When AI meets structural genome annotation

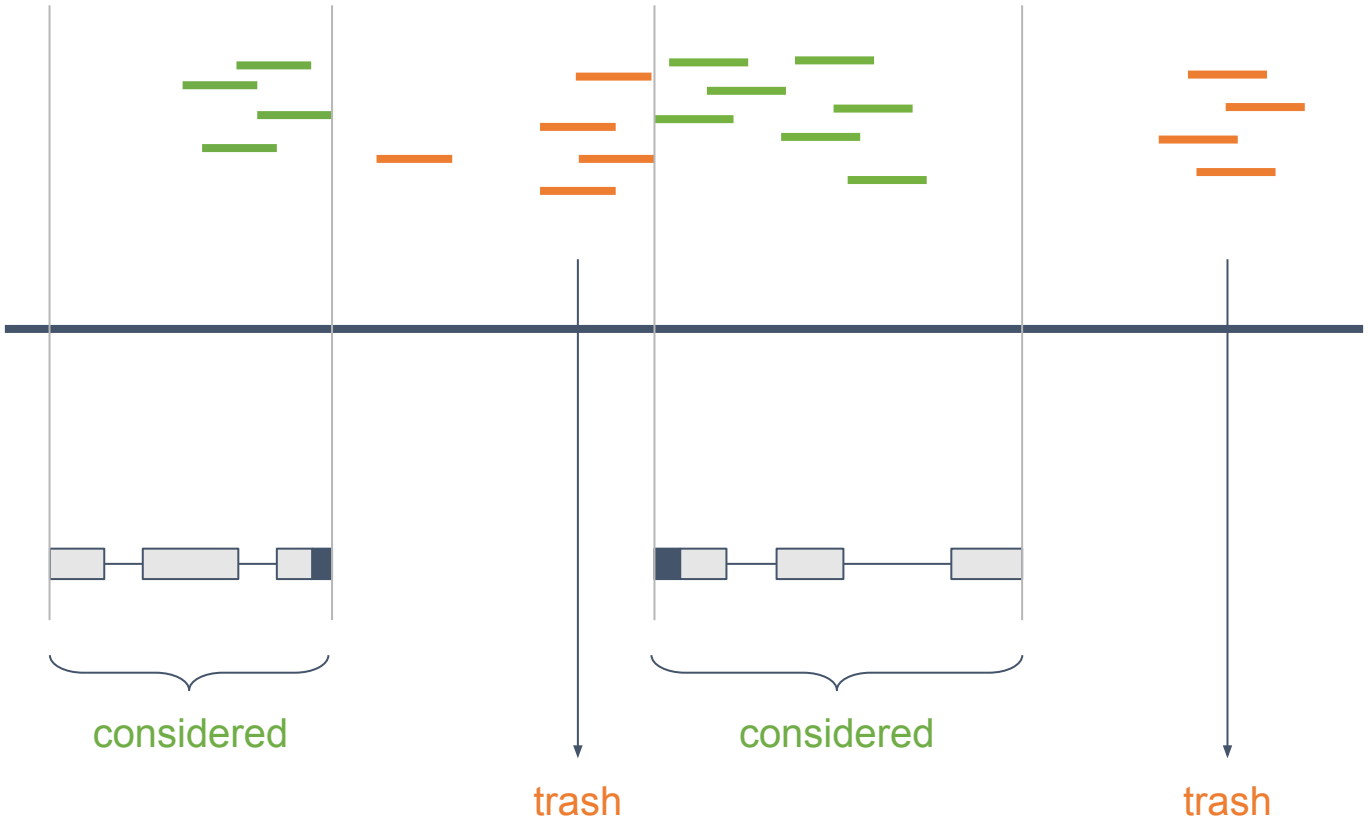
Lessons from Helixer in *Lepidoptera*

Audrey Onfroy[1], Romuald Marin[2], Manuela López Villavicencio[3,4], Joséphine Ledamoisel[3,4], Ludvine Brajon[4], Lili Babin[4], Violaine Llaurens[4] and Sophie Lemoine[1]

[1] GenomiqueENS, Institut de biologie de l'ENS (IBENS), École normale supérieure, CNRS, INSERM, PSL, Paris, France
[2] IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, Villejuif 94800, France ; PRABI, Rhône-Alpes Bioinformatics Center, Université Lyon 1, Villeurbanne CEDEX 69622, France
[3] Institut de Systématique, Évolution et Biodiversité (ISYEB), CNRS, MNHN, SU, EPHE, UA, Paris, France
[4] Centre Interdisciplinaire de Recherche en Biologie (CIRB), CNRS, INSERM, Collège de France, Paris, France

JOBIM 2026 - L'annotation structurale des gènes dans les génomes Eucaryotes est-elle toujours une problématique ?

Structural genome annotation



Sequencing reads (FASTQ)

Mapping (BAM)

Reference genome (FASTA)

Structural annotation (GFF)

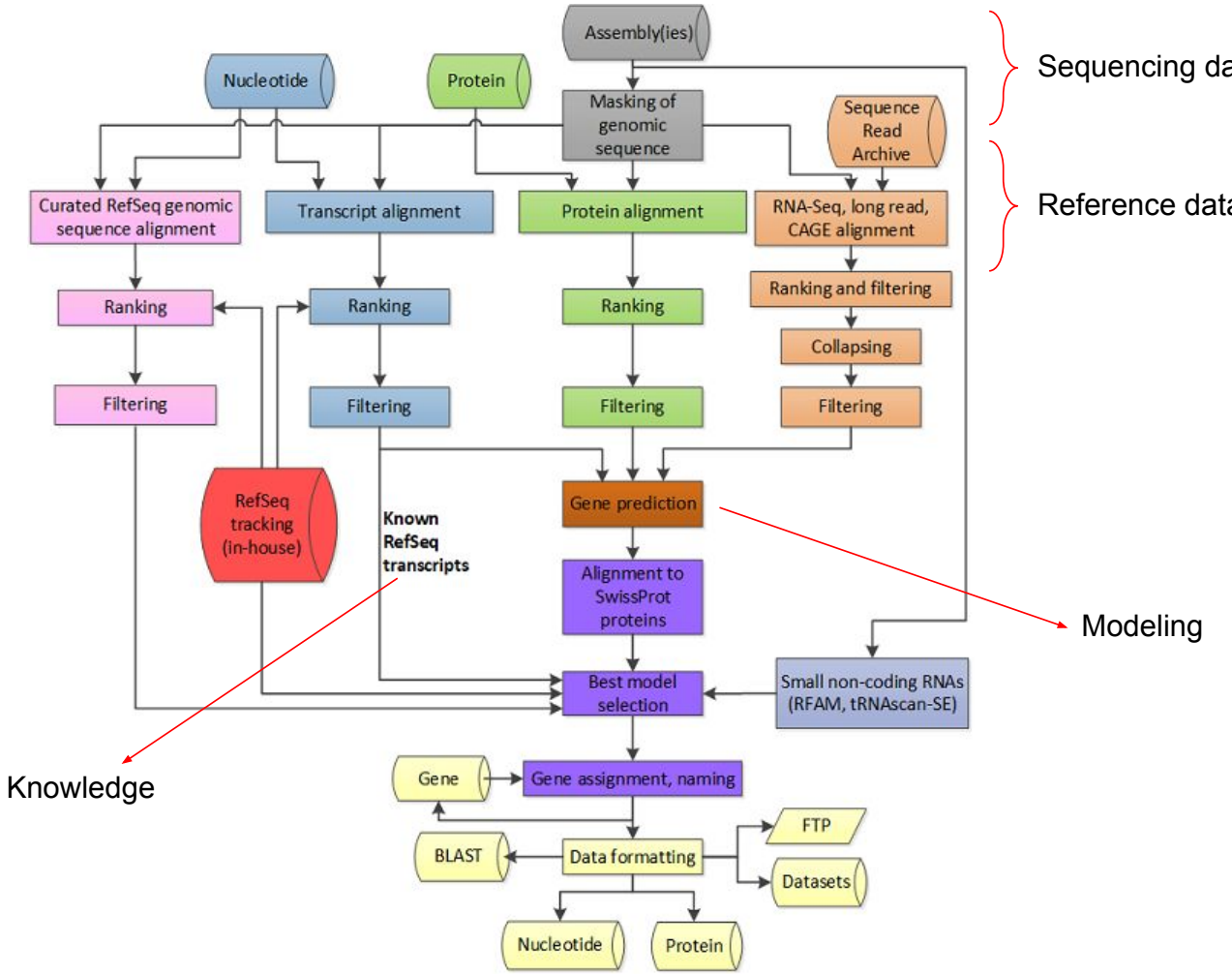
Quantification (CSV)

Most of the **downstream analyses** are based on the count data.

The quality of the structural annotation determines the relevance of the downstream analyses.

How do we obtain a structural genome annotation?

- **BRAKER** pipeline
  - ❖ Various input data, including the query FASTA file
  - ❖ Statistical models
  - ⇒ **Stat–based** annotation
- **Long-read sequencing data**–based approaches
  - ❖ Example: RNA-Bloom, IsoQuant, **Egzotek** pipeline
  - ❖ Various isoforms annotated
  - ❖ Selection bias based on the capture (eg. polyA)
  - ⇒ **LR–based** annotation
- **Deep learning** approaches
  - ❖ One input: FASTA file
  - ❖ Example: Tiberius, Annevo, **Helixer** …
  - ⇒ **Helixer–based** annotation

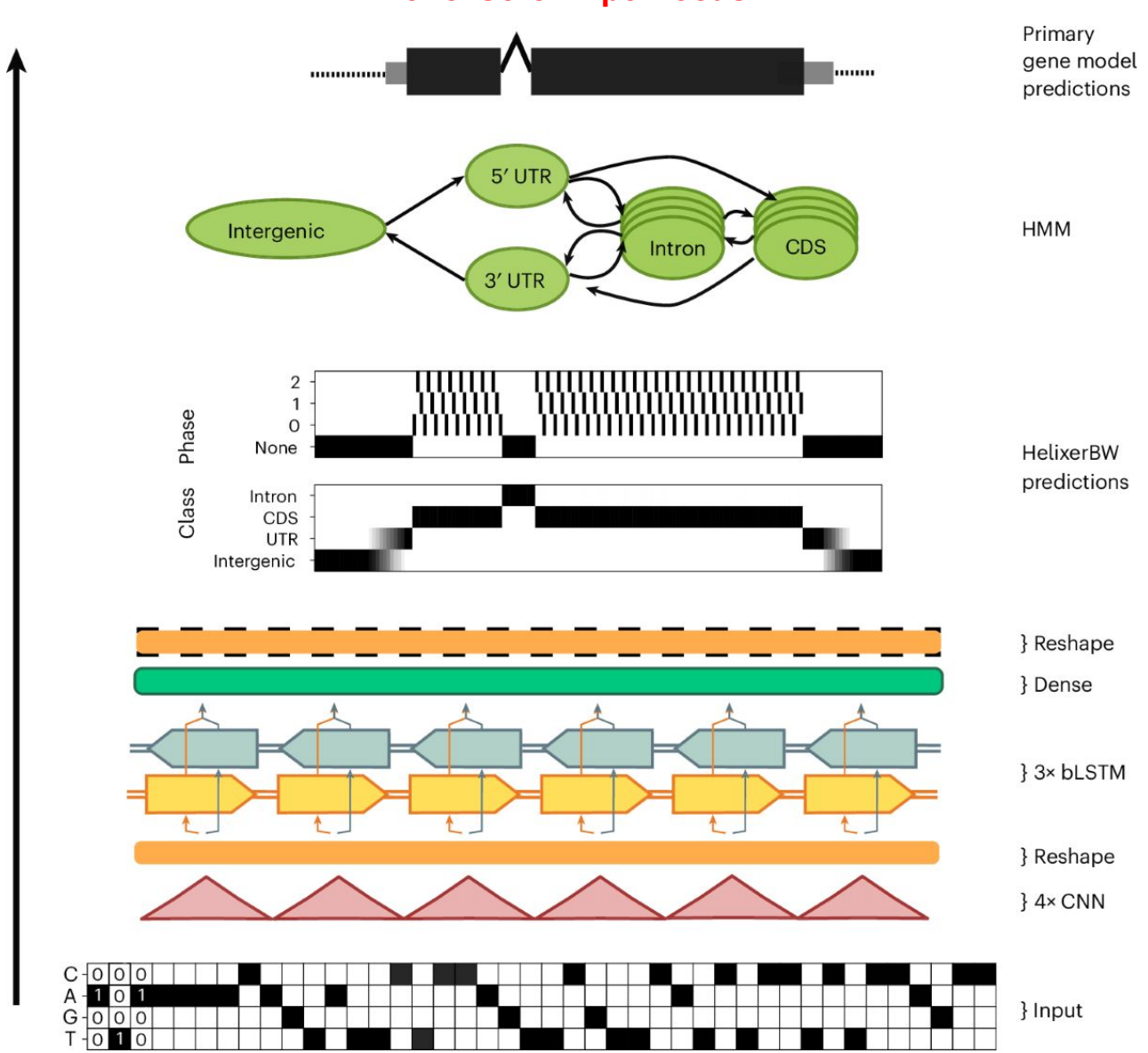


The NCBI Eukaryotic Genome Annotation Pipeline
https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/process/

How does Helixer work?

- **Default models** → can be fine-tuned
  - Fungi: 4 submodels
  - Invertebrate: 6 submodels
  - Land plant: 8 submodels
  - Vertebrate: 6 submodels

- **Process**
  - **Pattern detection**: Multiple stacked Convolutional Neural Network to detect local patterns in the DNA sequences
  - **Context integration**: Multiple stacked bidirectional Long Short-Term Memory to connect the local patterns together
  - **Annotation**: Base-wise predictions to assign each base to an initial annotation (CDS, introns, **UTR**, intergenic)
  - **Processing**: Hidden Markov Model post-processing to ensure coherent gene structure, enforcing biological rules



Helixer process
Holst *et al*. (2025, Nature Methods)

Can Helixer be used to improve the quantification of SR RNA-Sequencing data?

- **Species of interest**: *Morpho helenor*
  - Annotation provided by the lab: **Stat–based**
  - Annotation obtained from another tissue: **LR–based**
- Aim 1 – Predict annotations using Helixer: **Helixer–based**
  - **Fine-tune** a default model on *Lepidoptera*
    - ➢ Identify species for the training / validation
    - ➢ Set up a pipeline
  - **Prediction annotations**
    - ➢ Set up a pipeline
- Aim 2 – **Evaluate the relevance of the annotations**
  - **gffcompare** against the Stat–based annotation
  - **BUSCO** analysis
  - **Quantification** of SR RNA-Seq data

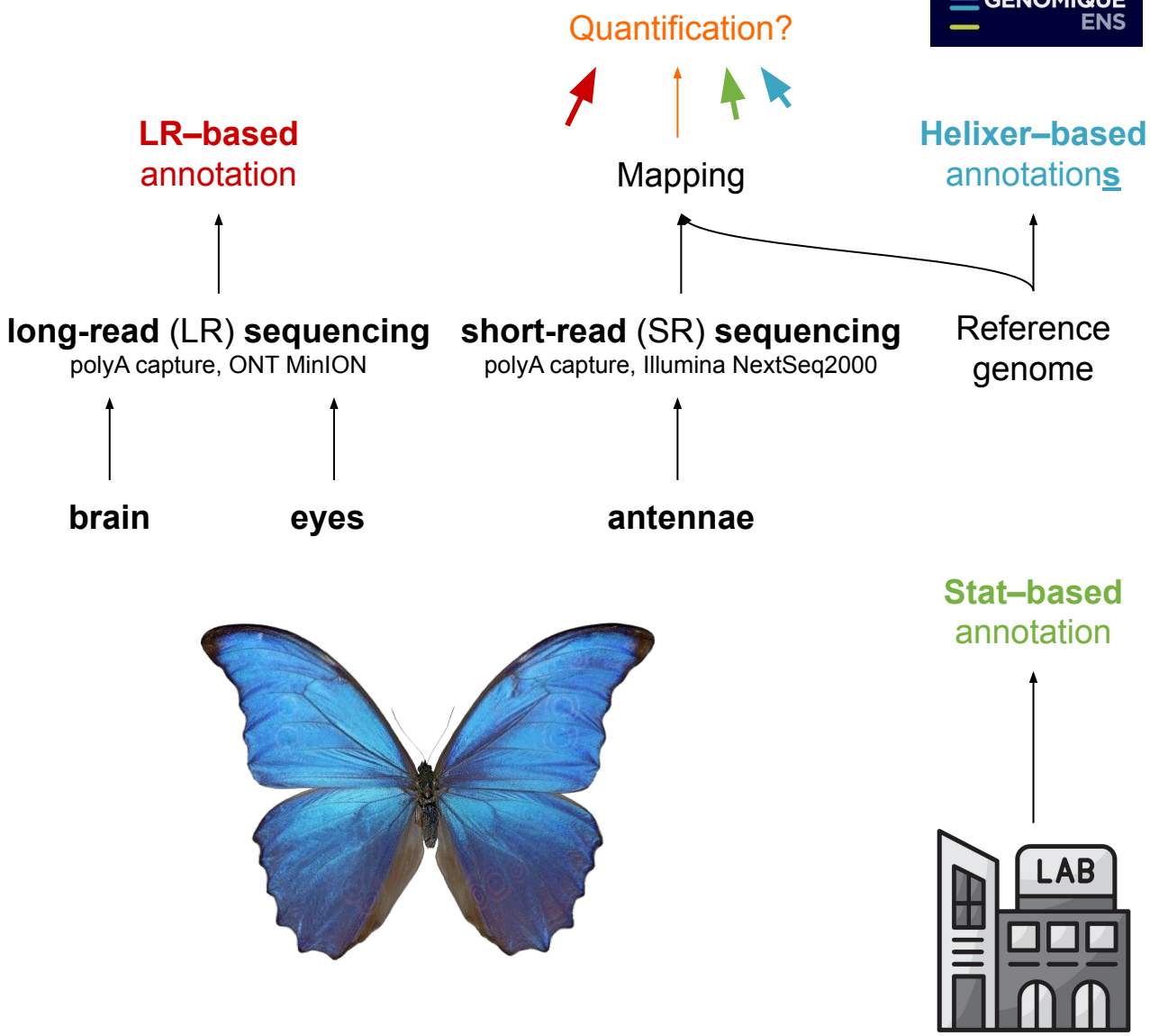


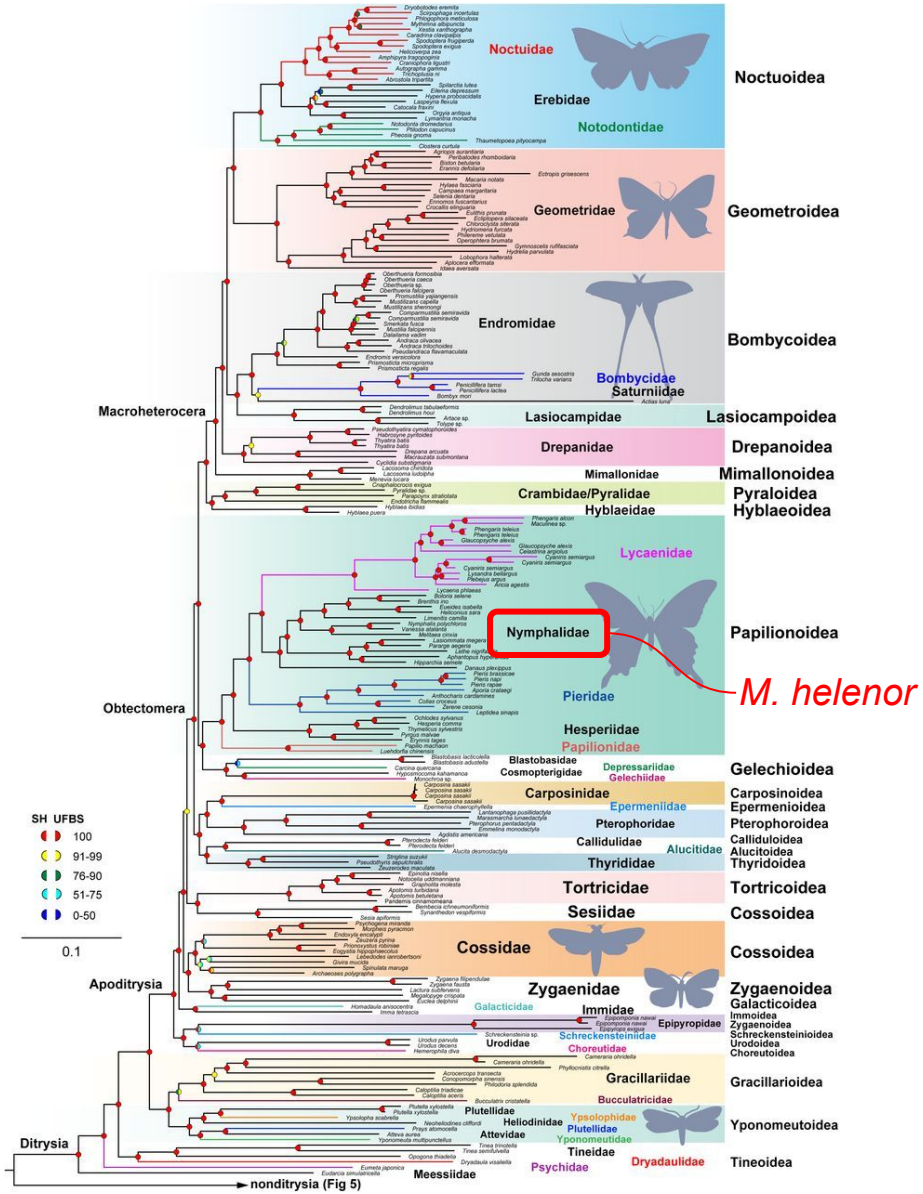
Table of Contents

- **Species of interest**: *Morpho helenor*
  - Annotation provided by the lab: **Stat–based**
  - Annotation obtained from another tissue: **LR–based**

- Aim 1 – Predict annotations using Helixer: Helixer–based
  - **Fine-tune** a default model on *Lepidoptera*
    - **Identify species for the training / validation**
    - Set up a pipeline
  - **Prediction annotations** using the default and fine-tuned models
    - Set up a pipeline

- Aim 2 – **Evaluate the relevance of the annotations**
  - **gffcompare** against the Stat–based annotation
  - **BUSCO** analysis
  - **Quantification** of SR RNA-Seq data

Identification of well-annotated *Lepidoptera* genomes



NCBI assemblies database

only *Lepidoptera*

4522 *Lepidoptera* assemblies

with an annotation
Based on the Busco/RefSeqAnnotationRelease element being non empty

64 *Lepidoptera* annotated assemblies

the assembly is of type “Chromosome”
Not Contig, not Scaffolds

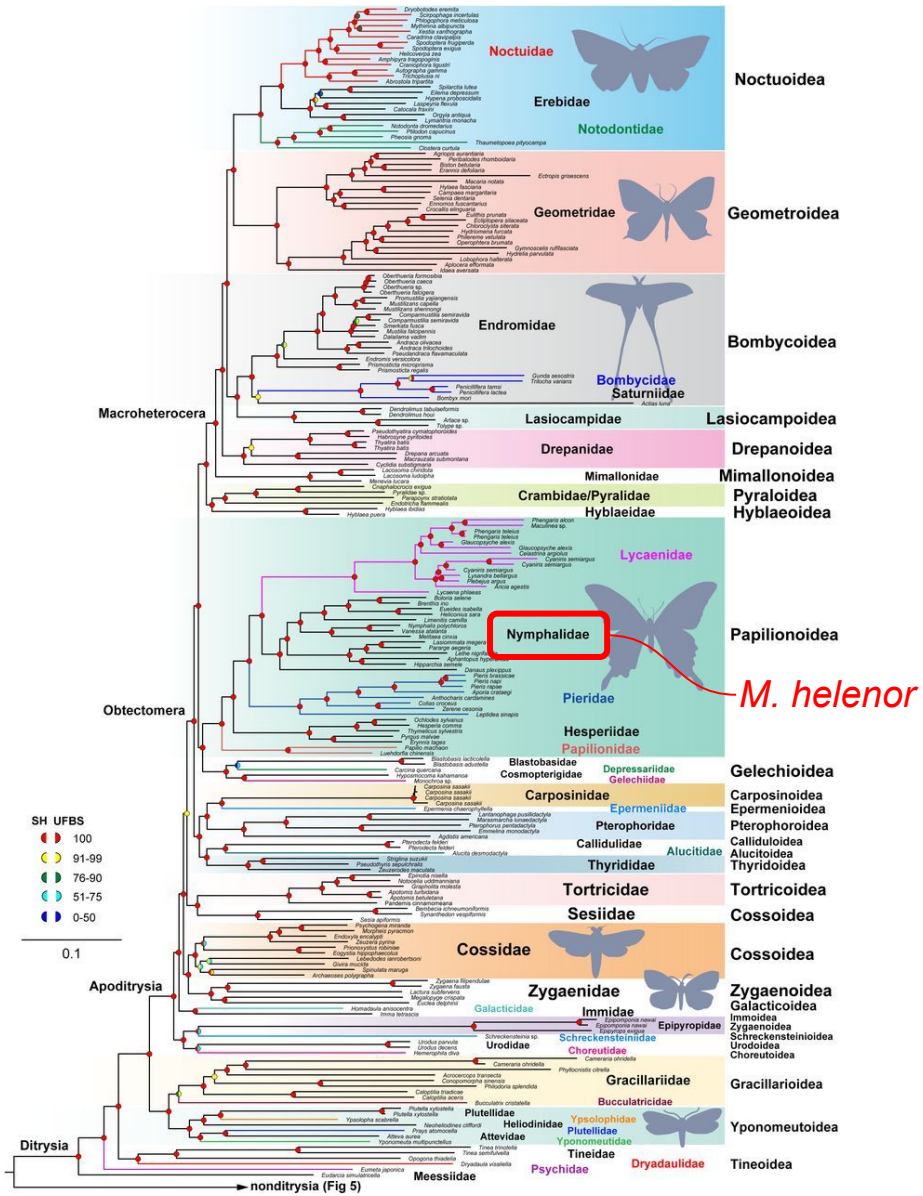
43 *Lepidoptera* annotated assemblies
of type “Chromosome”

keep the latest release per species

37 *Lepidoptera* assemblies

21 + 3 *Lepidoptera* assemblies

Overview of the 37 species of interest



M. helenor

Erebidae Anticarsia gemmatalis

Noctuidae
Helicoverpa armigera
Helicoverpa zea
Spodoptera frugiperda
Spodoptera litura
Trichoplusia ni

Bombycidae Bombyx mori
Sphingidae Manduca sexta
Crambidae Ostrinia nubilalis
Pyralidae Plodia interpunctella
Lycaenidae Aricia agestis*
Papilionidae Papilio machaon

Nymphalidae
Bicyclus anynana
Danaus plexippus
Maniola hyperantus*
Maniola jurtina
Melitaea cinxia
Nymphalis io
Pararge aegeria
Vanessa atalanta
Vanessa cardui
Vanessa tameamea

Family
Species
Available in the DTOL with two releases*

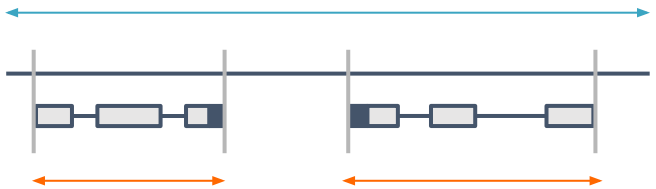
Pieridae
Colias croceus
Leptidea sinapis
Pieris brassicae
Pieris napi*
Pieris rapae
Zerene cesonia

Gelechiidae Pectinophora gossypiella

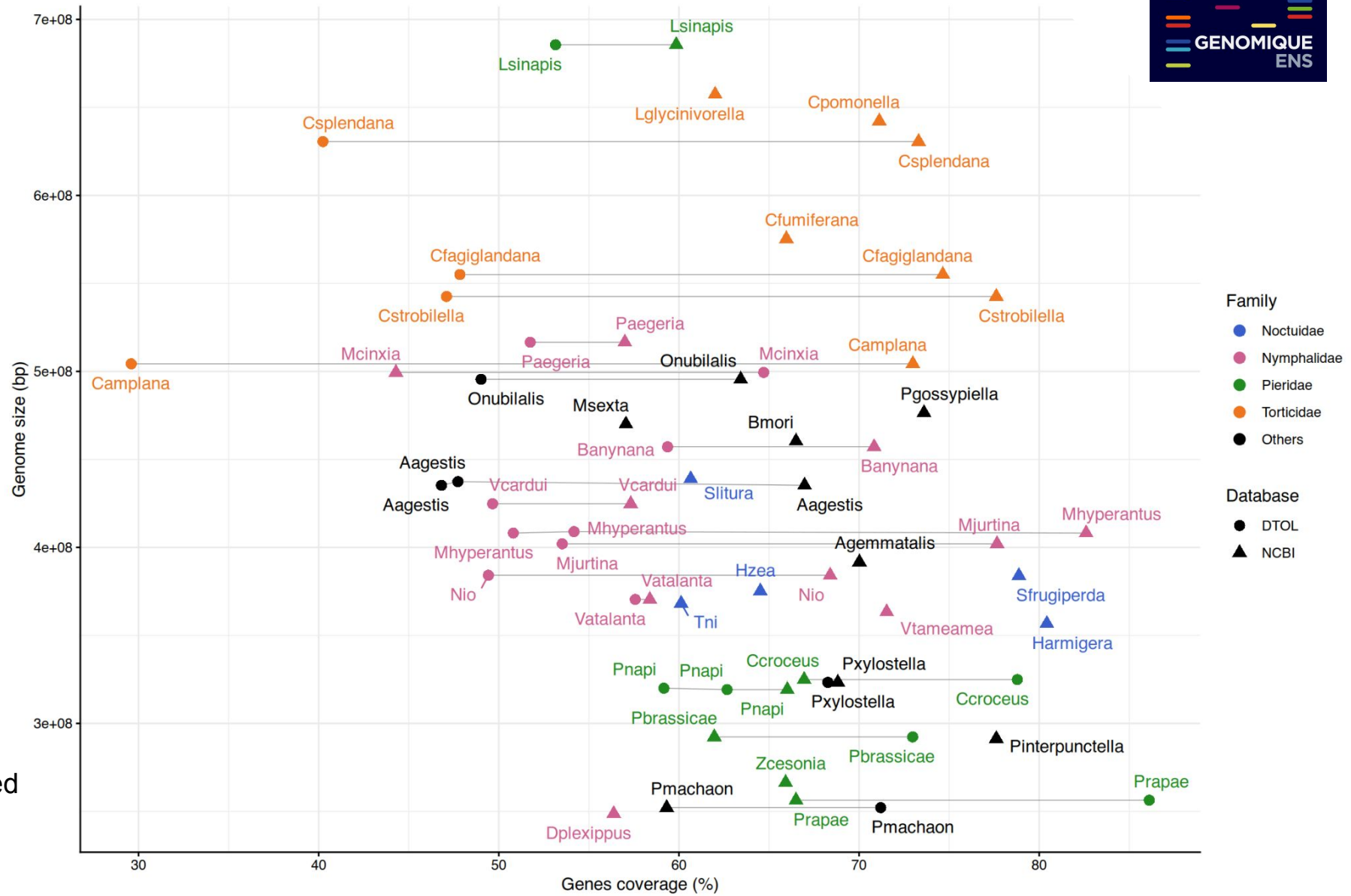
Tortricidae
Choristoneura fumiferana
Cydia amplana
Cydia fagiglandana
Cydia pomonella
Cydia splendana
Cydia strobilella
Leguminivora glycinivorella

Plutellidae Plutella xylostella

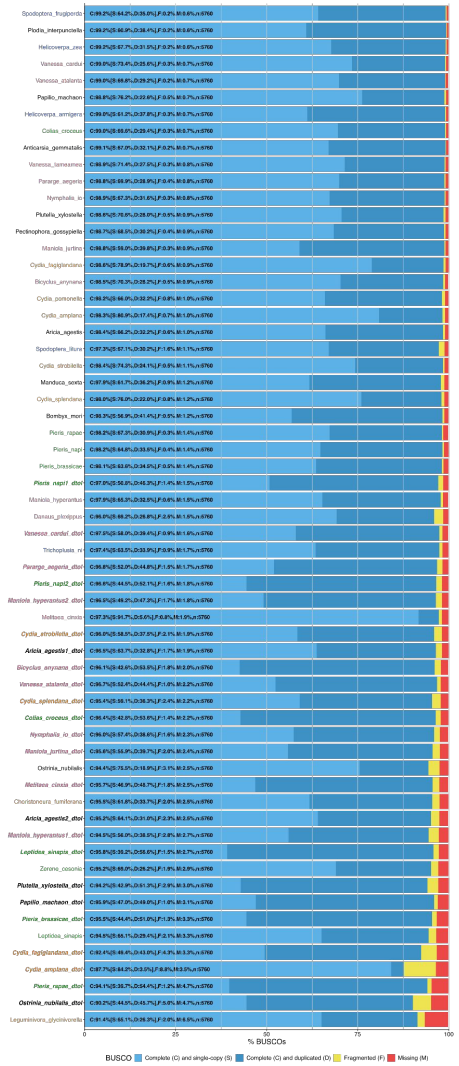
Part of the genome annotated as “gene”



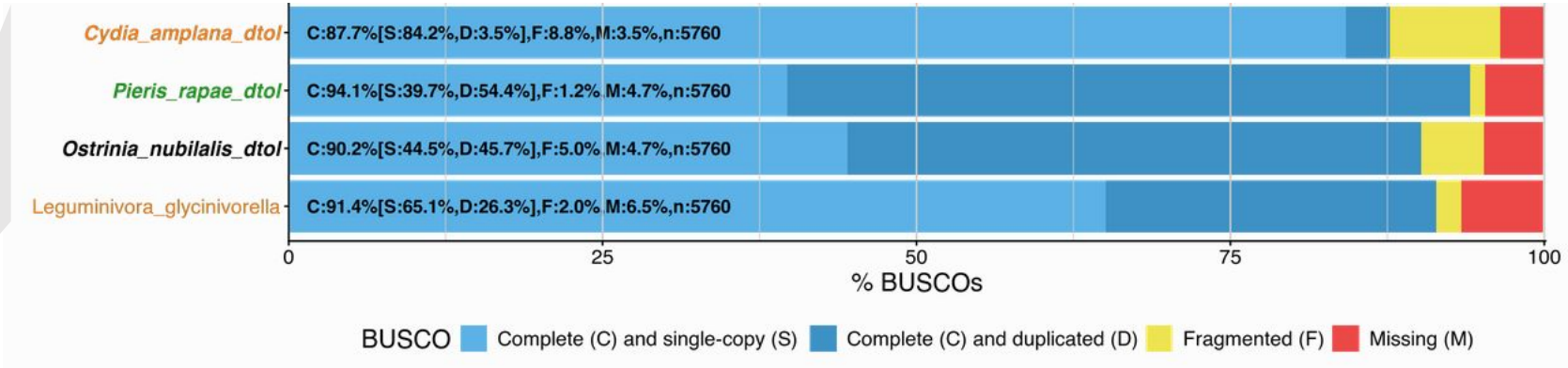
- Genome assemblies have the same size between the NCBI and the DTOL databases.
- In the DTOL compared to NCBI, for the species considered:
  - Either genes are shorter
  - Or there are less genes
- To fine-tune the Helixer model, we have decided to keep the annotated genomes from the **NCBI only**.



Identification of the best annotated species using BUSCO analysis



- BUSCO genes = orthologs present as a **single-copy** in more than 90% of the species (**universal**)
- We expect a good genome annotation to contain these genes.
- BUSCO analysis:
  - Subset the genome FASTA file for the sequences annotated in the GTF
  - Assess the presence of the BUSCO genes



BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs
Simão *et al.* (2019, Bioinformatics)

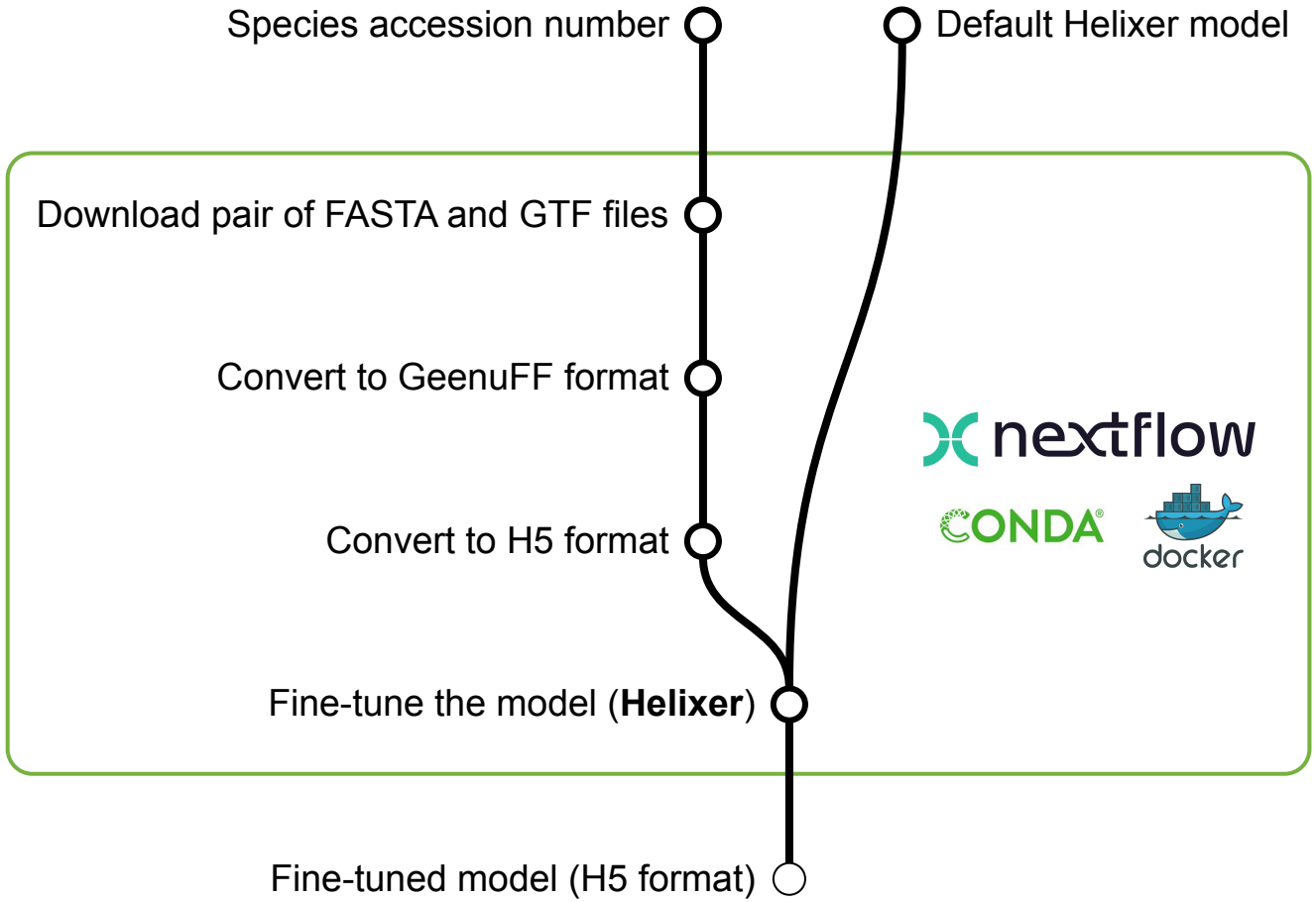
Table of Contents

- **Species of interest**: *Morpho helenor*
  - Annotation provided by the lab: **Stat–based**
  - Annotation obtained from another tissue: **LR–based**

- Aim 1 – Predict annotations using Helixer: Helixer–based
  - **Fine-tune** a default model on *Lepidoptera*
    - Identify species for the training / validation
    - **Set up a pipeline**
  - **Prediction annotations** using the default and fine-tuned models
    - Set up a pipeline

- Aim 2 – **Evaluate the relevance of the annotations**
  - **gffcompare** against the Stat–based annotation
  - **BUSCO** analysis
  - **Quantification** of SR RNA-Seq data

Fine-tuning of a default Helixer model



Obtention of **4 fine-tuned models**:

- Based on the invertebrate model **0100**
  - Validation with *A. agestis* and *P. rapae*
- Based on the invertebrate model **0200**
  - Validation with *A. agestis* and *P. rapae*
  - Validation with *A. agestis*, *B. mori and S. frugiperda*
  - Validation with *S. frugiperda*, *P. interpunctella*, *V. cardui*, *C. croceus* and *M. cinxia*

Resource usage for one execution:

- Duration: 3-4 days
- CPU hours: 122

Collaboration with **Romuald MARIN** (IFB-core, Lyon)

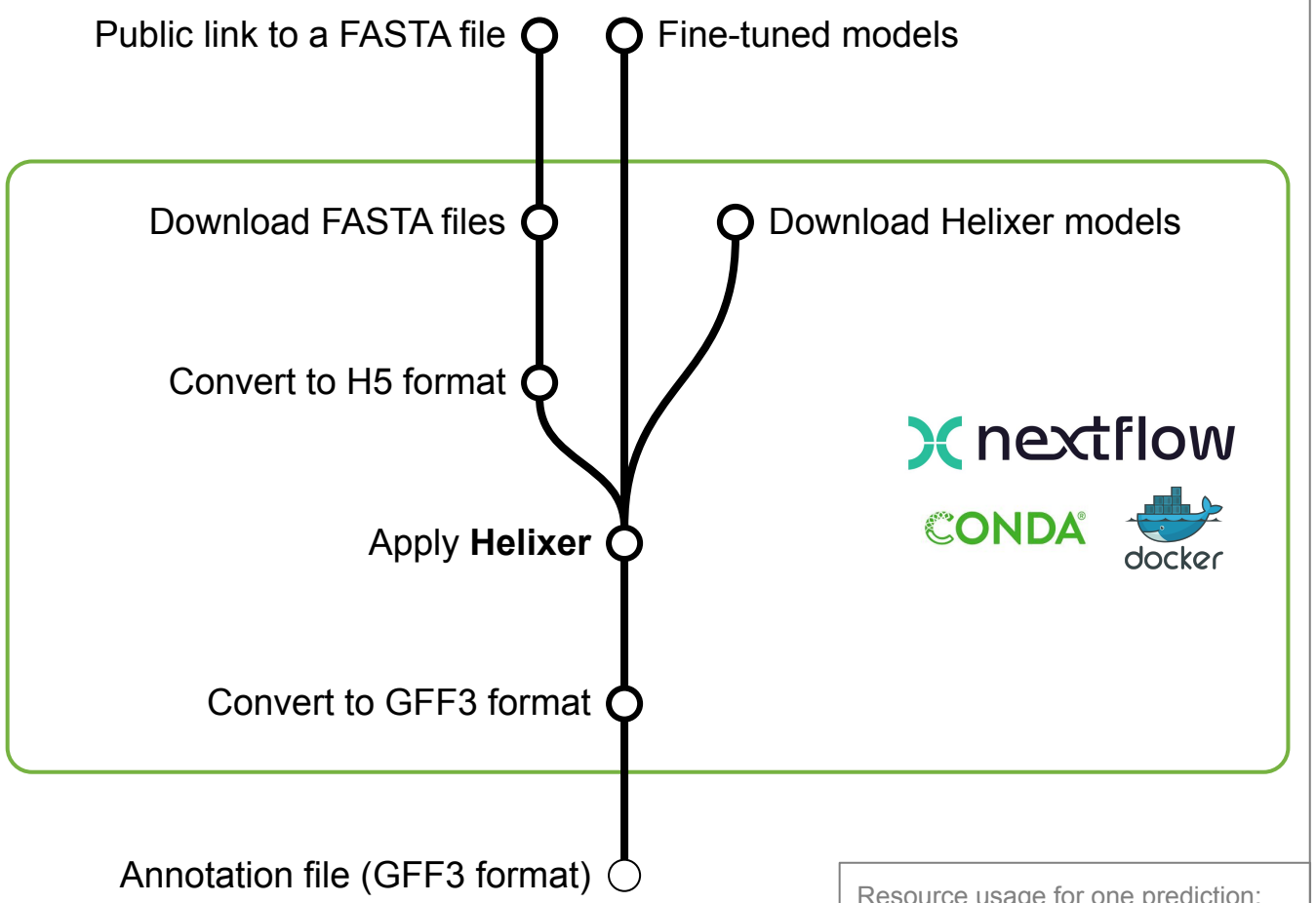
Table of Contents

- **Species of interest**: *Morpho helenor*
  - Annotation provided by the lab: **Stat–based**
  - Annotation obtained from another tissue: **LR–based**

- Aim 1 – Predict annotations using Helixer: Helixer–based
  - **Fine-tune** a default model on *Lepidoptera*
    - Identify species for the training / validation
    - Set up a pipeline
  - **Prediction annotations** using the default and fine-tuned models
    - **Set up a pipeline**

- Aim 2 – **Evaluate the relevance of the annotations**
  - **gffcompare** against the Stat–based annotation
  - **BUSCO** analysis
  - **Quantification** of SR RNA-Seq data

Set up a pipeline to predict annotations using Helixer



Resource usage for one prediction:

- Duration: ~ 2 hours
- CPU hours: 1.5

Prediction of **10 annotations for *M. helenor***:

- Using the default models, **6 annotations**

invertebrate_v0.3_a_0300
invertebrate_v0.3_a_0400
invertebrate_v0.3_a_0500
invertebrate_v0.3_a_0600
invertebrate_v0.3_m_0100
o invertebrate_v0.3_m_0200

- Using the fine-tuned models, **4 annotations**

lepidoptera_model_0100_Aagestia_Prapae
lepidoptera_model_0200_Aagestia_Bmori_Sfrugiperda
lepidoptera_model_0200_Aagestia_Prapae
i lepidoptera_model_0200_Sfrugiperda_Pinter_Vcardui_Ccroceus_Mcinxia

More annotations:

- Already available annotations, **2 annotations**

^ Morpho_helenor_OR
" Morpho_helenor_final_2601

- Combined annotations, **4 annotations**

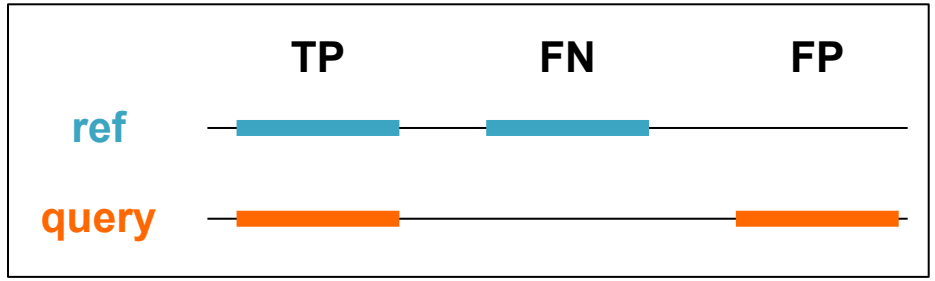
Ö Morpho_helenor_final_2601_AND_invertebrate_v0.3_m_0200
Ï Morpho_helenor_final_2601_AND_lepidoptera_model_0200_Sfrugiperda_Pinter_Vcardui_Ccroceus_Mcinxia
Ô Morpho_helenor_OR_AND_invertebrate_v0.3_m_0200
Î Morpho_helenor_OR_AND_lepidoptera_model_0200_Sfrugiperda_Pinter_Vcardui_Ccroceus_Mcinxia

Table of Contents

- **Species of interest**: *Morpho helenor*
  - Annotation provided by the lab: **Stat–based**
  - Annotation obtained from another tissue: **LR–based**
- Aim 1 – **Predict annotations using Helixer**: **Helixer–based**
  - **Fine-tune** a default model on *Lepidoptera*
    - Identify species for the training / validation
    - Set up a pipeline
  - **Prediction annotations** using the default and fine-tuned models
    - Set up a pipeline
- Aim 2 – Evaluate the relevance of the annotations
  - **gffcompare against the Stat–based annotation**
  - **BUSCO** analysis
  - **Quantification** of SR RNA-Seq data

Gffcompare – Principle

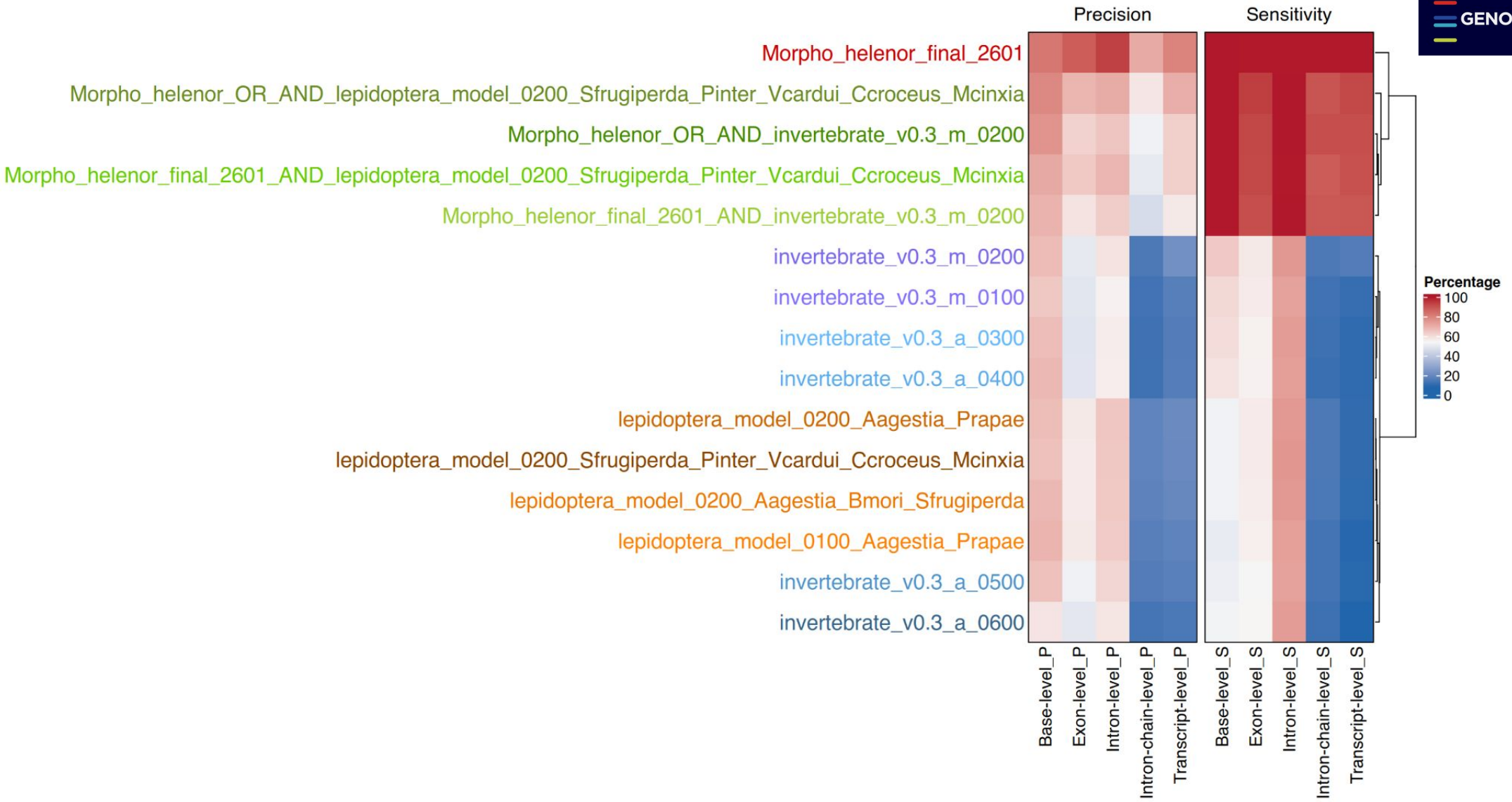
Sensitivity	Precision
TP / (TP + FN)	**TP / (TP + FP)**
% of reference seq found in the query	% of query seq described in the ref



- Evaluation at 5 levels:
  - **Base**: Exon overlapping at a nucleotide-level
  - **Exon**: Exon boundaries are correct
  - **Intron**: Individual splice junctions are correct
  - **Intron-chain**: Entire splice structure is correct
  - **Transcript**: Combine **Exon** and **Intron-chain**

```
/gff_comp_predictions_Morpho_helenor_invertebrate_v0.3_a_0600_VS_Morpho_helenor_OR$ cat *stats
# gffcompare v0.12.10 | Command line was:
#gffcompare predictions_Morpho_helenor_invertebrate_v0.3_a_0600.gff3 -r Morpho_helenor_OR.gff -o
 gff_comp_predictions_Morpho_helenor_invertebrate_v0.3_a_0600_VS_Morpho_helenor_OR/gffcomp -e 10
0 -d 100
#

#= Summary for dataset: predictions_Morpho_helenor_invertebrate_v0.3_a_0600.gff3
#     Query mRNAs :   16793 in   16793 loci  (15443 multi-exon transcripts)
#            (0 multi-transcript loci, ~1.0 transcripts per locus)
# Reference mRNAs :   29756 in   29666 loci  (17324 multi-exon)
# Super-loci w/ reference transcripts:    13214
#----------------| Sensitivity | Precision  |
        Base level:    52.7     |    58.1    |
        Exon level:    54.5     |    49.3    |
      Intron level:    74.3     |    57.8    |
Intron chain level:    11.7     |    13.1    |
  Transcript level:     7.9     |    14.0    |
       Locus level:     7.9     |    14.0    |

     Matching intron chains:    2029
       Matching transcripts:    2356
              Matching loci:    2356

         Missed exons:   27260/119380  ( 22.8%)
          Novel exons:   38141/131950  ( 28.9%)
        Missed introns:    8186/89624   (  9.1%)
         Novel introns:   23336/115157  ( 20.3%)
           Missed loci:   14946/29666   ( 50.4%)
            Novel loci:    2348/16793   ( 14.0%)

 Total union super-loci across all input datasets: 15562
16793 out of 16793 consensus transcripts written in gff_comp_predictions_Morpho_helenor_inverteb
rate_v0.3_a_0600_VS_Morpho_helenor_OR/gffcomp.annotated.gtf (0 discarded as redundant)
```

Gffcompare – Results



Gffcompare – Results

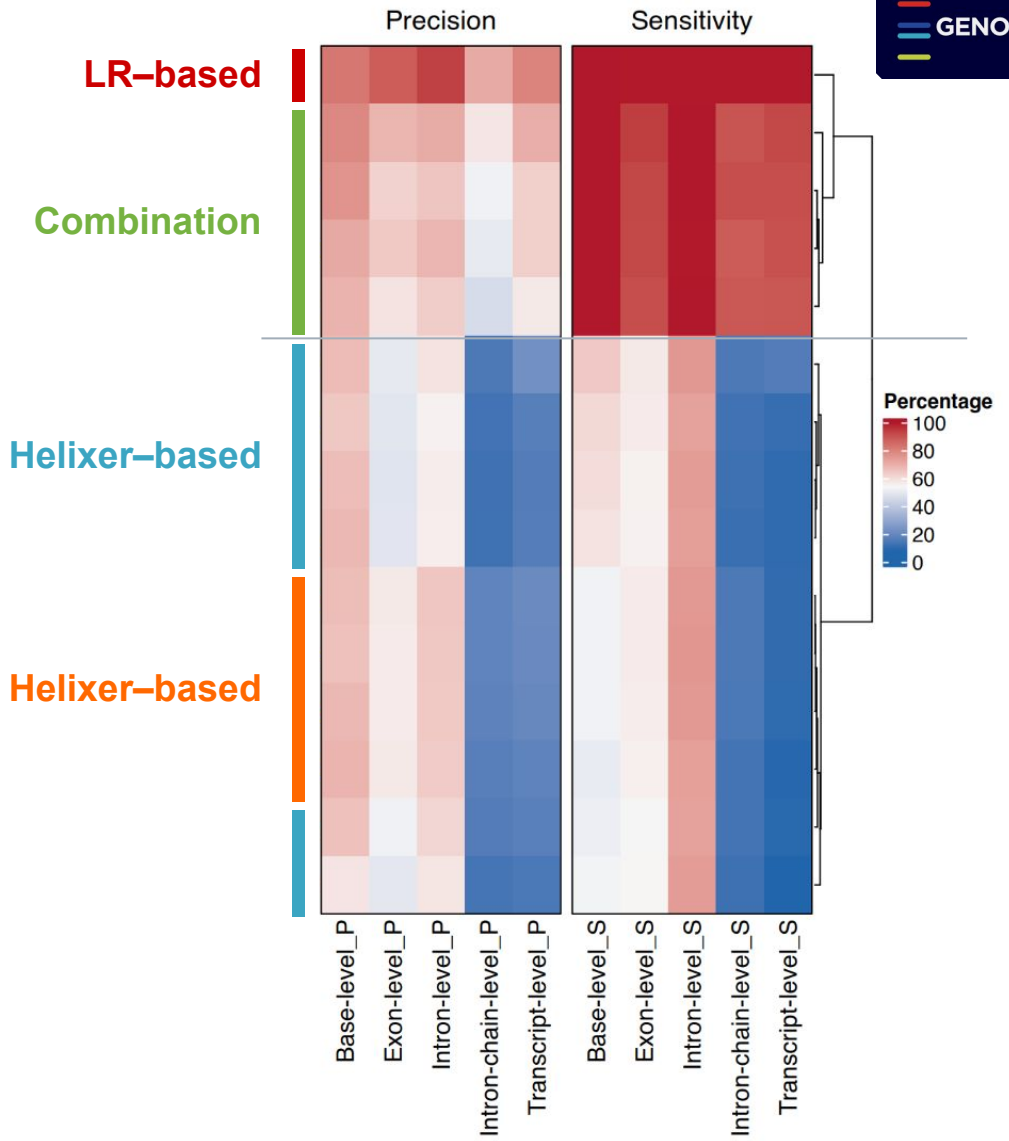
- Compared to the Stat–based annotation, manually curated:
  - The LR–based annotation is excellent.
  - Helixer–based annotations are good at a base–level but do not recapitulate the splice structure of transcripts.
  - Combining annotations does not improve the metrics.

- High **precision**: The annotation does not contain things that are not in the reference

  Related to false positive

- High **sensitivity**: The annotation contains things that are in the reference

  Related to false negative

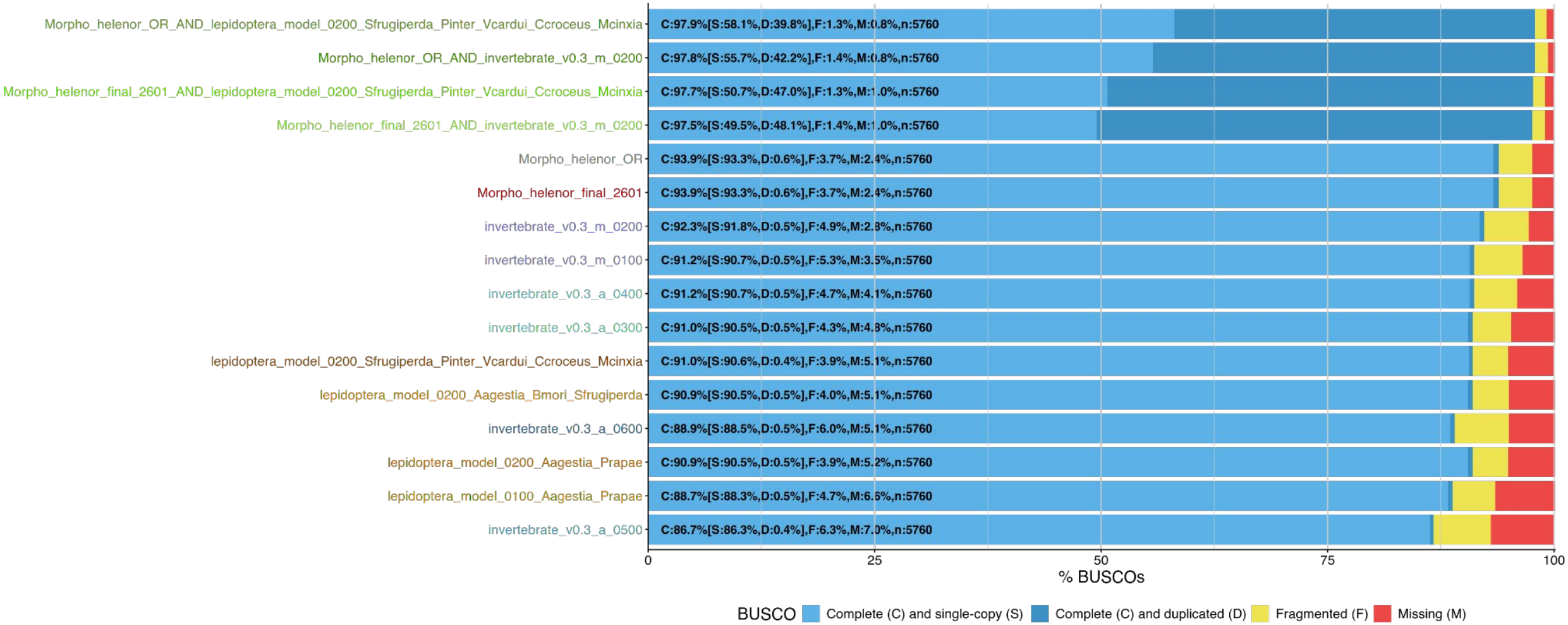


But maybe the reference itself is bad?

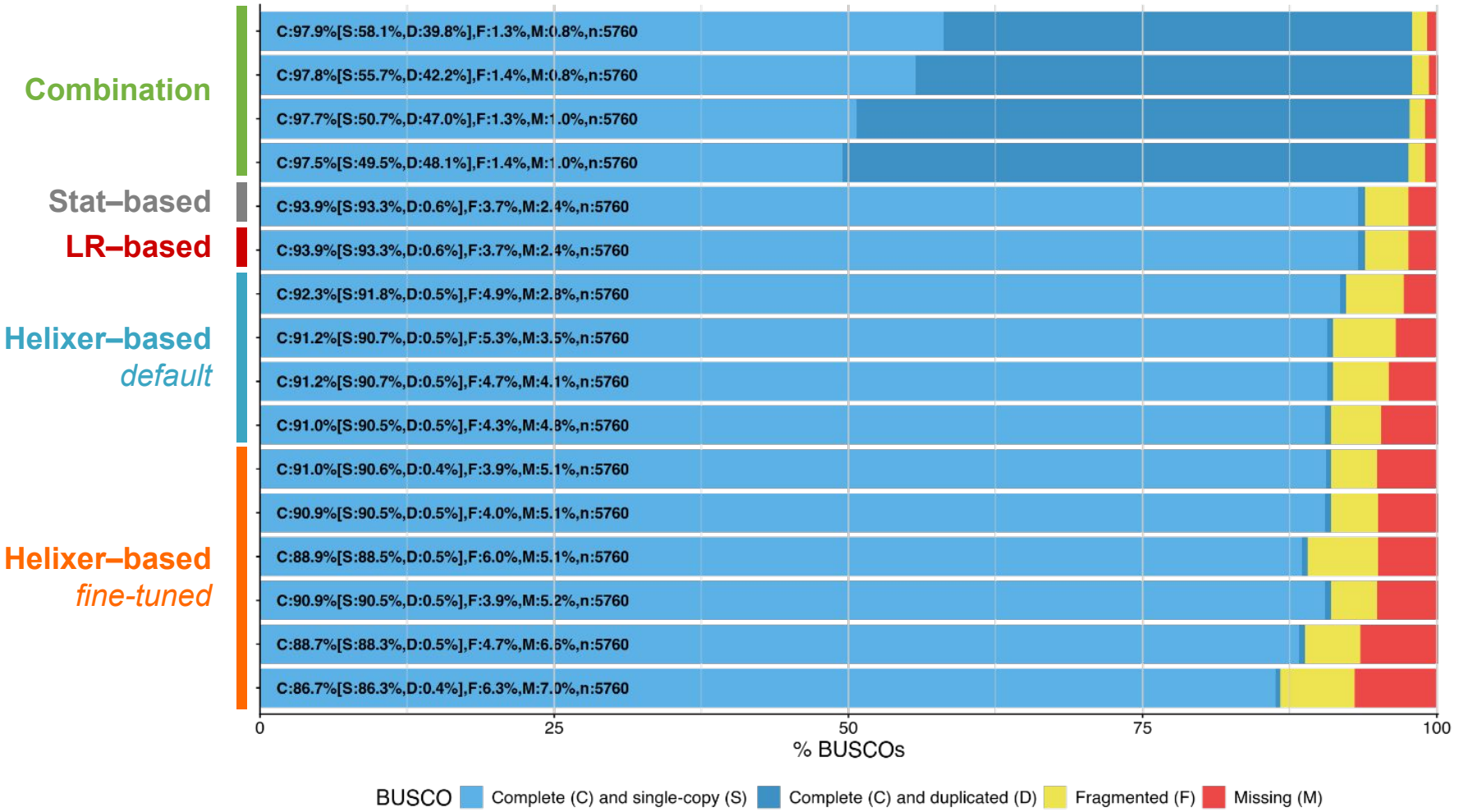
Table of Contents

- **Species of interest**: *Morpho helenor*
  - Annotation provided by the lab: **Stat–based**
  - Annotation obtained from another tissue: **LR–based**
- Aim 1 – **Predict annotations using Helixer**: **Helixer–based**
  - **Fine-tune** a default model on *Lepidoptera*
    - Identify species for the training / validation
    - Set up a pipeline
  - **Prediction annotations** using the default and fine-tuned models
    - Set up a pipeline
- Aim 2 – Evaluate the relevance of the annotations
  - **gffcompare** against the Stat–based annotation
  - **BUSCO analysis**
  - **Quantification** of SR RNA-Seq data

BUSCO analysis of the 16 annotations

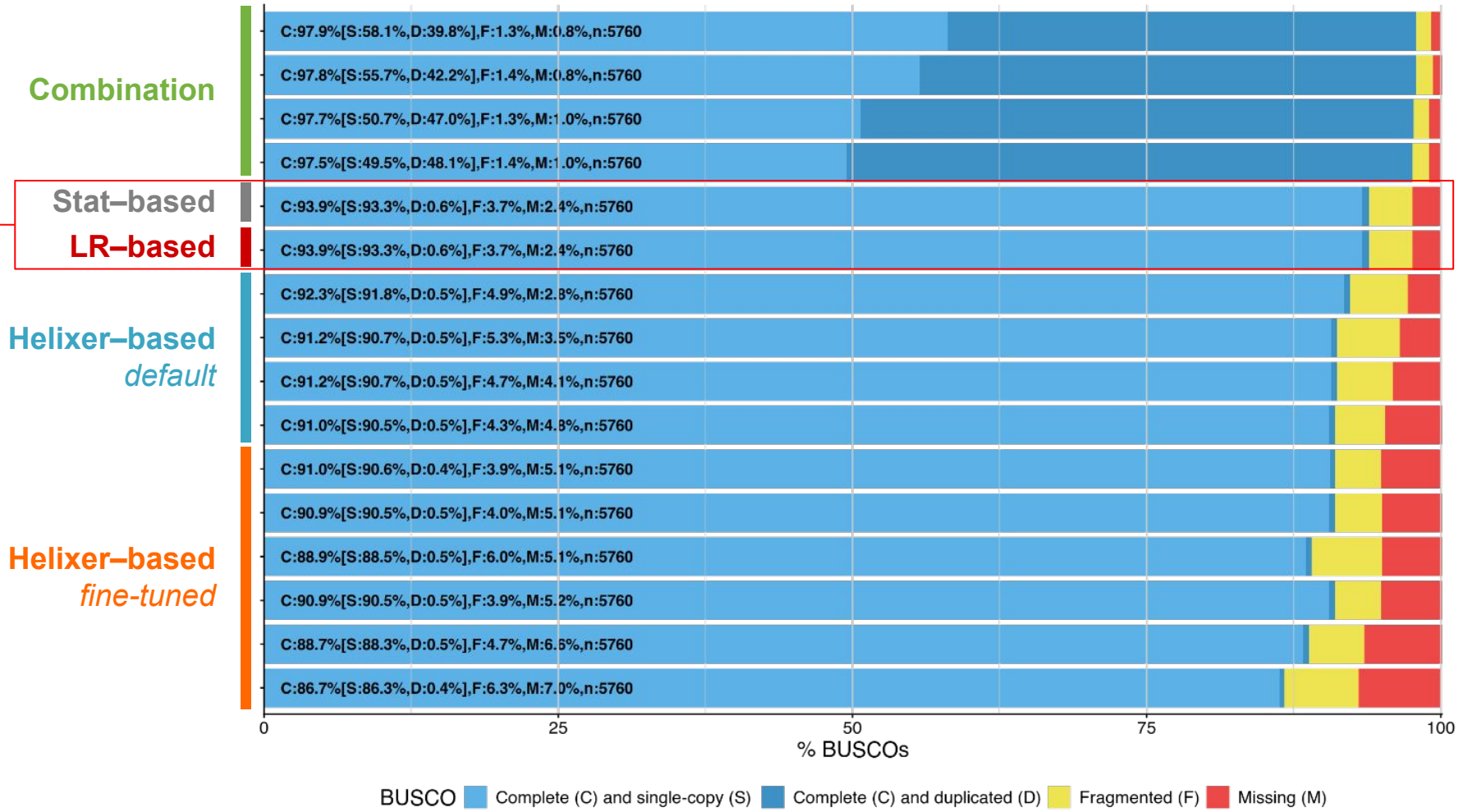


BUSCO analysis of the 16 annotations



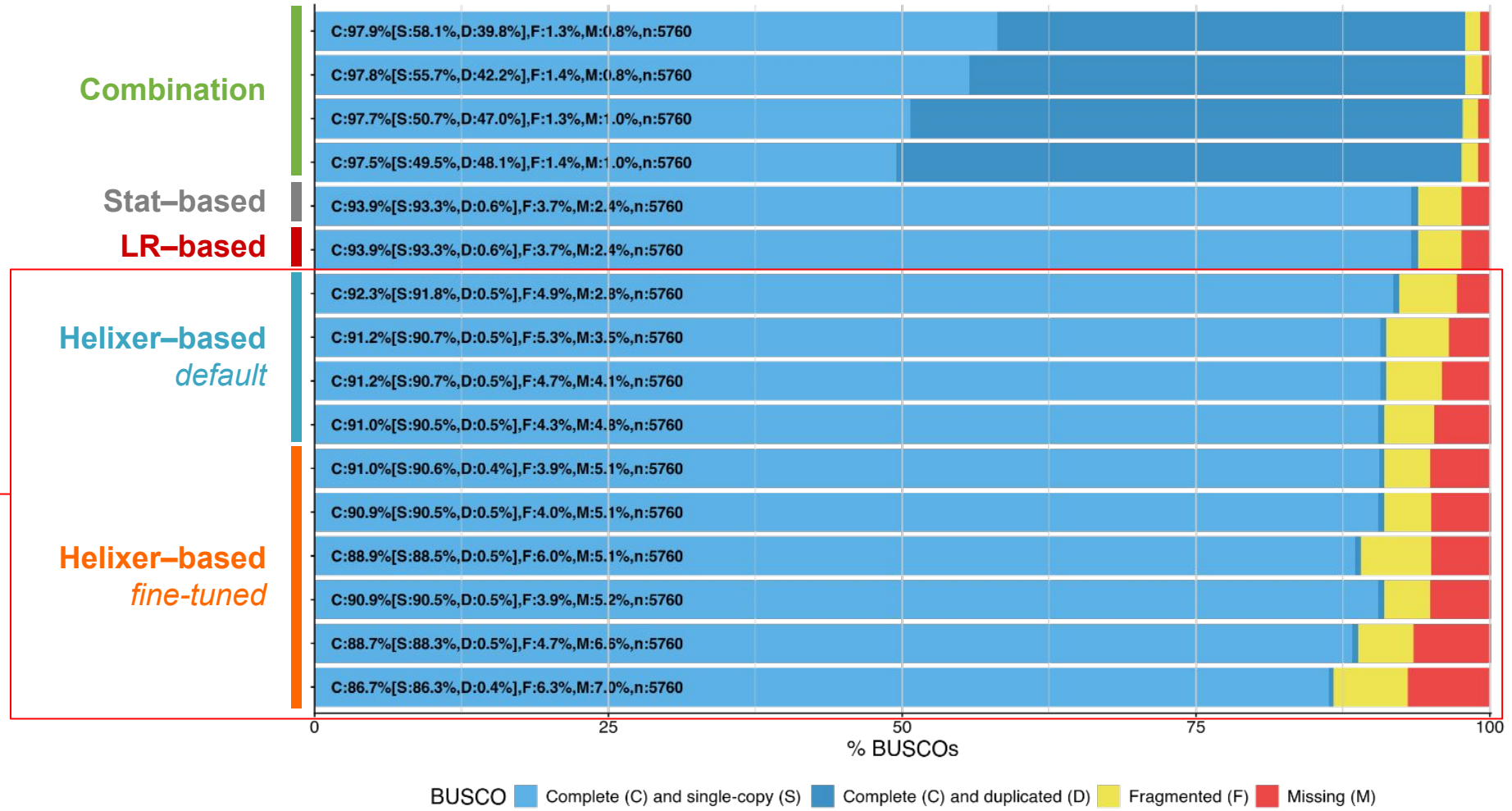
BUSCO analysis of the 16 annotations

More fragmented or missing BUSCO genes in the Helixer–based annotations compared to the data–based annotations.



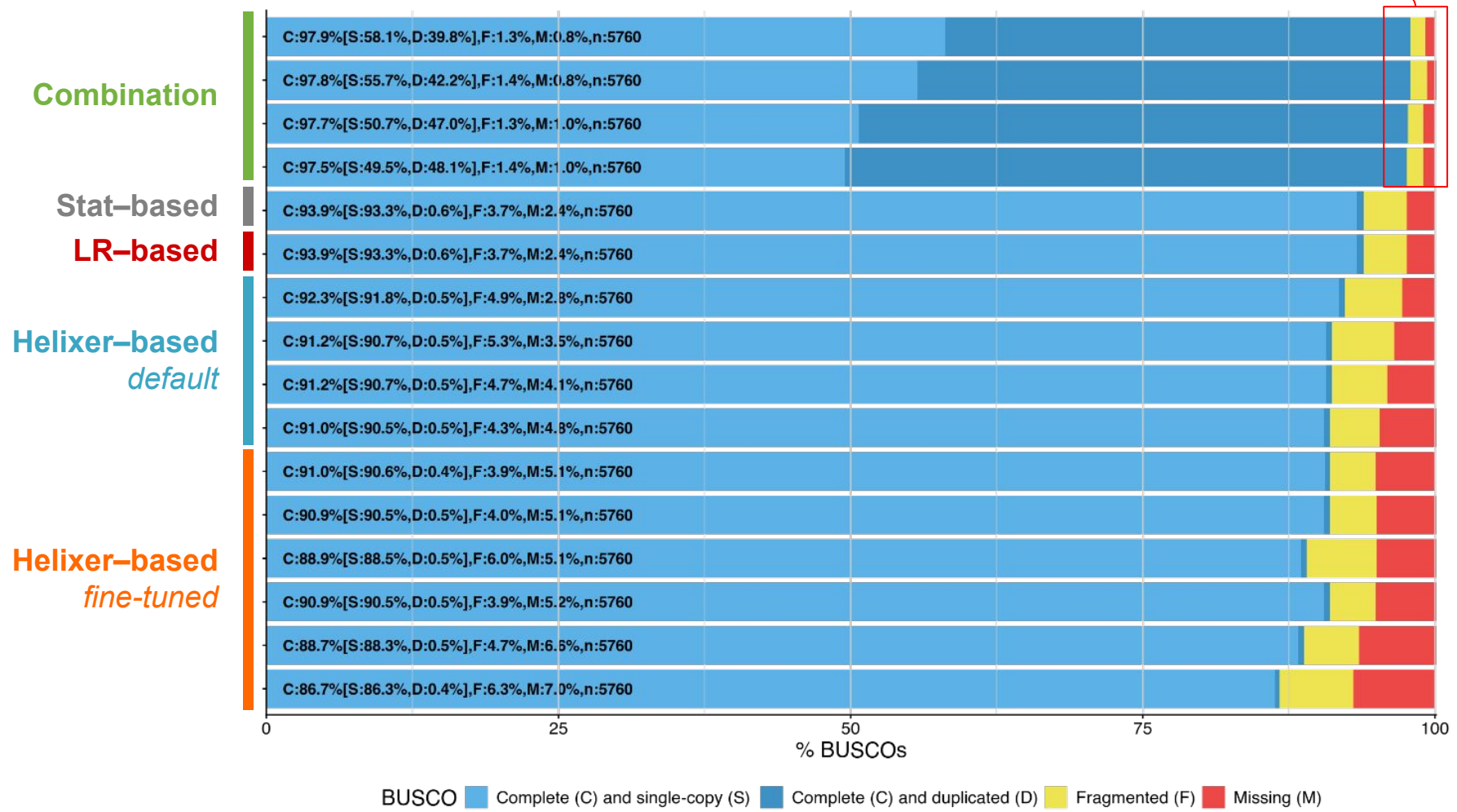
BUSCO analysis of the 16 annotations

Slightly more fragmented BUSCO genes in the annotations obtained with the fine-tuned models compared to the default ones.



BUSCO analysis of the 16 annotations

Combining an Helixer–based annotation to a data–based annotation identify more complete BUSCO genes…



BUSCO analysis of the 16 annotations

… but there are more duplicated information, whereas the other annotations are very clean.

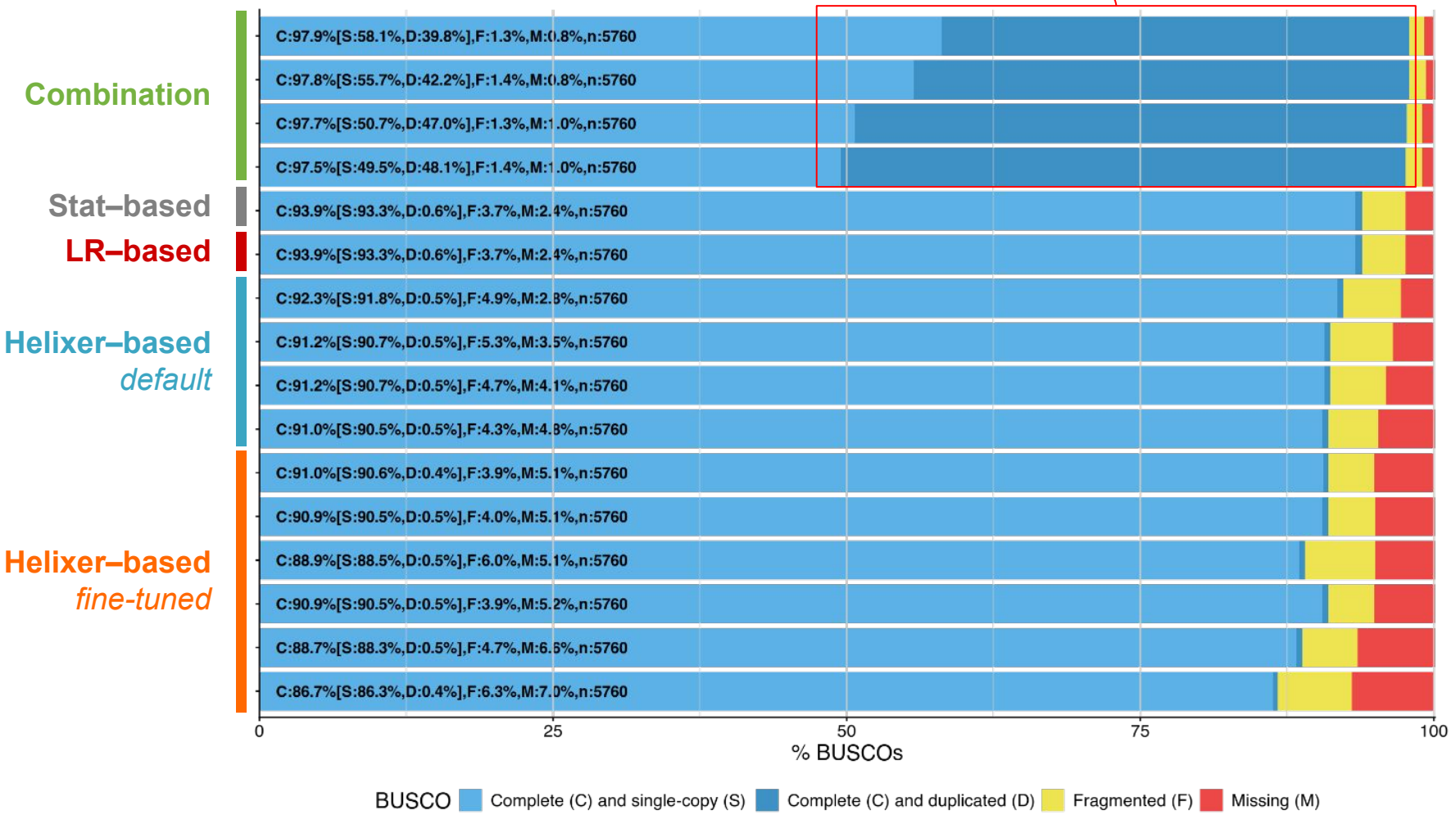
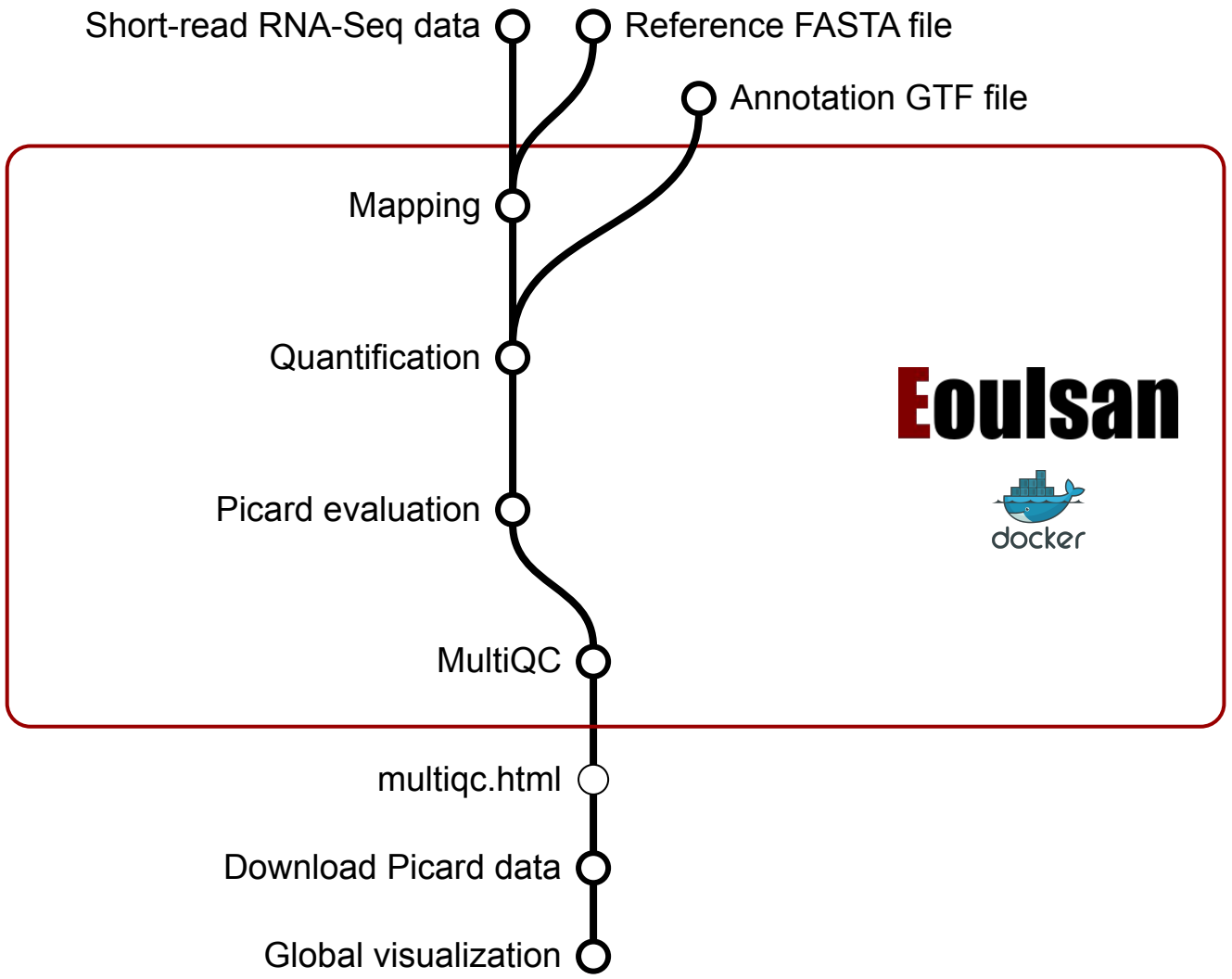


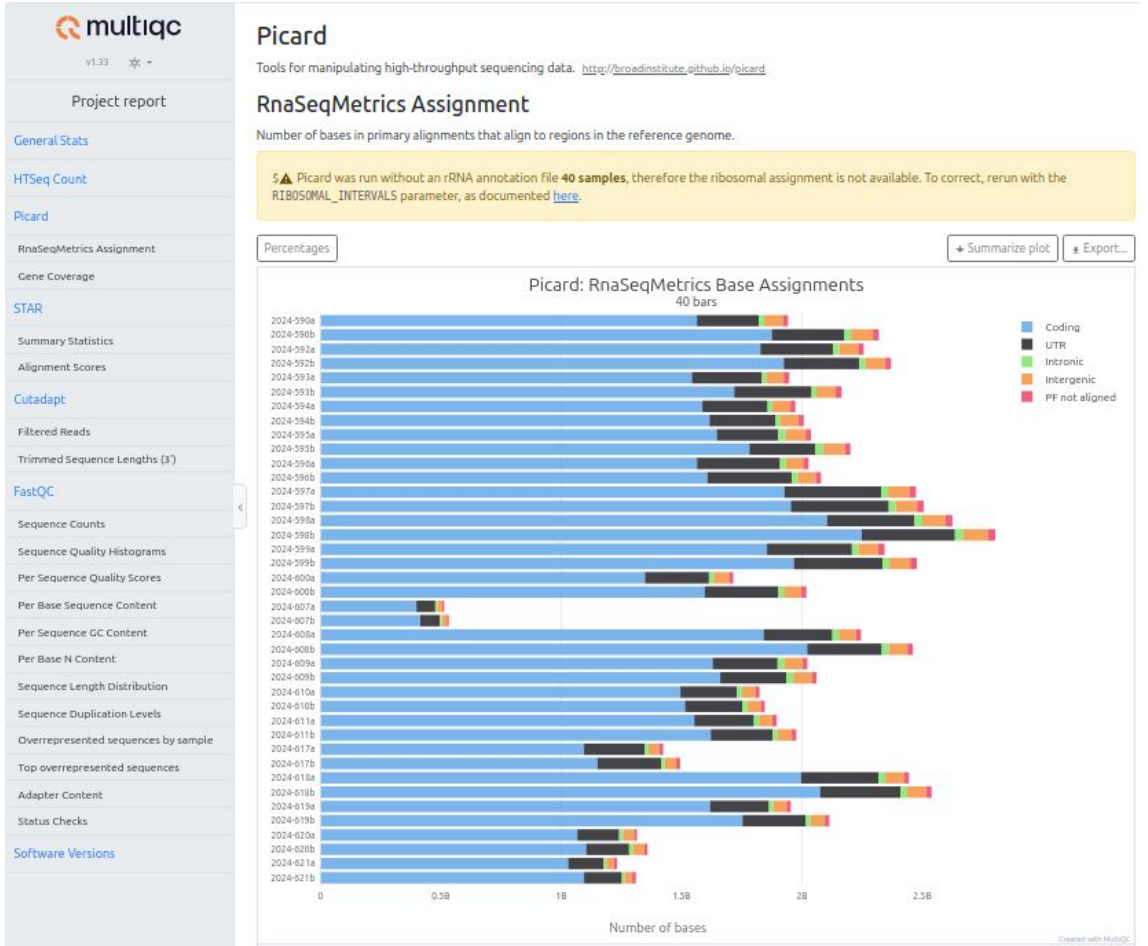
Table of Contents

- **Species of interest**: *Morpho helenor*
  - Annotation provided by the lab: **Stat–based**
  - Annotation obtained from another tissue: **LR–based**
- Aim 1 – **Predict annotations using Helixer**: **Helixer–based**
  - **Fine-tune** a default model on *Lepidoptera*
    - Identify species for the training / validation
    - Set up a pipeline
  - **Prediction annotations** using the default and fine-tuned models
    - Set up a pipeline
- Aim 2 – Evaluate the relevance of the annotations
  - **gffcompare** against the Stat–based annotation
  - **BUSCO** analysis
  - **Quantification of SR RNA-Seq data**

Quantification of the short-read data using the various annotations



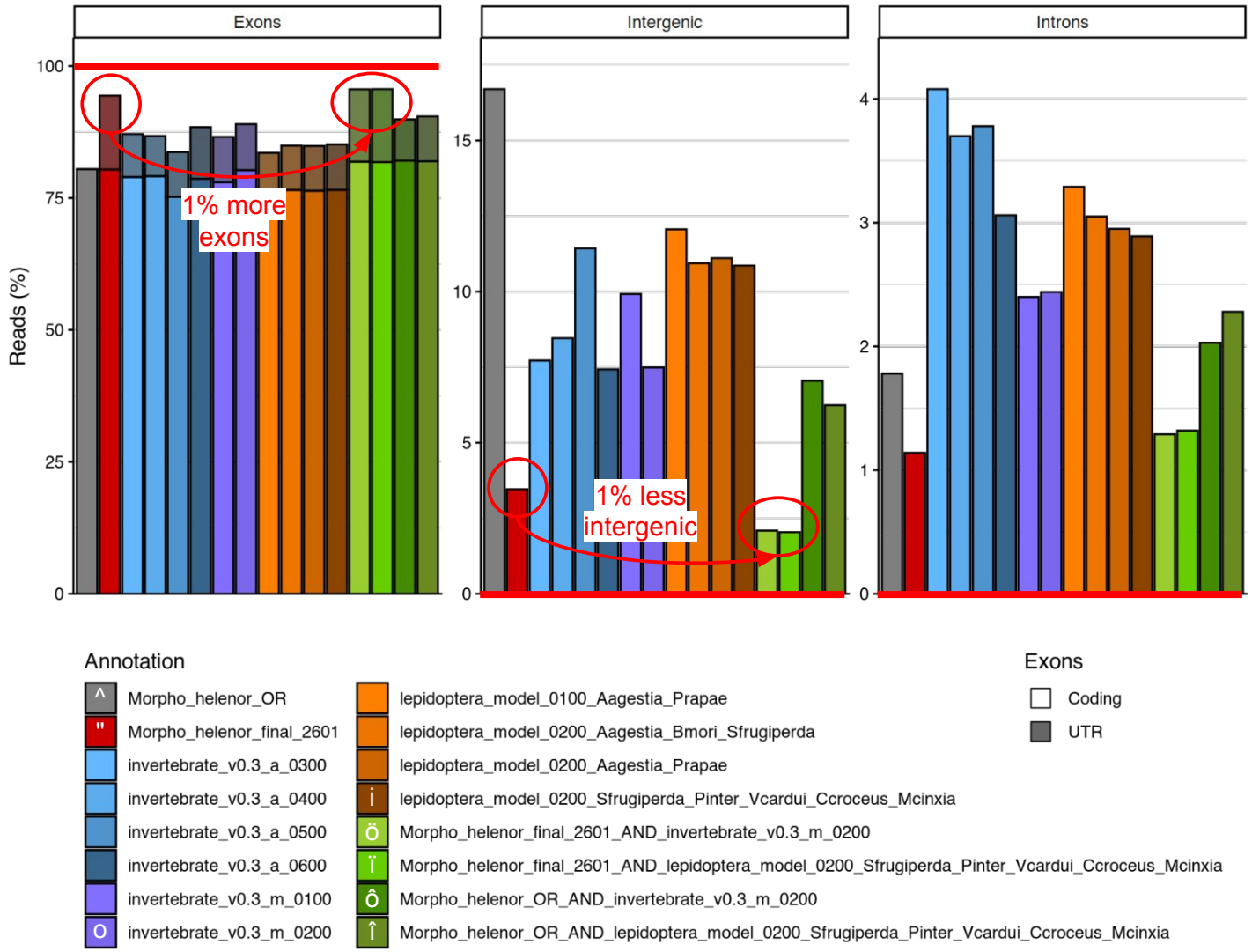
Screenshot of the Picard plot on a multiqc report



Quantification of the short-read data using the various annotations

- Short-read RNA-Seq data, we expect:
  - A lot of exons
  - Few introns
  - No intergenic

- Results:
  - Missing UTR in the **Stat–based** annotation
  - The **LR–based** annotation is the best one
  - **Helixer–based annotations** obtained with fine-tuned models have missing exons compared to the ones obtained from **default models**
  - **Combining annotations** slightly improves the quantification



Conclusion in the context of the *Lepidoptera* order and the *M. helenor* species

- **Using Helixer to predict structural annotation**
  - Enables to obtain an annotation using the genome assembly **only**.
  - Covers a wide range of **eukaryotes**.
  - Predicts the **UTR**.
  - The **default models** are already quite good. Fine-tuning a model can impede the *creativity* of the prediction pipeline.
  - Easy to use.

- Annotation obtained using **long-read RNA-Seq** data are excellent, but these data are not always available.

- **Perspectives**
  - Structural annotation **consensus**: duplicates, isoforms, artefacts…?
  - **Functional annotation**: Are the annotated structures biologically relevant?

Thank you!

Romuald Marin

Lili Babin
Ludvine Brajon
Joséphine Ledamoisel
Violaine Llaurens
Manuela Lopez Villavincencio

Corinne Blugeon
Laurent Jourdren
Sophie Lemoine
Tiphaine Marvillet
Catherine Sénamaud-Beaufort
Morgane Thomas-Chollier

The slides are available on HAL.