"iBOL Europe: A regional nexus for the DNA barcoding community"

Monday 28 October, 14:00 – 14:55 CET

BioDiversity Genomics Conference 2024 (BG24)

Skimming at scale

bringing historic collections into the genomic era

Ben Price

iBOL EUROPE

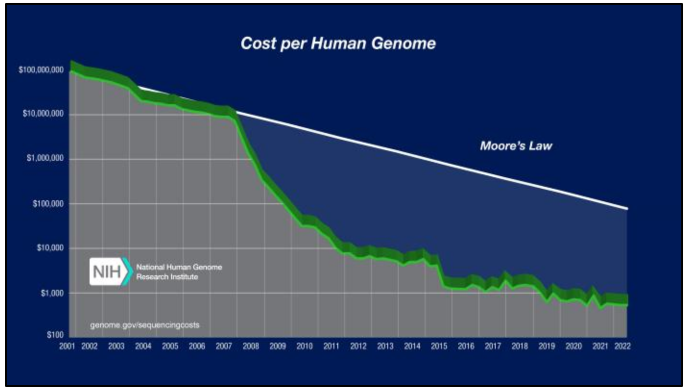
Why museums and why now?

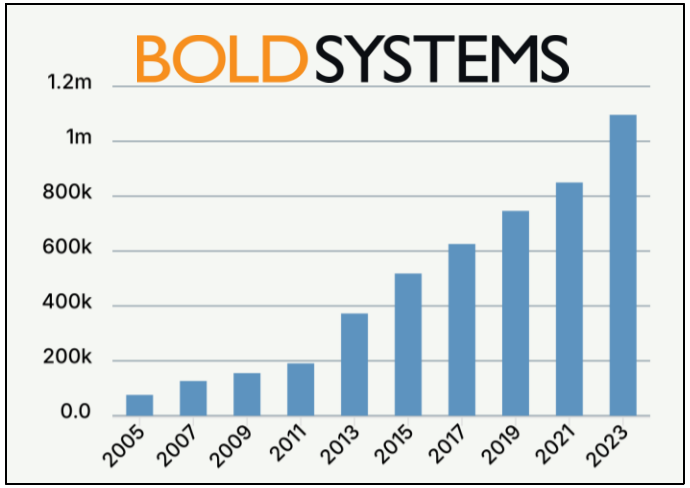
Museums:

- Time travel
- Taxonomic completeness
- Type specimens
- Frontiers of discovery (new species)

Timing:

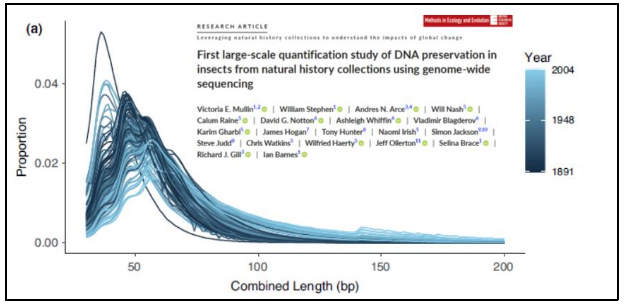
- Technology (cost per base, automation)
- BINs will outnumber named species by 2029
- Embed BINs into the Linnean framework





Museomic challenges

- Variety in sample quality
- Very little DNA
- DNA highly fragmented
- Contaminants (human, fungal, parasitoids)
- Unknown preservation history
- Balance current vs future use
- Complex Nagoya Protocol (ABS) landscape

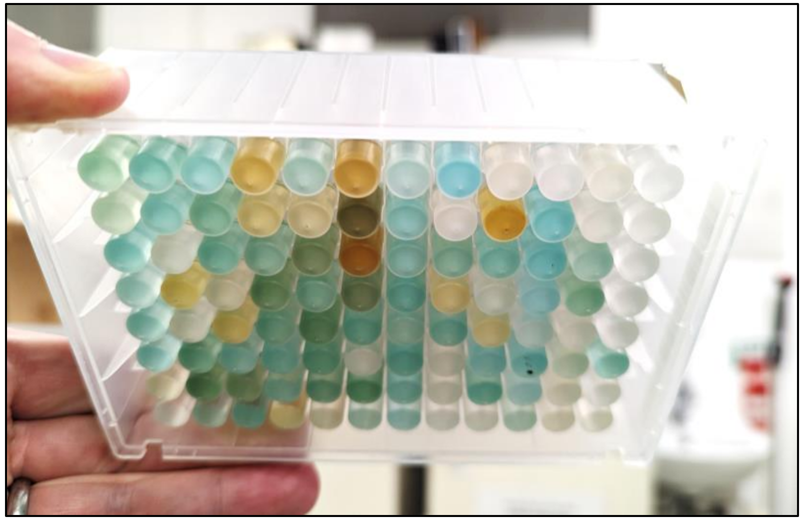


RESEARCH ARTICLE

Leveraging natural history collections to understand the impacts of global change

First large-scale quantification study of DNA preservation in insects from natural history collections using genome-wide sequencing

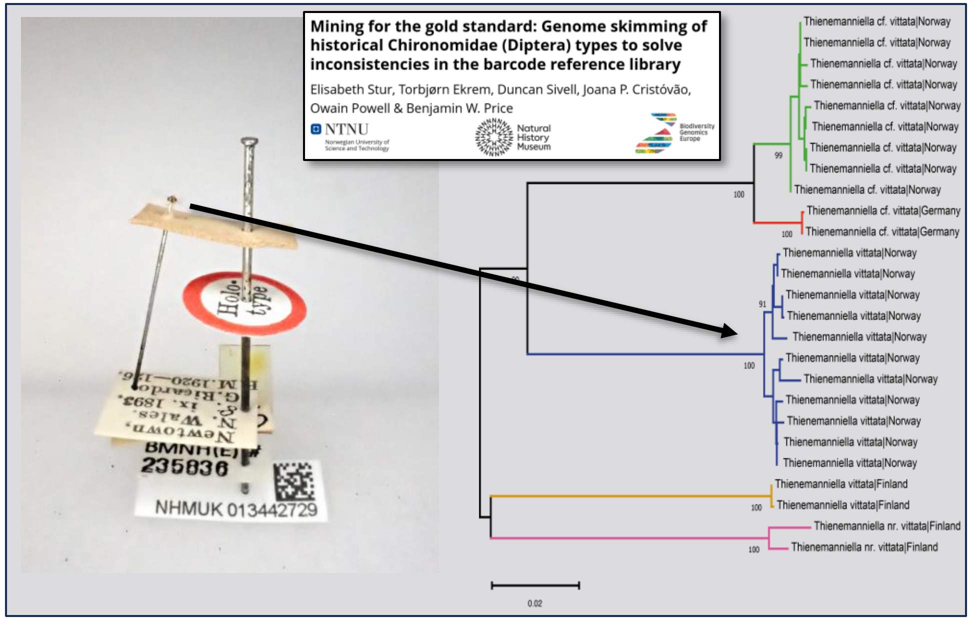
Victoria E. Mullin[1,2] | William Stephen[1] | Andres N. Arce[3,4] | Will Nash[5] | Calum Raine[5] | David G. Notton[6] | Ashleigh Whiffin[6] | Vladimir Blagderov[6] | Karim Gharbi[5] | James Hogan[7] | Tony Hunter[8] | Naomi Irish[5] | Simon Jackson[9,10] | Steve Judd[8] | Chris Watkins[5] | Wilfried Haerty[5] | Jeff Ollerton[11] | Selina Brace[1] | Richard J. Gill[3] | Ian Barnes[1]



Museomic benefits

Recent examples

- Ref. library gap filling (15k species)
- Resolving tiny type specimens
- New Aus. species found in UK!
- Integrating with AI for curation



Mining for the gold standard: Genome skimming of historical Chironomidae (Diptera) types to solve inconsistencies in the barcode reference library

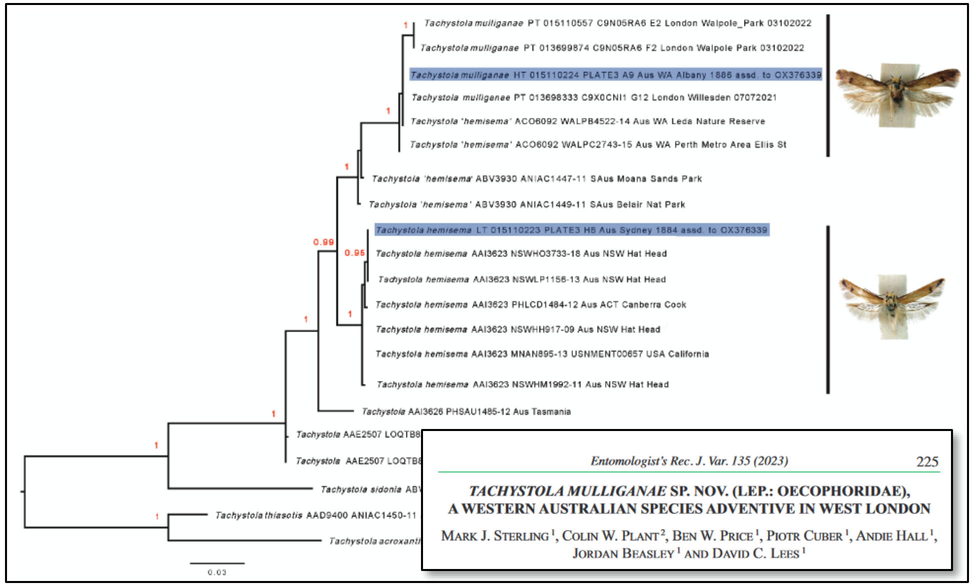
Elisabeth Stur, Torbjørn Ekrem, Duncan Sivell, Joana P. Cristóvão, Owain Powell & Benjamin W. Price

NTNU Norwegian University of Science and Technology · Natural History Museum · Biodiversity Genomics Europe

Methods in Ecology and Evolution

A computer vision method for finding mislabelled specimens within natural history collections

Jack Hollister[1,2,3], Geoff Martin[1], Xaiohao Cai[2], Tammy Horton … Cristovao[1], Owain Powell[1], Mark Sterling[1], Glory Turnbull[1], Ben Price[1], Phillip Fenbe…



Entomologist's Rec. J. Var. 135 (2023) 225

TACHYSTOLA MULLIGANAE SP. NOV. (LEP.: OECOPHORIDAE), A WESTERN AUSTRALIAN SPECIES ADVENTIVE IN WEST LONDON

MARK J. STERLING[1], COLIN W. PLANT[2], BEN W. PRICE[1], PIOTR CUBER[1], ANDIE HALL[1], JORDAN BEASLEY[1] AND DAVID C. LEES[1]

Approaches – amplicons

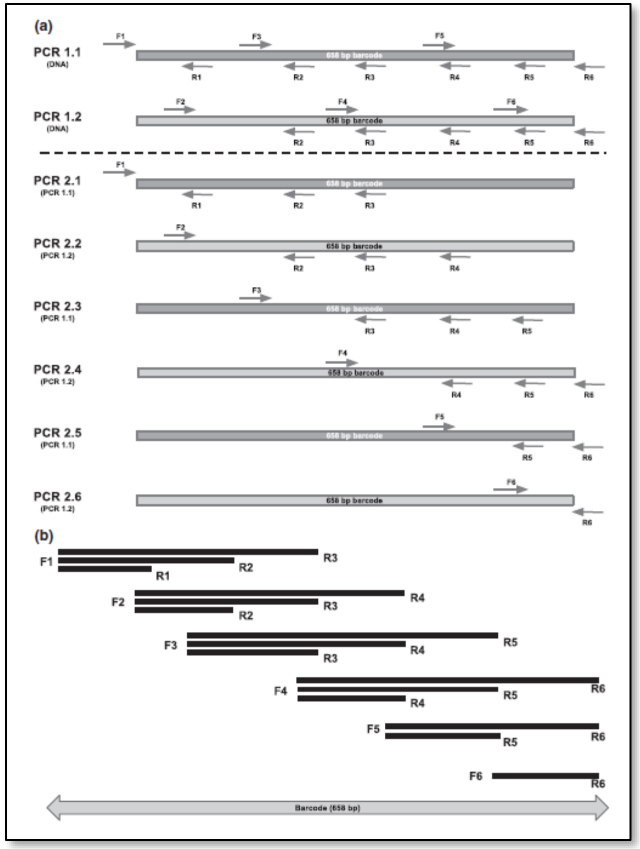
- Multiple PCRs target short fragments
- Varied success (50 – 80%)
- Costs €25 per sample (incl. lab staff)
- Recovers a single gene

MOLECULAR ECOLOGY RESOURCES

Molecular Ecology Resources (2016) 16, 487–497 doi: 10.1111/1755-0998.12474

DNA barcodes from century-old type specimens using next-generation sequencing

SEAN W. J. PROSSER,* JEREMY R. DEWAARD,* SCOTT E. MILLER† and PAUL D. N. HEBERT*



A SMRT approach for targeted amplicon sequencing of museum specimens (Lepidoptera)—patterns of nucleotide misincorporation

Jacopo D'Ercole[1,2,*], Sean W.J. Prosser[1,*] and Paul D.N. Hebert[1,2]

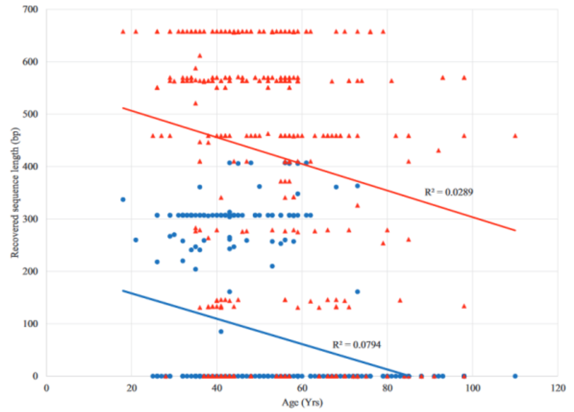
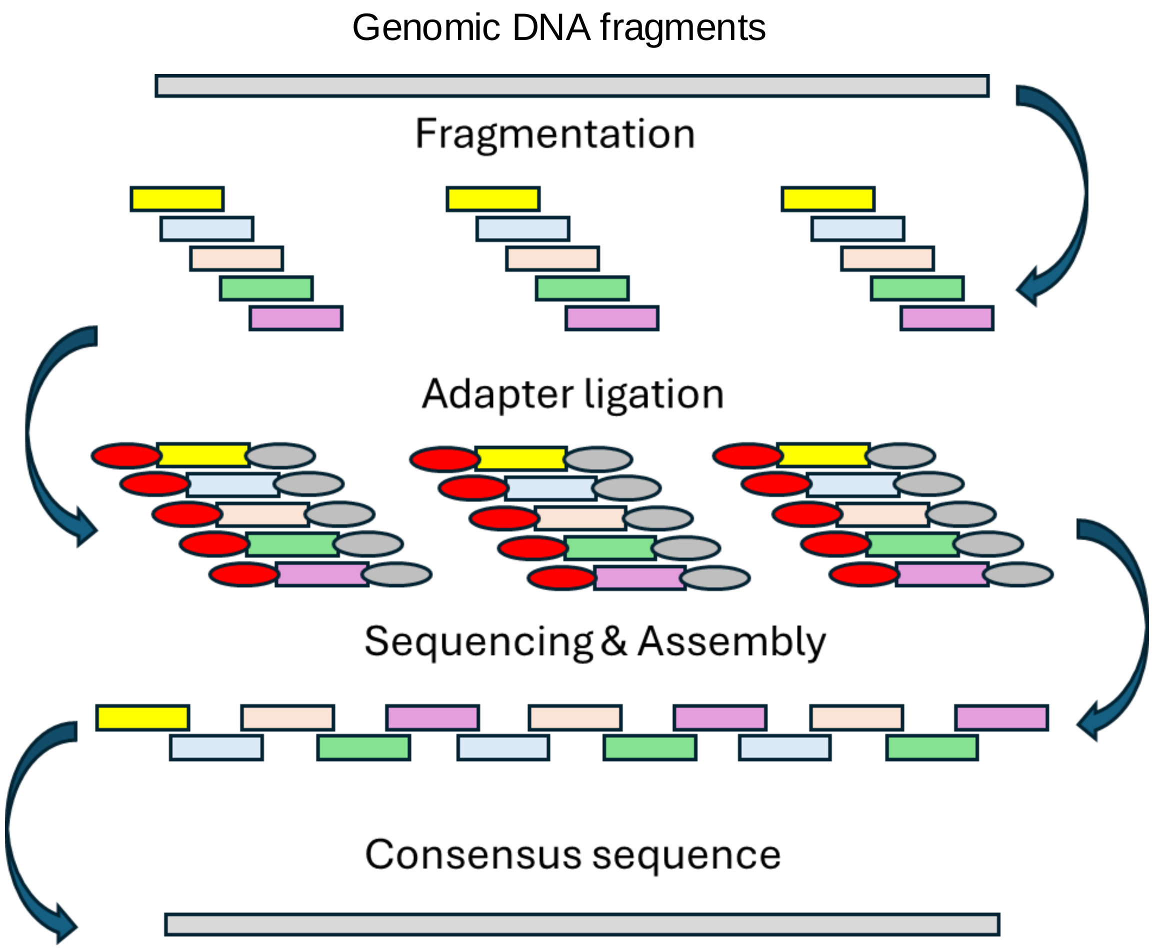


Figure 2 **Relationship between length of the COI sequence recovered and age of the source specimen with Sanger (blue) and SMRT sequencing (red).** Trend lines and R^2 values are shown.

Full-size DOI: 10.7717/peerj.10420/fig-2

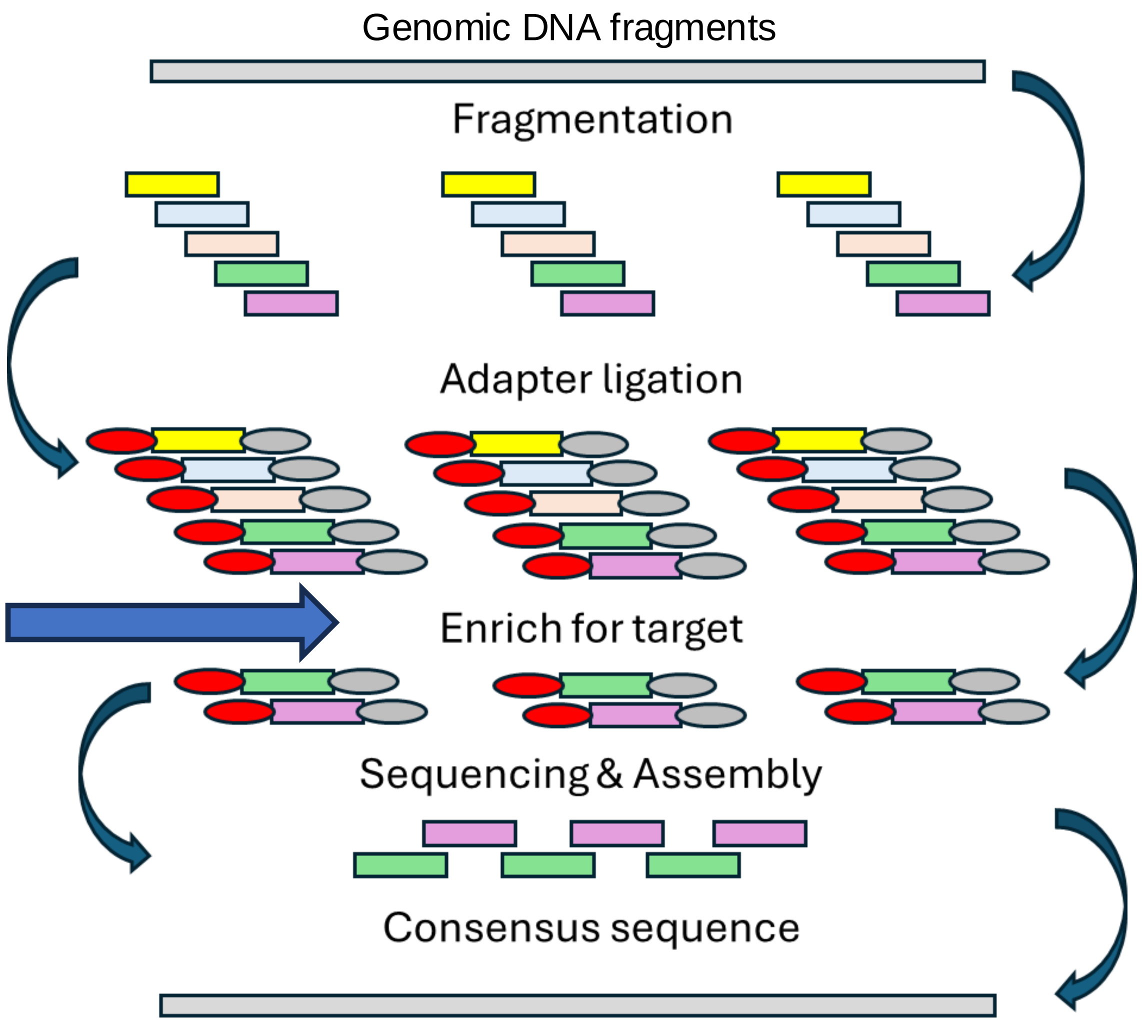
Approaches – genome skimming

- Low coverage whole genome sequencing
- €15-30 depending on genome size
- ~70% successful barcode recovery
- Sequence all DNA (good & bad)
- High copy nuclear/organelle genes
- Revisit DNA library



Approaches – target enrichment

- Targeted genome sequencing
- Similar prep cost to WGS
- Needs capture probes
- Less sequencing required
- Better target recovery
- Not representative of the genome



Current approach

Reduce reagents and increase throughput

- Tissue sampling:
    - Minimal tissue used (balance future use)
    - Prevent cross contamination

- Extraction
    - Modified aDNA (Kingfisher APEX)
    - Reduced volumes
    - Non-destructive
    - Cost = € 0.70
    - Protocols.io

Current approach

Reduce reagents and increase throughput

- Library building:
  - Reduced volume kit (NEB)
  - DIY forensic approach (SCR)
  - Cost = €8 - 12

- Sequencing:
  - Dual unique indexing
  - 380 samples per NovaSeq 25B lane
  - Cost = € 8 (€1.50 per M reads; aim: 5-10M)

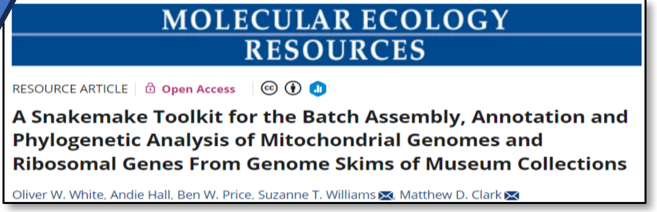
Current approach

Reduce reagents and increase throughput

- Bioinformatics:
  - Testing several pipelines
    - MGE
    - skim2mito
    - Barcode validator (bot for QC)
  - Upscaling target references
    - “Gene Fetch” & “Go Fetch” scripts
  - HPC resource & storage
    - Upscale NCBI / ENA upload

iBOL EUROPE

Published today!!

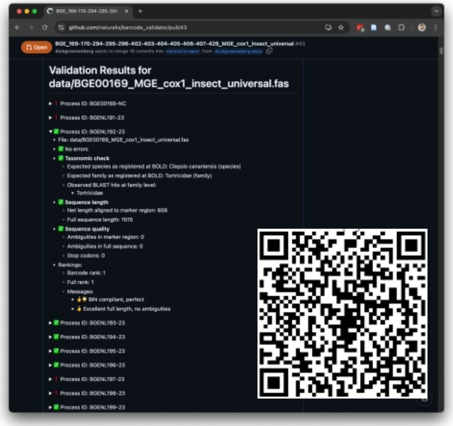


MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE | Open Access

A Snakemake Toolkit for the Batch Assembly, Annotation and Phylogenetic Analysis of Mitochondrial Genomes and Ribosomal Genes From Genome Skims of Museum Collections

Oliver W. White, Andie Hall, Ben W. Price, Suzanne T. Williams, Matthew D. Clark

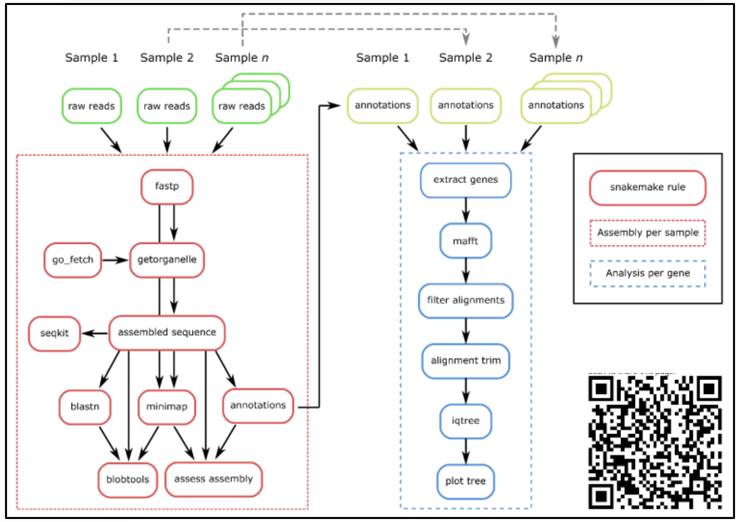


BGE_169-170-294-295-296-402-403-404-405-406-407-426_MGE_cox1_insect_universal #43

dickgroenenberg wants to merge 16 commits into naturalis:main from dickgroenenberg:main

Validation Results for data/BGE00169_MGE_cox1_insect_universal.fas

- ▶ ! Process ID: BGE00169-NC
- ▶ ! Process ID: BGENL191-23
- ▼ ✅ Process ID: BGENL192-23
  - File: data/BGE00169_MGE_cox1_insect_universal.fas
  - ✅ No errors
  - ✅ **Taxonomic check**
    - Expected species as registered at BOLD: Clepsis canariensis (species)
    - Expected family as registered at BOLD: Tortricidae (family)
    - Observed BLAST hits at family level:
      - Tortricidae
  - ✅ **Sequence length**
    - Net length aligned to marker region: 656
    - Full sequence length: 1515
  - ✅ **Sequence quality**
    - Ambiguities in marker region: 0
    - Ambiguities in full sequence: 0
    - Stop codons: 0
  - Rankings:
    - Barcode rank: 1
    - Full rank: 1
    - Messages:
      - 🥇🌟 BIN compliant, perfect
      - 🥇 Excellent full length, no ambiguities
- ▶ ✅ Process ID: BGENL193-23
- ▶ ✅ Process ID: BGENL194-23
- ▶ ✅ Process ID: BGENL195-23
- ▶ ✅ Process ID: BGENL196-23
- ▶ ! Process ID: BGENL197-23
- ▶ ! Process ID: BGENL198-23
- ▶ ✅ Process ID: BGENL199-23



APPLICATION

Methods in Ecology and Evolution

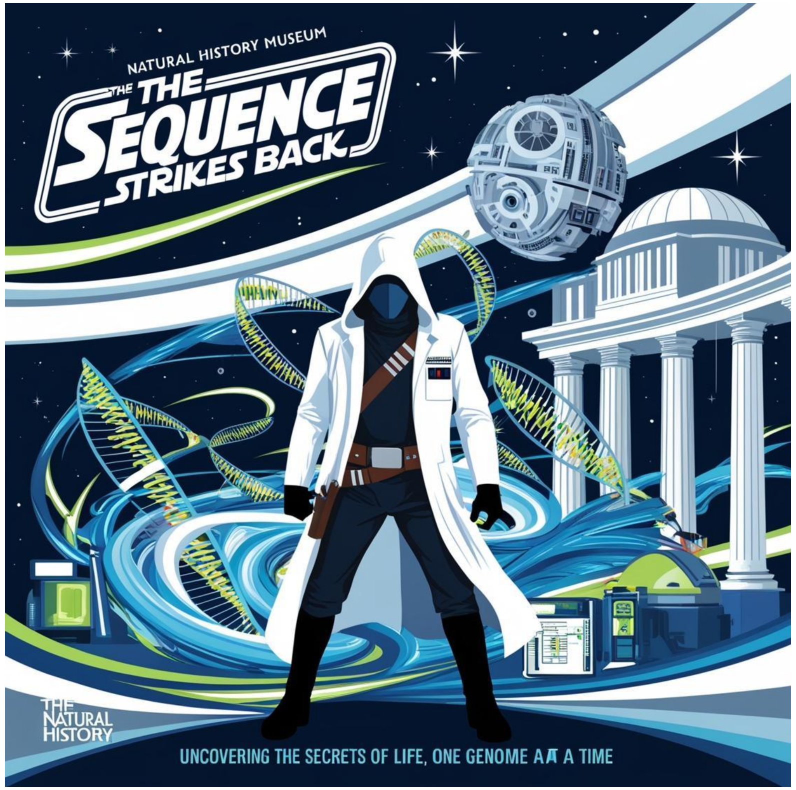
MitoGeneExtractor: Efficient extraction of mitochondrial genes from next-generation sequencing libraries

Marie V. Brasseur | Jonas J. Astrin | Matthias F. Geiger | Christoph Mayer

Biodiversity Genomics Europe

Lessons learnt

- Difficult to predict success
- Dedicated facilities if possible
- Robotic implementation takes time
- Don’t always need robots →
- Quick wins = reduce volume / DIY
- Current cost = €20 excl. staff
- 2-person team = 8 plates per month
  - Adds €20 per sample!



Looking ahead

- Refine protocols (consumable costs)
- Reduce plastic waste (cost & environment)
- Increase throughput (reduce staff costs)
- Develop assembly / mapping pipelines
  - No single best method at the moment

- Explore the rest of the genomic data

Thanks

- Andie Hall
- Owain Powell
- Stacey Dubbeldam
- Ian Barnes
- Elza Duijm
- Joana Cristovao
- Alyssa Paul
- Carla Gustave
- Jordan Beasley
- SciLifeLab
- **All curators and BGE tissue providers**

iBOL EUROPE

European Commission

UK Research and Innovation

Biodiversity Genomics Europe

Thank you

iBOL EUROPE