 2022).
97. Shapiro, B. et al. A Bayesian phylogenetic method to estimate unknown sequence ages. *Mol. Biol. Evol.* **28**, 879–887 (2011).
98. Feyling-Hanssen, R. W. A remarkable foraminiferal assemblage from the Quaternary of northeast Greenland. *Bull. Geol. Soc. Denmark* **38**, 101–107 (1989).
99. Huang, D. I., Hefer, C. A., Kolosova, N., Douglas, C. J. & Cronk, Q. C. B. Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, *Populus balsamifera* and *P. trichocarpa* (Salicaceae). *New Phytol.* **204**, 693–703 (2014).
100. 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).

Q31

Q32

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

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. designed 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.W. M.W.P., T.S.K. and C.S.M. conceived, designed and performed the DNA damage estimates. K.K.S. and S.J., conceived and designed the DNA–mineral aspects of the study, interpreted, and wrote about the DNA–mineral data, and participated in the thermal age calculations. K.H.K., M.W.P., A.H.R., A.R., I.G.A. and E.W. undertook the floristic analysis and interpretations. K.K.K. performed cartography and GIS analysis. I.S. designed and carried out palaeomagnetic analysis and interpreted the results. J.C.G. prepared and analysed eight samples for cosmogenic ²⁶Al and ¹⁰Be 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.Á.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 the contribution to interpretation of marine metagenomic signals. M.J.C. contributed to the thermal age and DNA modelling. M.E.A. contributed to the DNA decay rate estimates. K.H.K., M.W.P., A.H.R. and E.W. interpreted the results and wrote the manuscript with contributions from K.K.S., S.J., A.R., B.D.S., B.D.C., I.G.A., J.C.G., I.S. and N.K.L., with inputs from the other authors.

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>.

B *Explainable ML and Anaemia*

The following 10 pages contain the draft paper:

Christoffer C. Jørgensen, **Christian Michelsen**, Troels C. Petersen, Henrik Kehlet (2022), “*Gender-specific haemoglobin thresholds in relation to preoperative risk assessment in fast-track total hip and knee arthroplasty*”. Unpublished paper draft.

Based on the same data as used on Paper II, the paper uses the SHAP curves to understand the machine learning model. In particular, it compares the preoperative haemoglobin level in the patient with the risk-score for being resubmitted to the hospital within 30 days after the operation, stratified by sex and operation type (knee vs. hip replacement).

Type of article: Science Letter

Submitting author: Christoffer C Jørgensen
Department of Anaesthesia
Hospital of Northern Zealand - Hillerød
Dyrehavevej 29, 3400 Hillerød, Denmark

Gender-specific haemoglobin thresholds in relation to preoperative risk assessment in fast-track total hip and knee arthroplasty.

C. C. Jørgensen,¹ C. Michelsen², T. C. Petersen,³ H. Kehlet⁵

1 Anaesthetist,

Department of Anaesthesia, Hospital of Northern Zealand, Hillerød, Denmark
The Centre for Fast-track Hip and Knee Replacement, Copenhagen, Denmark

2, PhD-student 3, Ass. Professor,

The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

4 Professor,

Section for Surgical Pathophysiology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark

The Centre for Fast-track Hip and Knee Replacement, Copenhagen, Denmark

Correspondence to:

Dr. C.C Jørgensen

Email: christoffer.calov.joergensen@regionh.dk

Short title:

Preoperative anaemia and recovery in fast-track THA and TKA

Within the last decade there has been an increasing focus on anaemia, iron deficiency and transfusion strategies leading to the concept of “patient blood management” (PBM), aiming at reducing the need for blood transfusions by preoperative optimisation of haemoglobin (Hb) and iron-status and use of intra- and postoperative restrictive transfusion protocols [1].

When diagnosing preoperative anaemia most practitioners have adhered to the WHO guidelines which were developed in 1968 and use gender specific criteria of a Hb of $< 130 \text{ g.l}^{-1}$ for men and $< 12 \text{ g.l}^{-1}$ for women [2]. However, these thresholds are based on studies with less sophisticated laboratory and epidemiological techniques than presently available and are consequently under current revision [3].

Furthermore, it has been argued that the WHO definitions of anaemia may not apply to surgical patients, as the relative blood-loss is larger in women, potentially leading to increased risk of allogenic blood transfusions and morbidity when using a gender specific lower preoperative anaemia threshold [4-6].

In total hip (THA) and knee arthroplasty (TKA) it is internationally acknowledged that preoperative iron deficiency anaemia should be corrected by treatment with intravenous (i.v.) iron [7]. However, detailed knowledge of the Hb threshold to increase the risk of postoperative morbidity, indications for treatment and whether it differs in men and women is sparse. The aim of this secondary analysis was to investigate the influence of preoperative Hb level in a comprehensive machine-learning model aimed at identifying patients at “high-risk” of medical complications leading to either a length of hospital stay of >4 days or 30-days readmission after an established fast-track THA and TKA [8]. While the primary study focused on comparing potential benefits of an overall machine-learning model in preoperative risk-prediction [9], this secondary analysis focus specifically on the influence of preoperative Hb level per se and potential differences according to gender and age.

We used a well-defined cohort of elective fast-track THA and TKA patients and evaluated the effect of preoperative Hb-level on the machine-learning model by SHAP-analysis which evaluates the individual effect of the variables included in a machine-learning model [10]. Furthermore, we assessed the distribution of Hb-levels and increases in risk-profile according to gender and age.

From January 2017 to August 2017, we included 3913 patients with a median length of stay of 1 day. Mean preoperative Hb was 154.8 (SD:15.12) but lower in women (149.4 vs. 162 g.l⁻¹; p<0.001) and there were 30.5% of women vs. 12.0% of men with a Hb of <130 g.l⁻¹ (p<0.001). SHAP-analysis demonstrated an immediate steep increase in the risk-score for medical complications with a preoperative Hb below 147.6 g.l⁻¹, and irrespective of gender and age (figure 1). Finally, the median SHAP-value of Hb-level was 0.35 (IQR:) in the patients with a Hb-level below 147.6 g.l⁻¹. These results remained consistent regardless of analysing THA and TKA separately (online Supporting Information Figure S1a+b).

Our analysis demonstrates that in a comprehensive machine-learning risk-model, the preoperative Hb threshold was the same in men and women for an increased risk of prolonged length of stay or readmissions due to medical issues after fast-track THA and TKA. The threshold value of 147.6 g.l⁻¹ is remarkably close to the 130 g.l⁻¹ suggested for men in the current WHO guideline. Thus, the results of our study support the current WHO threshold for anaemia in men, but importantly also for removing gender specific Hb criteria for preoperative anaemia in women, at least in elective THA and TKA. Furthermore, the influence of preoperative Hb level < 147.6 g.l⁻¹ was consistent regardless of age, supporting that the removal of gender specific criteria should apply to all patients. Finally, the effect of Hb level on the accumulated risk-score was clinically meaningful. Thus, figure 1, illustrates that preoperative Hb level contributed with SHAP-values of approximately 0.4 in patients with a Hb of

<147.6 g.l¹. This corresponds with about 50% increased odds of being a high-risk patient. In contrast, in those with Hb-levels >147.6 g.l¹ the odds of being high-risk patients decreased with about 15%.

That gender specific Hb criteria may be inappropriate and need further consideration, has also been demonstrated in cardiac surgery, where women with a preoperative Hb of 120-129 g.l¹ received more blood transfusions and had increased length of hospital stay compared to those with a Hb of >129 g.l¹ [11]. That women with a preoperative Hb level of < 130 g.l¹ may potentially benefit from iron-treatment prior to surgery was illustrated by a large study investigating preoperative Hb levels and iron deficiency in major elective surgery and finding similar incidence of iron deficiency in women with Hb < 130 g.l¹ and < 120 g.l¹ [12]. Our study has some limitations, including lack of information on perioperative blood-transfusions and potential use of preoperative i.v. iron. However, preoperative optimisation with i.v. iron was not standard in the participating departments, and even if some of the outcomes was due to transfusion-related morbidity it would not change the finding of similar SHAP-curves between men and women. Study strengths include well-established fast-track protocols, detailed data on comorbidity and patient outcomes, a complete follow-up, and use of a sophisticated machine-learning model.

In conclusion, from a machine-learning model in fast-track THA and TKA, a Hb threshold of 146.7 g.l¹ was found to increase risk of impaired recovery, regardless of gender or age, thus calling for re-evaluation of preoperative anaemia risk criteria in the elective surgical setting.

Competing Interests

The study was sponsored by a grant from the Novo Nordisk Foundation.

Acknowledgements

The authors would like to acknowledge the members of the Fast-track Hip and Knee Replacement Centre Collaborative group.

Frank Madsen M.D. Consultant, Department of Orthopedics, Aarhus University Hospital, Aarhus, Denmark

Torben B. Hansen M.D., Ph.D., Prof. Department of Orthopedics, Holstebro Hospital, Holstebro, Denmark

Kirill Gromov, M.D., Ph.D., Ass.Prof. Department of Orthopedics Hvidovre Hospital, Hvidovre Denmark

Thomas Jakobsen, M.D., Ph.D., DM.Sci., Ass. Prof. Department of Orthopedics, Aalborg University Hospital, Farsø, Denmark

Claus Varnum, M.D., Ph.D., Ass. Prof. Department of Orthopedic Surgery, Lillebaelt Hospital - Vejle, University Hospital of Southern Denmark, Denmark

Soren Overgaard, M.D., DM.Sci., Prof, Department of Orthopedics, Bispebjerg Hospital, Copenhagen, Denmark

Mikkel Rathnach, M.D., Ph.D., Ass. Prof. Department of Orthopedics, Gentofte Hospital, Gentofte, Denmark

Lars Hansen, M.D., Consultant, Department of Orthopedics, Sydvestjysk Hospital, Grindsted, Denmark

References

1. Goodnough LT, Shander A Patient blood management. *Anesthesiology* 2012; **116**: 1367-76.
2. Nutritional anaemias. Report of a WHO scientific group. *World Health Organ Tech.Rep.Ser.* 1968; **405**: 5-37.
3. Pasricha SR, Colman K, Centeno-Tablante E, Garcia-Casal MN, Peña-Rosas JP Revisiting WHO haemoglobin thresholds to define anaemia in clinical medicine and public health. *Lancet Haematol* 2018; **5**: e60-e2.
4. Munoz M, Gomez-Ramirez S, Kozek-Langeneker S, et al. 'Fit to fly': overcoming barriers to preoperative haemoglobin optimization in surgical patients. *Br.J Anaesth.* 2015; **115**: 15-24.
5. Butcher A, Richards T, Stanworth SJ, Klein AA Diagnostic criteria for pre-operative anaemia-time to end sex discrimination. *Anaesthesia* 2017; **72**: 811-4.
6. Gombotz H, Rehak PH, Shander A, Hofmann A The second Austrian benchmark study for blood use in elective surgery: results and practice change. *Transfusion* 2014; **54**: 2646-57.
7. Gómez-Ramírez S, Maldonado-Ruiz M, Campos-Garrigues A, Herrera A, Muñoz M Short-term perioperative iron in major orthopedic surgery: state of the art. *Vox Sang* 2019; **114**: 3-16.
8. 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.
9. Michelsen C, Jorgensen C, Heltberg M, et al. Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty - a machine-learning based approach. In revision, 2022.
10. Lundberg SM, Erion G, Chen H, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* 2020; **2**: 56-67.
11. Blandszun G, Munting KE, Butchart A, Gerrard C, Klein AA The association between borderline pre-operative anaemia in women and outcomes after cardiac surgery: a cohort study. *Anaesthesia* 2018; **73**: 572-8.
12. Muñoz M, Laso-Morales MJ, Gómez-Ramírez S, Cadellas M, Núñez-Matas MJ, García-Erce JA Pre-operative haemoglobin levels and iron status in a large multicentre cohort of patients undergoing major elective surgery. *Anaesthesia* 2017; **72**: 826-34.

Figure legend:

Figure 1a+b

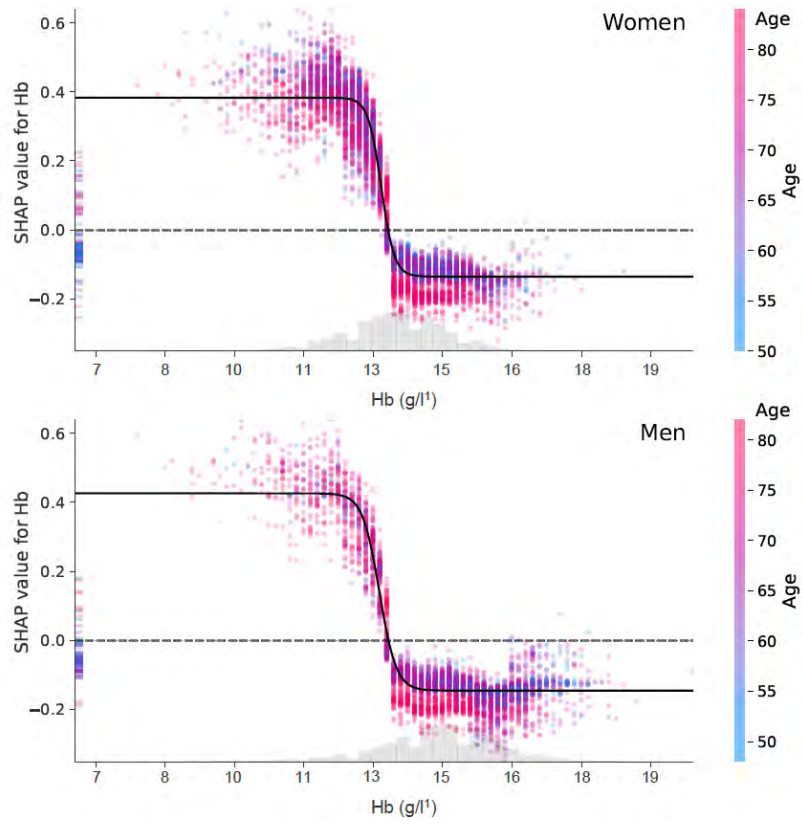
SHapley Additive exPlanations (SHAP) curves for preoperative haemoglobin level in relation to preoperative risk-stratification according to the machine-learning algorithm. Each dot indicates a patient with the colour indicating age (increasing from blue to red). Increasing SHAP values indicate increasing risk-score and decreasing values a decreased risk-score. The cut-off for going from a negative to a positive SHAP-value is indicated by the dotted line at a preoperative Hb level of 147.6 g.l¹

Supplemental material

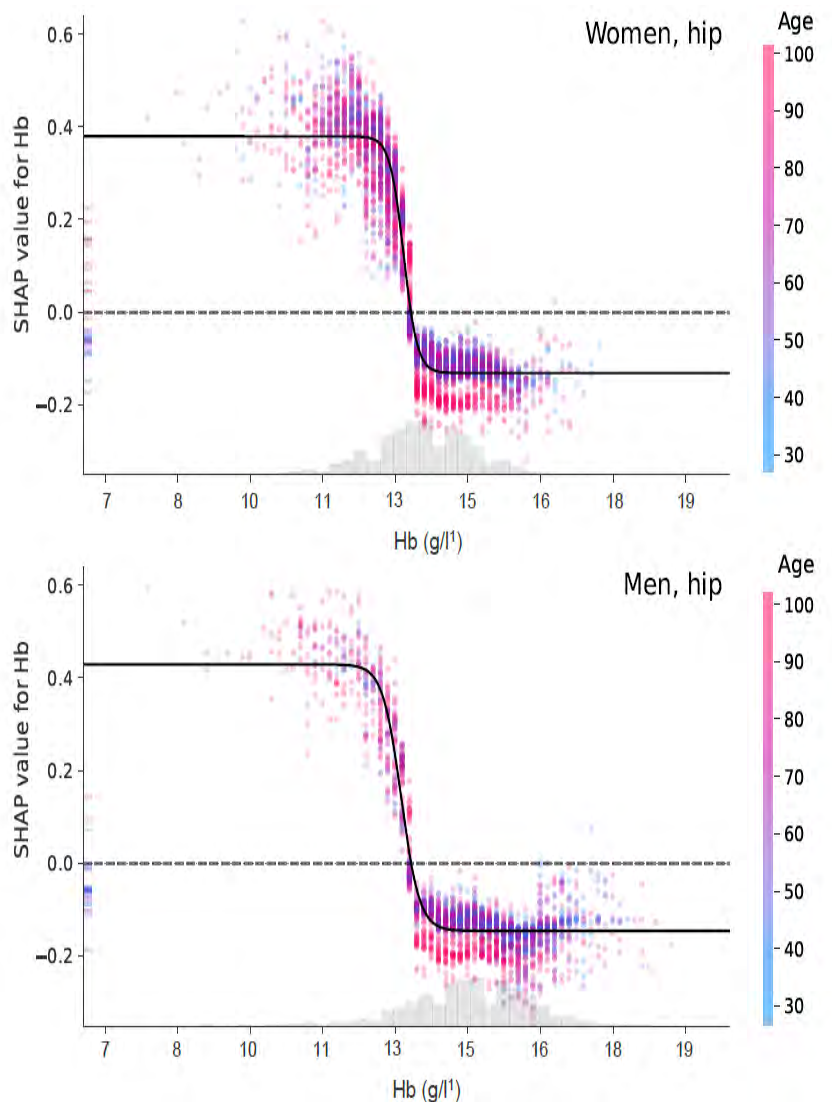
Figure 1a+b

SHapley Additive exPlanations (SHAP) curves for preoperative haemoglobin level in relation to preoperative risk-stratification according to the machine-learning algorithm for total hip (1a) and total knee arthroplasty (1b), respectively. Each dot indicates a patient with the colour indicating age (increasing from blue to red). Increasing SHAP values indicate increasing risk-score and decreasing values a decreased risk-score. The cut-off for going from a negative to a positive SHAP-value is indicated by the dotted line at a preoperative Hb level of 147.6 g.l¹

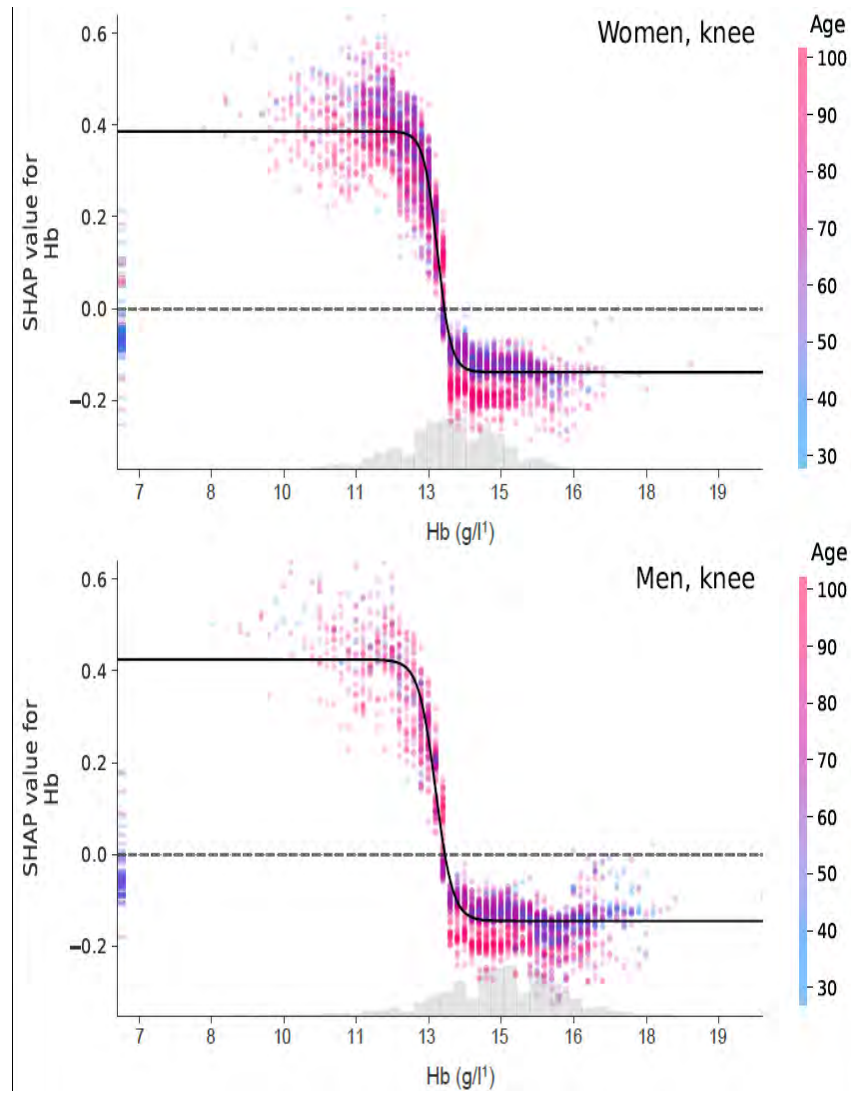
Figure 1 a+b



Supplemental figure 1a



Supplemental figure 1b



C *SSI Ekspertrapport*

The following 28 pages contain the report from Statens Serum Institut:

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.

Ekspertreport af den 10. december 2020

Effekten af kontaktopsporing



Indhold

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



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 belyses efterhånden som yderligere data frembringes. Hvor der ikke er specifikke eller komplette data, vil en agentbaseret model have unøjagtigheder eller risikere at være baseret på antagelser, som ikke nødvendigvis er retvisende. I modellerne anvendes der endvidere ens ventetidsfordelinger for alle agenter, selvom der i realiteten kan være lokale udsving i ventetider.

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

Der er i rapporten heller ikke taget højde for den stigende brug af private antigen test. Coronaopsporingen under STPS foretager også opsporing af nære kontakter, for primærttilfælde 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 primærttilfælde. Gevinsten i form af en reduktion i kontakttallet er således større, såfremt der sikres effektiv opsporing for samtlige primærttilfælde, relativt til reduktionen i kontakttallet, som kan opnås ved at nedbringe ventetiden til test og testsvar for primærttilfældet.

Ventetiden til test og testsvar for et primærttilfælde 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 primærttilfælde 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 primærttilfældet, som er forudsætningen for at opsporingen af nære kontakter kan initieres.

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



2. Formål og baggrund

2.1 Formål og baggrund for modelgruppen

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

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

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

2.2 Formål med rapporten

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

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



3. Opsporing og håndtering af nære kontakter i Danmark

3.1 Forudsætninger for en effektiv kontaktopsporing

Den vigtigste forudsætning for, at kontaktopsporing er et effektivt redskab til at nedbringe smitten med COVID-19 er, at der til hver en tid identificeres flest mulige smittede personer, som der derved kan udføres smitteopsporing for. Jo lavere mørketallet er, desto flere vil kunne smitteopspores. Det er derfor afgørende, at der sikres nem og hurtig adgang til test, først og fremmest for personer med COVID-19 lignende symptomer, men også for øvrige personer, der kunne have mistanke om at være smittet med COVID-19. Den Nationale Prævalensundersøgelse i Danmark har vist, at op mod 40-50% af dem, som havde antistoffer mod SARS-CoV-2 i blodet, ingen erindring havde om at have haft COVID-19 lignende sygdom¹. Ved at udbyde adgang til test for flest mulige personer, vil man også finde flere asymptomatiske smittebærere.

3.2 Definition af en nær kontakt

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

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

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

Definitionen af en nær kontakt er beskrevet i Sundhedsstyrelsens rapport om smitteopsporing af nære kontakter².

En nær kontakt er defineret som en af følgende personer:

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

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

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



3.3 Periode for smitteopsporing

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

3.4 Opsporing af nære kontakter

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

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

Nære kontakter opsporet af Coronaopsporingen

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

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

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

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

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

Nære kontakter der ikke opspores af Coronaopsporingen

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



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

I oktober måned blev der i alt testet 1.091.966 personer. Heraf havde 62% (n = 675.623) bestilt tid på coronaprøver.dk. Blandt disse svarede 58% (n = 391.146) på spørgeskemaet på coronaprøver.dk, hvoraf 25% (n = 99.389) anførte, at de blev testet fordi, de var nær kontakt til en smittet (herunder personer som er adviseret af Smitte|Stop app). Kun 13% (n = 12.706) af dem som svarede, at de blev testet fordi de var nær kontakt til en smittet, var testet på én af de rekvisitionskoder, som der anvendes i Coronaopsporingen (Tabel 1). Samlet set blev 45.616 personer testet på én af de rekvisitionskoder som anvendes i Coronaopsporingen i oktober måned, hvor test-positivprocenten var ca. 4%. Til sammenligning var positivprocenten for de personer, der svarede, at de var nær kontakt til en smittet på Coronaprøver.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		
	N	Testpositive (1. test)	
		n	%
Testede personer	1.091.966	14.723	1,35
Total antal tests rekvireret via Coronaopsporingen	45.616	1.941	4,26
Bestilt på coronaprøver.dk	675.623	10.335	1,53
Svaret på spørgeskema	391.146	5.387	1,38
Ja, nær kontakt til smittet (herunder adviseret på Smitte Stop app)	99.389	2.544	2,56
Rekvireret test via Coronaopsporingen	12.706	524	4,12



4. Agentbaserede modeller

4.1 Om agentbaserede modeller

I indeværende rapport er resultaterne for effekten af kontaktopsporing frembragt fra to forskellige agentbaserede modeller, som er udviklet på henholdsvis Danmarks Tekniske Universitet (DTU) og Niels Bohr 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 infektiøs (I), hvorefter agenten ikke længere er smitsom og betragtes som rask (R). De gennemsnitlige tider i hvert sygdomsstadie kan findes i bilag 1 og 2.

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

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

4.2 Forbehold

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



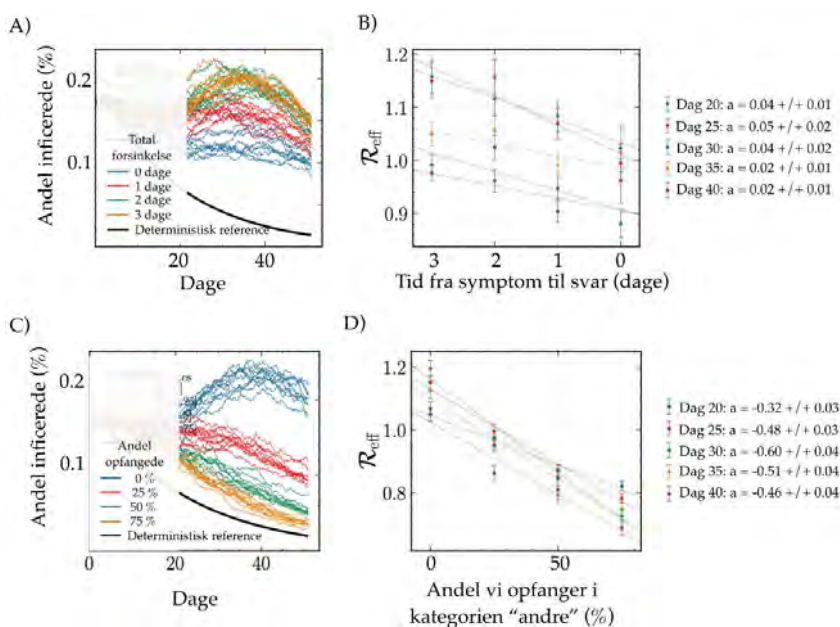
5. Resultater

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

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

Ventetiden fra 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)).



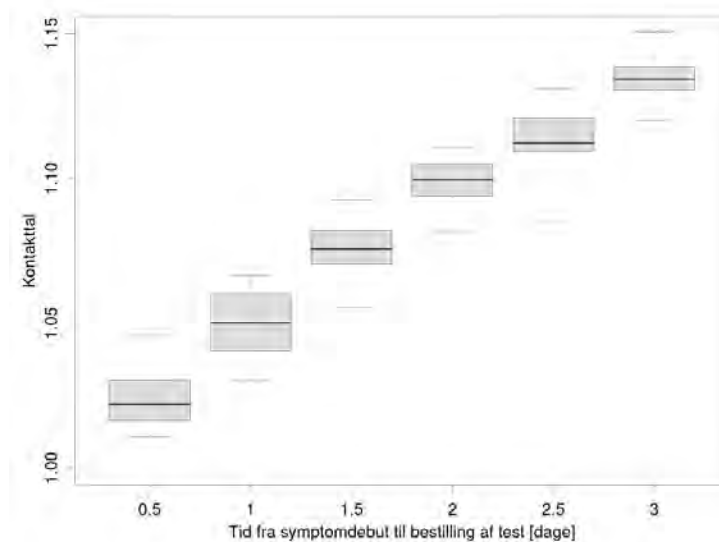


Figur 1: A) Simuleret model, hvor hver kørsel (markeret med samme farve) gentages 10 gange for forskellige værdier af tiden fra symptom til svar. B) Værdien af kontakttallet estimeret på forskellige tidspunkter i simulationen vist i A). Den lineære sammenhæng giver en værdi for hvor mange procent kontakttallet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

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

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

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

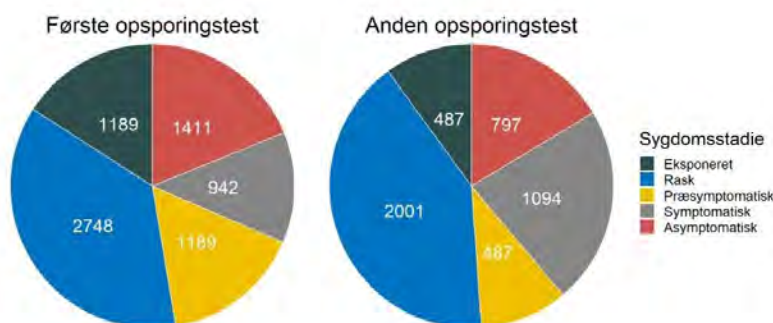


Figur 2: Kontakttallets afhængighed af den gennemsnitlige tid fra at primærtillfældet har symptomdebut til der bestilles en test og personen går i en grad af isolation. For hver parameter værdi er der foretaget 40 simulationer, og boxplottet viser median, de indre kvartiler samt minimum og maksimum af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



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

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



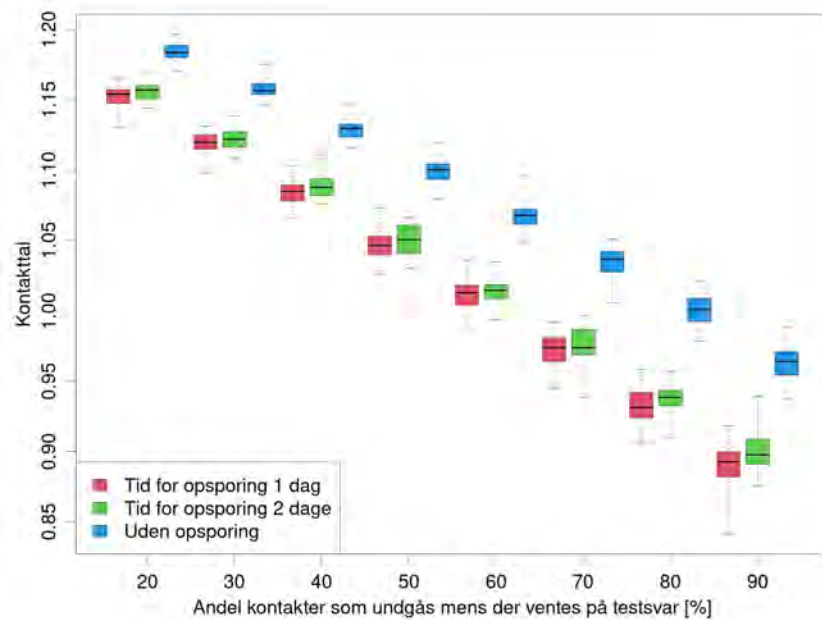
Figur 3. Antal opsporede og smittede i hvert sygdomsstadie, når de får foretaget hhv. første og anden test i opsporingsprocessen. Der er flere, som ikke kommer til anden test, bl.a. fordi de tester positiv i første test eller efter negativt testsvar vælger ikke at få taget den opfølgende test. Derudover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

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

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



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



Figur 4. Kontakttallets afhængighed af andelen af kontakter et primærttilfælde og opsporede nære kontakter reducerer, i ventetiden fra der bestilles en test til at testsvar foreligger, samt betydningen af ventetiden til at en nær kontakt opspores og går i tilsvarende isolation. For hver parameter værdi 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., Wiggert, 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>

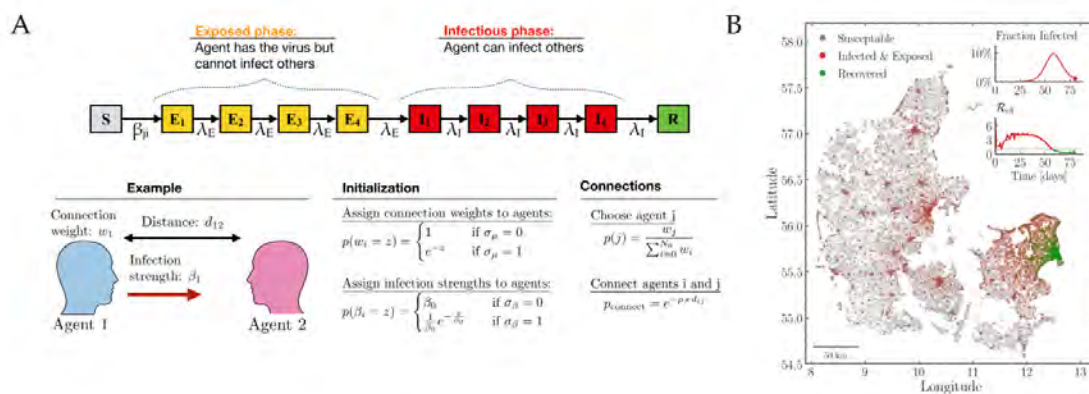
Plank, M., James, A., Lustig, A., Steyn, N., Binny, R. & Hendy, S. (2020). Potential reduction in transmission of COVID-19 by digital contact tracing systems. *MedRxiv*. Lokaliseret: <https://doi.org/10.1101/2020.08.27.20068346>



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 flokkimmunitet i København.

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

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

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



Tabel 2: Parametre i modellen.

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



Bilag 2. Beskrivelse af den agentbaserede model fra DTU

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

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

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

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

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

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

³ FAM122N: <https://www.statistikbanken.dk/FAM122N>

FAM133N: <https://www.statistikbanken.dk/FAM133N>

FAM55N: <https://www.statistikbanken.dk/FAM55N>

PEND100: <https://www.statistikbanken.dk/PEND100>

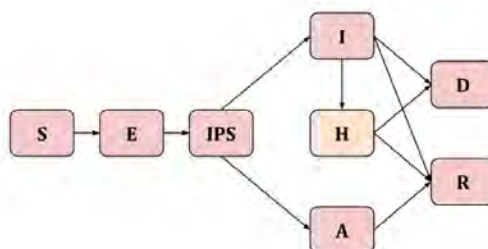
ERHV6: <https://www.statistikbanken.dk/ERHV6>

UVM (Normering grundskoler): <https://uddannelsesstatistik.dk/Pages/Reports/1577.aspx>

UVM (Normering gymnasier): <https://uddannelsesstatistik.dk/Pages/Reports/1851.aspx>

UVM (Normering erhvervsuddannelse): <https://uddannelsesstatistik.dk/Pages/Reports/1850.aspx>

Daginstitutioner: <https://www.institutioner.dk/>



Figur 6. Flowdynamisk diagram af bevægelse gennem sygdomsstadier.

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

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

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

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



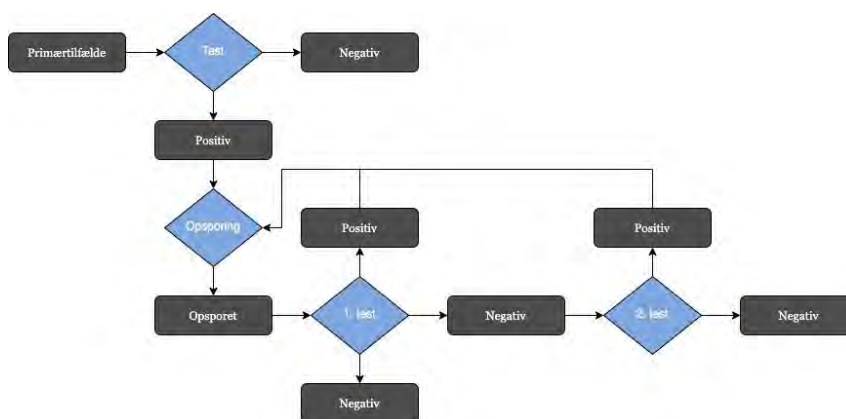
Hospitaliseret (H) Under 60 år	2	3	2	3	5	Linelisten SSI
Hospitaliseret (H) 60 år og derover	2	5	3	5	7	Linelisten SSI
Ventetider						
timeSymp-ToOrderTest	5	1	1	1	1	Antagelser - afventer STPS data
timeOrderTo-Test	2	2	1	2	3	Antagelser - afventer STPS data
timeTestToResult	6	1,5	2	2	2	Ventetider fra samfundssporet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

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

- Når en agent får symptomer, er der en sandsynlighed ($p_{\text{TestGivenSymptoms}} = 80\%$) for, at de bestiller en test efter en gammafordelt ventetid ($\text{timeSympToOrderTest}$). 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 test-tider relativt til 48 timer før den positive fik symptomer - eller blev testet i et asymptomatisk tilfælde. Hvis muligt, gives test på dag 4 og dag 6, ellers dag 5 og 7, og ellers én test hurtigt muligt.
 - Personer, som er i et igangværende opsporingsforløb, får kun tildelt test, hvis de venter på mindre end to testsvar.
 - Den opsporede person har samme ventetider på testsvar, som symptomatiske personer.



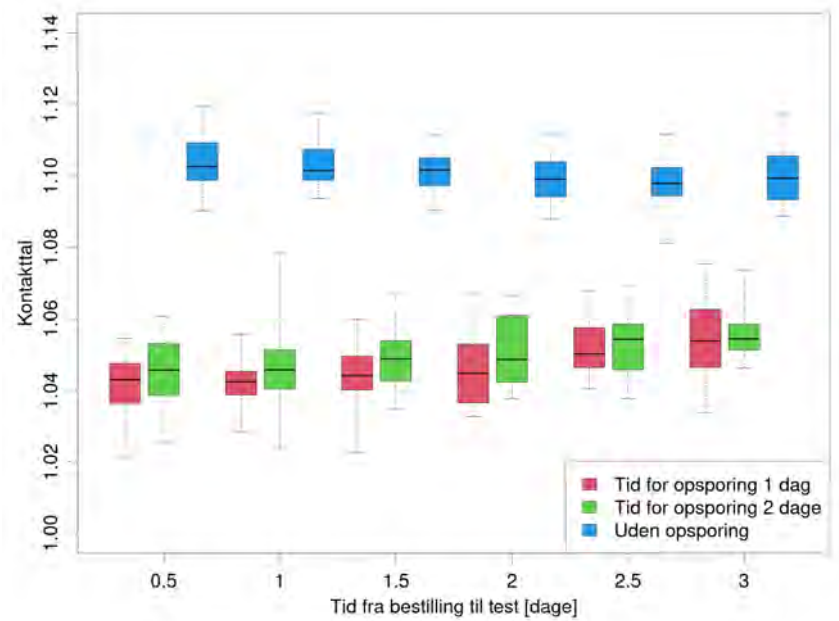
- o Mens der ventes på test og testsvar, isoleres den opsporede person på samme måde som en symptomatisk, der venter på svar.
- o Hvis en opsporet person får negativt svar på den første test, vil der være en sandsynlighed for ($p_{\text{NoShow2ndTest}} = 40\%$) at de ikke tager test nummer 2.
- o Efter et negativt svar på test nummer 1, vil isolationen brydes. Hvis der fås et positivt svar, inden test nummer 2 er taget, annulleres test nummer 2, og personens egne netværk opspores.
- For alle tests – om det er en opsporet person eller ej – antages der en sandsynlighed på 20% for en falsk negativ test (Kucirka et al., 2020).



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

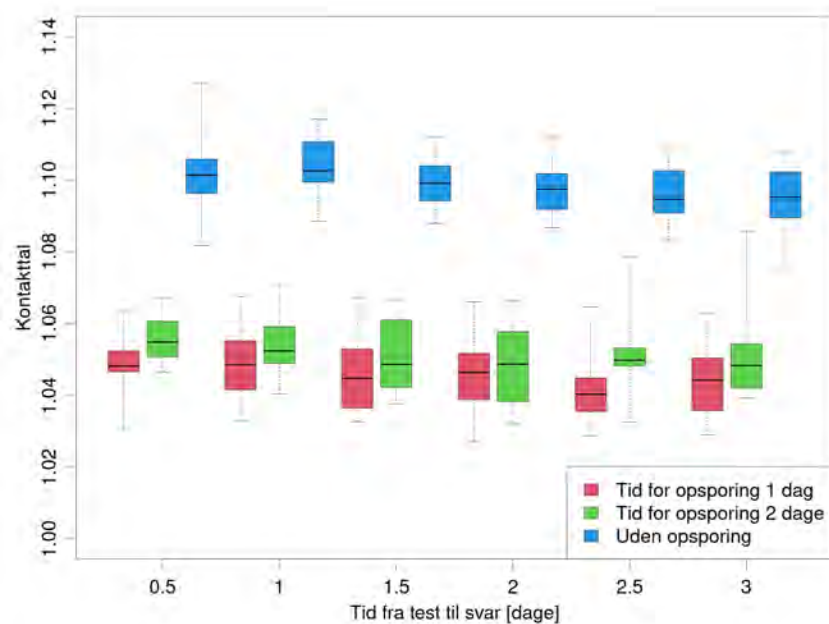


Yderligere resultater



Figur 8. Kontakttallets afhængighed af ventetiden på at få taget en test hos primærtillæddet, samt betydningen af hvor lang tid der går før nære kontakter opspores og går i tilsvarende isolation. For hver parameter værdi 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 primærtillæddet testes, i forhold til at reducere ventetiden til opsporing af nære kontakter.



Figur 9. Kontakttallets afhængighed af tiden fra der testes til at der foreligger et testsvar, samt betydningen af hvor lang tid der går indtil nære kontakter opspores og går i tilsvarende isolation. For hver parameter værdi 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 primært tilfældet og opsporede kontakter testes til der foreligger et testsvar, i forhold til at reducere ventetiden til opsporing af nære kontakter. En årsag kan være, at ventetiden til testsvar gør, at en masse opsporede og modtagelige kontakter er isoleret i længere tid og derfor ikke bliver smittet. Det er ikke undersøgt om dette kun ses når kontakttallet er nær 1.



Bilag 3. Regneeksempel

Følgende er et illustrativt regneeksempel på den agentbaserede model fra Niels Bohr 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 smittede 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 primærttilfældet. Denne person kan enten findes tilfældigt, eller ved at primærttilfældet testes, og at sekundærttilfældet opspores efter en tidsperiode (d for delay). Denne ventetid, er tiden fra at primærttilfældet testes til at sekundærttilfældet kontaktes, og afspejler således både ventetid til test samt ventetid til opsporing. Igen antages det, at sekundærttilfældet går i isolation øjeblikkeligt. Ved igen at trække tal tilfældigt fra de relevante fordelinger fås en eksponeringsperiode, hvori sekundærttilfældet 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 smittedetilfælde, findes via kontaktopsporing, og derved reduceres smitten, idet eksponeringsperioden for opsporede 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 for korte eksponeringsperioden for kontakter.

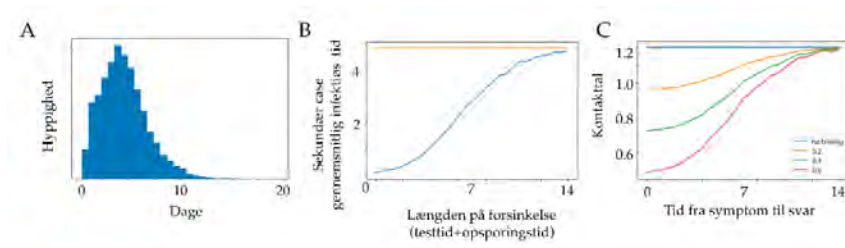
I modellen indgår 4 mulige eksponeringsperioder. 1: Kontakter opspores ikke, hvorved eksponeringsperioden ikke afkortes (blå graf), 2: 20% af kontakter opspores (gul graf), 3: 40% af kontakter opspores (grøn graf) og 4): hvis 80% af kontakter opspores (rød graf).

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

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



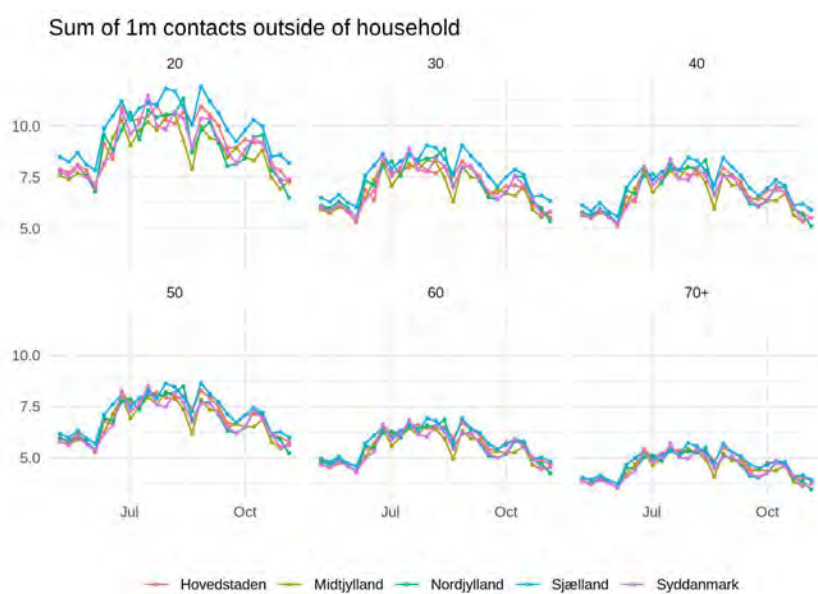
Det skal i øvrigt bemærkes, at det i eksemplet antages, at opsporede kontakter går i isolation, indtil de får svar på deres test.



Figur 10: A) Fordeling af eksponeringstiden, gennemsnit = 4.9 dage. B) Gennemsnitlig eksponeringstid for sekundære tilfælde (blå), som funktion af den samlede ventetid til test og opsporing. Den orange graf viser gennemsnittet i ventetiden til test og opsporing for 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



Bilag 5. Beskrivelse af parametre brugt i rapporten

Modellerne i rapporten bygger på en række parametre. Estimerne, som parametre er baseret på er udvalgt af den relevante institution, der har udarbejdet modellerne. Begrundelsen for valg af estimerne 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. Estimer over befolkningens adfærd i forbindelse med covid-19 bygger på en række danske undersøgelser fra i år, samt på data over danskernes rejsemønstre.

Estimer for latensperiode, inkubationsperiode og infektiøs periode fra litteraturen:

Særligt relevant for simuleringerne over effekten af kontaktopsporing er estimerne bag sygdomsforløbet, herunder hvor lang tid der går fra eksponering til, at vedkommende kan smitte, og derefter til, at vedkommende vises symptomer. Estimerne 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 infektiøse periode*, hvor det bedste estimat er 5 dage, mens det bedste interval er mellem 3-7 dage, har Ekspertgruppen gennemgået følgende artikler:

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

HOPE rapporter og data:

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

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

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

Trafik data:

Antagelser om befolkningens adfærd bygges ligeledes på trafikdata, hvorudfra man kan bestemme danskernes rejsemønstre. Efter aftale med Trafik-, Bygge- og Boligstyrelsen får Ekspertgruppen løbende adgang anonymiserede data over danskernes bevægelse rundt i landet. Data er bl.a. brugt til at bestemme den typiske afstand mellem kontakter og afstanden mellem 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 Lufthavn
- Overblik over biltrafikken på Statsvejsnettet og cykeltrafikken (samlet ud fra tællestationer) 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 hustanden var smittet og lignende.

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

Den nationale prævalensundersøgelse for covid-19:

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



Bilag 6. Medlemmer af ekspertgruppen

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

Danmarks Tekniske Universitet, Institut for Matematik og Computer Science

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

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

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

Københavns Universitet, Institut for Folkesundhedsvidenskab

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

Københavns Universitet, Niels Bohr Institutet

- 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

D *SSI Notat*

The following 9 pages contain the report from Statens Serum Institut:

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 estimater for den forventede udbredelse af cluster B.1.1.7 i den kommende periode, dels ved logistisk regression af udviklingen i forekomsten af varianten, og dels ud fra simuleringer af spredningen af varianten i en agentbaseret model.

Sammenfatning

- Den observerede udvikling i forekomsten af cluster B.1.1.7 i Danmark, svarer til en ugentlig vækstrate for forholdet mellem cluster B.1.1.7 og de øvrige virusvarianter på 72% (95% CI: [37, 115] %).
- Med udgangspunkt i den aktuelle situation hvor 2,3% af virusvarianterne i den rutinemæssige helgenomsekventering tilhører cluster B.1.1.7, estimeres det, at varianten vil udgøre halvdelen af de cirkulerende virusstammer i Danmark om 40-50 dage såfremt ovennævnte stigning fortsætter.
- Det nuværende niveau af restriktioner forventes ikke at være tilstrækkeligt til at få kontakttallet for cluster B.1.1.7 under 1. Derfor vil denne vokse eksponentielt påagt 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 tryktest, og i samme periode er der udført et stigende antal antigen test. Derimod så vi i juledagene, at kun ganske få har ladet sig teste. Disse ændringer i testdynamikker gør det svært at følge udviklingen i covid-19, idet de vanlige indikatorer såsom incidenser, positivprocenter og kontakttallet påvirkes af den ændrede fordeling af covid-19-positive blandt de testede. Et lignende mønster forventes i dagene op til og efter nytår. Desuden har vi endnu ikke set effekten af de sidst indførte tiltag, herunder lukning af detailhandlen og liberale erhverv. Samlet set giver dette usikkerhed omkring det aktuelle kontakttal. Analysen er baseret på 76 isolater med cluster B.1.1.7 fordelt på de fem regioner. Den lille stikprøve giver relativt store statistiske usikkerheder. Der vil derfor være behov for at løbende 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. Estimer 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



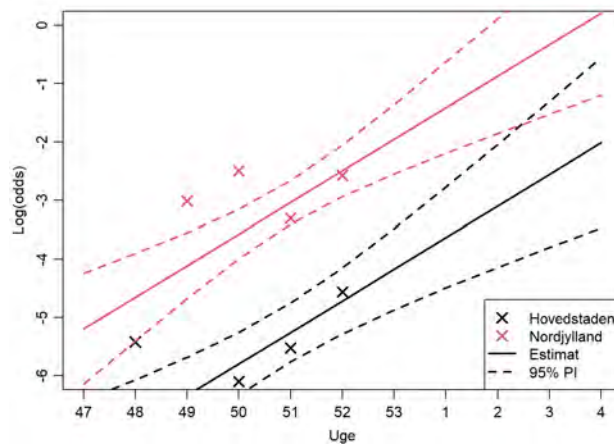
Det ses, at log(odds) for at detektere cluster B.1.1.7 er 2.2 højere i Region Nordjylland end i Region Hovedstaden. Det svarer til odds på 9.2. Det mest interessante er den tidlige udvikling, hvoraf det ses at log(odds) øges med 0.54 for hver uge. Dette svarer til at cluster B.1.1.7 har en ugentlig vækstrate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

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

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

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

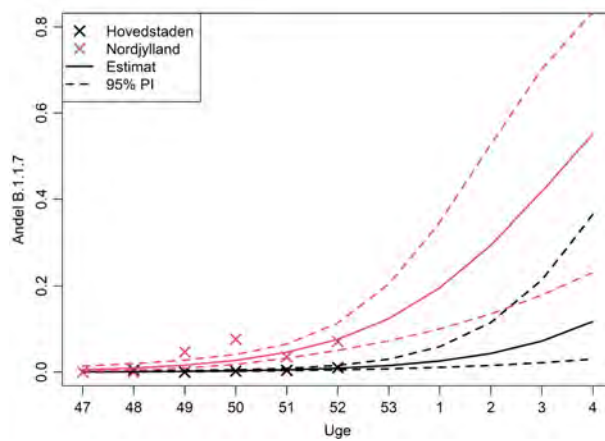
Figur 1 viser en fremskrivning af log(odds) for B.1.1.7 mod andre virusvarianter baseret på ovenstående logistiske regression. Estimateret er, at cluster B.1.1.7 allerede i uge 4 vil udgøre halvdelen af alle positive test i Region Nordjylland. Dette er dog behæftet med stor usikkerhed på baggrund af 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 Hovedstaden¹

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



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



Figur 2. Udviklingen i forekomsten af cluster B.1.1.7 i de kommende uger. Fremskrivningen viser, at halvdelen 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 udtale sig om niveauet i de andre tre regioner, men særlig Region Sjælland synes at have oplevet en hurtig stigning, om end det er baseret på meget lidt data. De næste par uger vil forbedre estimatet af niveauet i alle regioner.

Hen over julen har der været et nyt toppunkt i antal indlagte og der er endnu kun set små fald. Det er først i uge 1, at vi kan forvente at se eventuelle indlæggelser som følge af smitte i julen. Alt andet lige må dette forventes at give en yderligere kortvarig pukkel i antal nye indlæggelser.

På nuværende tidspunkt er prognosen, at vi har omkring en måned før det samlede kontakttal for alle virusvarianter hurtigt vil stige på grund af øget udbredelse af cluster B.1.1.7. Hvis restriktionerne skærpes i den kommende tid, vil det give en reduktion i smittetal og indlæggelser og dermed et lavere udgangspunkt for den forøgede smitte og stigende kontakttal, som vi må forvente.

Et første estimat af kontakttallet for cluster B.1.1.7 for perioden uge 47 til 52 og baseret på observationer fra Region Hovedstaden og Region Nordjylland er 1.5 (95% CI [1,2 ; 1,7]) - dette er estimeret vha. Poisson regression med offset lig med $0.7 \cdot \log(\text{antal sekventerede})$. 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 har. Hver agent repræsenterer således en person, som er knyttet til en lokation i Danmark, svarende til deres bopæl. Agenterne indgår i flere forskellige netværk, f.eks. husstand, job og skole hvor de har kontakt til andre personer. Derudover har de kontakt til tilfældige personer i samfundet i den tid, hvor personen ikke er hjemme, på job eller i skole. Hvis en agent bliver smittet med SARS-CoV-2, er forløbet for den enkelte agent beskrevet således, at agenten først er eksponeret (E) og derefter infektiøs (I), hvorefter agenten ikke længere er smitsom og betragtes som rask (R). De gennemsnitlige tider i hvert sygdomsstadie kan findes i bilag 1. Hver kontakt, som en agent eksponeres for, tildeles en sandsynlighed for at blive smittet af en anden agent, såfremt denne er smitsom. For en detaljeret beskrivelse af den agentbaserede model, herunder de inkluderede parametre, henvises til bilag 1.

Forbehold



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

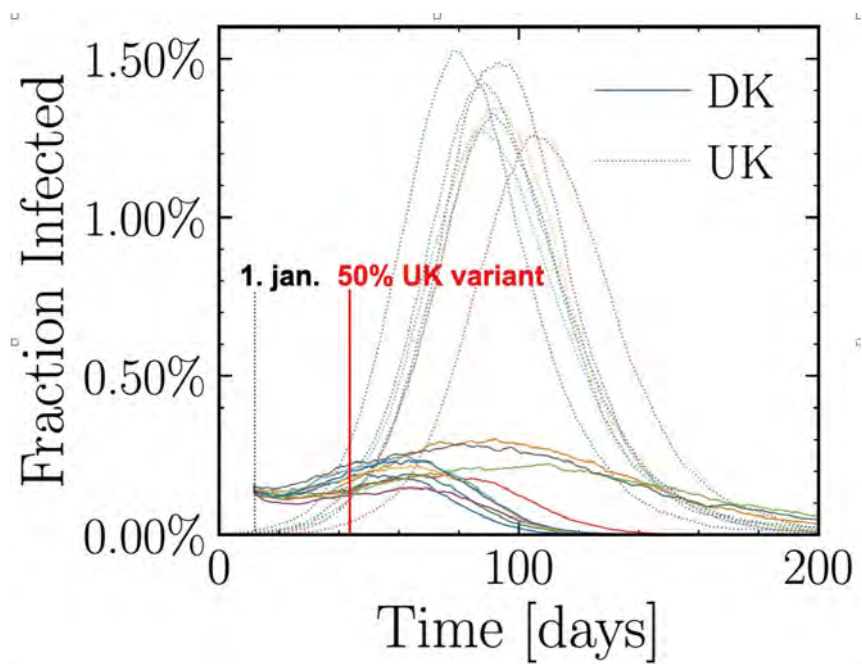
Resultater

I det følgende er udviklingen simuleret i en model, hvor udgangspunktet er 1/10 af Danmarks befolkning, og hvor cluster B.1.1.7 fra starten udgør omkring 5% af de cirkulerende virusvarianter. Epidemien simuleres ud fra et kontakttal på omkring 1,0, samt en antagelse om, at cluster B.1.1.7 smitter 50% mere, som rapporteret fra England²

Figur 3 viser, hvordan en epidemi vil udvikle sig i tid, forudsat at det simulerede scenarie ikke ændres. Der opdeles i hhv. de nuværende virusvarianter (DK, fulde linjer) og det engelske cluster B.1.1.7 (UK, stiplede linjer). Simulationen er gentaget flere gange (forskellige farver) for at se, hvor store variationer der forekommer. Som det kan ses, så udfases DK-versionen af smitten, mens UK-versionen B.1.1.7 giver ophav til en eksponentiel vækst, idet kontakttallet for denne er væsentligt over 1.

Af figuren fremgår det, at cluster B.1.1.7 ca. 35-40 dage fra simulationens start ("1. jan.") udgør omkring 50% af de cirkulerende virusvarianter. Da simulationen er startet med en større andel UK-varianter (5%) end det aktuelle landsgennemsnit (2.3%), så bliver estimeret 40-50 dage til at halvdelen af de sekventerede varianter tilhører cluster B.1.1.7. I de viste simulationer er de første smittede 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.

²2020_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.

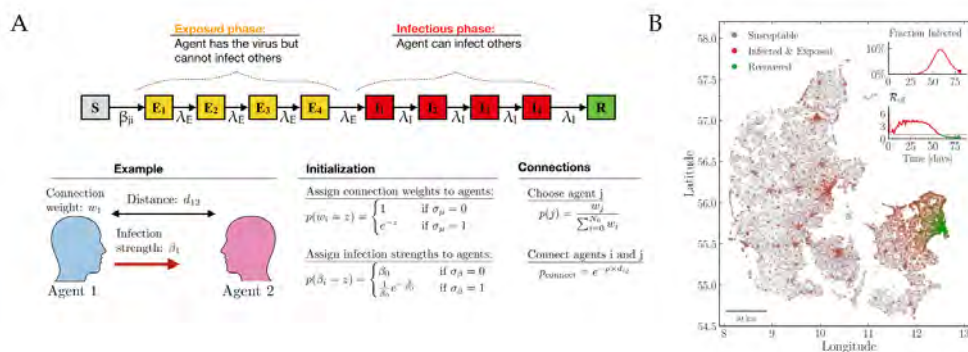


Bilag 1. Beskrivelse af den agentbaserede model

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 afstandsuaafhængige kontakter (%)	3-5	Trafik data
Tid fra symptom til test (Dage)	0-2	Fordeling fra spørgeskemaundersøgelse i foråret 2020 (ikke offentliggjort)
Sandsynlighed for at få symptomer og blive testet (%)	20-60 %	Prævalensundersøgelsen
Sandsynlighed for at kontakte husstand (%)	100%	Antagelse
Sandsynlighed for at kontakte kollegaer (%)	40-80	Antagelse
Sandsynlighed for at kontakte andre (%)	0-75	Antagelse

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