Museomics: Genetic Exploration for the Past and Future

Niklas Wahlberg
Lund University

Some slides courtesy of Victoria Twort (University of Helsinki)

Museomics

- Combination of genomics + museum specimens
- Allows the utilisation of important/interesting taxa
  - Eg. Extinct, difficult to collect
- Wide variety of applications
- Applied to a variety of taxa

Historical vs Ancient DNA

hDNA

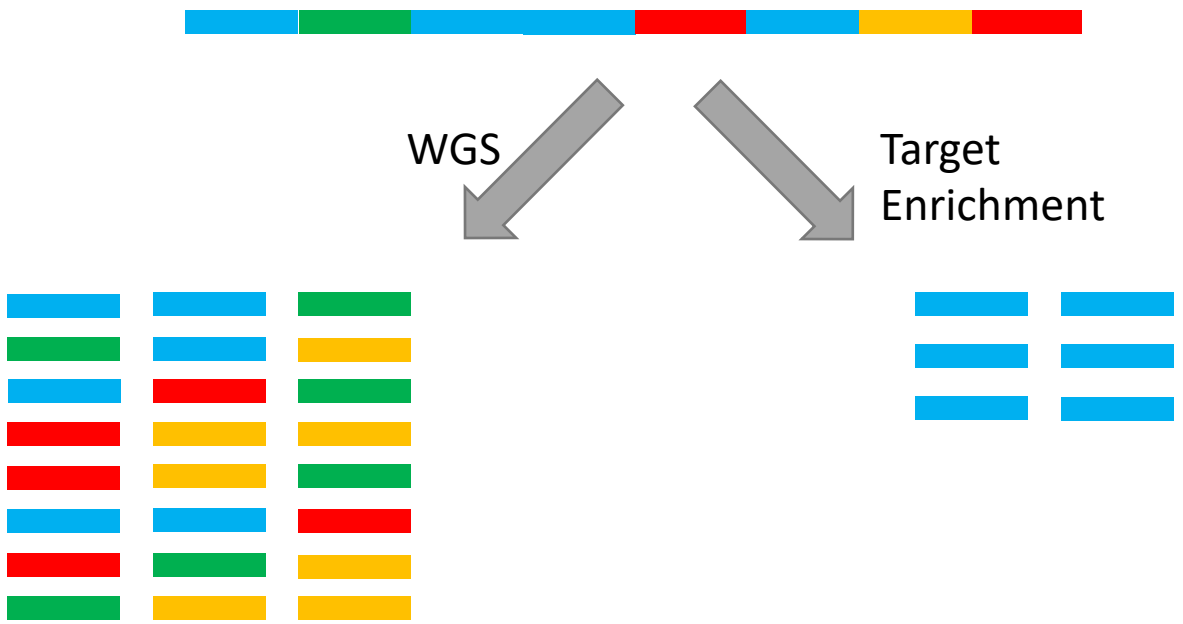
- Derived from specimens archived in museums
- Typically up to hundreds of years old

aDNA

- Naturally preserved specimens
- Usually thousands – 1 million years old
- Typically contain trace amounts of DNA that is highly degraded
- Requires specialist facilities

What can one get?

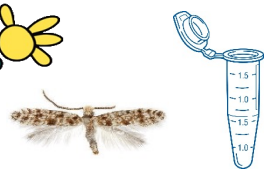
- Whole genome sequencing
- Genome reduction methods
  - Target enrichment
  - HyRAD and HyRAD-x
- Genome sized important consideration
  - <1 Gb → WGS
  - >1 Gb → Genome reduction



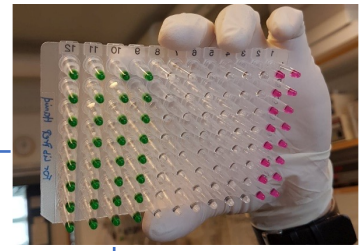
Museomics!

- Whole Genome Sequencing from museum specimens
- Whole Genome Sequencing from existing DNA extracts

DNA Extraction



Fragment DNA (Sonication)



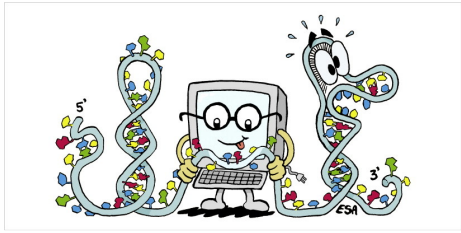
Library Preparation

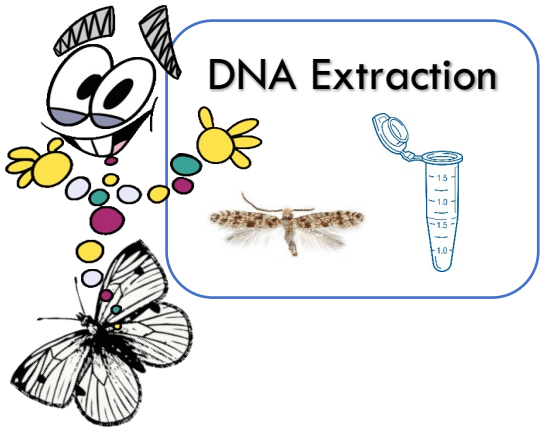
Pooling and Sequencing

illumina®

Quality Control

Downstream Analysis

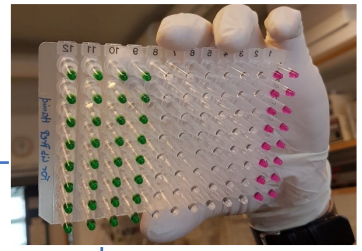


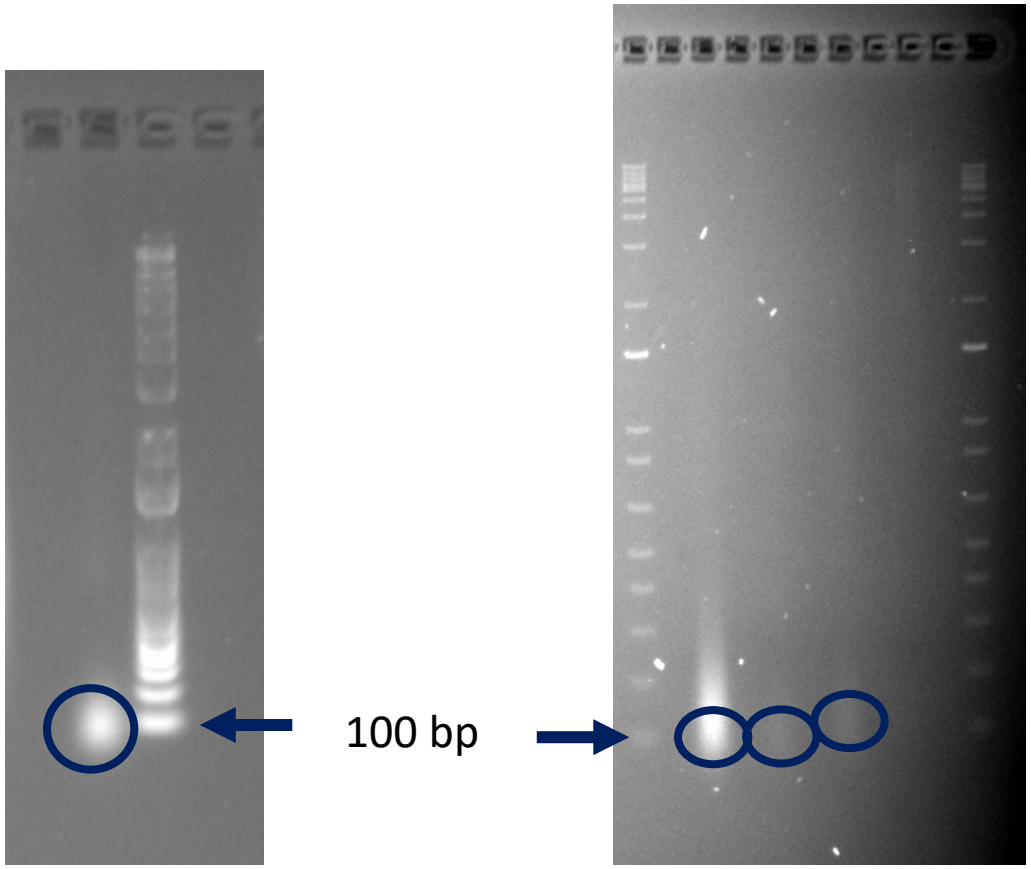


- DNA is highly fragmented (usually)
- Using normal kits for DNA extractions risks losing a lot of DNA
- Several options
  - Phenol-chloroform extraction
  - Kit for short fragments of DNA (e.g. Qiagen's QIAamp DNA Micro Kit)
  - Bead extraction (e.g. Sera-mag SpeedBeads)
- Even small amounts of tissue possible, "non-destructive sampling"
- Best to have a dedicated Museomics room for extraction and library prep

Fragment DNA (Sonication)

Library Preparation



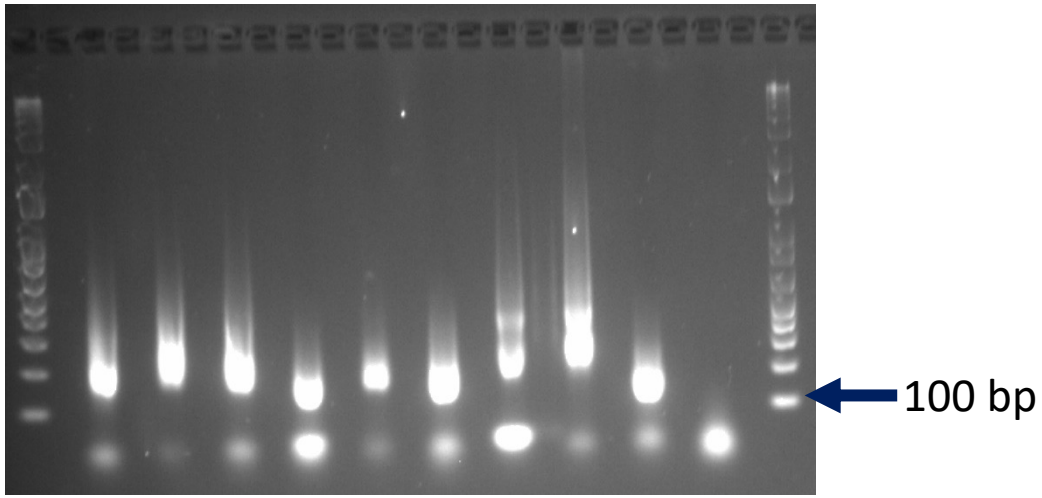


- If working with old DNA extracts, sonication might be necessary
- Otherwise, library prep can be done directly on museum DNA extract
- DNA concentration is usually very low –> companies do not like working with them
- We do our own library prep with a modified protocol
- doi:10.6084/m9.figshare.12927500
- Library prep about 20-30 EUR per sample

Pooling and Sequencing

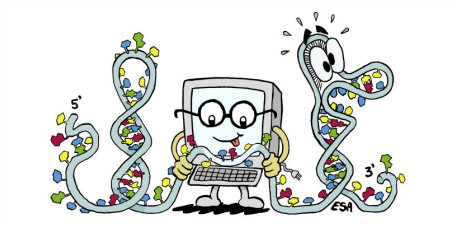
illumina

- Pooling for Illumina sequencing, depends on genome size!
- We pool up to 60 individuals in one Illumina NovaSeq S4 cell, cost was about 6 kEUR
- New sequencer at SciLifeLab, cuts costs further, now 4 kEUR for same number of genomes



Quality Control

Downstream Analysis



- Low coverage, short read sequences
- Cleaning is easy, adapters cut off, NNNNs removed
- Assembling genomes is more difficult, easier with a reference genome
- For de novo assembly we have used spADES
- Expect very low N50s, very many contigs
- In our experience exons for protein coding genes are assembled quite well

An example

- Sequence museum specimens of species with a reference genome available
  - *Pieris napi*
- How much butterfly DNA are we sequencing?
- Can we get good coverage of the genome?

Pieris napi WGS resequencing

Year	Average size (bp)	Map to the reference (%)	Nuclear coverage (X)	Mitochondrial coverage (X)
1885	49.58	80.23	6.90	771.24
1900	59.63	80.39	11.32	522.53
1906	57.79	81.52	5.30	1145.31
1909	56.29	81.98	8.32	1028.67
1918	58.48	81.09	13.81	1350.54
1922	59.57	82.49	14.21	1165.48
1941	52.99	75.94	10.84	1399.68
1941	54.02	81.44	18.36	1188.90
1947	53.70	81.07	31.57	1516.03
1954	52.36	81.60	15.77	1937.56
1954	56.32	80.19	24.49	3213.14
1985	77.06	85.06	15.60	4389.44
1989	81.10	85.20	25.95	5889.47

Pieris napi WGS resequencing

Year	Average size (bp)	Map to the reference (%)	Nuclear coverage (X)	Mitochondrial coverage (X)
1885	49.58	80.23	6.90	771.24
1900	59.63	80.39	11.32	522.53
1906	57.79	81.52	5.30	1145.31
1909	56.29	81.98	8.32	1028.67
1918	58.48	81.09	13.81	1350.54
1922	59.57	82.49	14.21	1165.48
1941	52.99	75.94	10.84	1399.68
1941	54.02	81.44	18.36	1188.90
1947	53.70	81.07	31.57	1516.03
1954	52.36	81.60	15.77	1937.56
1954	56.32	80.19	24.49	3213.14
1985	77.06	85.06	15.60	4389.44
1989	81.10	85.20	25.95	5889.47

Pieris napi WGS resequencing

Year	Average size (bp)	Map to the reference (%)	Nuclear coverage (X)	Mitochondrial coverage (X)
1885	49.58	80.23	6.90	771.24
1900	59.63	80.39	11.32	522.53
1906	57.79	81.52	5.30	1145.31
1909	56.29	81.98	8.32	1028.67
1918	58.48	81.09	13.81	1350.54
1922	59.57	82.49	14.21	1165.48
1941	52.99	75.94	10.84	1399.68
1941	54.02	81.44	18.36	1188.90
1947	53.70	81.07	31.57	1516.03
1954	52.36	81.60	15.77	1937.56
1954	56.32	80.19	24.49	3213.14
1985	77.06	85.06	15.60	4389.44
1989	81.10	85.20	25.95	5889.47

Pieris napi WGS resequencing

Year	Average size (bp)	Map to the reference (%)	Nuclear coverage (X)	Mitochondrial coverage (X)
1885	49.58	80.23	6.90	771.24
1900	59.63	80.39	11.32	522.53
1906	57.79	81.52	5.30	1145.31
1909	56.29	81.98	8.32	1028.67
1918	58.48	81.09	13.81	1350.54
1922	59.57	82.49	14.21	1165.48
1941	52.99	75.94	10.84	1399.68
1941	54.02	81.44	18.36	1188.90
1947	53.70	81.07	31.57	1516.03
1954	52.36	81.60	15.77	1937.56
1954	56.32	80.19	24.49	3213.14
1985	77.06	85.06	15.60	4389.44
1989	81.10	85.20	25.95	5889.47

How old can one go?

Drosophila approximata type specimen collected in the early 1800s (described in 1847 by Zetterstedt, kept in Biological Museum Lund)

before extraction

after extraction

Whole genome sequenced and mapped on to reference genome with good success (project of Hamid Ghanavi)

How small can one go?

Mark Blaxter: if an organism has 1000 cells, we can sequence its genome from a single specimen

DNA extractions from parasitoid wasps using “non-destructive” methods (project of Emma Kärrnäs)

De novo sequencing - *Whalleyana*

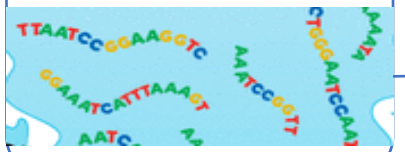
- Own superfamily
- Endemic to Madagascar
- 2 species in genus
- Biology and phylogenetic position unknown
- Where does the taxon fit in Lepidoptera?

Twort, V. G., Minet, J., Wheat, C. W., Wahlberg, N. 2021: Museomics of a rare taxon: placing Whalleyanidae in the Lepidoptera Tree of Life. Systematic Entomology 46: 926-937. doi:10.1111/syen.12503

Testing with museum specimens

- Extractions from four specimens collected in 1960s and 1970s
- One extraction used previously in a Sanger sequencing project
- WGS using Illumina platform

Raw Data



Quality Control

De novo Genome Assembly

Extract Genes of Interest

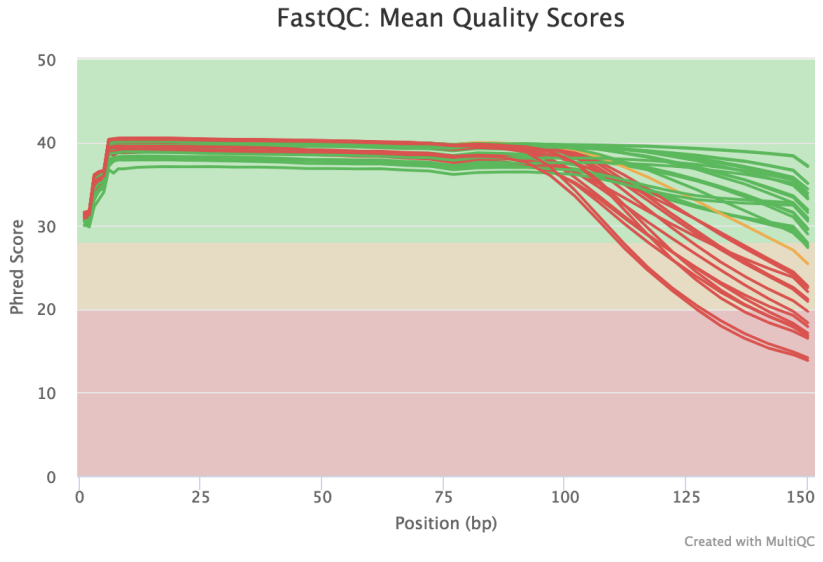
Identify Genes of Interest

Alignment & Quality Control

Phylogeny

De novo Sequencing Results

Species	G bp Raw Data	Mreads
Whalleyana vroni	139	462.34
Helicomitra pulchra	21	68.28
Hyblaea madagascariensis	25	81.67
Griveaudia vieui	22	73.45
Hyblaea puera	20	68.57



De novo Genome Assembly

Species	No. Contigs	Total Length	Av. Length	N50 (bp)
Whalleyana vroni	1 639 567	484 Mbp	296 bp	478
Helicomitra pulchra	1 155 426	321 Mbp	279 bp	361
Hyblaea madagascariensis	746 054	246 Mbp	330 bp	376
Griveaudia vieui	700 194	198 Mbp	284 bp	317
Hyblaea puera	985 209	411 Mbp	418 bp	2078

Fragmented Genome Assembly

BUSCO completeness

Identify Genes of Interest

Genes of Interest (332 curated gene set)

Fragmented Genome Assembly

Reference Protein Set (331 Genes)

Identify Genes of Interest

BUSCO completeness

MESPA (Mining exons in scaffolding-poor assemblies)

Genes of Interest (CDS & Protein Sequences)

Neethiraj, R, Hornett, EA, Hill, JA, Wheat, CW. Investigating the genomic basis of discrete phenotypes using a Pool-Seq-only approach: New insights into the genetics underlying colour variation in diverse taxa. Mol Ecol. 2017; 26: 4990–5002. https://doi.org/10.1111/mec.14205

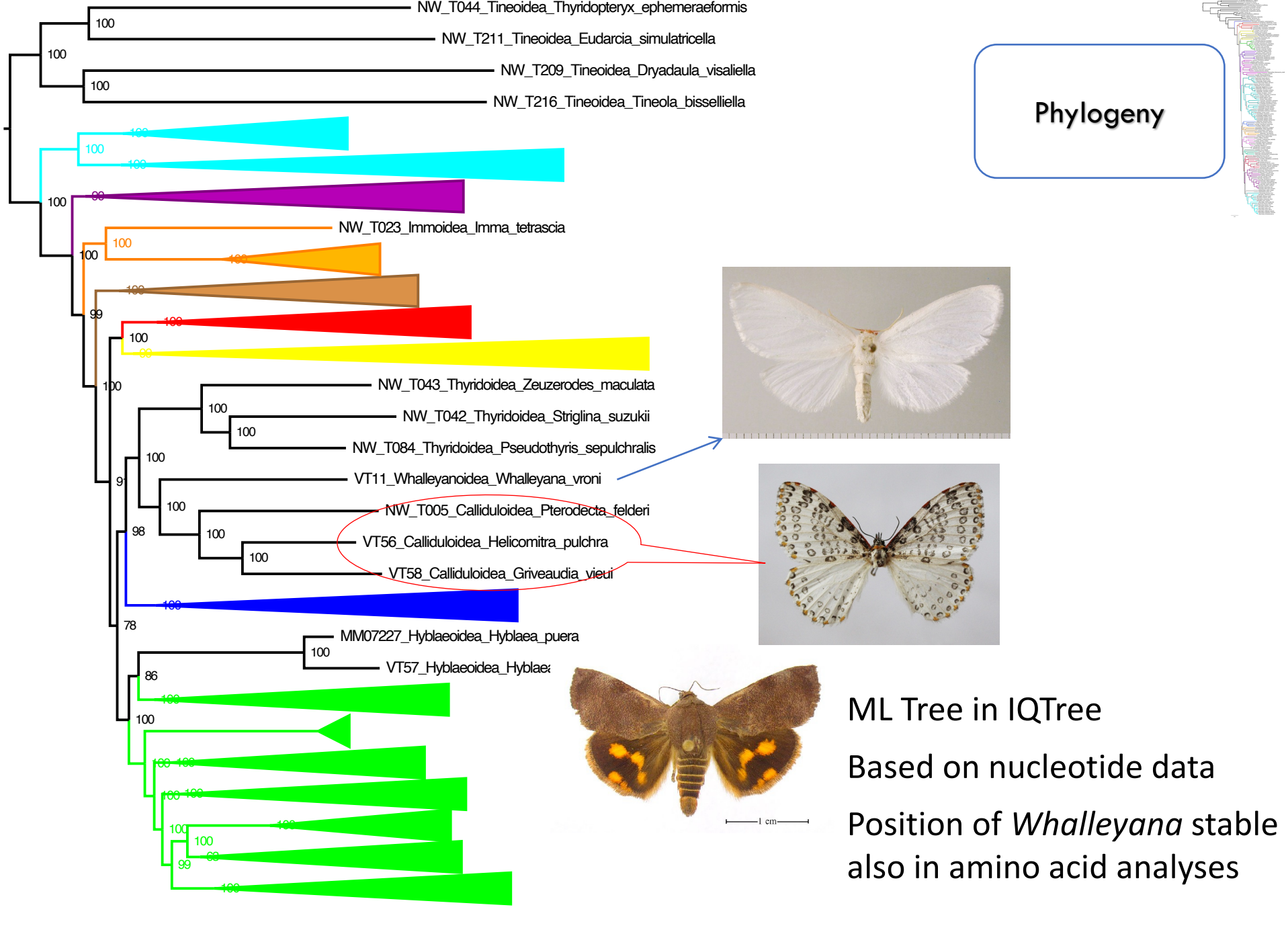
Identify Genes of Interest

Species	BUSCO	No. Genes	N50 (bp)
Whalleyana vroni	17.5% (38%)	255	478
Helicomitra pulchra	3% (12%)	244	361
Hyblaea madagascariensis	4.5% (16%)	244	376
Griveaudia vieui	1.5% (6.34%)	215	317
Hyblaea puera	73% (87%)	304	2078

Identify Genes of Interest

Species	BUSCO	No. Genes	N50 (bp)
Whalleyana vroni	17.5% (38%)	255	478
Helicomitra pulchra	3% (12%)	244	361
Hyblaea madagascariensis	4.5% (16%)	244	376
Griveaudia vieui	1.5% (6.34%)	215	317
Hyblaea puera	73% (87%)	304	2078

Phylogeny



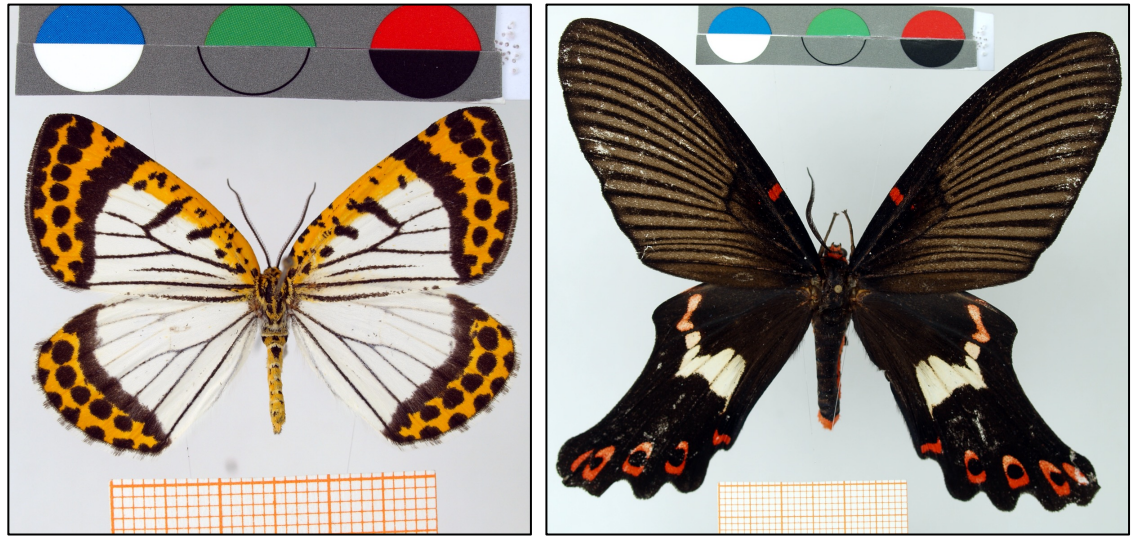
ML Tree in IQTree

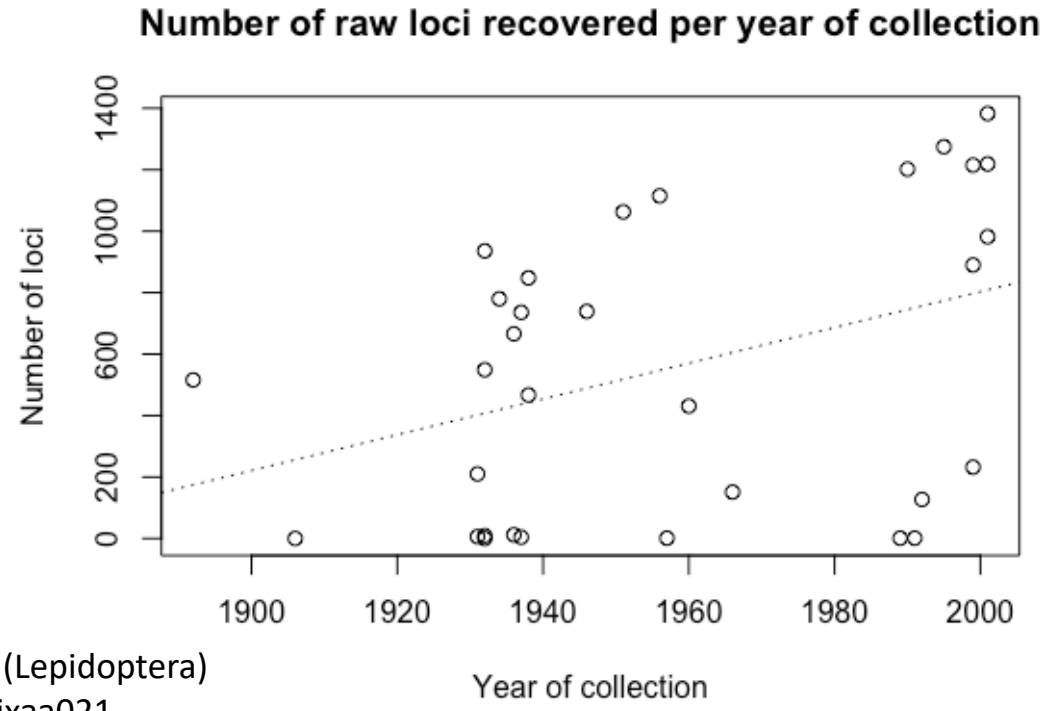
Based on nucleotide data

Position of *Whalleyana* stable also in amino acid analyses

An example of Target Enrichment

- Probes made for more than 2000 genes
- Tested on a set of species from the small family Epicopeiidae (Lepidoptera)

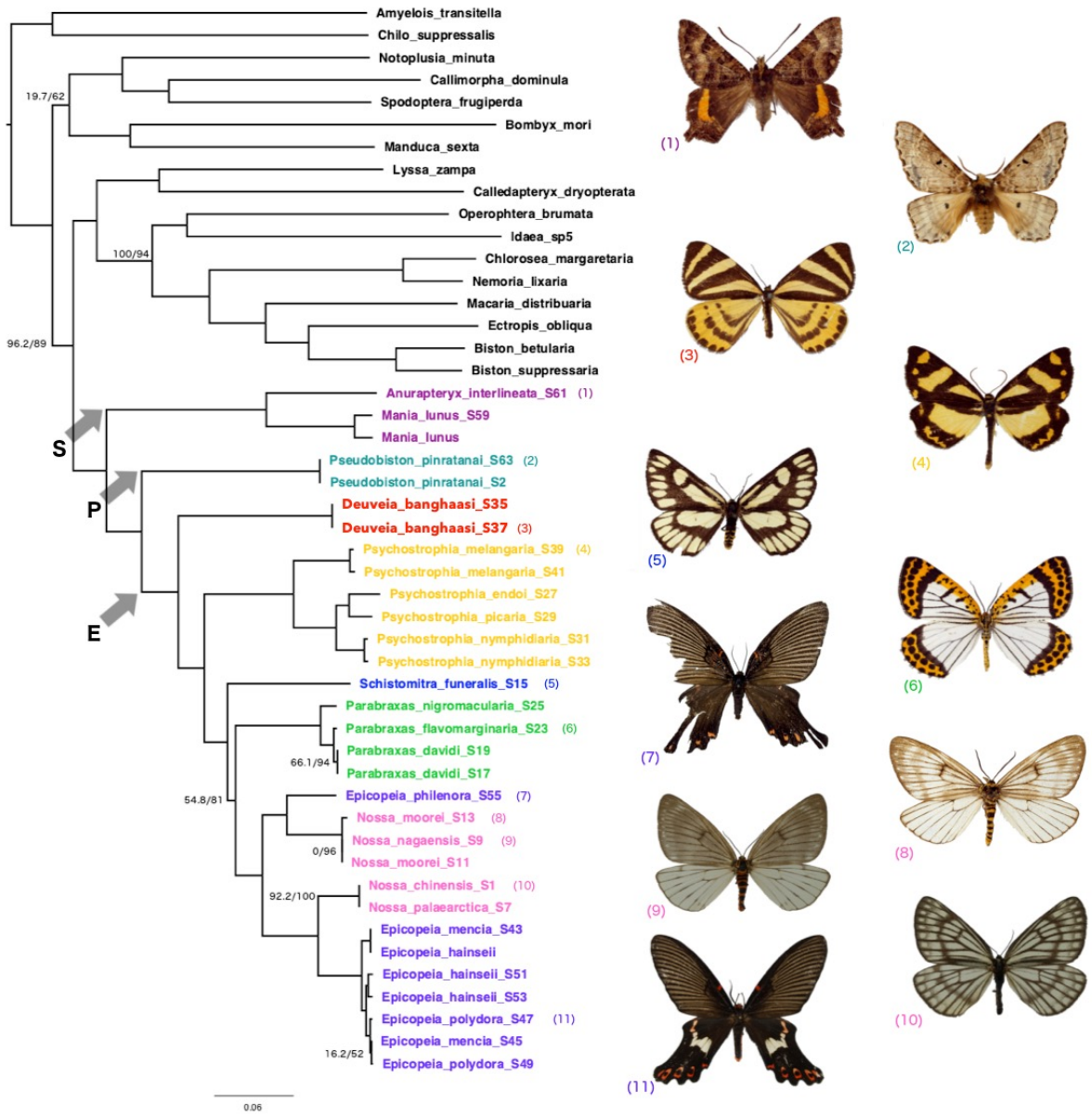




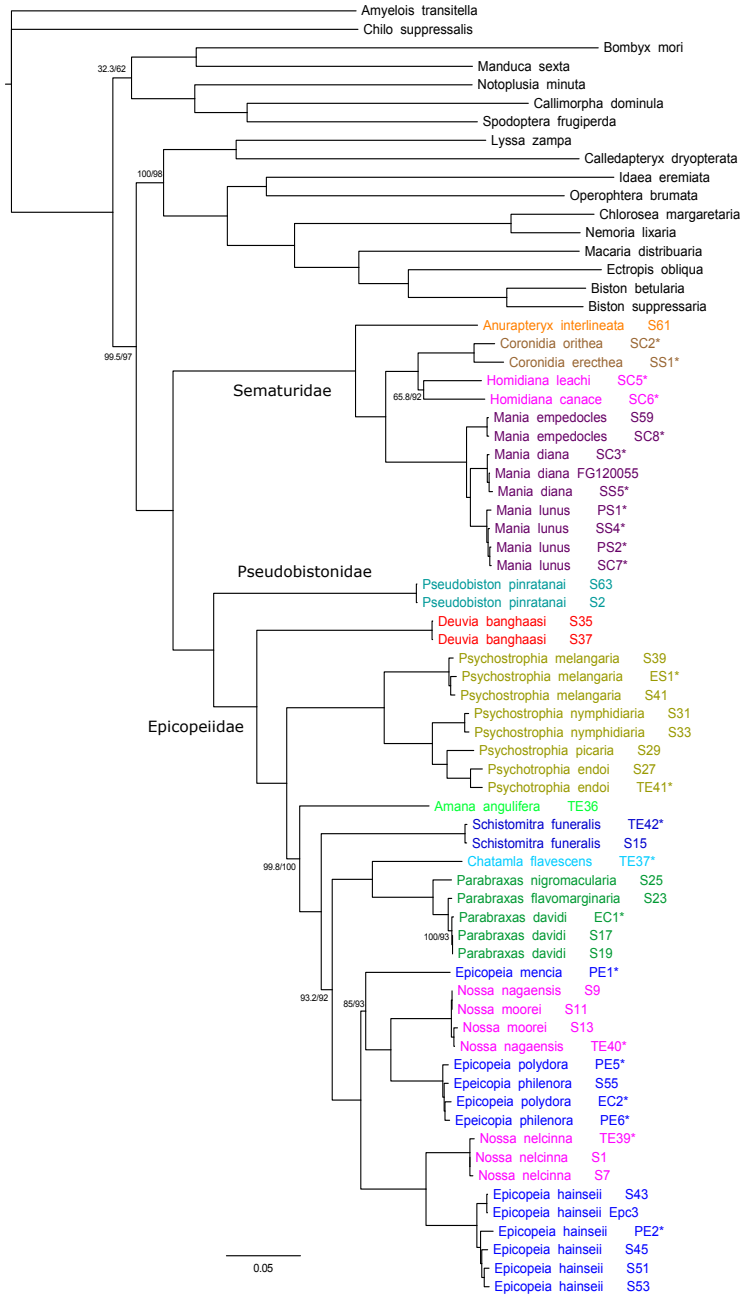
Call, E. et al. 2021: Museomics: phylogenomics of the moth family Epicopeiidae (Lepidoptera) using target enrichment. Insect Systematics & Diversity 5(2): 6 doi:10.1093/isd/ixaa021

Resulting tree

378 genes chosen with maximal coverage



Whole genome sequences added to target enrichment data



Pitfalls

- Highly fragmented genomes!
- Functional genomics maybe not possible
  - If only exons are assembled properly, assigning correct exons to a single gene in a gene family might be impossible
- What looks like a good library prep might turn out to be a good library prep of a contaminant