

Shotgun barcode baiting: capturing barcoding genes from environmental samples for species identification

Kevin Nota¹, Ludovic Orlando², Alexis Marchesini³, Matteo Girardi³, Cristiano Vernesi^{3,#}, Laura Parducci^{1,4,#}

1)Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden, 2) Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, CNRS UMR 5288, Université de Toulouse, France, 3) Section for Geogenetics, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark, 4) Research and Innovation Centre, Fondazione Edmund Mach, S. Michele all'Adige (TN), Italy, and 5) Department of Environmental. # joint senior authorship

Introduction

(background, aims)

Four order bait-set

(specificity and sensitivity)

Ancient DNA damage

(in-silico)

Universal bait design

Taxonomic resolution

(in-silico)

Take home messages

(and work in progress)

Introduction

Pros and **cons** for metabarcoding and shotgun sequencing are more or **less counter poles**. Target capture is a method to **enrich** shotgun libraries for specific **targets**. This method has shown to work in **capturing** several targets such as whole mitochondrial genomes from **ancient** environmental samples. Capturing barcoding genes retains some of the pros of metabarcoding by focusing sequencing efforts on taxonomically informative markers.

Our study is aimed at addressing the following four main questions:

- How to efficiently design baits without introducing biases?
- What is the taxonomic resolution of capturing short fragments of barcoding genes?
- Are ancient DNA damage patterns detectable?
- Are there any new biases introduced during the capture process such as GC% bias?

We present here the first results we obtained in answering these questions.

Metabarcoding

Pros

- Robust and well established
- Highly scalable (high number sample low cost)
- Low sequence depth
- Targeted
- Overall good amount of reference sequences

Cons

- Primer biases (amplification, binding, etc.)
- In some cases low taxonomic resolution
- Ancient DNA damage is not retained
- Only one target at the time
- Amplicon size bias

Shotgun sequencing

Cons

- Thus far little used
- Not easily scalable (low number sample high cost)
- High sequence depth
- Untargeted
- Overall low amount of reference sequences

Pros

- No primer bias
- Might increase taxonomic resolution
- Ancient DNA damage retained!
- Untargeted
- Even fragments of 30 bases can be sequenced

Shotgun barcode baiting

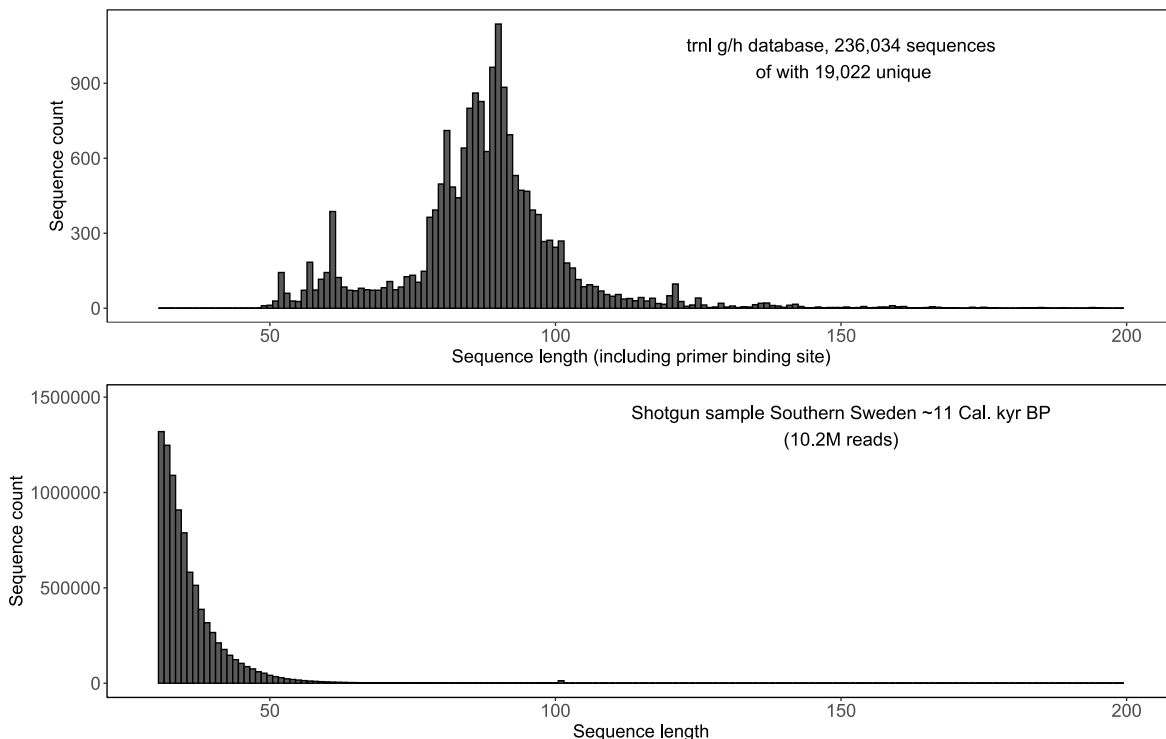
Cons

- Thus far little used
- Not easily scalable (low number sample high cost)
- Low sequence depth
- Targeted
- Overall good amount of reference sequences

Pros

- No primer bias
- Might increase taxonomic resolution
- Ancient DNA damage retained!
- Targeted multi loci at the same time
- Even fragments of 30 bases can be sequenced

Introduction (II)



We aimed to **produce a bait set** for plants as universal as possible. Because there is little information about designing baits for a short gene region for a large number of species, we started designing a bait set that includes four key plant orders as a proof of concept of designing universal baits, investigating biases in wet-lab experiments, and explore taxonomic resolution. The selected orders are Asterales, Fagales, Poales, and Pinales. Which combined represent ~16% of the described vascular plant species.

The p6 loop of the chloroplast gene trnL has been the most commonly used metabarcoding target for characterising flora in ancient environmental samples. This powerful locus is amplified with the g/h primer set (Taberlet *et al.*, 2007). Due to a large amount of size variation, the taxonomic resolution is good, considering the small size of the amplicon. However, in the figure above, a histogram of trnL sequences including 39 nucleotide primer binding regions and the barplot below provides the size of fragments typically obtained with shotgun sequencing. This highlights that the majority of DNA in ancient samples are unable to be utilised by metabarcoding, due to the need for intact primer binding sides on top of the insert.



Asterales
Herbaceous daisy
like flowers



Fagales
Common Trees
such as Oak

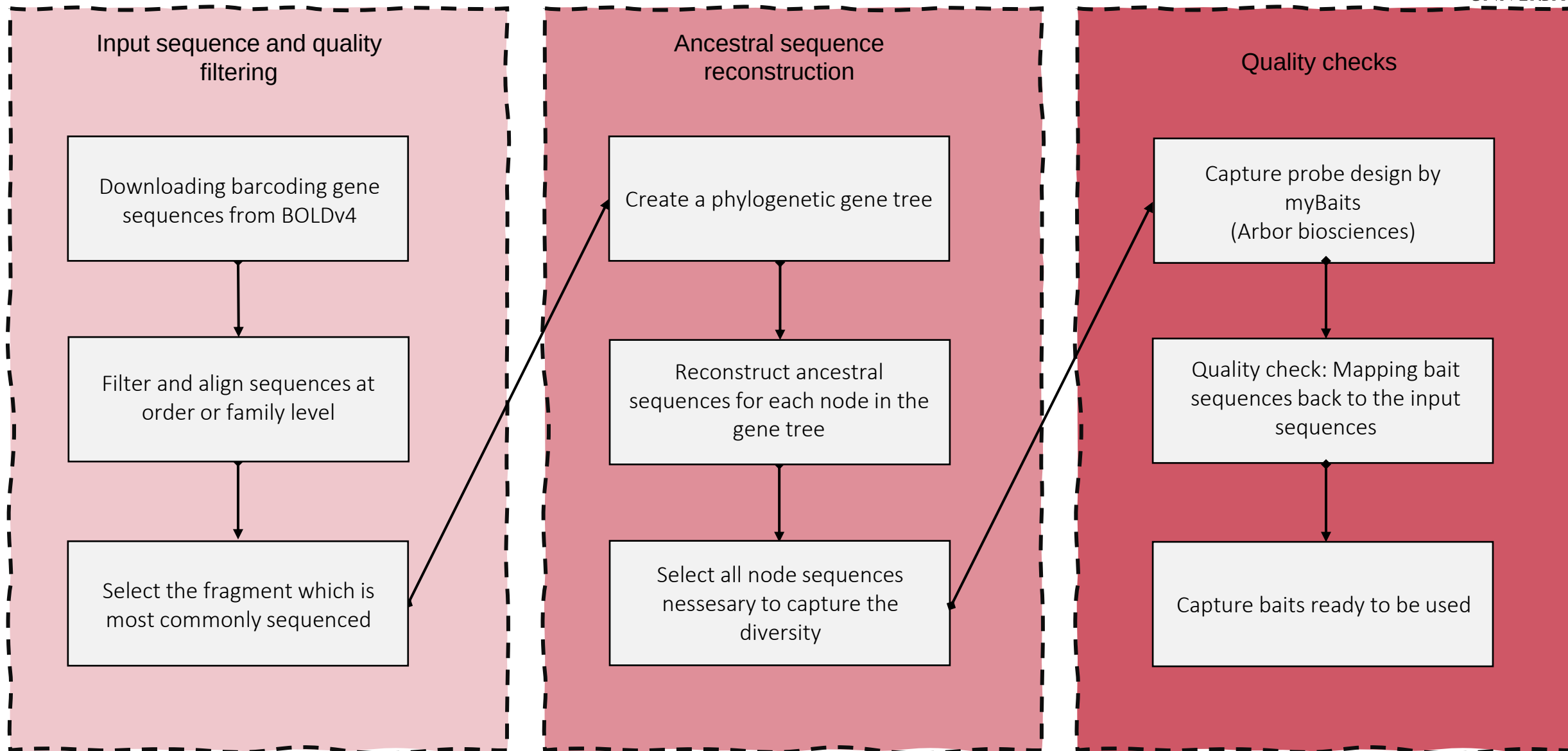


Pinales
Conifer trees
such as Pine



Poales
Grasses and sedges

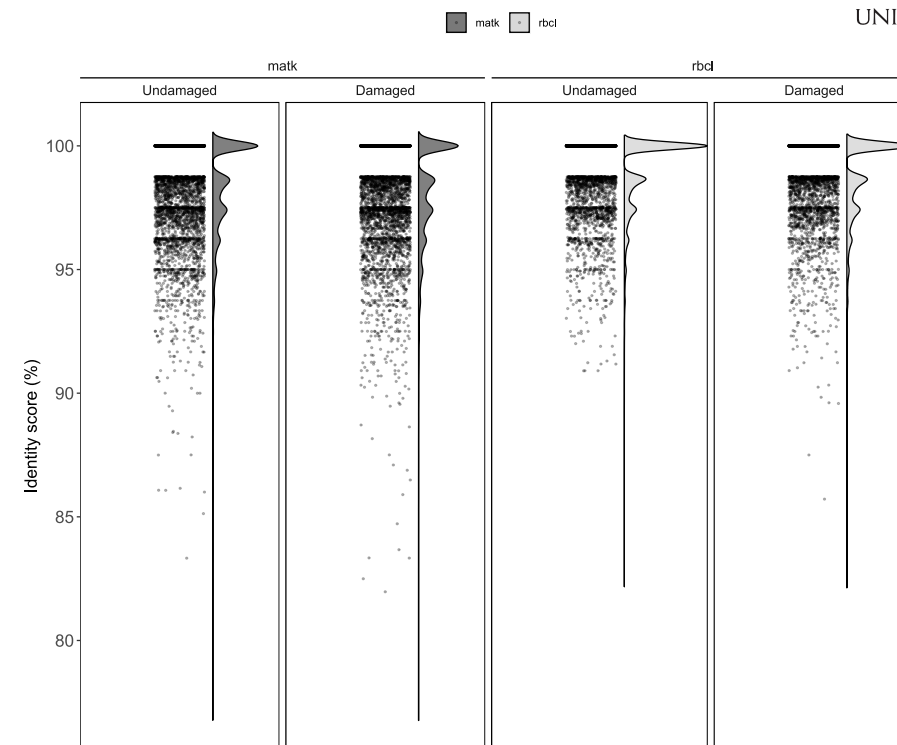
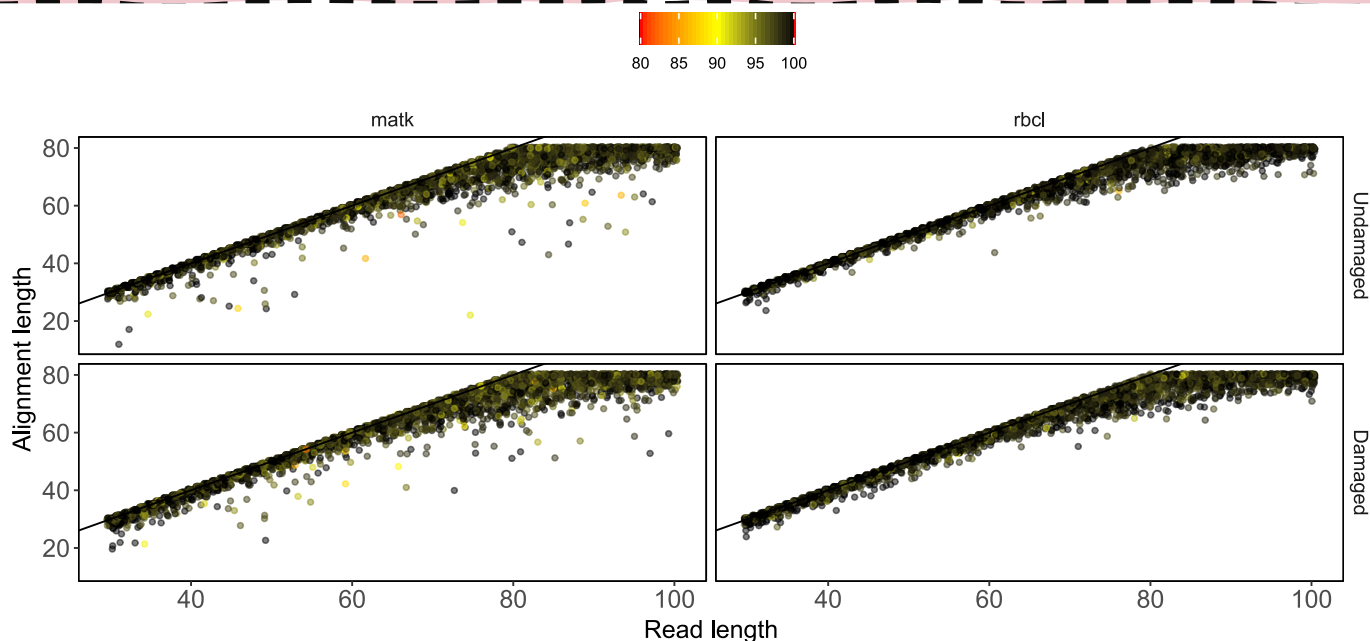
Universal bait design



The Bait set

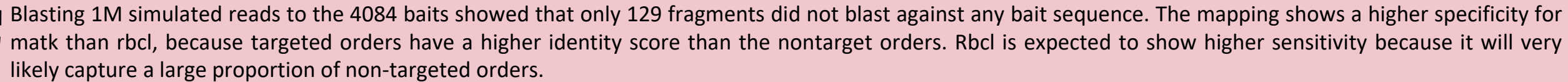
In total, our four order capture set contains 4084 baits with a length of 80 nucleotide targeting ~500 and ~650-1300 nucleotides of the *rbcl* and *matk* gene respectively. The baits are designed based on [~18,000 *rbcl*](#) and [~13,000 *matk*](#) reference sequences and contain ~7000 unique taxa.

To test the effectiveness of the baits, we simulated 1M fragments of 30-100 nucleotides using ~250,000 reference sequences downloaded from v4BOLD (these include all available *rbcl* and *matk* sequences that matched the fragment for which our bait set is designed). Of the simulated fragments 184,549 belonged to the four targeted orders, only 7 fragments did not blast against any bait (0.00379%). The figure below shows 15,000 randomly selected reads with read and alignment length plotted on the axes. The colour indicates the identity score. The plot shows that the majority of the reads map to the full length of the baits, and that there is no difference in mapping between damaged and undamaged reads.



The plot above shows the identity score of all the simulated reads from the four targeted orders. Only 0.43% of the fragments showed an identity score below 91%. Overall, *matk* shows more taxonomic resolution and slightly lower identity scores. There is no difference between the mapping of undamaged and damaged reads to the baits.

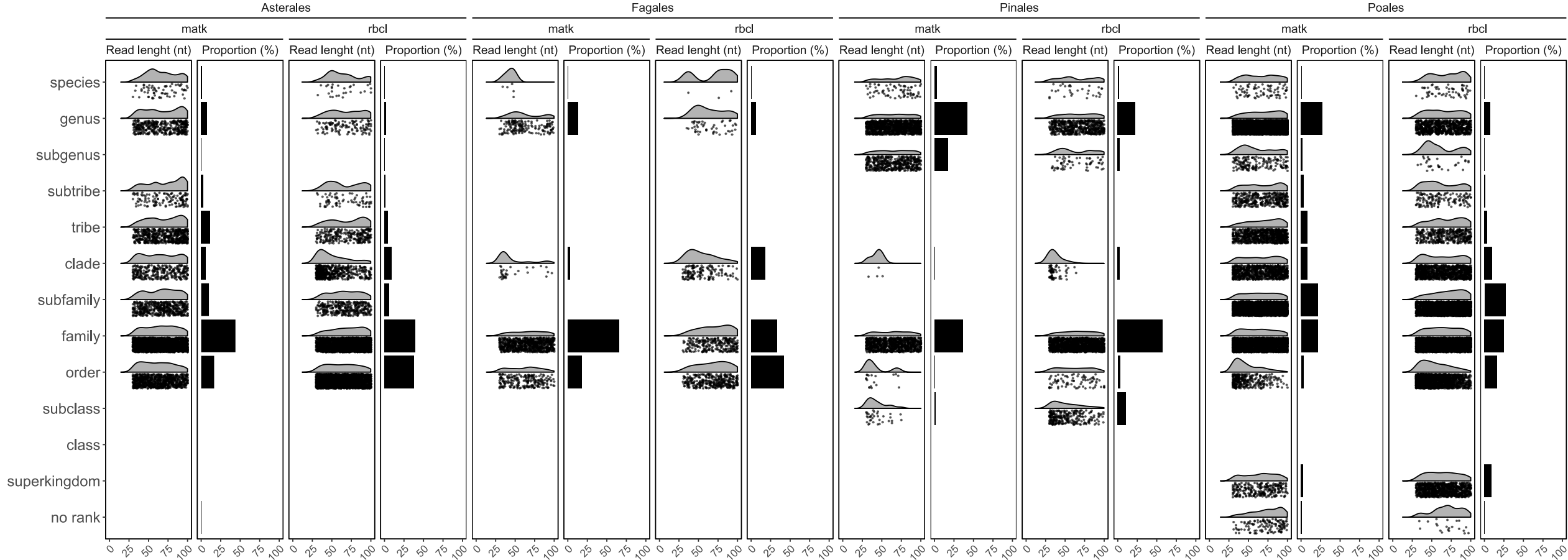
These results show that only 4084 baits can in theory capture the vast majority of sequences belonging to the four targeted orders.

UPPSALA
UNIVERSITET

What taxonomic resolution can you expect?

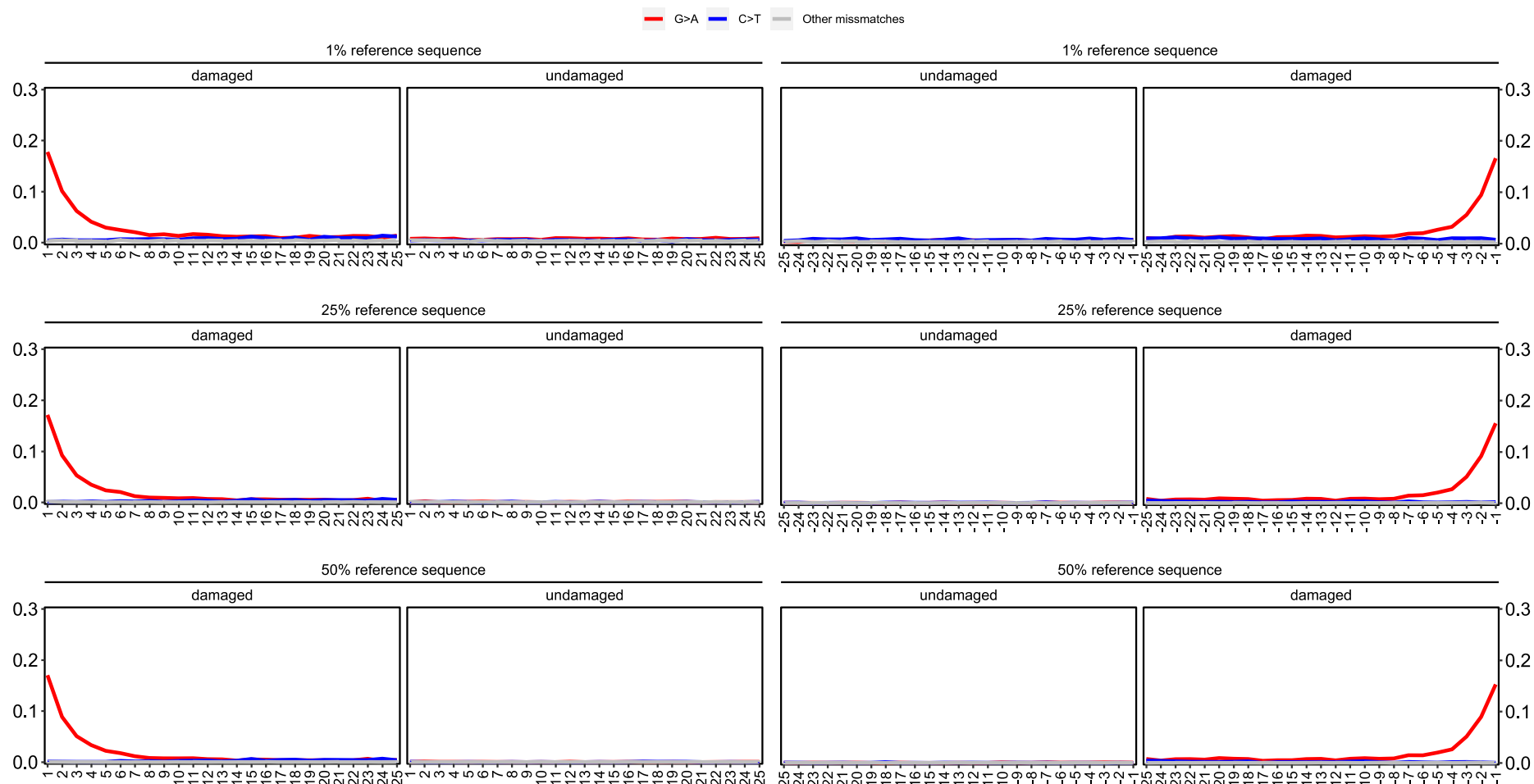


UPPSALA
UNIVERSITET



A simulated sequence data file was created containing 40,000 reads with 30 to 100 nucleotides. The taxonomic level to which fragments are assigned in MEGAN6 is plotted in the figure above. The read length columns show the taxonomic level plotted against the read length of the fragment. The proportion column shows the proportion of reads assigned to a taxonomic level. In most cases, reads are assigned to order or family level, but reads, even 30 nucleotides long, can map to genus or even species level. The databases used for assigning has 12,923 matk and 17,936 rbcL sequences without correcting for local floras. Therefore, the taxonomic resolution presented here is minimal and it will only improve when local flora is considered. There is also a visible difference between the orders, a higher proportion of reads map to genus level in the Pinales and Poales orders. In general matK shows a better taxonomic resolution than rbcL.

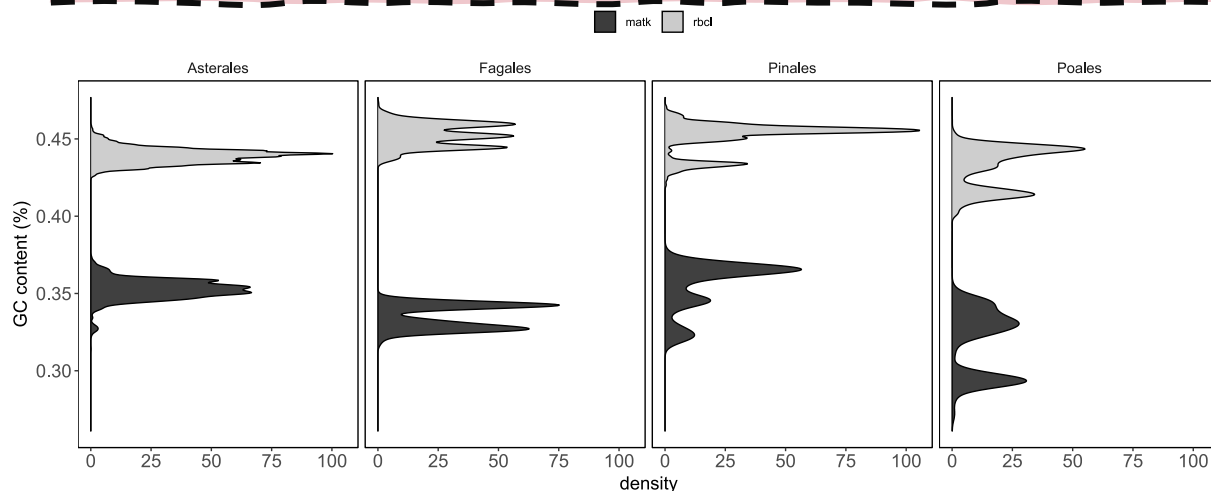
Ancient DNA damage patterns



Retaining ancient DNA damage is one of the main advantages of using a target capture approach. The 40,000 sequences that were simulated for the taxonomic resolution were duplicated but with the incorporation of ancient DNA damage. We mapped the simulated reads with *bwa-mem* to reference sequences. To make sure the damage patterns are also observed when there are gaps in the database used for mapping, we randomly reduced the reference database to 50% (15,434 sequences), 25% (7701 sequences) and 1% (304 sequences) of the original sequences. The number of reads was kept the same. MapDamage2 was used to detect the ancient DNA damage pattern. Even when only 1% of the original reference sequences is used for mapping, the DNA damage patterns are retained. This is likely because a large proportion of the fragments have a low taxonomic resolution, and will therefore always map even if there are only a few reference sequences available. However, we expect that larger numbers of reads will be necessary to recover genuine misincorporation profiles when post-mortem damage is more limited.

In progress

- Wet lab experiments to test whether GC-content introduces a capture bias. The figure below shows the GC content for both targeted markers and grouped by order. GC content is relatively variable. Overall, matk shows the lowest GC content. Species with low GC content might be less efficiently captured.
- We used so far only true matk and rbcL sequences to assign taxonomy, however, there is a possibility that short nucleotide fragments map to bacteria, fungi or nuclear DNA from other organisms, which might in real setting introduce false positives. We will, therefore, investigate potential false positives.
- Test the bait design method for other chloroplast barcode regions such as trnL, and barcodes for other groups of such as COI for mammals. And making a universal python package for bait design



Take home message

- We created a universal pipeline to reduce redundancy of homologues barcoding genes and create the least number of sequences necessary to capture the full diversity of targeted organism.
- We show that the taxonomic resolution highly depends on the sequenced fragment. Even short 30 base fragments can be identified to species in the best case. A large proportion of fragments, however, are not informative beyond family level, but these reads are important for the recovery of ancient DNA damage patterns. This, also might be a key indicator for creating confidence in presence of rare species, because we expect a decreasing number of reads mapping to lower taxonomic levels.
- We presented in-silico support that ancient DNA patterns are retained even when reference databases for mapping the reads are incomplete, which is one of the most important of features this method in *sedaDNA* context.

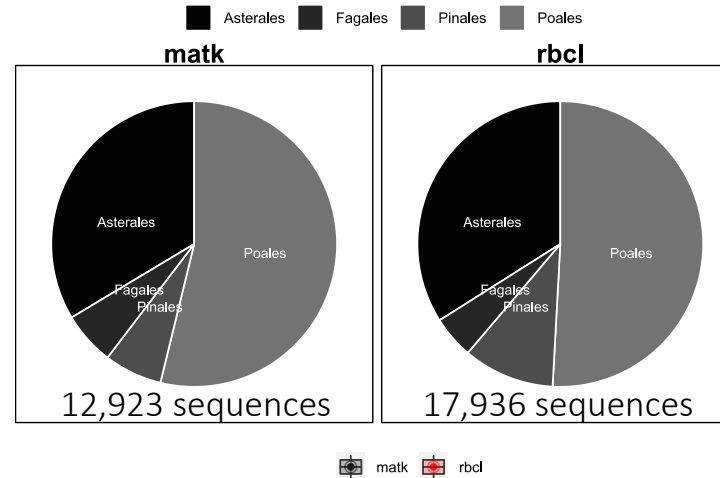
For more information and questions contact me at
kevin.nota@ebc.uu.se

Database used for bait design and taxonomic assessment



UPPSALA
UNIVERSITET

The **boxplots** show the families within the four orders for which sequences are available. The points show the sum of species within genera. The large crossed circles indicate the total number of species in the family.



The two **pie charts** show the proportions of sequences belonging to the four orders. Poales and Asterales are the most species-rich orders.

