Insect Systematics and Diversity, 7(3), 5; 2023, 1–9
https://doi.org/10.1093/isd/ixad011
Research

ENTOMOLOGICAL SOCIETY OF AMERICA
SHARING INSECT SCIENCE GLOBALLY

OXFORD

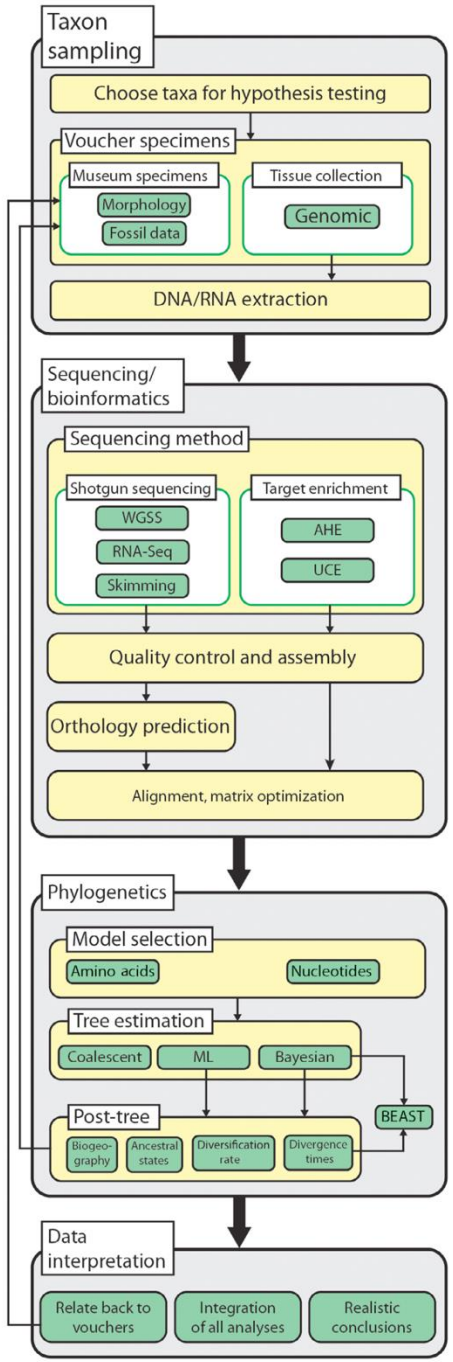
Molecular phylogenetics, phylogenomics, and phylogeography

Assessment of targeted enrichment locus capture across time and museums using odonate specimens

Aaron Goodman[1,2,*,†], Ethan Tolman[1,2,†], Rhema Uche-Dike[1,2,†], John Abbott[3], Jesse W. Breinholt[4,5], Seth Bybee[6], Paul B. Frandsen[7], J. Stephen Gosnell[2,8], Rob Guralnick[5], Vincent J. Kalkman[9], Manpreet Kohli[1,8], Judicael Fomekong Lontchi[6], Pungki Lupiyaningdyah[6,10], Lacie Newton[1], Jessica L. Ware[1]

Objectives for this study

- 1. Assess the effectiveness of targeted enrichment techniques in capturing genomic data from Odonata specimens preserved in museums
- 2. Evaluate the impact of specimen age and preservation methods on the success of locus capture and DNA quality
- 3. Determine whether high-quality genomic data can be generated from both recent and very old museum specimens, including those over 100 years old
- 4. Explore the potential for using historical museum specimens in phylogenetic and evolutionary studies, particularly for studying Odonata



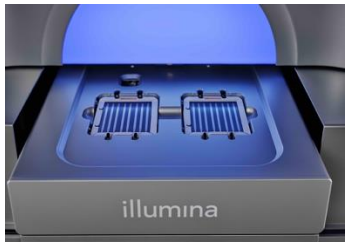
Museum specimen

94 specimen
64 Dragonflies
30 Zygoptera

48 AMNH
46 RMNH

DNA extraction

DNA Quantification

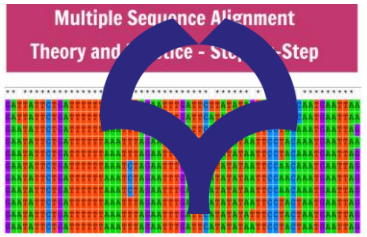


AHE illumina sequencing (Bybee et al.2021)

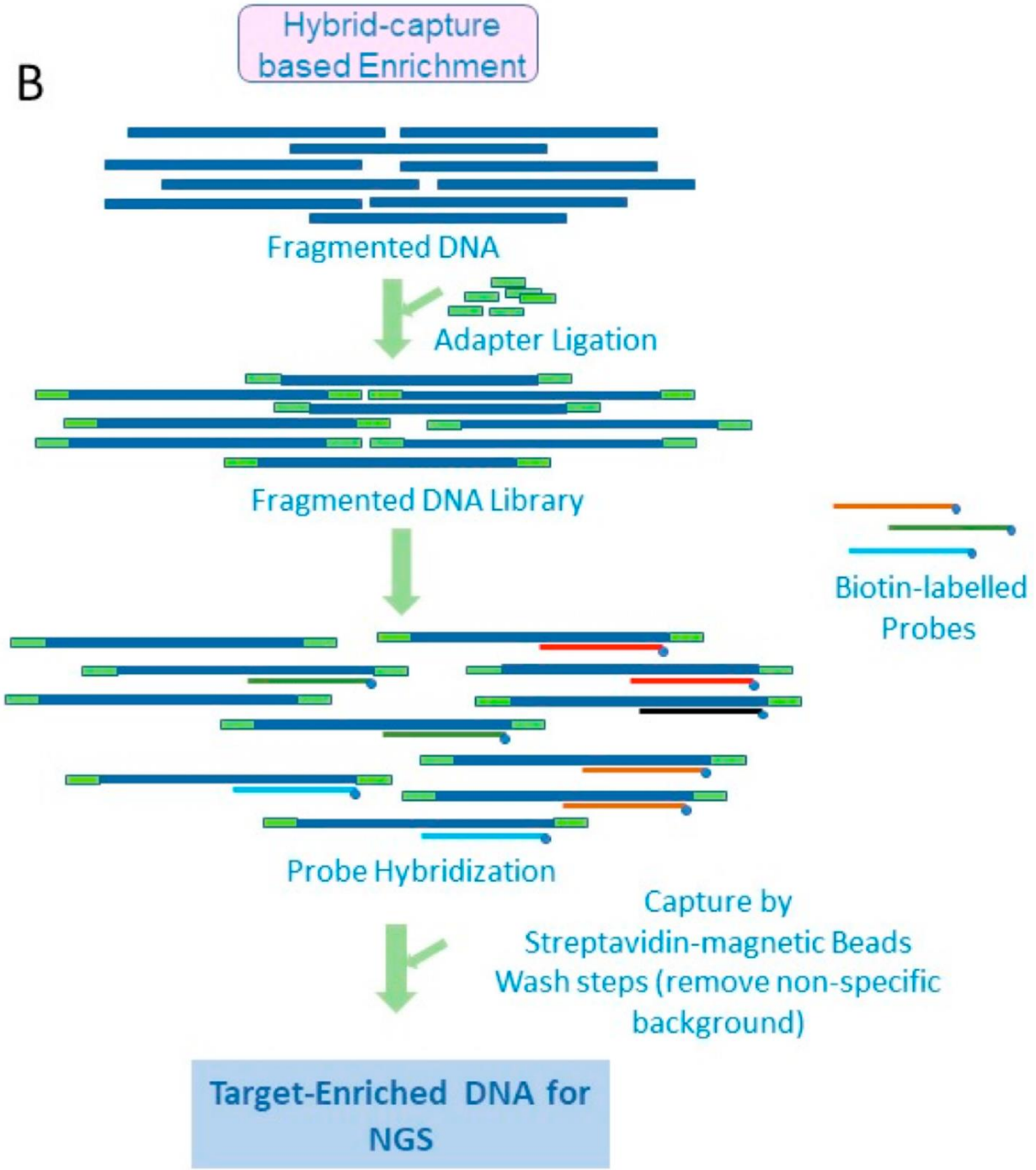
GenQA
Genomics Quality Assessment

(Wingett & Andrew, 2018)

Assembly with spades (Bankevitch,2012)



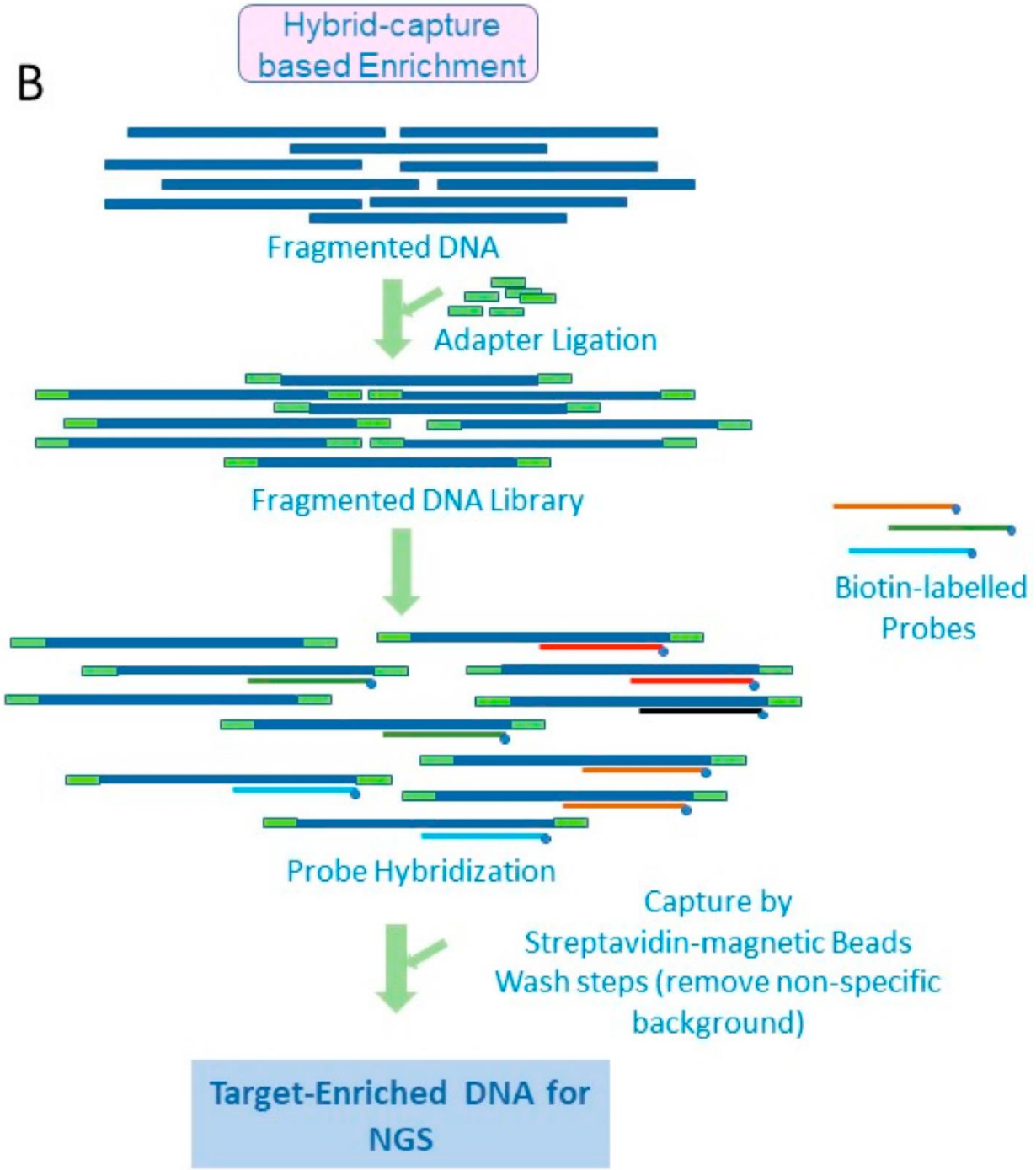
Multiple Sequence Alignment with MAFFT (Katoh,2002)



Targeted enrichment Overview

- **Why use AHE?** Sometimes, scientists don’t need all the DNA from an organism—just specific regions that are informative for their study. TAE allows us to focus on capturing and sequencing these targeted regions

- **What is the “anchored” part?**: The method uses short pieces of DNA called "probes" designed to match and "anchor" to certain genome regions across many species. These probes "enrich" or increase the target regions' presence during sequencing, making them easier to study

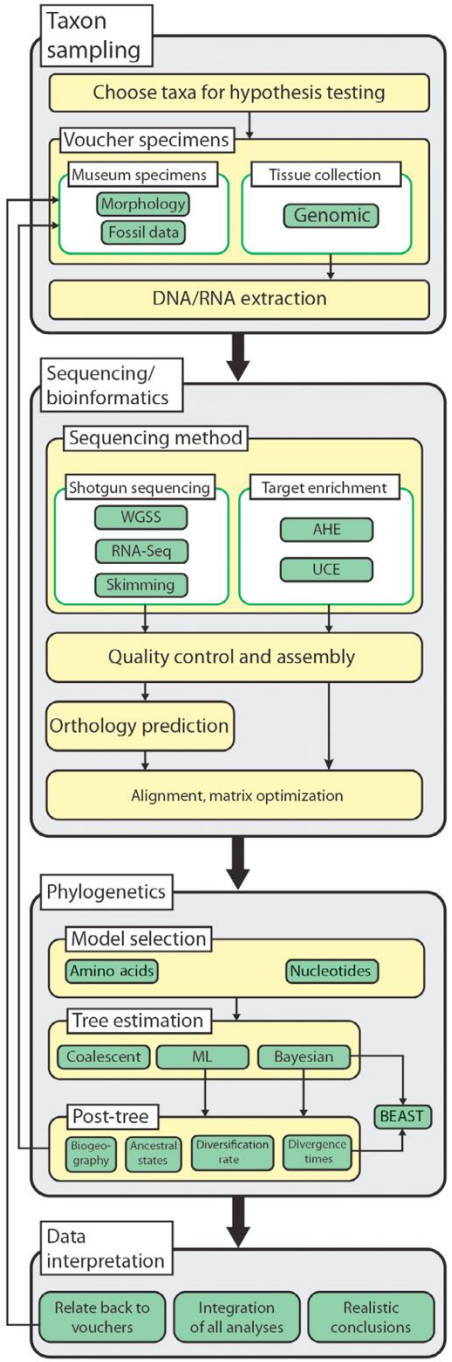


Targeted enrichment Overview

How the Process Works

- **Designing Probes**: Scientists design DNA probes that will stick to the regions of the genome they are interested in (the target regions)
- **Capturing the Targets**: DNA is extracted from the samples (e.g., dragonflies or damselflies), and the probes are added. The probes latch onto the target regions in the DNA
- **Sequencing**: Once the targeted regions are captured, they can be sequenced. This means scientists can read the DNA letters (A, T, C, G) from these regions
- **Analyzing the Data**: The sequenced DNA is used to compare different species, build phylogenetic trees, and explore evolutionary relationships.

Overview of the Bioinformatics Workflow for AHE



1. Data Preparation
2. Read Alignment
3. Assembly or Reference Mapping
4. Quality Control
5. Phylogenetic Analysis

Step 1: Data Preparation

- **Input Data**: The sequencing process provides raw DNA sequences, often in a file format called FASTQ. These files contain millions of short DNA sequences (reads) generated from the specimens

  - **FastQC**: A tool for checking the quality of the raw sequence data. It looks for errors or biases in the data.
  - **Trimmomatic**: A tool used to trim low-quality sequences or adapters that may have been attached during sequencing.

Install fastp with bioconda:

fastp is a program that performs several pre-processing quality control steps for fastQ files (e.g., trim adapters, filter out bad reads, trim ends)

```
## create a conda environment (-n fastp)
## install fastp with bioconda (-c bioconda fastp)
conda create -n fastp -c bioconda fastp

## may have to reset shell settings before activating environment
source ~/.bashrc
```

Log onto the supercomputer:

```
## logging onto the Mendel AMNH cluster
ssh <username>@mendel.sdmz.amnh.org
```

Download and install Miniconda on the cluster:

```
## download the script to install miniconda
wget
https://repo.anaconda.wgetcom/miniconda/Miniconda3-latest-Linux-x86_6
4.sh

## install miniconda
bash Miniconda3-latest-Linux-x86_64.sh

## reset shell settings; this will allow us to install and activate
## conda environments without logging off and back on
source ~/.bashrc

## lists info about current conda install, check that install worked
conda info
```

Step 2: Read Alignment (Mapping Reads to the Target Regions)

- **What's Happening?**: After cleaning the data, the next step is to align the reads to a reference genome or the regions that the AHE probes targeted.
- **Why?**: We need to see where each small DNA sequence (read) belongs in the genome to reconstruct the target regions.
- **Tools**:
  - **BWA (Burrows-Wheeler Aligner)**: Aligns the reads to a reference genome or target regions.
  - **Bowtie2**: Another alignment tool, faster in some cases, for mapping reads to the reference.

To view multiple fastp results in one summary report, we will install and run multiqc:

```
## deactivate fastp environment
conda deactivate

## create a new environment to install multiqc; installing multiple
## programs into one environment can result in conflicting
## dependencies, thus new environment prevents that
conda create -n multiqc -c conda-forge -c bioconda multiqc python=3

## make directory for multiqc in directory above FASTQ file directory
mkdir ../multiqc

## copy json files to multiqc directory
cp *.json ../multiqc

## change to multiqc directory
cd ../multiqc

## Make sure that all files end in "fastp.json"
rename .json _fastp.json *json

## activate multiqc environment
conda activate multiqc
```

Step 3: Assembly or Reference Mapping

- **De novo assembly**: If there is no reference genome, you can assemble the short reads into longer sequences (contigs)
- **Reference mapping**: If you have a reference genome, you can map your reads to it to reconstruct the target regions
- **Tools**:
  - **SPAdes**: A popular tool for assembling genomes from scratch (de novo assembly)
  - **SAMtools**: A tool for working with aligned reads, helping organize them into a usable format (BAM files)

Assembly & Orthology Filter Steps:

In the assemblies folder: tanypteryx genome database files, unique assembly list (FASTQ names coupled with sample IDs, e.g. GEODE# or RMNH#), assembly slurm array job file, python script that does assembly steps: ASS_F4_fast.py, and the program usearch.

***** genome and database files unique across datasets; *Tanypteryx* is a dragonfly genome *****

NOTE TO SELF: include information about making a blast database for a genome file

```
## create new environment and install assembly and alignment programs
conda create -n ahe_pipeline -c conda-forge -c bioconda python=3
mafft blast spades=3.15.4 biopython

## go to assemblies directory and list contents
cd ../assemblies
ls
```

Step 4: Quality Control

- **What's Happening?**: After assembling or mapping, it's important to check if the process worked well and if the data is accurate.
- **Why?**: We want to ensure that the sequences we've reconstructed are reliable and represent the actual regions of DNA we targeted.
- **Tools**:
  - **QUAST**: A tool that checks how well the assembly or mapping performed.
  - **Coverage analysis**: Determines how well each region of the genome is represented by reads
  - Check Contamination

Post-assembly steps to clean out contamination and choose the best sequence for each taxon:

In the after_assembly folder: singleline.pl, split.py, SELFBLAST_array.slurm, contamination_filter.py, getlist.py, removelist.py, usearch program

```
## change into after assembly directory
cd ../after_assembly

## concatenate all orthologs from each sample into one fasta file
cat ../assemblies/*targetsFULL_ORTHO.fasta > ALL_FULL_ORTHO.fa
```

Step 5: Phylogenetic Analysis

```
## submit alignment job
sbatch mafft_array.job

## NOTE TO SELF BEFORE ALIGNMENT
## write script to remove loci with less than 4 taxa or other cutoff
## to avoid including loci with no sequence data (will get alignment
## errors in array job and with FASconCAT)
## for loop to count number of taxa
## might have to type '>' in terminal; weird symbol issue?
for i in `cat locus_names.txt`; do grep '>' $i | wc -l >>
taxa_number.txt; done

##REMEMBER TO DELETE taxa_number.txt each time, the code does not
overwrite
```

- Once you have your sequences, the next step is to compare them across different species or samples to infer evolutionary relationships.
  Tools:
  MAFFT: A tool for aligning multiple DNA sequences to see where they match or differ.
  IQTREE: A tool for building phylogenetic trees, which show evolutionary relationships.

Concatenating our alignments

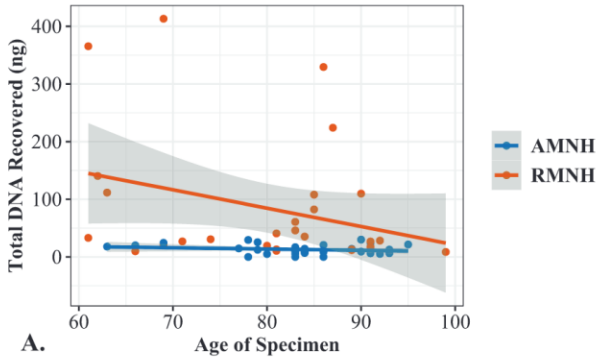
The next step in the process will be to concatenate our single locus files into a "concatenated supermatrix".

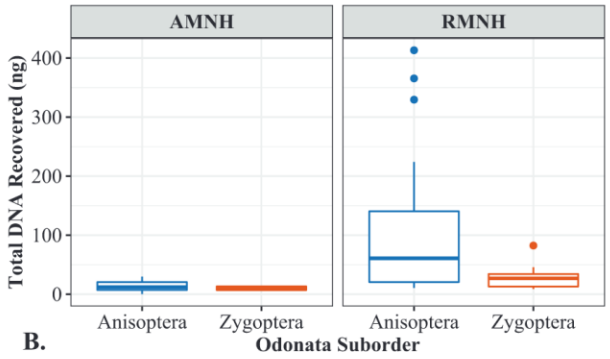
```
## change directory
cd ../concatenated_loci_full

## copy alignments
```

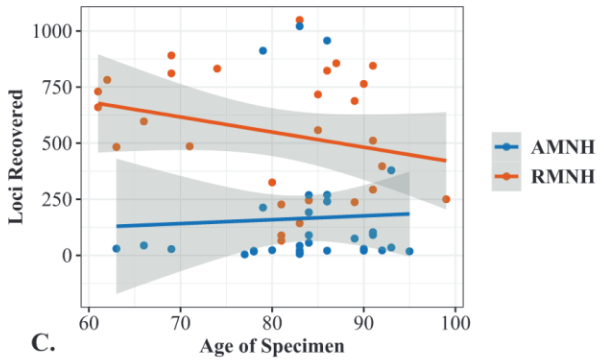
Select models and tree search with IQtree:

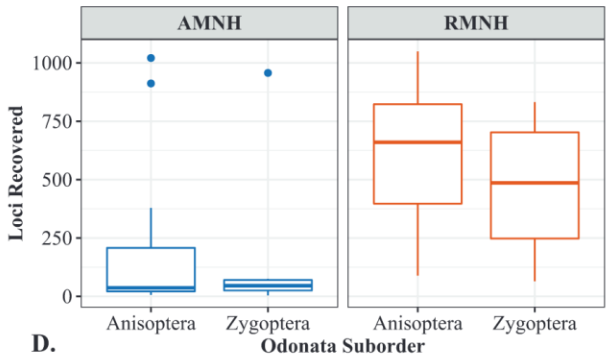
```
## edit job file with email
nano iqtree_model_selection_and_tree_search.job

## run job
sbatch iqtree_model_selection_and_tree_search.job
```

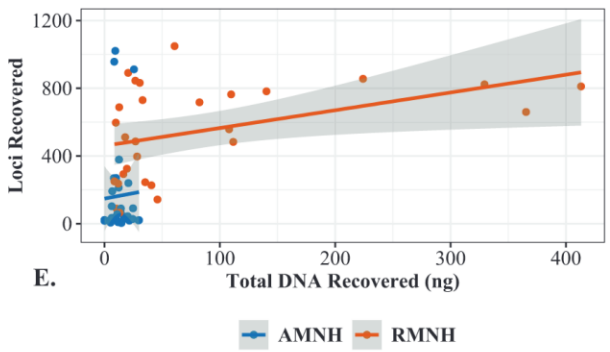
RESULTS

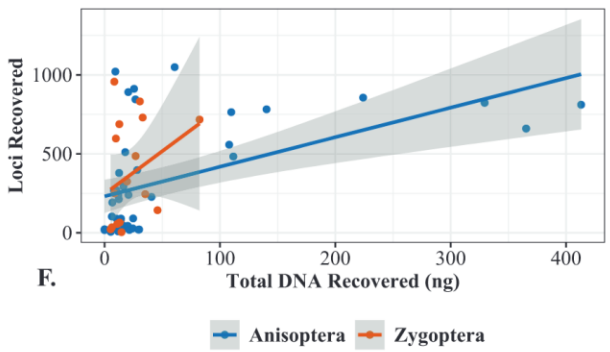

Total DNA Recovered (ng)
Age of Specimen
AMNH
RMNH
A.


AMNH
RMNH
Total DNA Recovered (ng)
Anisoptera
Zygoptera
Odonata Suborder
B.


Loci Recovered
Age of Specimen
AMNH
RMNH
C.


AMNH
RMNH
Loci Recovered
Anisoptera
Zygoptera
Odonata Suborder
D.


Loci Recovered
Total DNA Recovered (ng)
E.
AMNH
RMNH


Loci Recovered
Total DNA Recovered (ng)
F.
Anisoptera
Zygoptera


OD1074_Hemiphlebiidae_Hemiphlebia_mirabilis
GEODE_7630_Synlestidae_Megalestes_dislans
OD1120_Perilestidae_Perissolestes_remotus
OD81_Synlestidae_Synlestes_weyersii
OD1305_Synlestidae_Chlorolestes_umbratus
OD106_Lestidae_Archilestes_grandis
OD1321_Lestidae_Lestes_disjunctus
OD122_Lestidae_Sympecma_paedisca
OD1127_Lestidae_Orolestes_octomaculatus
GEODE_7636_Lestidae_Indolestes_luxatus
GEODE_7665_Lestidae_Indolestes_albicaudus
OD63_Lestidae_Austrolestes_annulosus
OD1278_Lestidae_Indolestes_cyaneus
OD131_Platystictidae_Protosticta_sanguinostigma
OD103_Coenagrionidae_Neoneura_maria
OD105_Coenagrionidae_Protoneura_capillaris
OD178_Coenagrionidae_Argia_nahuana
OD136_Coenagrionidae_Papuagrion_occipitale
OD137_Coenagrionidae_Pyrrhosoma_nymphula
OD1299_Coenagrionidae_Xanthocnemis_tuanuii
OD66_Coenagrionidae_Pseudagrion_aureofrons
OD15_Coenagrionidae_Ischnura_barberi
OD1116_Coenagrionidae_Cyanallagma_sp
OD50_Coenagrionidae_Agriocnemis_pygmaea
OD1320_Coenagrionidae_Agriocnemis_sp
OD1131_Heteragrionidae_Oxystigma_petiolatum
OD1132_Platycnemididae_Onychargia_atrocyana
OD1143_Platycnemididae_Mesocnemis_singularis
OD1123_Platycnemididae_Coeliccia_pyriformis
GEODE_7673_Platycnemiidae_Calicnemia_erythromelas
GEODE_7668_Platycnemiidae_Calicnemia_pulverulans
GEODE_7651_Platycnemiidae_Calicnemia_miles
GEODE_7670_Synthemistidae_Idionyx_carinata
OD138_Platycnemididae_Copera_marginipes
OD148_Platycnemididae_Platycnemis_pennipes
GEODE_7653_Platycnemiidae_Esme_mudiensis
GEODE_7624_Platycnemiidae_Disparoneura_apicalis
OD1122_Platycnemididae_Prodasineura_croconota
OD1306_Platycnemididae_Elattoneura_glauca
GEODE_7637_Platycnemiidae_Nososticta_insignis
GEODE_7656_Platycnemiidae_Nososticta_selysii
OD1317_Platycnemididae_Nososticta_solida
GEODE_7660_Platycnemiidae_Nososticta_exul
OD123_Platycnemididae_Nososticta_sp
OD55_Isostictidae_Rhadinosticta_simplex
OD159_Isostictidae_Isosticta_robustior
OD1277_Isostictidae_Neosticta_fraseri
OD1312_Isostictidae_Labidiosticta_vallisi
OD1079_Philoganga_vetusta
OD1138_Chlorogomphidae_Watanabeopetalia_usignata
OD1119_Philosinidae_Rhinagrion_sp
OD1130_Philosinidae_Philosina_buchi
OD1281_Megapodagrionidae_Teinopodagrion_meridionale
OD1129_Platycnemididae_Mesocnemis_singularis
OD1309_Argiolestidae_Nesolestes_sp
GEODE_7678_Argiolestidae_Luzonargiolestes_baltazarae
OD162_Argiolestidae_Caledargiolestes_uniseries
OD61_Argiolestidae_Griseargiolestes_albescens
OD142_Argiolestidae_Austroargiolestes_icteromelas
OD1083_Argiolestidae_Austroargiolestes_icteromelas2
OD976_Calopterygidae_Ormenophlebia_imperatrix
OD1170_Calopterygidae_Hetaerina_americana
OD126_Calopterygidae_Vestalis_apicalis
OD102_Calopterygidae_Matrona_basilaris
OD1125_Calopterygidae_Neurobasis_ianthinipennis
GEODE_7672_Chlorocyphidae_Cyrano_unicolor
GEODE_7675_Chlorocyphidae_Disparocypha_biedermanni
GEODE_7667_Chlorocyphidae_Libellago_greeni
OD134_Chlorocyphidae_Libellago_semiopaca
GEODE_7627_Chlorocyphidae_Sundacypha_petiolata
OD35_Chlorocyphidae_Rhinocypha_tincta
GEODE_7666_Chlorocyphidae_Heliocypha_bisignata
OD120_Chlorocyphidae_Heliocypha_biseriata
OD1147_Amanipodagrionidae_Amanipodagrion_gilliesi
OD1283_Pseudolestidae_Pseudolestes_mirabilis
OD43_Lestoideidae_Diphlebia_coerulescens
OD157_Lestoideidae_Lestoidea_conjuncta
GEODE_7696_Euphaeidae_Epallage_fatime
GEODE_7647_Euphaeidae_Bayadera_continentalis
OD99_Euphaeidae_Bayadera_brevicauda
GEODE_7683_Euphaeidae_Anisopleura_lestoides
GEODE_7687_Euphaeidae_Cyclophaea_cyanifrons
OD132_Euphaeidae_Euphaea_ochracea
GEODE_7686_Euphaeidae_Euphaea_cardinalis
GEODE_7679_Euphaeidae_Euphaea_aspasia
GEODE_7694_Euphaeidae_Euphaea_refulgens
OD1142_Platycnemididae_Onychargia_atrocyana
OD1133_Incertae_sedis_group_2_Amanipodagrion_gilliesi
GEODE_7641_Synlestidae_Sinolestes_edita
OD151_Devadattidae_Devadatta_podolestoides
OD1145_Incertae_sedis_group_3_Heteropodagrion_sanguinipes
OD1148_Amphipterygidae_Amphipteryx_agricoides
OD96_Incertae_sedis_group_1_Rhipidolestes_aculeatus
OD1085_Incertae_sedis_group_1_Agriomorpha_xinglongensis
OD1121_Incertae_sedis_group_1_Agriomorpha_fusca
OD1141_Thaumatoneuridae_Thaumatoneura_inopinata
OD1134_Chlorogomphidae_Chlorogomphus_papilio
OD1135_Thaumatoneuridae_Paraphlebia_sp
OD1280_Incertae_sedis_group_5_Priscagrion_kiautai
OD1124_Philogeniidae_Philogenia_carrillica
OD1118_Heteragrionidae_Heteragrion_majus
OD1126_Incertae_sedis_group_3_Dimeragrion_percubitale
OD1137_Incertae_sedis_group_7_Tatocnemis_malgassica
OD1139_Pentaphlebiidae_Pentaphlebia_mangana
OD1308_Incertae_sedis_group_7_Protolestes_kerckhoffae
OD1146_Hypolestidae_Hypolestes_sp
OD1140_Philogangidae_Philoganga_vetusta
OD1183_Incertae_sedis_group_3_Mesagrion_leucorhinum
OD1157_Polythoridae_Polythore_procera
OD1180_Polythoridae_Miocora_aurea
OD1181_Polythoridae_Euthore_fasciata
0.5

CONCLUSION

Targeted enrichment successfully captures genomic data from recent and historical Odonata specimens, including those over 100 years old

Museum collections are valuable resources for phylogenetic and evolutionary studies, even with degraded DNA

The technique can be applied to other taxa, expanding research opportunities beyond Odonata

Museum specimens hold great potential for biodiversity, conservation, and evolutionary research