Comparative genomics using reference genomes

Introduction to
Biodiversity Genomics
Tena, Ecuador
2024

Comparative Genomics

NHGRI FACT SHEETS
genome.gov

Researchers choose the appropriate time-scale of evolutionary conservation for the question being addressed.

Common features of different organisms such as humans and fish are often encoded within the DNA evolutionarily conserved between them.

Looking at **closely related species** such as humans and chimpanzees shows which genomic elements are unique to each.

Genetic differences **within one species** such as our own can reveal variants with a role in disease.

NIH National Human Genome Research Institute

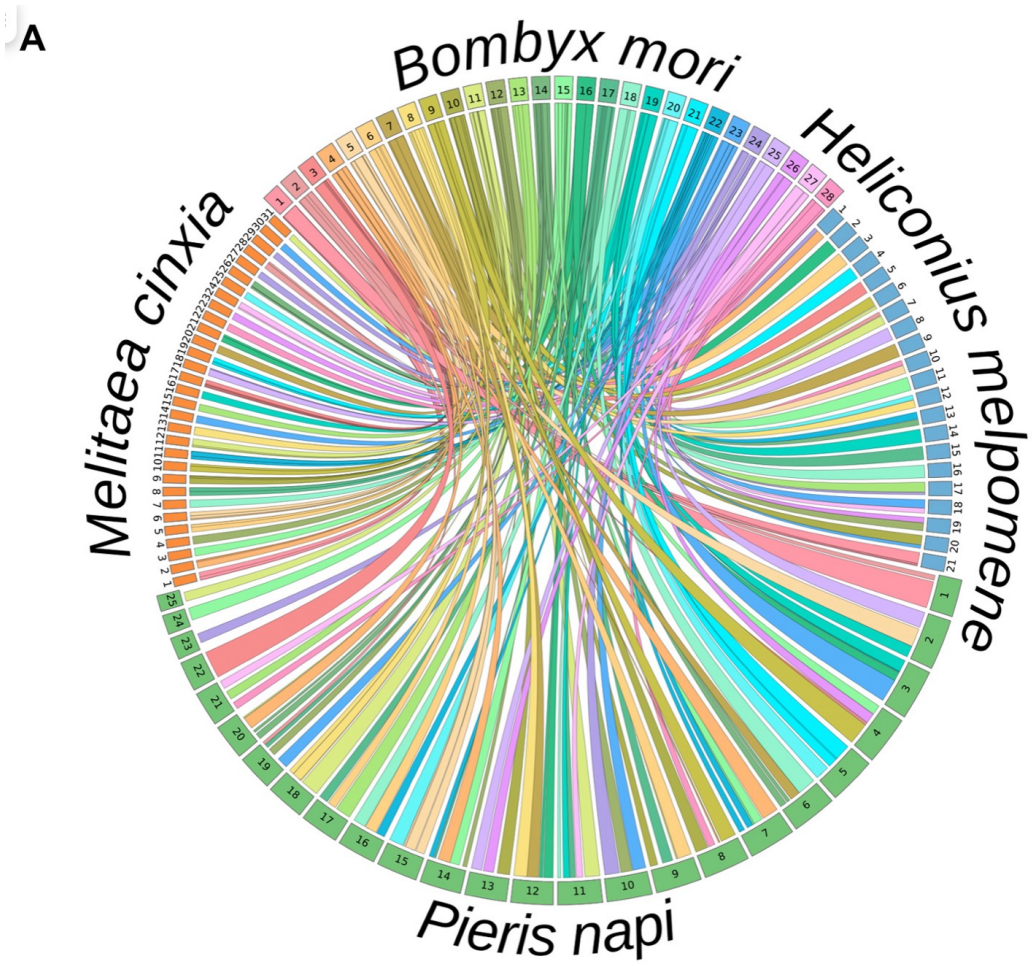
Comparative genomics using reference genomes

- Genome assembly
  - Reconstruction of the genome of one or more individuals

- Reference genome
  - Genome assembly that is the point representation of the structure and organisation of the genome of a species

Comparative genomics – common applications

- Genome structure and synteny
- Feature dynamics
  - Gene family dynamics
  - Conserved non-coding elements
  - Repeat exploration
- CG, codon usage and tRNA dynamics
- Heterozygosity

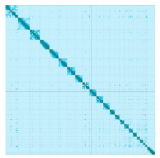


Hill et al. 2019

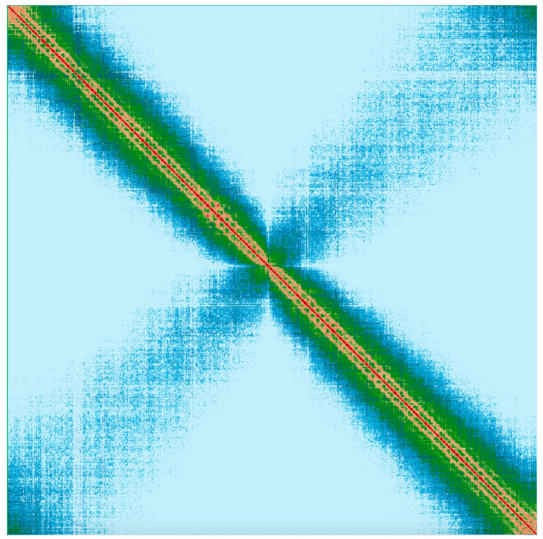
Genome structure

- Genome size

Human genome



Mistletoe Chr 1

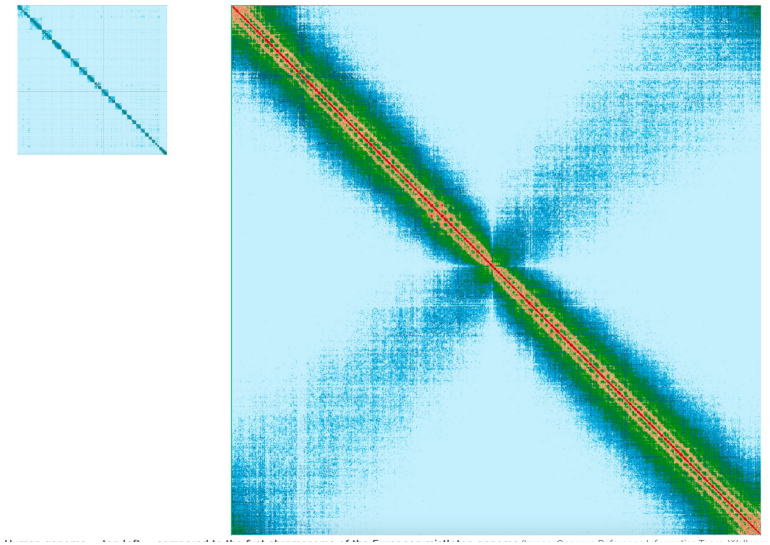


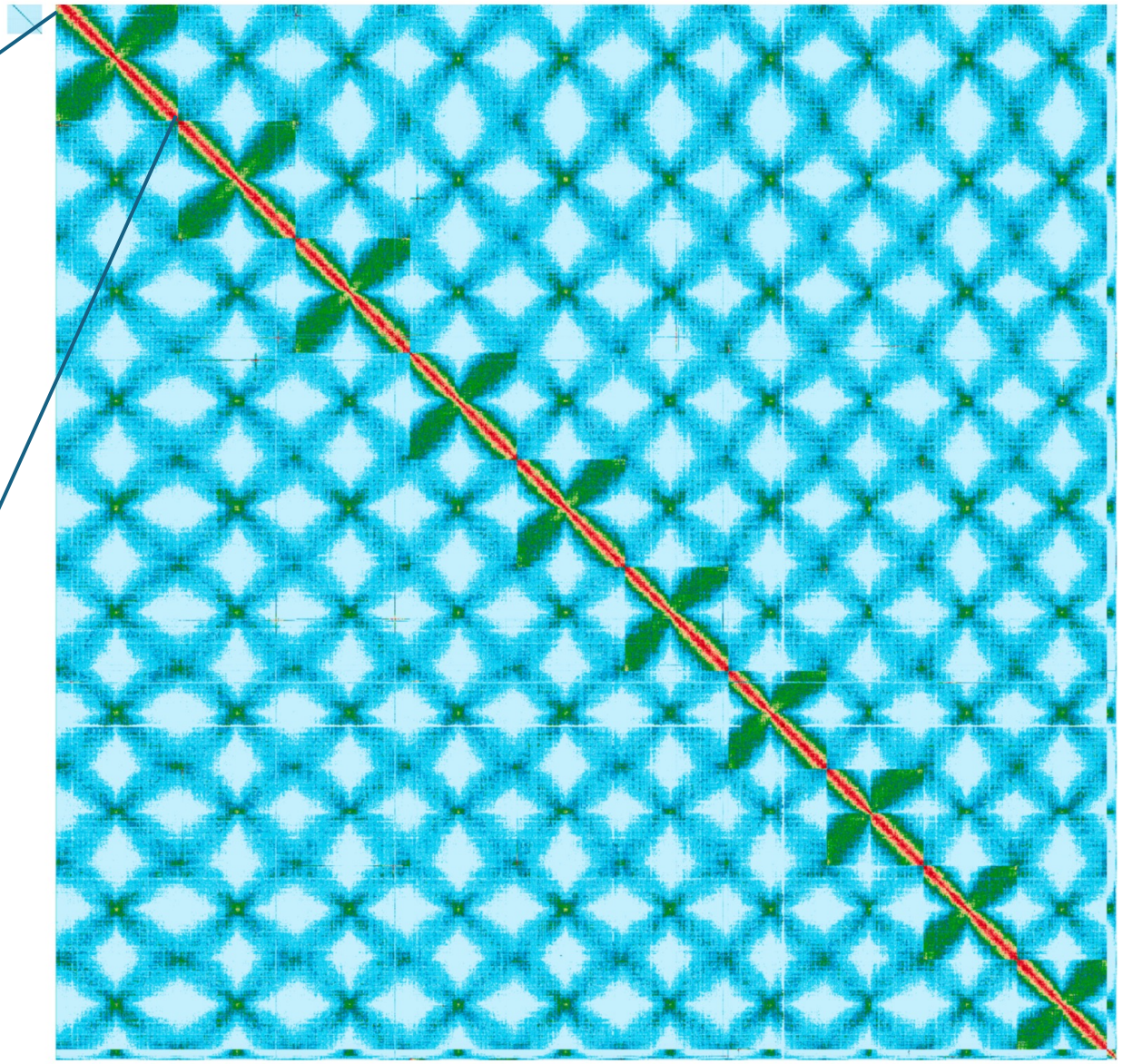
Genome structure

- Genome size

Human genome

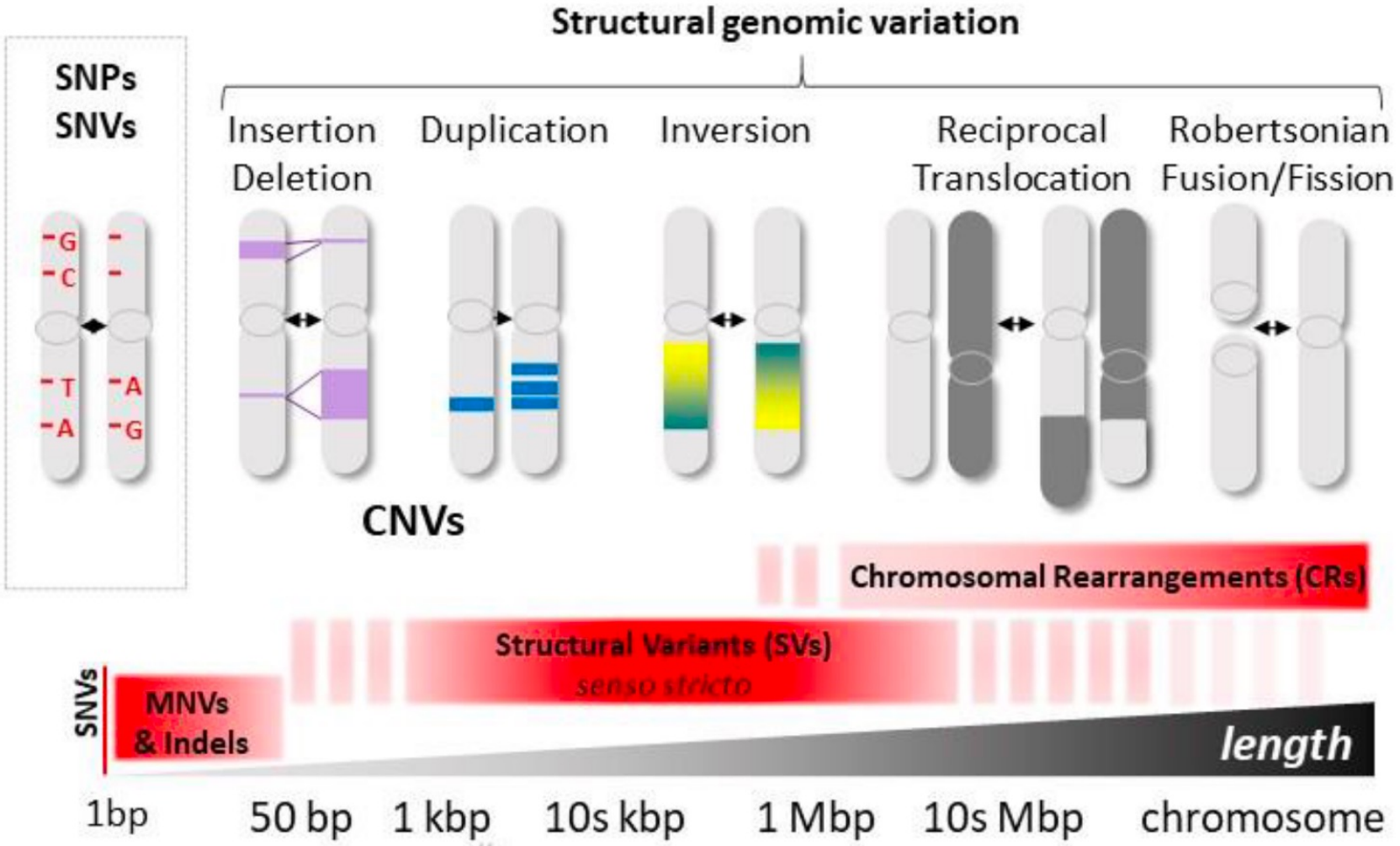
Mistletoe Chr 1





Human genome — again, top left — compared to the entire European mistletoe genome (Image: Genome Reference Informatics Team, Wellcome Sanger Institute)

Structural variation



Short read data

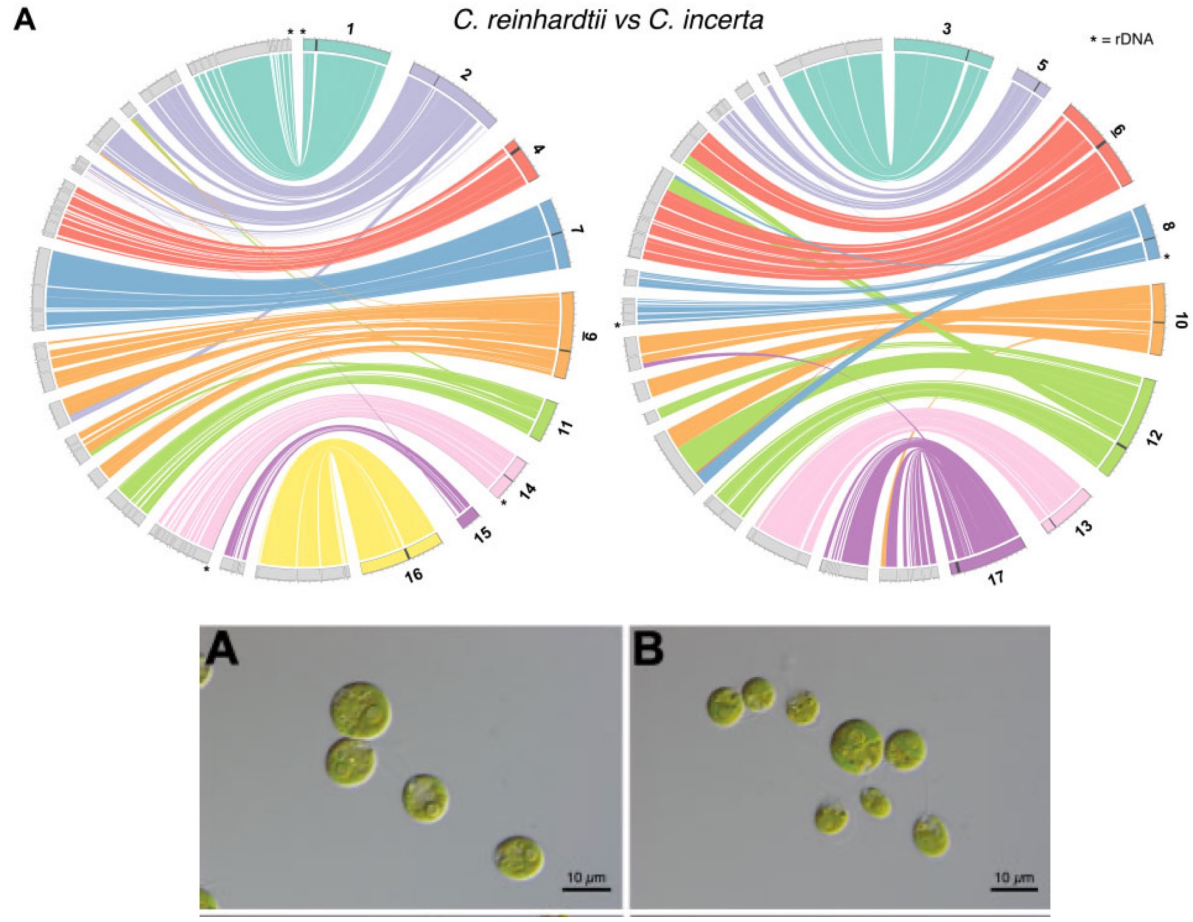
Long read data

Chromosome level genomes

Berdan et al 2023

Whole genome alignment

- Closely related taxa
  - MUMMER
  - Minimap2
  - Cactus
  - LastZ
  - D-Genies

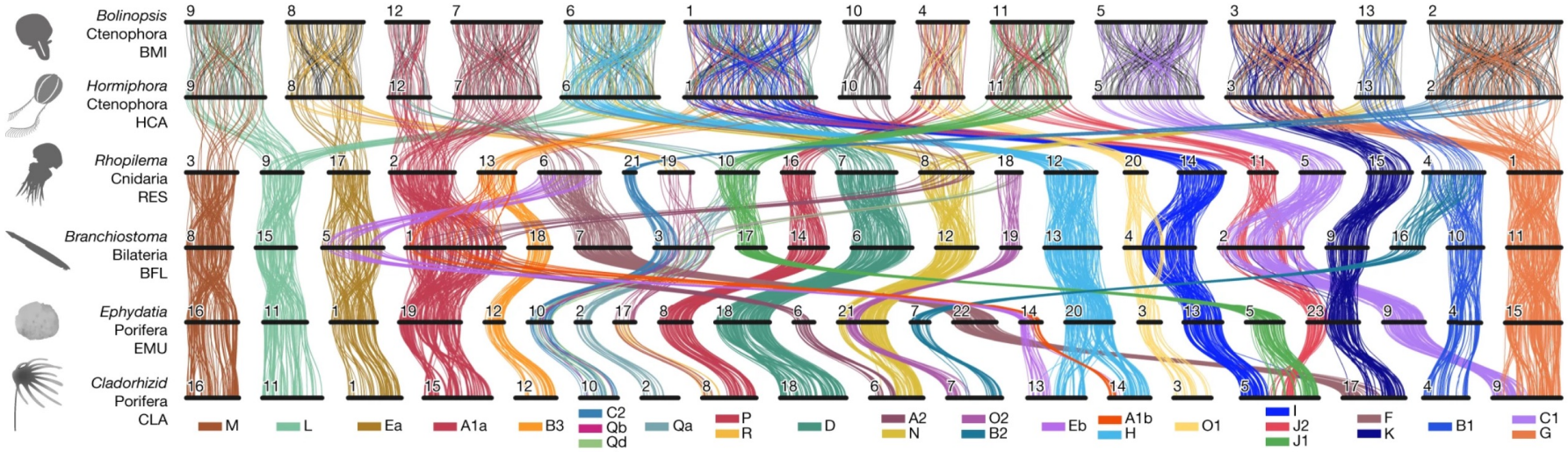


Green algae Chlamydomonas

Craig et al 2012

Orthologous markers

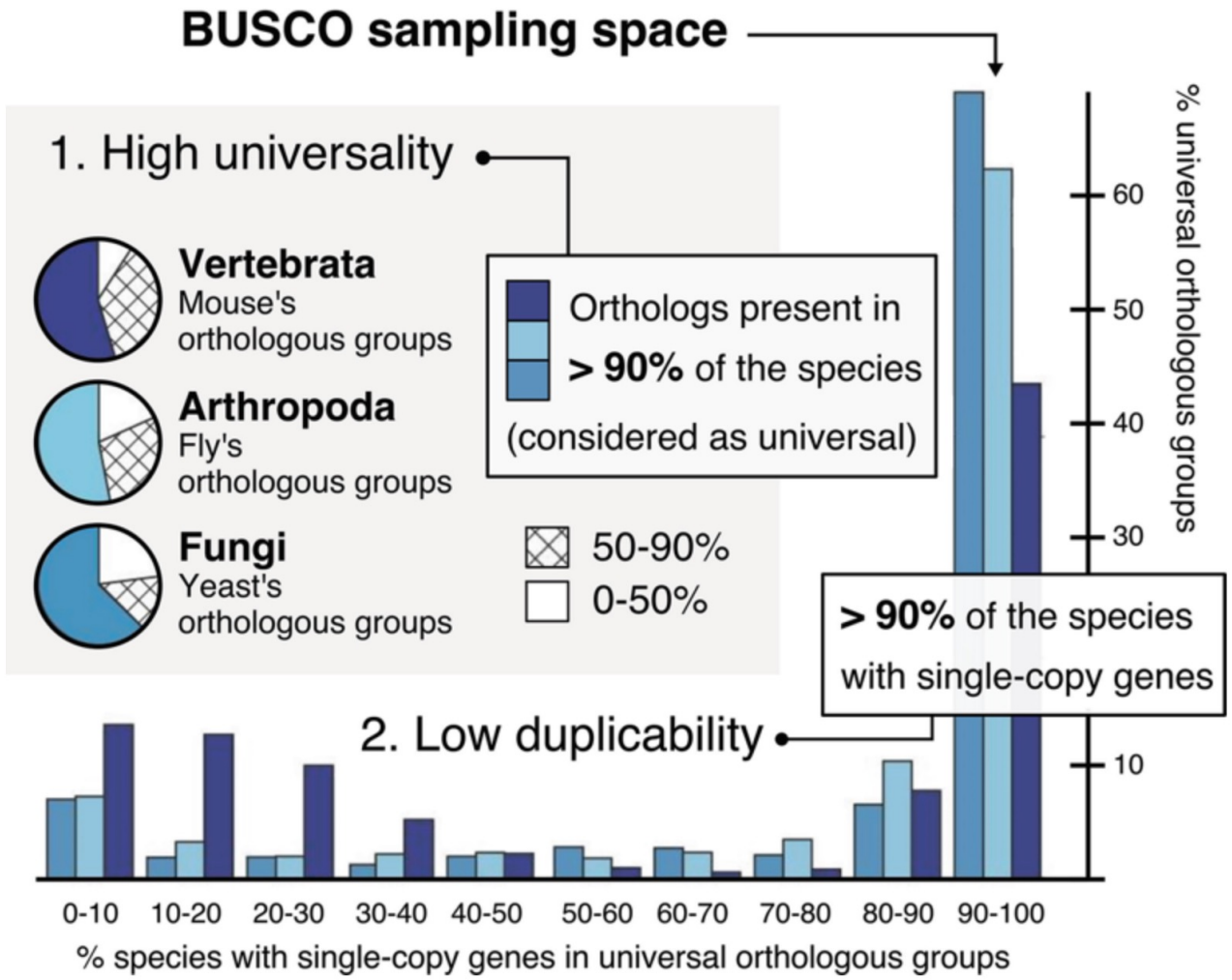
- Use conserved sequences as anchor markers
- Also for very divergent taxa



Schultz et al. 2023

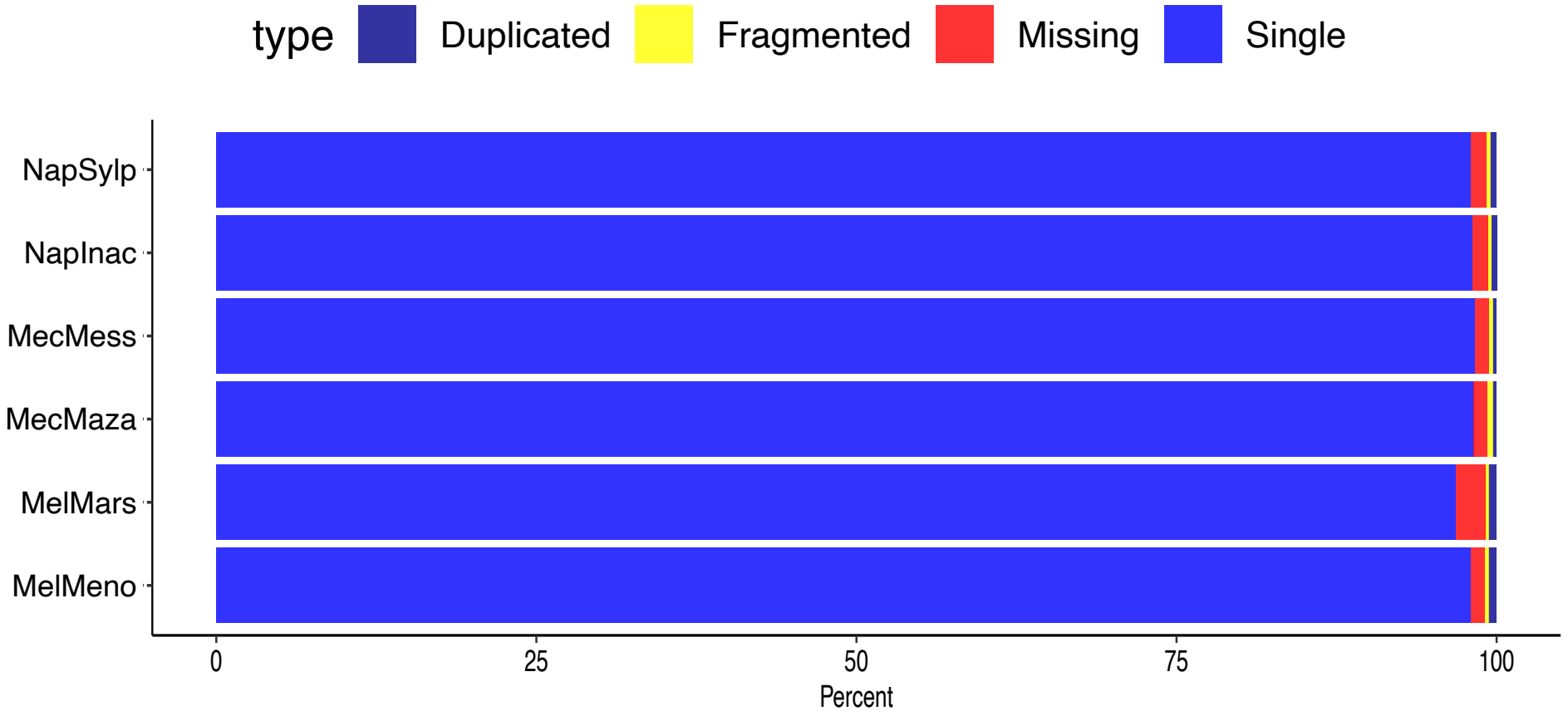
Orthologous markers

- BUSCO (Benchmarking Universal Single Copy Orthologues)
  - BLAST https://blast.ncbi.nlm.nih.gov/Blast.cgi
  - Finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.



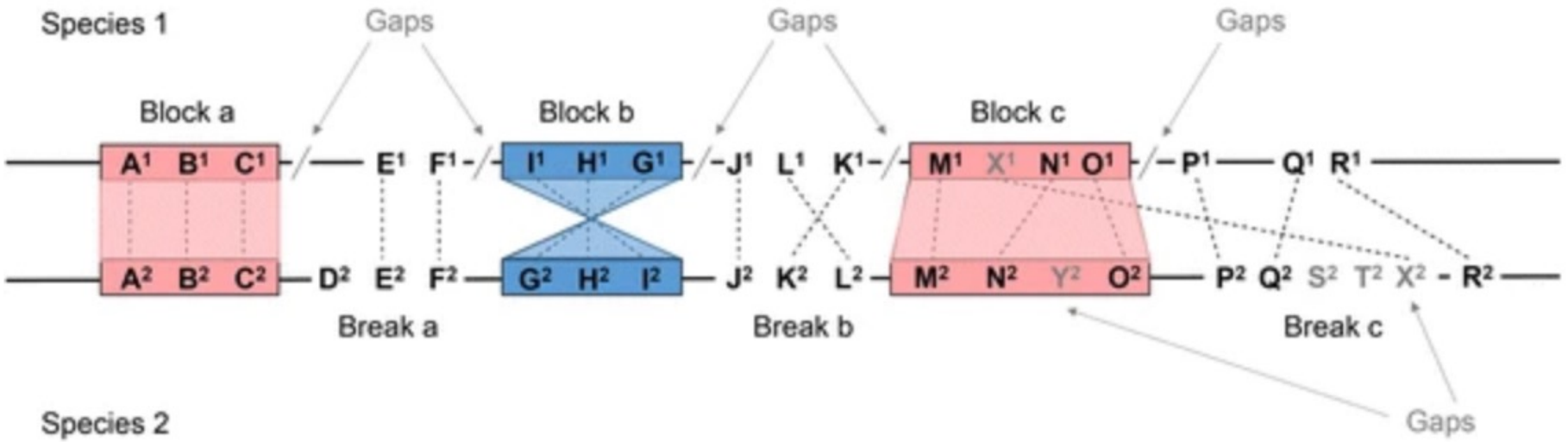
https://busco.ezlab.org

BUSCO completeness

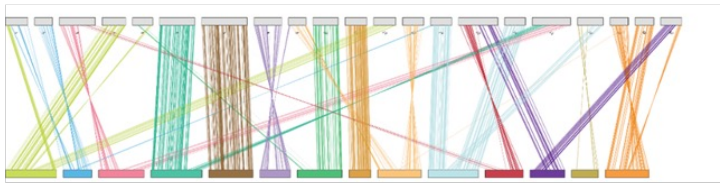


Visualisation

- Synteny



Ribbon plot

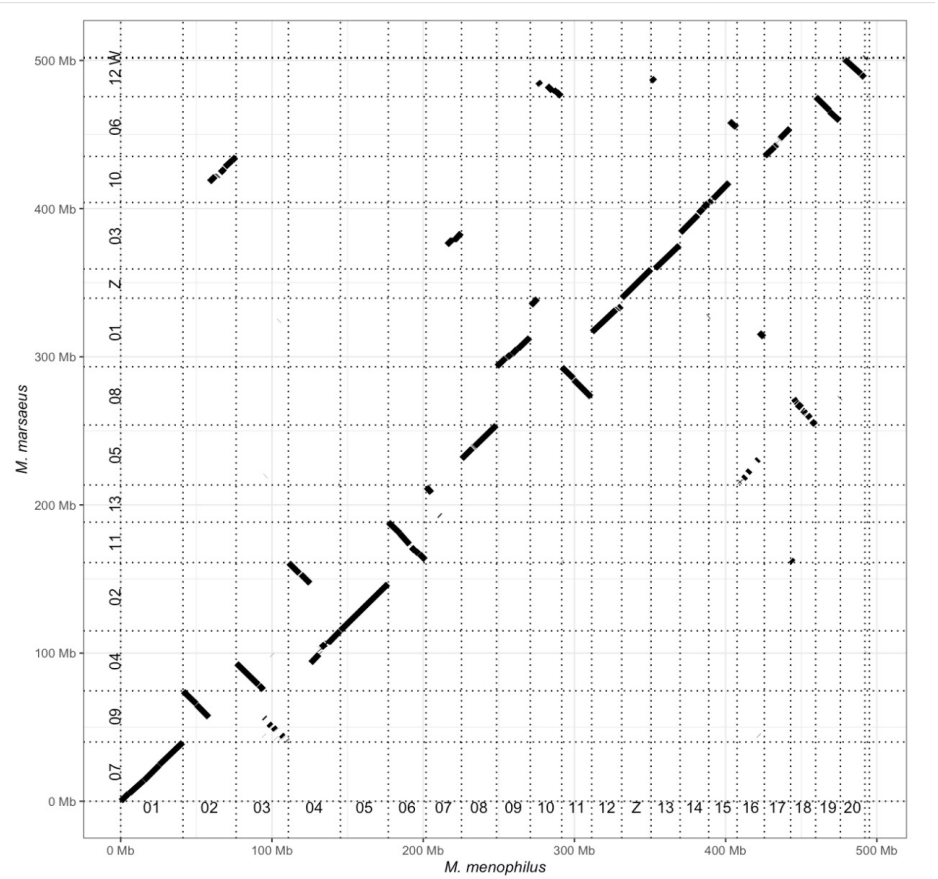


Melinaea marsaeus (n=14)

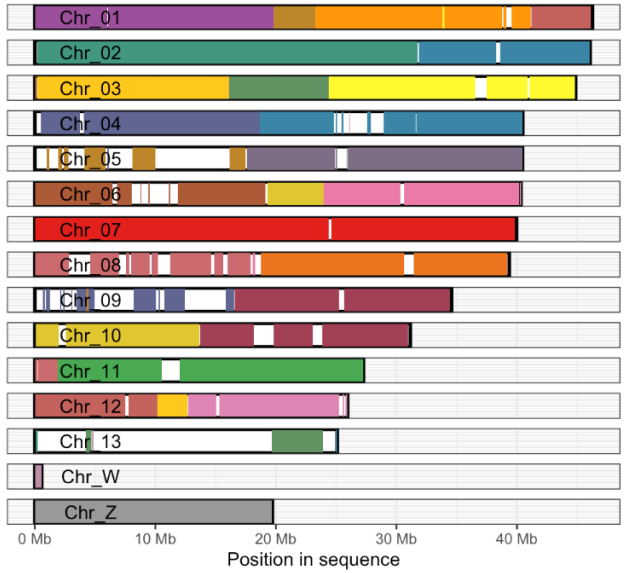
vs

M. menophilus (n=21)

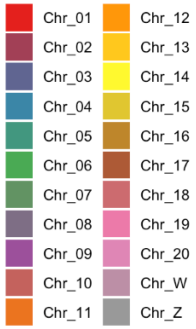
Dotplot



M. marsaeus



M. menophilus

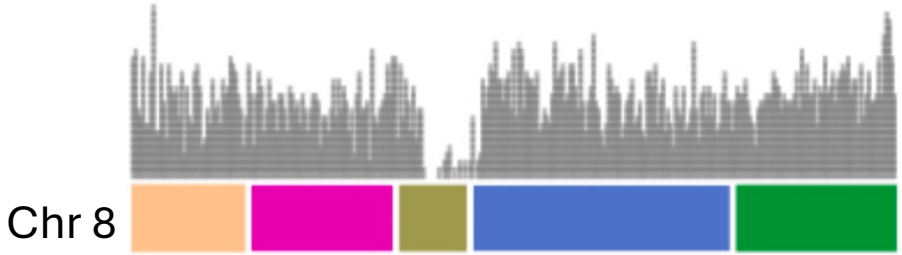


Comparative genomics - genomic features

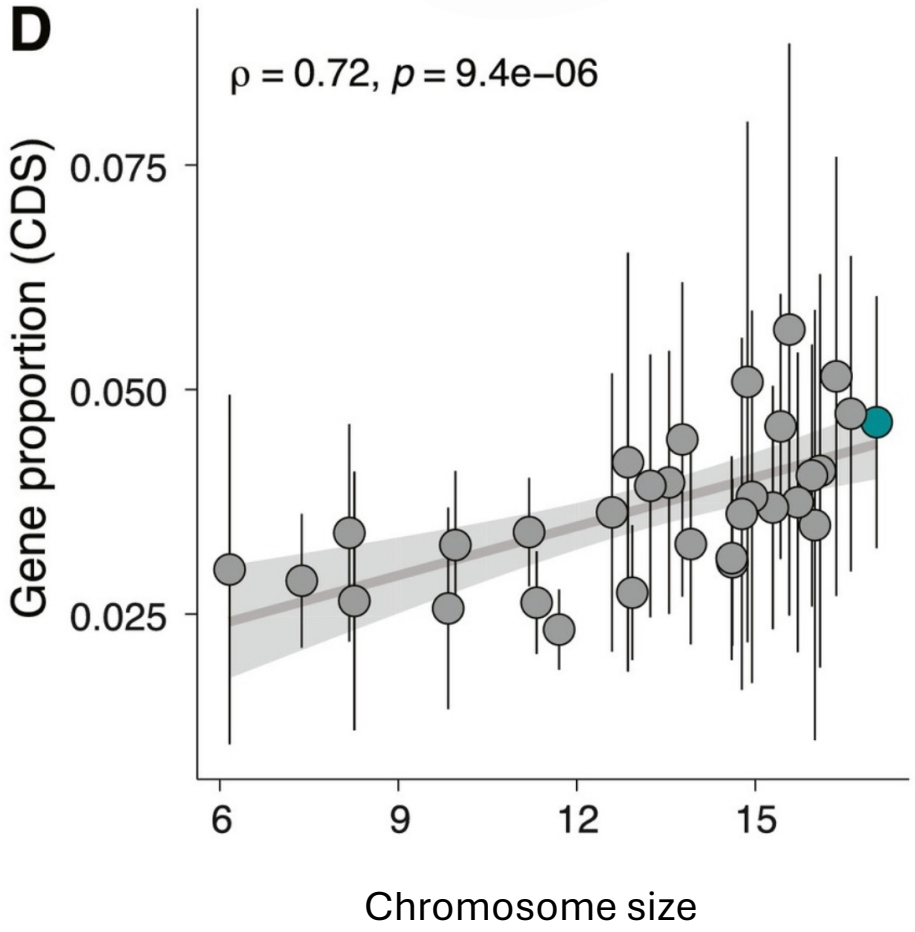
- Gene family expansion and contraction
- Regulatory element evolution
- Transposable element dynamics
- Genome composition evolution

Genes

- Gene content: protein coding genes and non-coding RNA within the genome



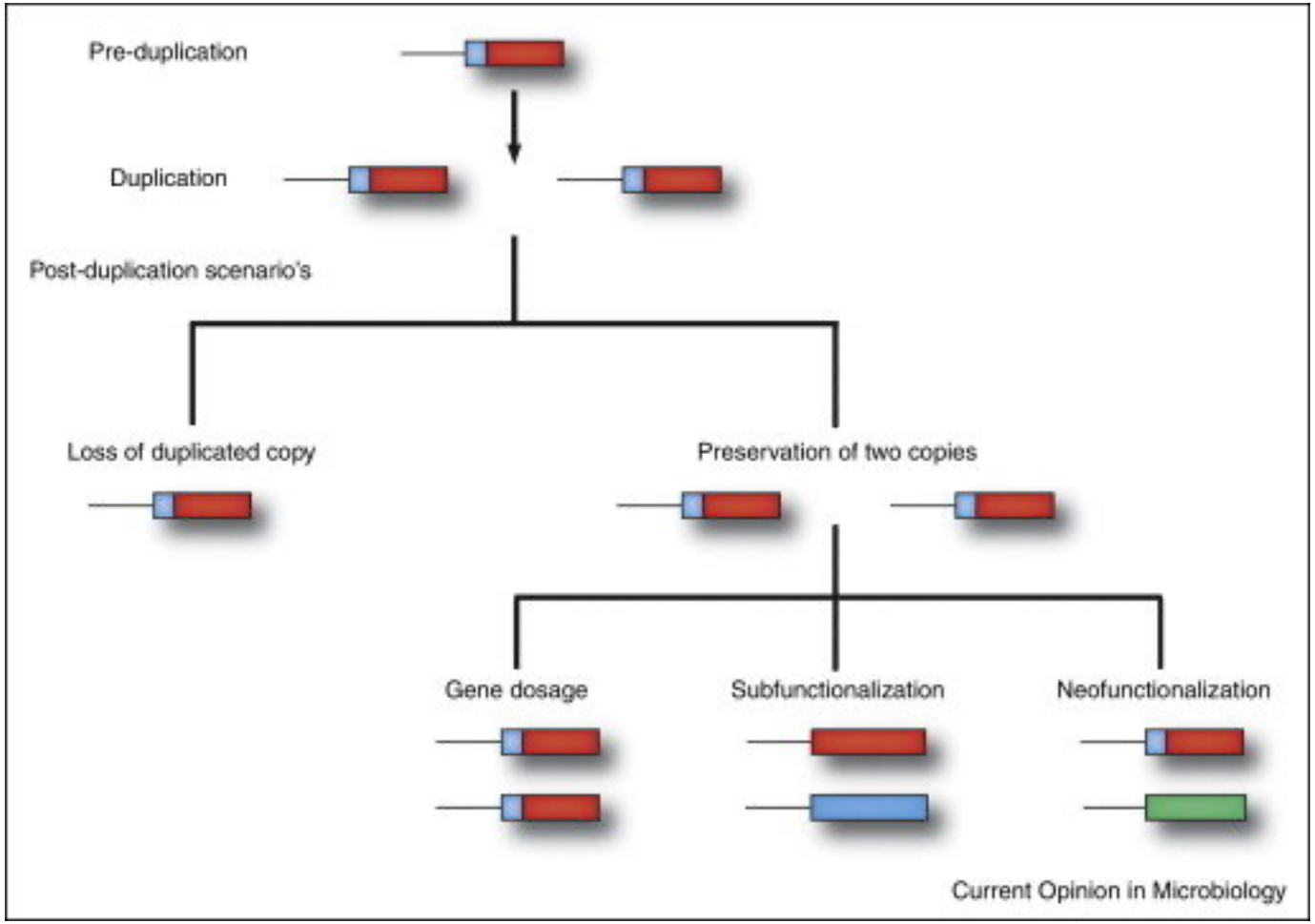
Arachis duranensis
Ren et al 2018



Shiplina et al 2022

Gene family expansion or contraction

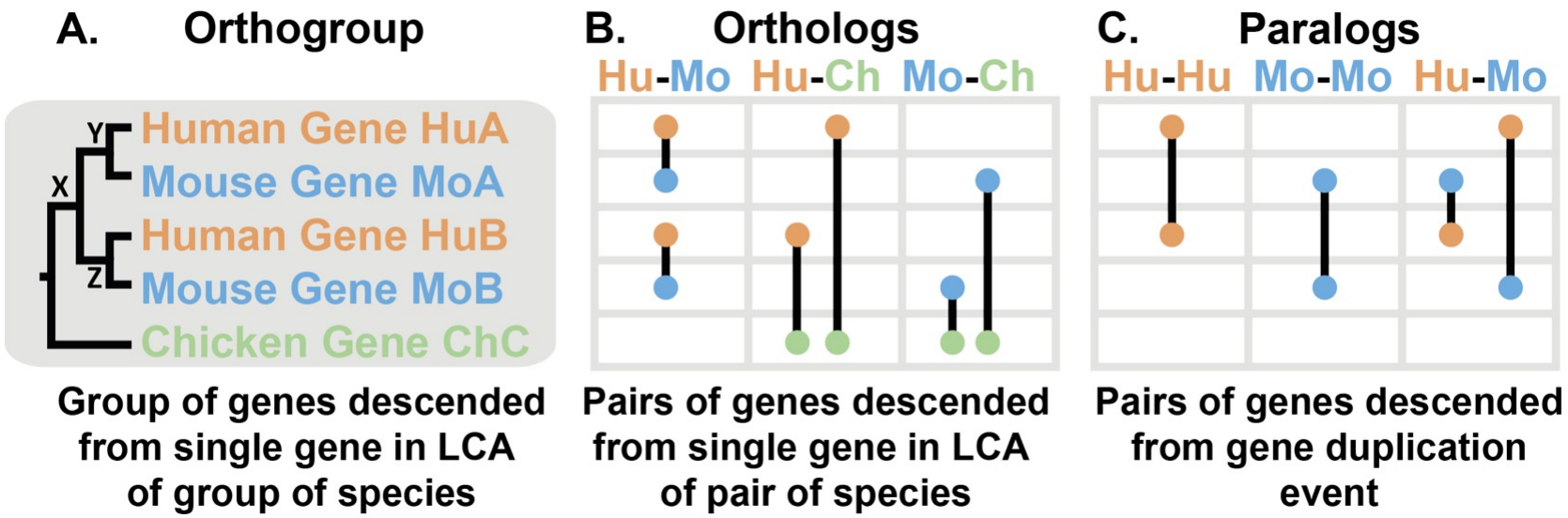
- Gene gain
- Gene loss



Voordeckers & Verstrepen 2015

Gene family expansion or contraction

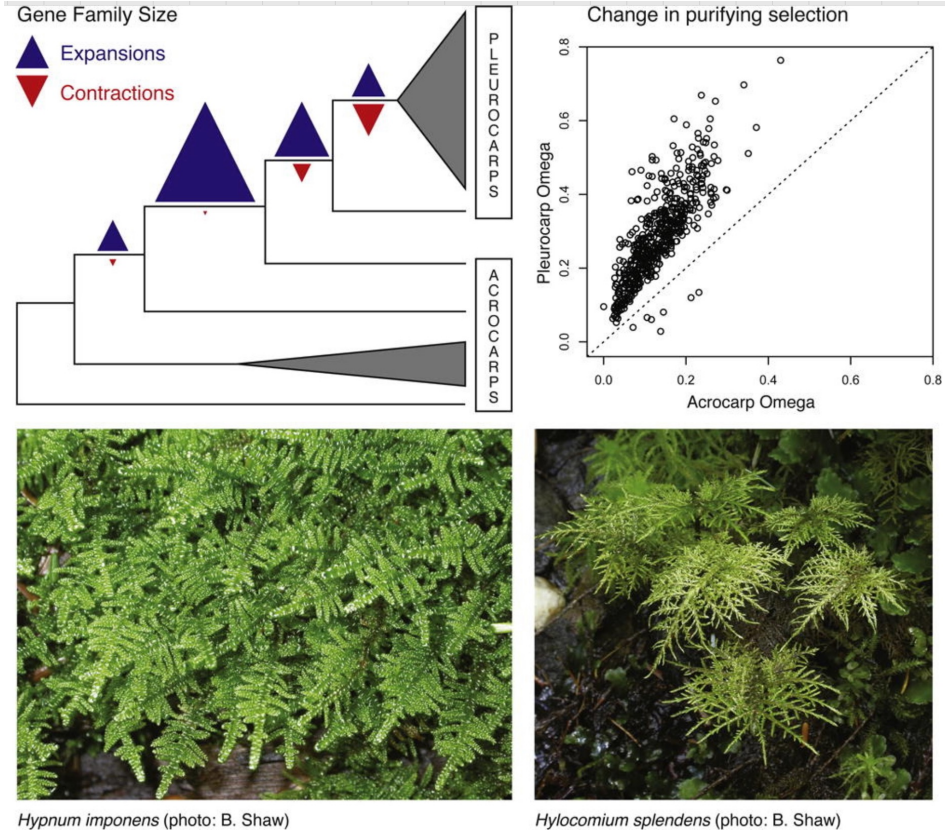
- Cluster in gene families by sequence homology
  - Orthofinder
    - Orthologs
    - Paralogs



Emms and Kelly (2019)

Gene family expansion or contraction

- Cluster in gene families by sequence homology
  - Orthofinder
- Infer phylogenetically informed rates of gain and loss
- Birth-death model for evolutionary inferences
- Maximum-likelihood estimation of a global or local gene family evolutionary rates
  - Café https://github.com/hahnlab/CAFE5
  - Badirate https://github.com/fgvieira/badirate



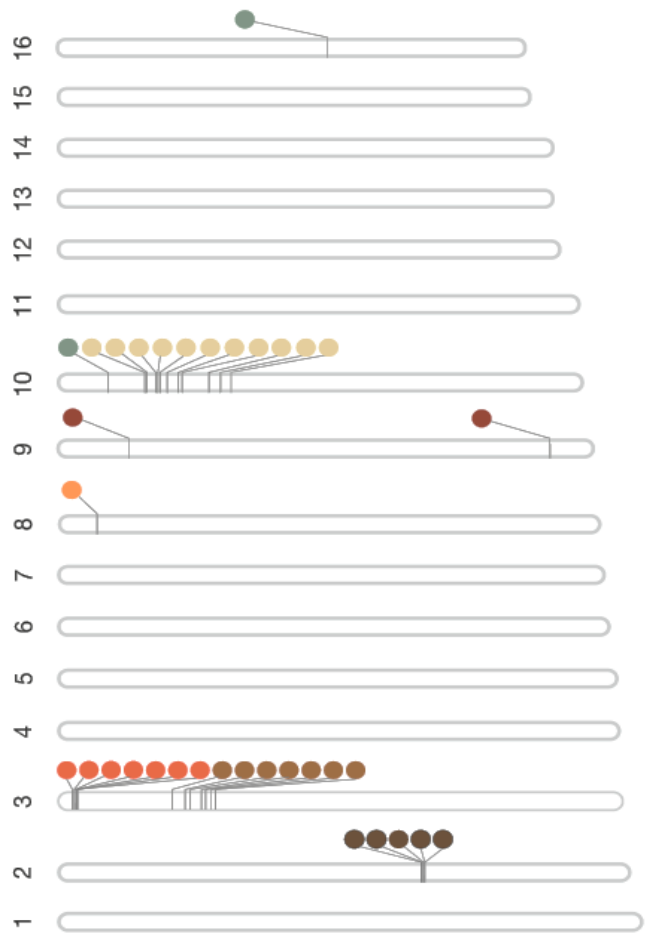
Johnson et al 2016

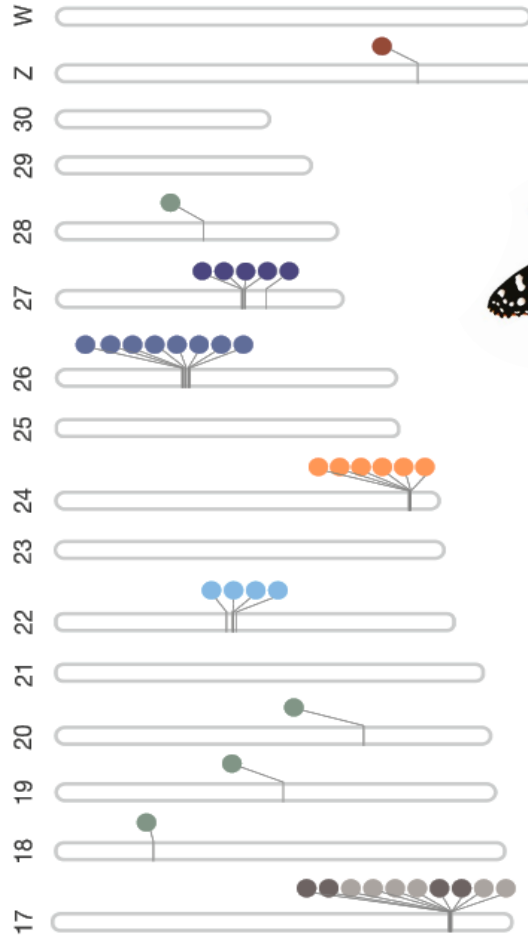
Gene family expansion or contraction

- Gene ontology analysis, approximation of gene function
  - Panther
  - TopGO
  - Domain analysis
- Functional validation

Candidate gene families

- ATPase
- Carboxylesterases
- Papain protease
- Moricin
- LDL-receptor
- Lipophilic transfer protein
- Ubiqutin-like protease
- Chorion protein
- Calpain protease
- UDP glycosyltransferase
- Trypsin-like serine protease
- Trypsin-like serine protease-2





Shiplina et al 2022

Regulatory elements

- Promoters binding site for polymerase
- Proximal and distal enhancers, silencers, insulators

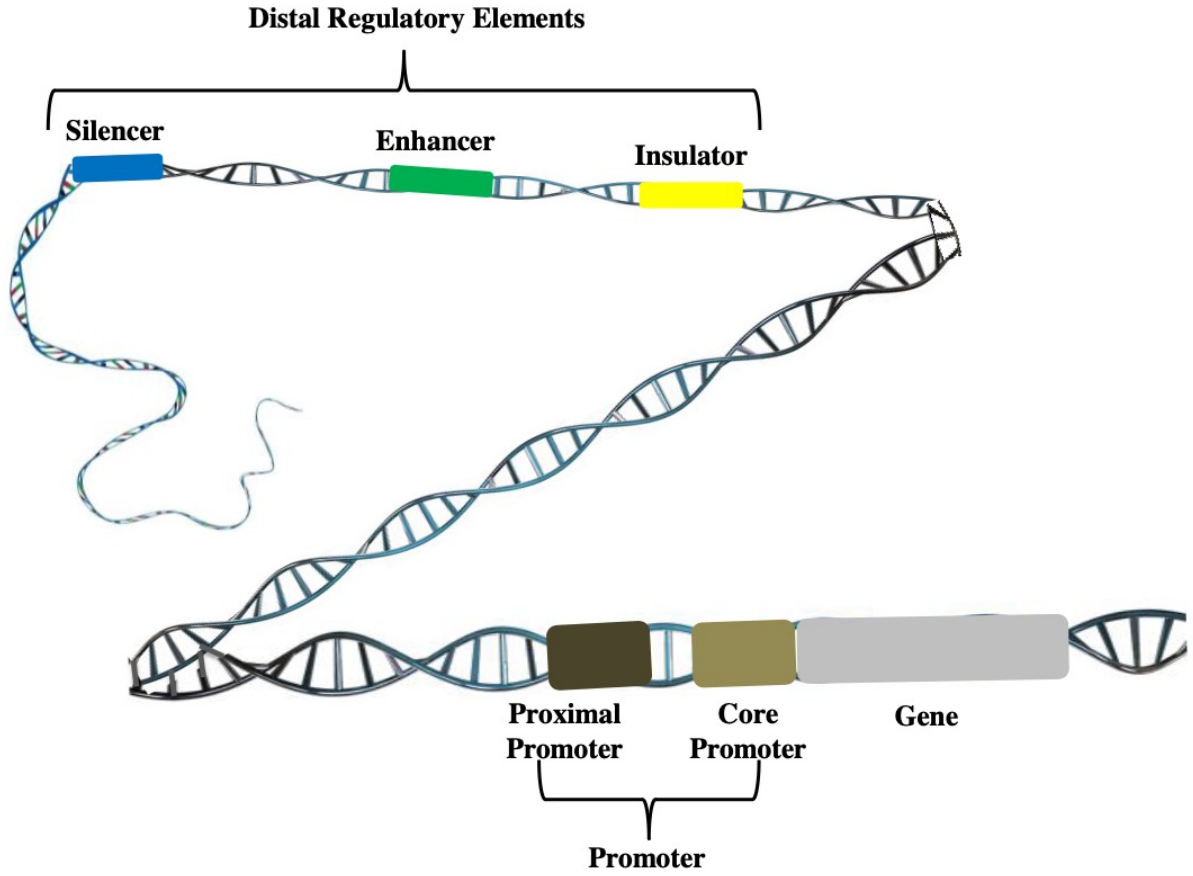
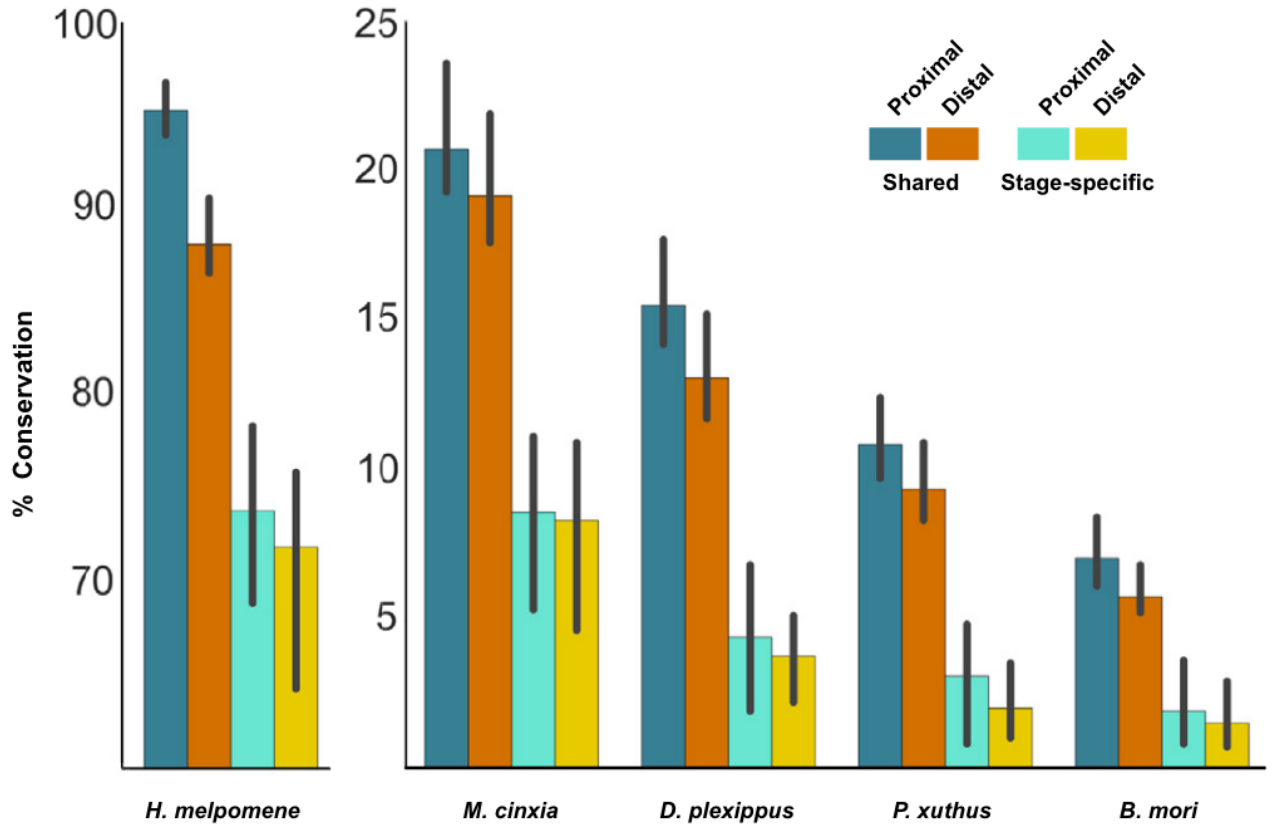


Fig. 1: Schematic representation of typical gene regulatory region.

Jadav et al 2016

Regulatory elements

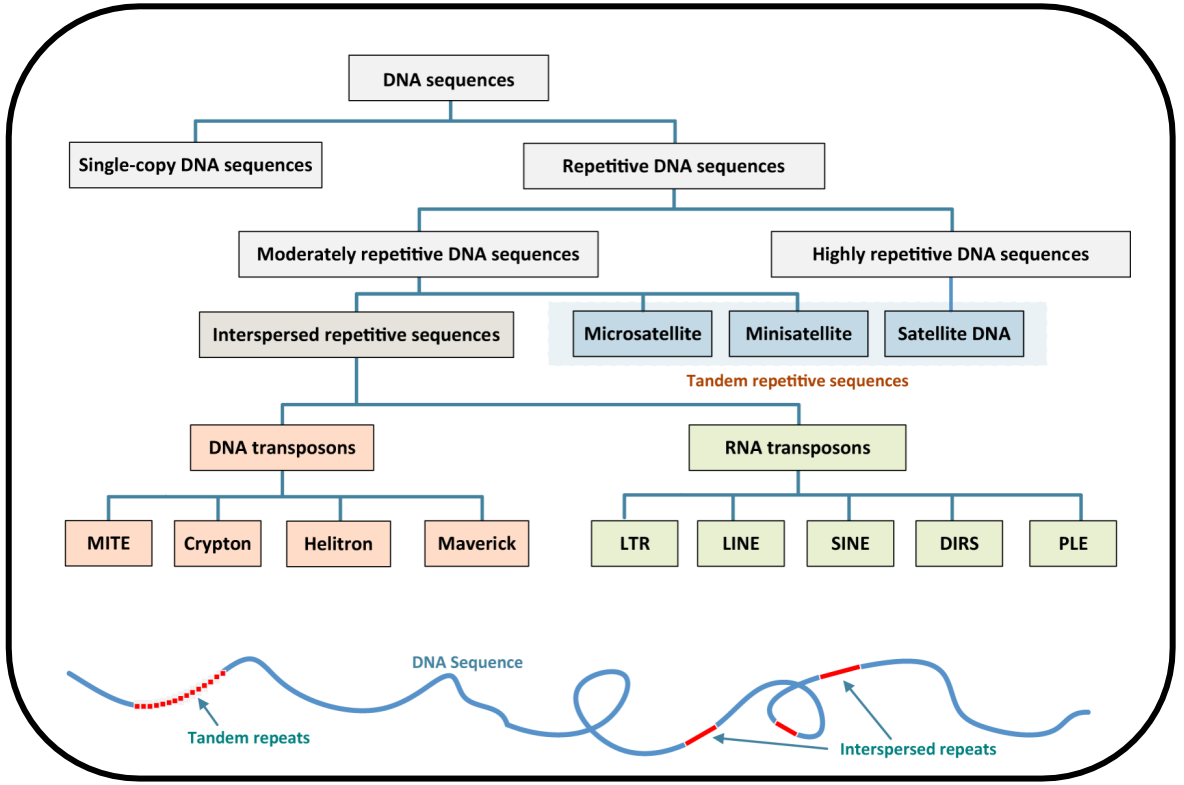
- Mutations, single nucleotide mutations, copy number variation (CNV) or TE insertions
- Infer by sequence homology
  - Whole genome alignment – conserved non-coding elements
  - Database search
- Prediction tools, specific motif
- Predict promoter - enhancer interactions in HiC-data PSYCHIC (Ron et al 2017)
- Experimental approaches
  - ChIP-seq
  - ATAC-seq



Lewis et al. 2016

Repetitive sequences

- Transposable elements (mobile elements)
  - Move (or duplicate) from one location in the genome to another
  - Initiates their own transcription, polymerase or retrotranscription and insertion using a wide range of endonucleases
- Simple repeats
  - satellite sequences
  - microsatellites
- Multi-copy RNA genes
  - rRNA, tRNA, snRNA



Liao et al 2023

Transposable elements (mobile elements)

- Class I: Replication via retrotranspostion (RT)
  - LTR (Long Terminal Repeats) retrotransposons
    - Copia, Gypsy, BEL and endogenous retroviruses (ERV), DIRS
  - Non-LTR retrotransposons
    - L1 (LINE) , SINE (non-autonomous), Penelope
- Class II: No RT
  - DNA-transposons
    - Harbiner, Mariner, Helitron

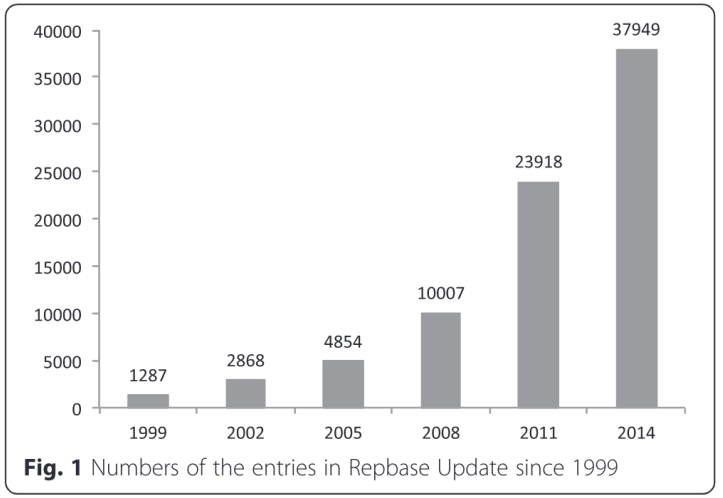
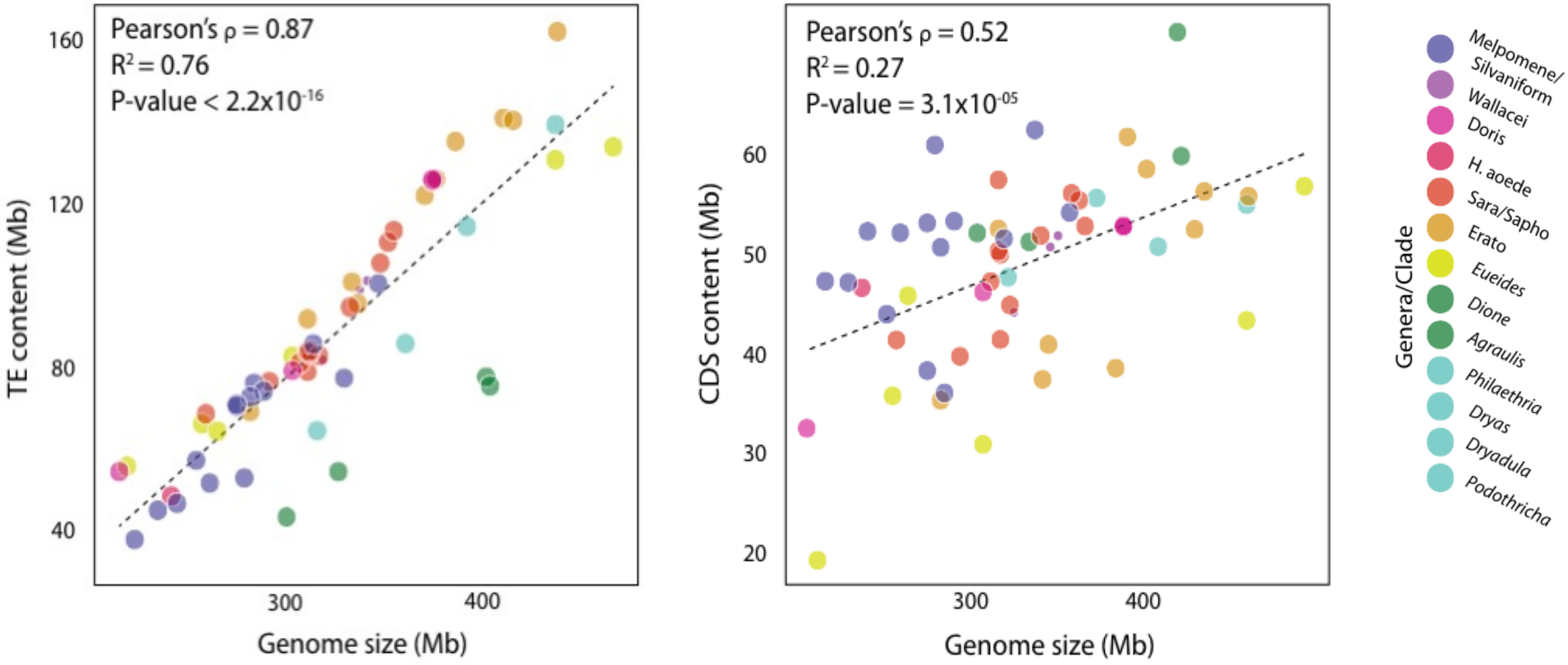


Fig. 1 Numbers of the entries in Repbase Update since 1999

Bao et al. 2015, Kojima 2019

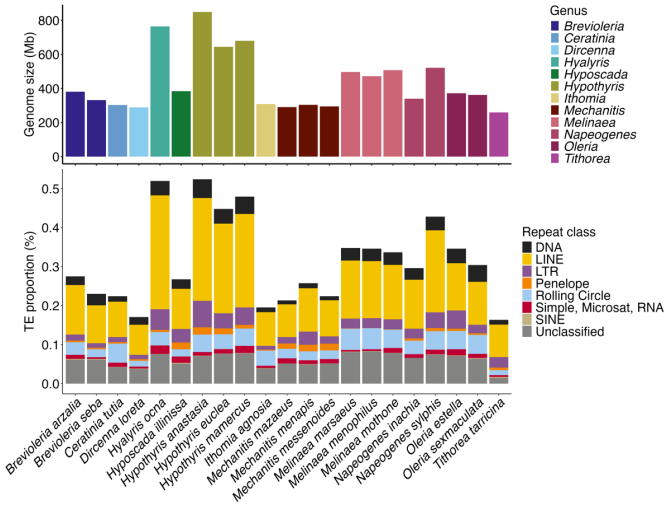
Transposable elements – genome size

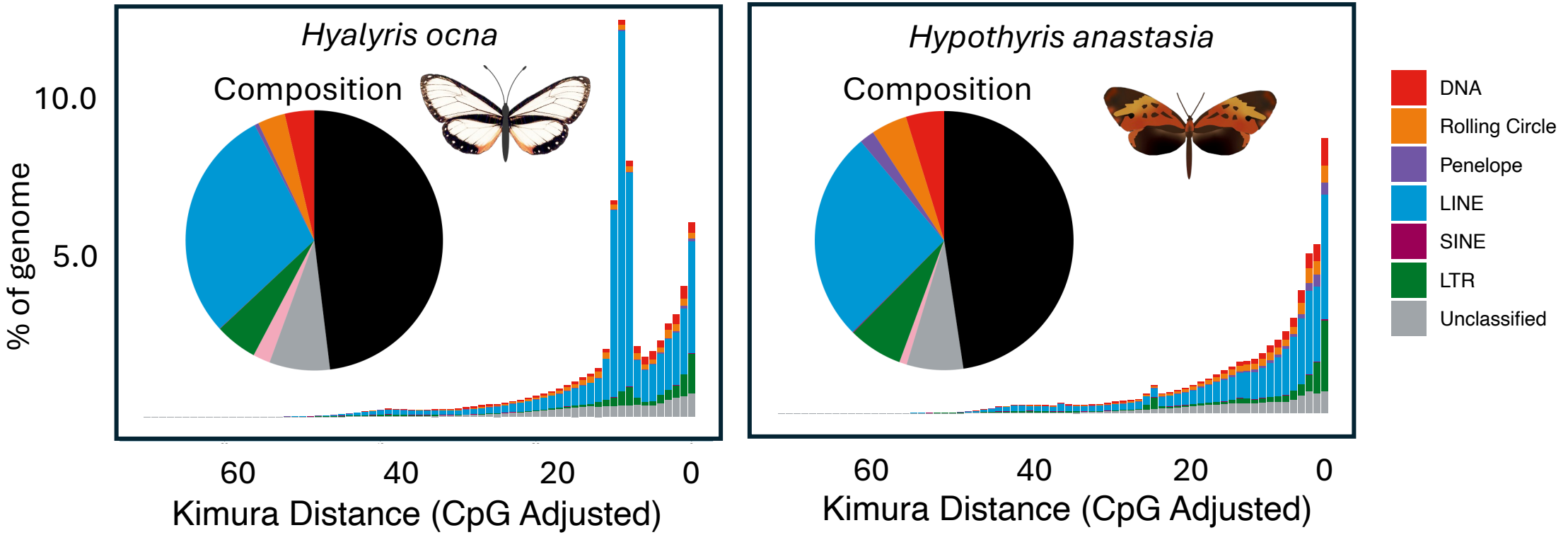


Ciccionardi et al. 2023

Dynamics through time

- Infer TE divergence time





Satellite repeats

- Centromere-associated repeats
  - Associated tandem repeat arrays
- Telomeric repeats
  - TeloBase https://doi.org/10.1093/nar/gkad672
  - $(T_xA_yG_z)_n$
  - Many insects $(TTAGG)_n$ $(TTCGGG)_n$
  - Length, sequence shift

- Microsatellites

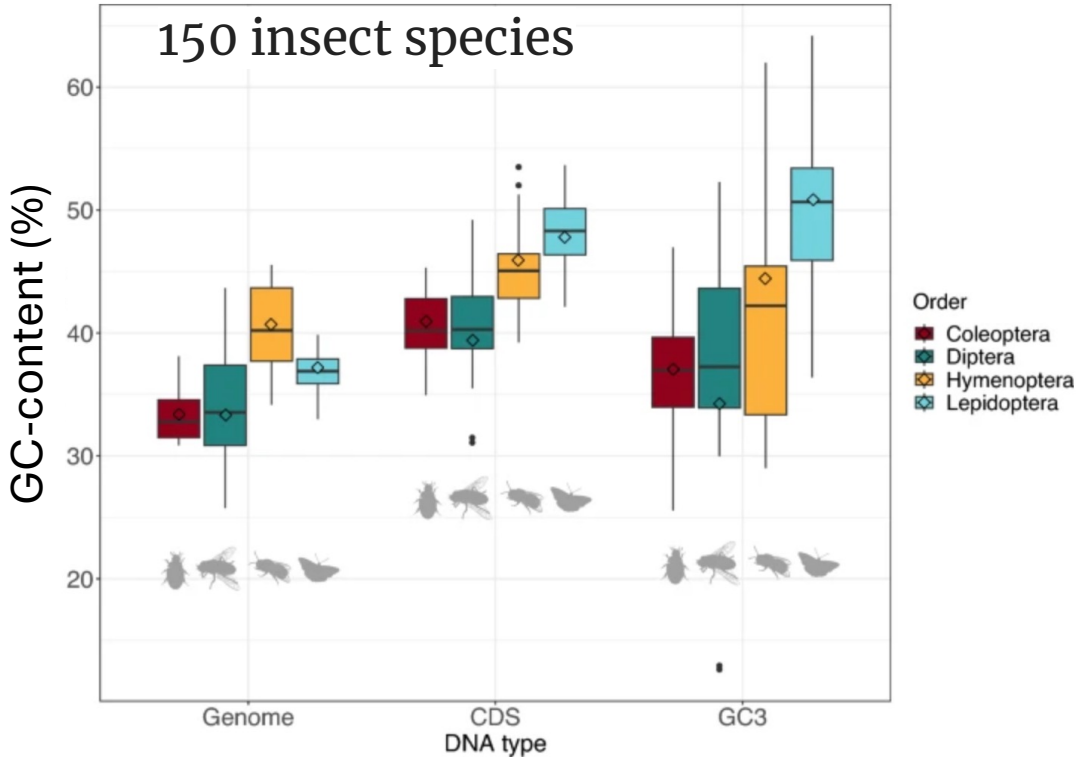
```
Ind1 ATATATATATAT
Ind1 ATATATATAT—
Ind2 ATATATATAT--
Ind2 --ATATATAT--
```

Transposable elements - annotation

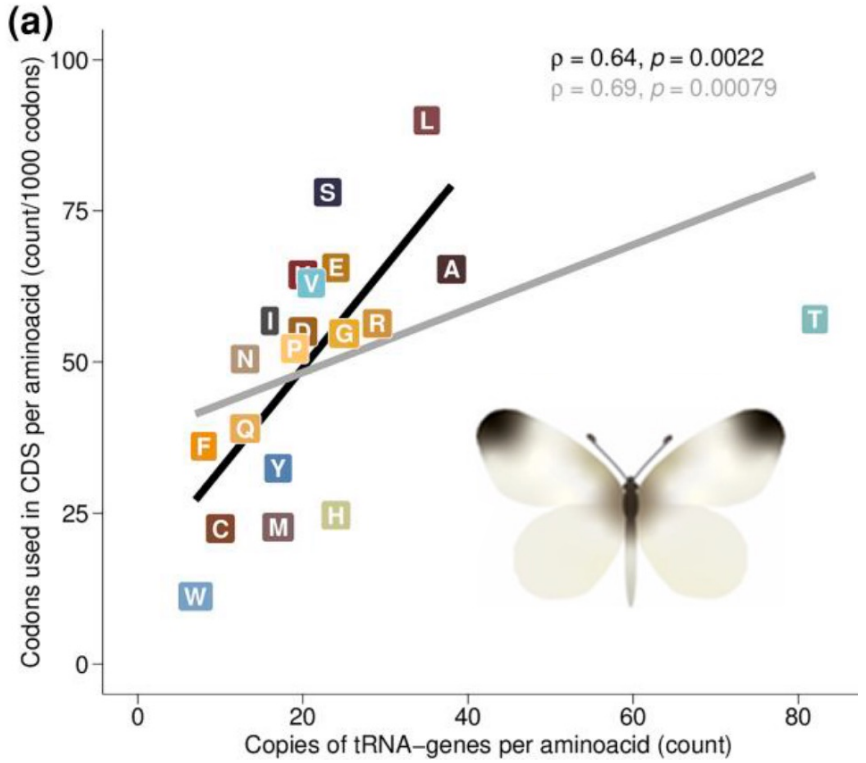
- TE prediction algorithms using structure and sequence homology
  - RepeatScout (Price et al. 2005)
  - RECON (Bao 2002)
  - LTRharvest (Ellinghaus et al. 2008)
- Consensus building and classification steps
  - RepeatModeler2 (includes the predictors above)
- Manual curation

- EarlGrey https://github.com/TobyBaril/EarlGrey
- Pantera

GC-content, codon usage and tRNA dynamics



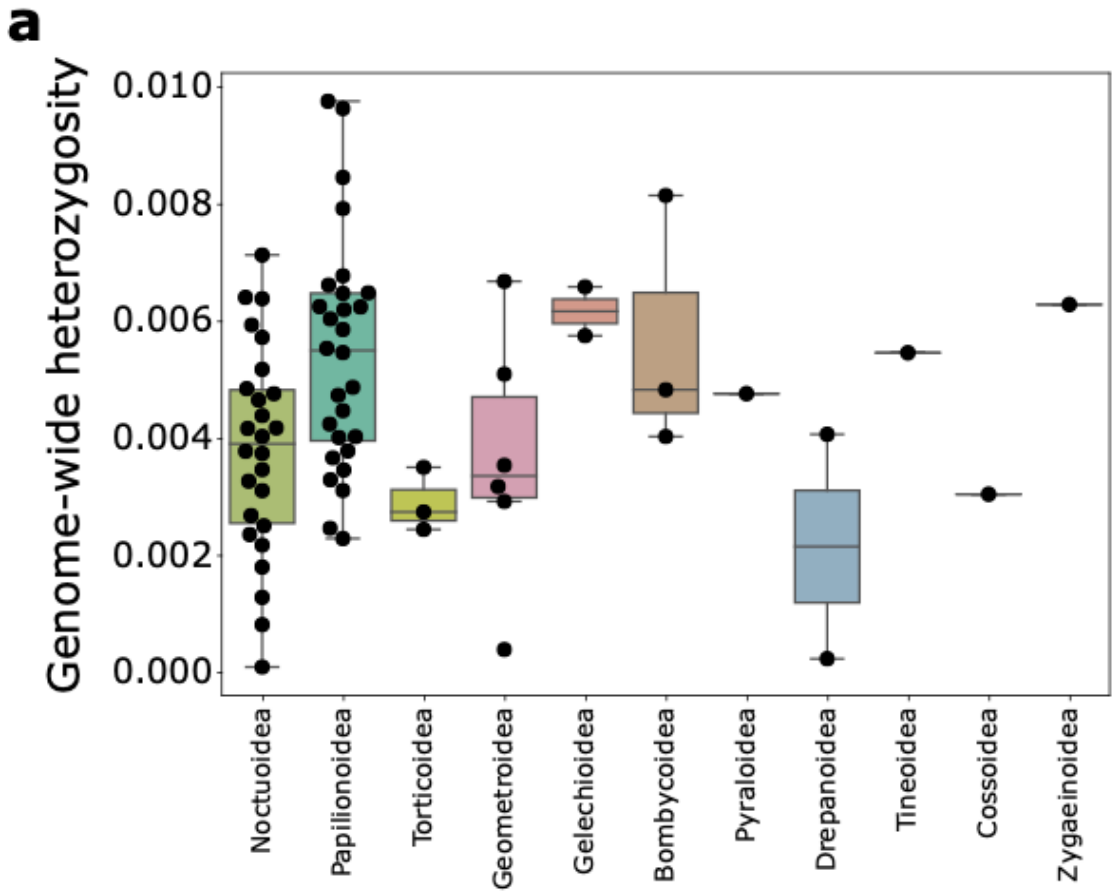
Kyriacou et al. 2024



Näsvall et al. 2023

Heterozygosity

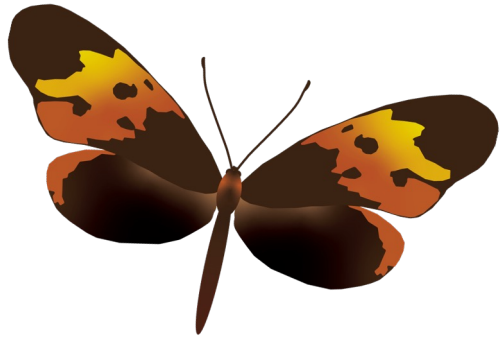
- Number of segregating sites (polymorphisms)
- Runs of homozygosity
  - Long uninterupted homozygous regions indicate low number of haplotypes segregating in the population
- Caveat only one individual although 2 two chromosomes (diploids)



Bortoluzzi et al 2023 bioRxiv

Hands-on synteny analysis

- Two species of Ithomiini
  - *Mechanitis marsaeus*
  - *Mechanitis messenoides*

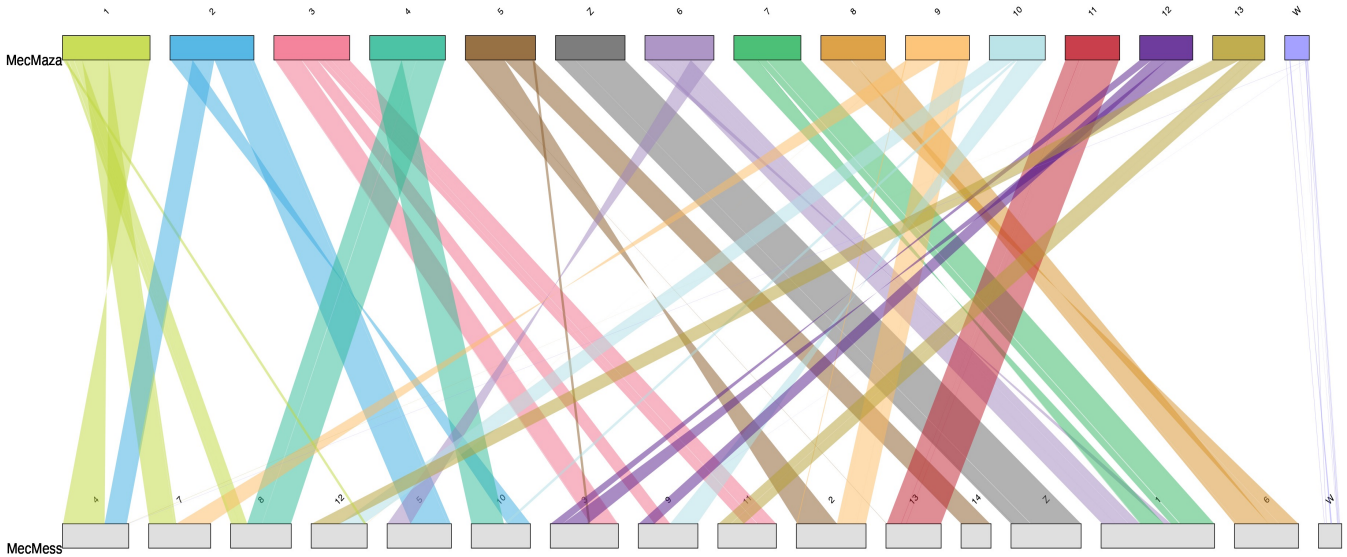


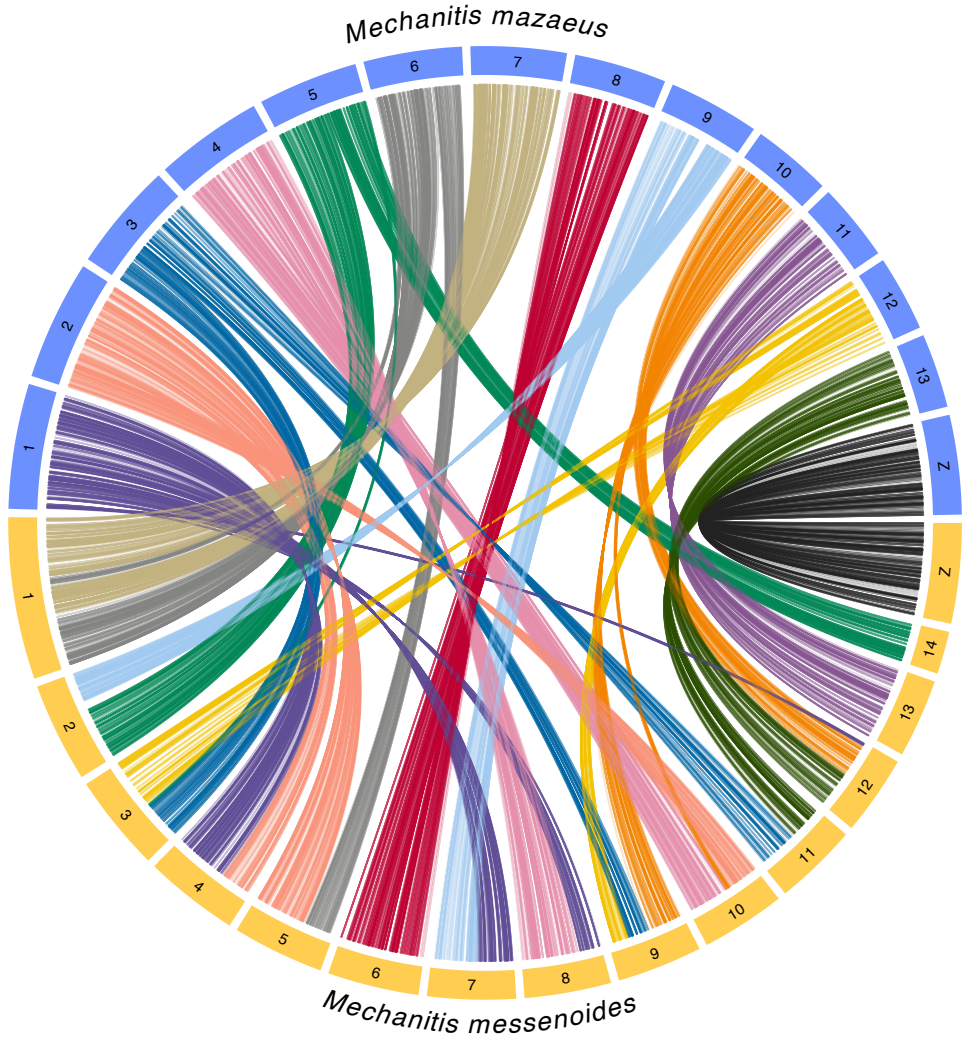
- Data type available
  - Reference genomes at NCBI

Hands-on synteny analysis

- Question:
  - Are there any chromosomal rearrangementsbetween the taxa?
- Methods
  - Whole genome alignment
    - MiniMap2
    - Ribbon plot
  - Marker orthology
    - BUSCO
    - Orthofinder
    - Circos

    - Extra: dotplot https://dgenies.toulouse.inra.fr


MecMaza
MecMess


Mechanitis mazaeus
Mechanitis messenoides