Introduction to Biodiversity Genomics

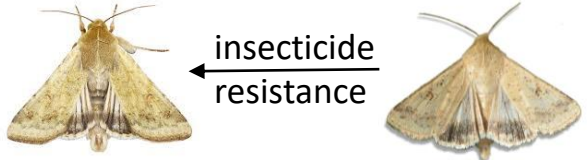
Using genomics to understand and preserve biodiversity from genetic diversity, populations, to species and ecosystems

Resolving the taxonomy
- Placing potentially new species
- Species delineation

Adaptation and speciation
- Identifying genomic regions involved in speciation
- Identifying genes underlying traits

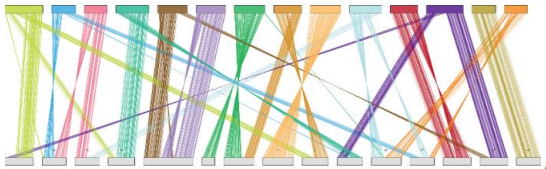
Are the species declining?
- Detecting past inbreeding
- Assessing genetic diversity

Studying gene flow
- Are populations/species hybridising or have in the past?
- Finding regions of adaptive introgression

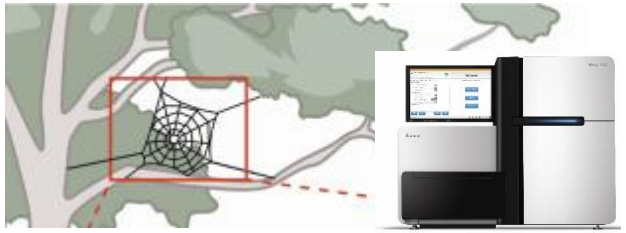


Helicoverpa zea *H. armigera*

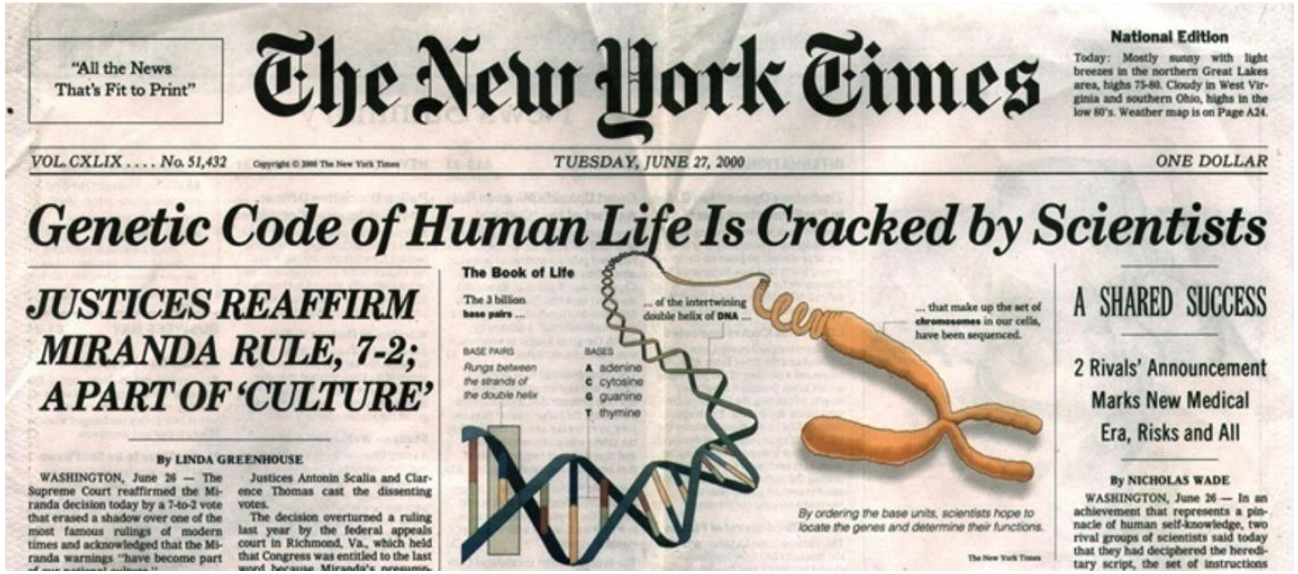
Studying genome evolution
- Gene expansions, e.g. olfactory
- Chromosomal rearrangements
- Genome size evolution (TEs, etc)



Which species occur here?
- Identifying biodiversity hotspots
- Monitoring effectiveness of conservation strategies



Human Genome Project



"All the News That's Fit to Print"

The New York Times

National Edition

Today: Mostly sunny with light breezes in the northern Great Lakes area, highs 75-80. Cloudy in West Virginia and southern Ohio, highs in the low 80's. Weather map is on Page A24.

VOL. CXLIX . . . . No. 51,432 TUESDAY, JUNE 27, 2000 ONE DOLLAR

Genetic Code of Human Life Is Cracked by Scientists

JUSTICES REAFFIRM MIRANDA RULE, 7-2; A PART OF 'CULTURE'

By LINDA GREENHOUSE

WASHINGTON, June 26 — The Supreme Court reaffirmed the Miranda decision today by a 7-to-2 vote that erased a shadow over one of the most famous rulings of modern times and acknowledged that the Miranda warnings "have become part

Justices Antonin Scalia and Clarence Thomas cast the dissenting votes.

The decision overturned a ruling last year by the federal appeals court in Richmond, Va., which held that Congress was entitled to the last

The Book of Life

The 3 billion base pairs ...

... of the intertwining double helix of DNA ...

... that make up the set of chromosomes in our cells, have been sequenced.

BASE PAIRS
Rungs between the strands of the double helix

BASES
A adenine
C cytosine
G guanine
T thymine

By ordering the base units, scientists hope to locate the genes and determine their functions.

The New York Times

A SHARED SUCCESS

2 Rivals' Announcement Marks New Medical Era, Risks and All

By NICHOLAS WADE

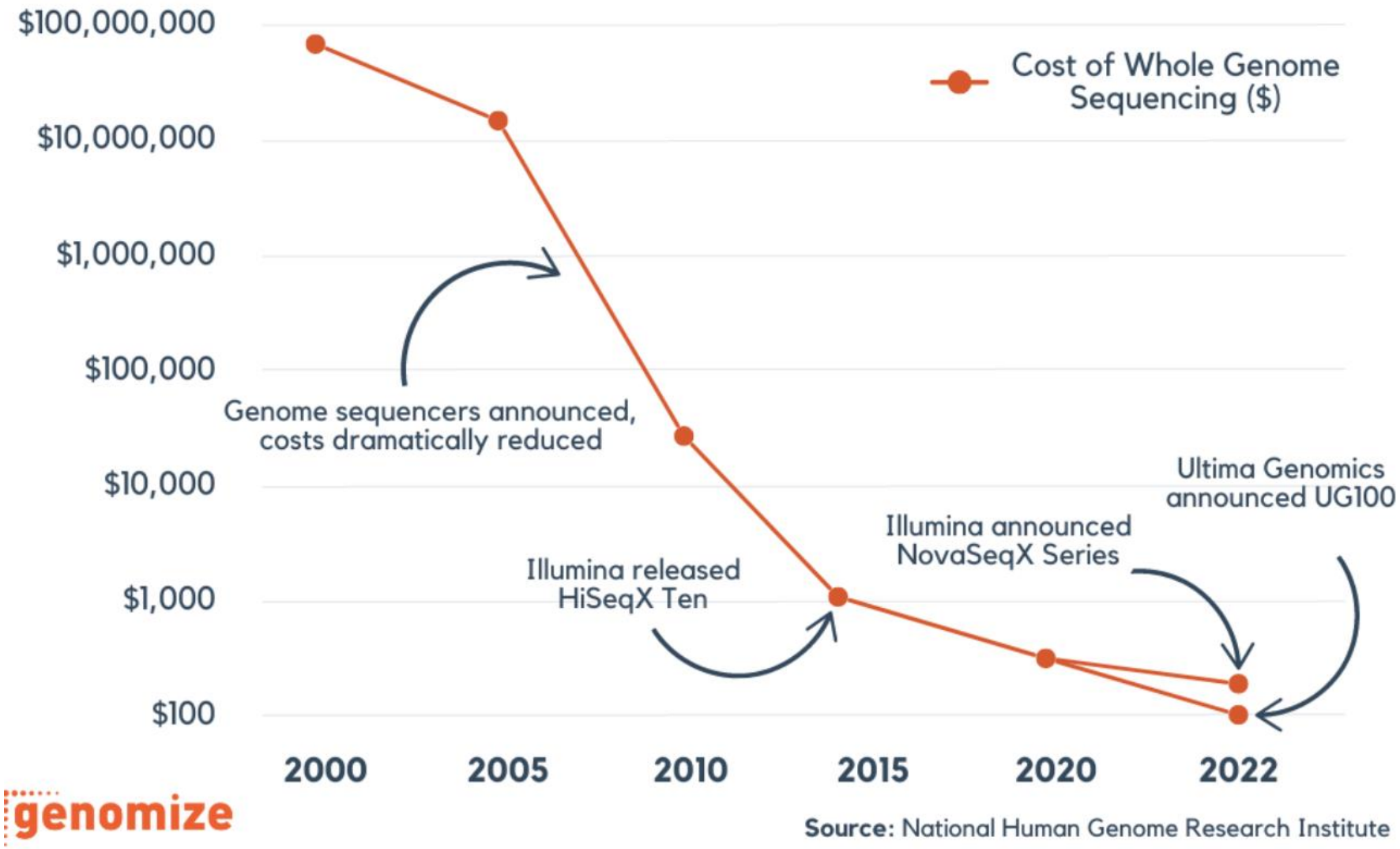
WASHINGTON, June 26 — In an achievement that represents a pinnacle of human self-knowledge, two rival groups of scientists said today that they had deciphered the hereditary script, the set of instructions

- Human genome project – started in 1990, completed in 2003
- Sequenced across ~20 institutions worldwide
- Cost an approximate $5 billion US dollars

The first human reference genome transformed modern medicine and understanding of human evolution and physiology

- Comparing populations e.g. to study how humans spread across the globe
- Finding introgression with neanderthals and denisovans
- Identifying genes under selection, like the laktase gene
- Finding genes causing diseases like breast cancer
- Understanding how cancer develops
- Personalised medicine
- Single-cell sequencing to understand which genes are active in which cells

Sequencing costs are decreasing rapidly

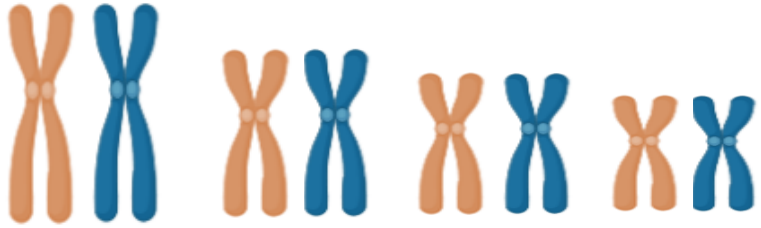


Since Oct 2023 PacBio Revio (66 Gbp per lane in 15 kb reads)


EARTH
BIOGENOME
PROJECT
Darwin
TREE
of
LIFE
Project Psyche
African BioGenome Project
Genomics for the future of biological diversity across Africa
ATLASea
Atlas des génomes marins
VGP
VERTEBRATE
GENOMES
PROJECT
A PROJECT OF THE G10K CONSORTIUM
BAT 1K
ERGA
CREATING A NEW FOUNDATION FOR BIOLOGY
Sequencing Life for the
Future of Life

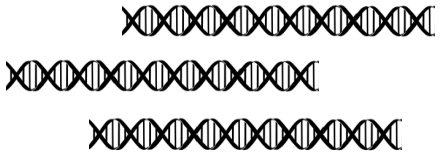
Why do we need a reference genome for whole-genome sequencing projects?

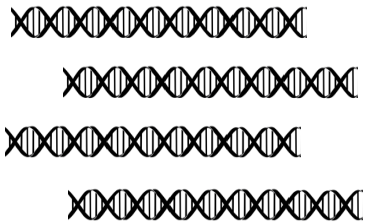
Genome = set of all chromosomes



Problem:
We do not know which of these sequences to compare

Break into many million short DNA fragments





Short-read sequencing machine (e.g. Illumina)

Millions of reads, 150 bp long

GATGCT
ATAGTG
GTGTGG

GTGTAG
GATGCT
CTGAGT
TGCTGA

How do we make a reference genome?

Genome = set of all chromosomes

Break into many million
long (10-50 kbp) DNA fragments

Long-read sequencing machine (e.g. PacBio)

more expensive than short-read sequencing

Millions of reads, 10-20 kbp long

GATGCTGAGTA
ATAGTGTGGAT
GTGTGGATGTG
TGCTGAGTTCG
TGGATGCTGAT
CTGAGTTCTCG

puzzling them together (aligning them to each other)

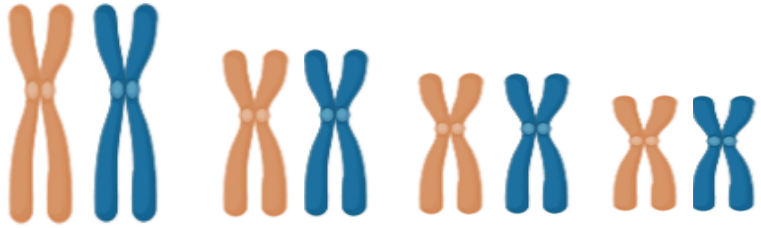
Reference genome

ATAGTGTGGATGCTGAGTTCGT

ATAGTGTGGATG
GTGTGGATGCT
TGGATGCTGAGT
GATGCTGAGTTC
TGCTGAGTTCG
CTGAGTTCGT

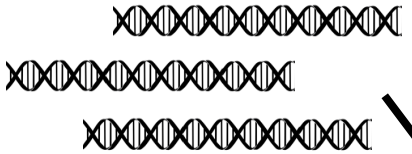
How do we use this reference genome?

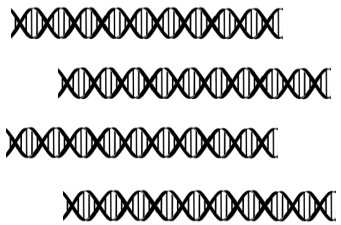
Genome = set of all chromosomes



Solution:
The reference genome allows us to place the reads so that we can compare them across individuals, populations or species

Break into many million short DNA fragments





Short-read sequencing machine (e.g. Illumina)

GATGCT
ATAGTG
GTGTGG

GTGTAG
GATGCT
CTGAGT
TGCTGA

Reference genome

ATAGTGTGGATGCTGAGTTCGT

GATGCT

ATAGTG
GTGTGG

GTGTAG

GATGCT

CTGAGT

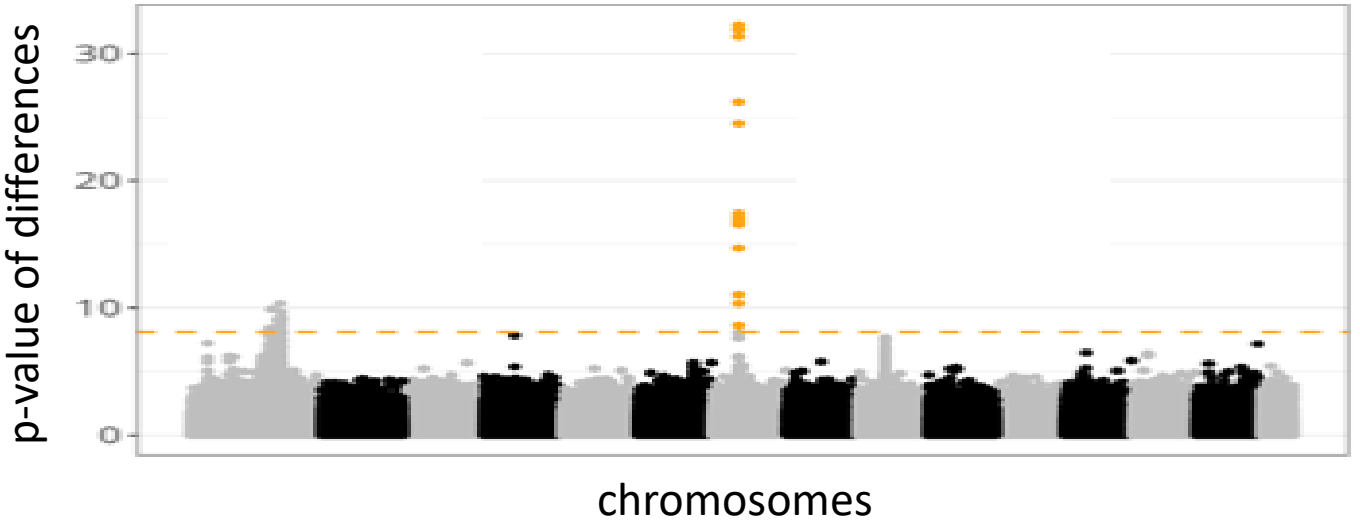
TGCTGA

Do these two butterflies belong to different species?

Eva van der Heijden

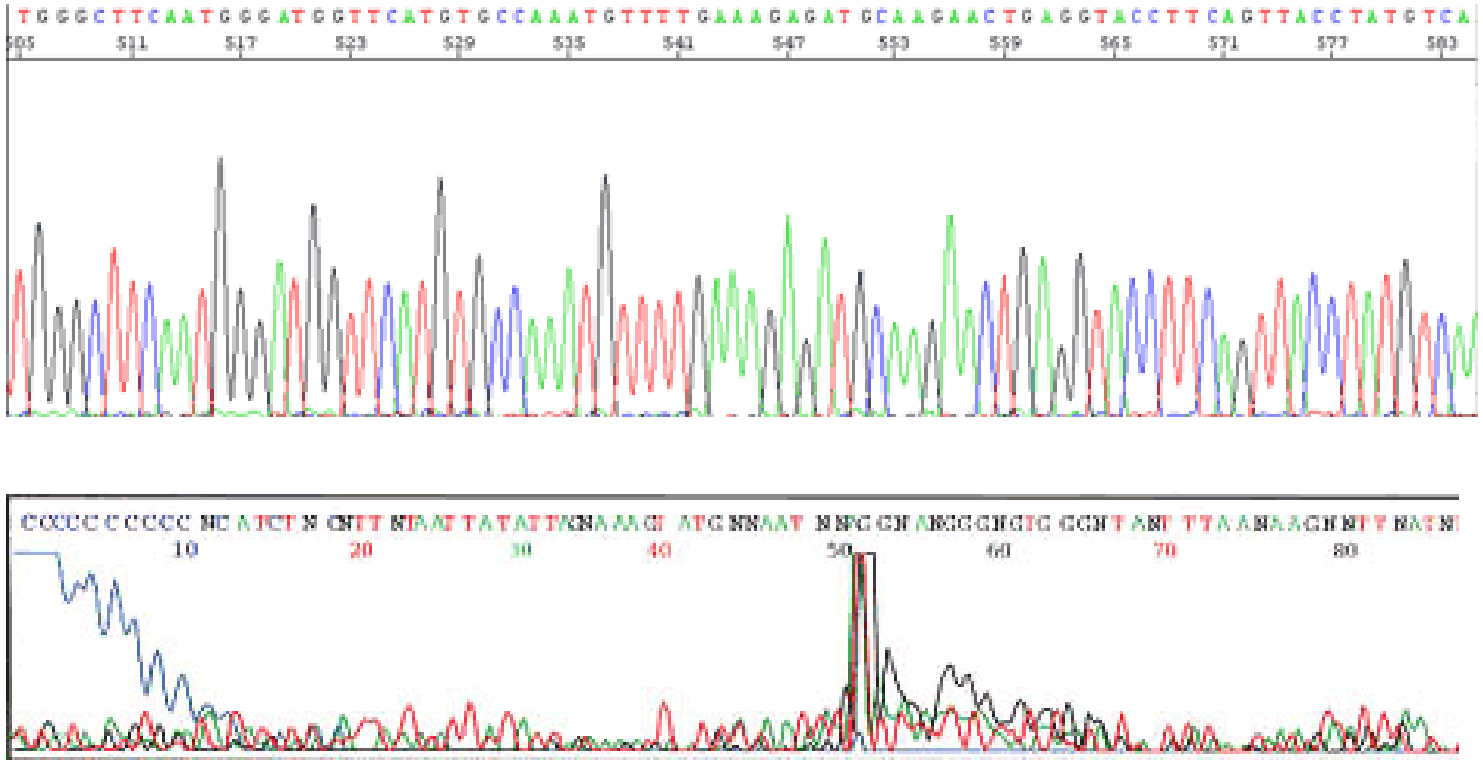
CRISPR butterflies with *cortex/ivory* knocked-out

They are only significantly different in one region in the genome, right next to *cortex/ivory,* which are also known to affect colour patterns in other butterflies and moths



Introduction to high-throughput or next-generation sequencing

Sanger Sequencing (since 1980s)



- Possible to manually check each sequence and resequence failed sequences
- Requires primer sequences and has very low throughput (expensive per bp)

Two main types of high-throughput sequencing

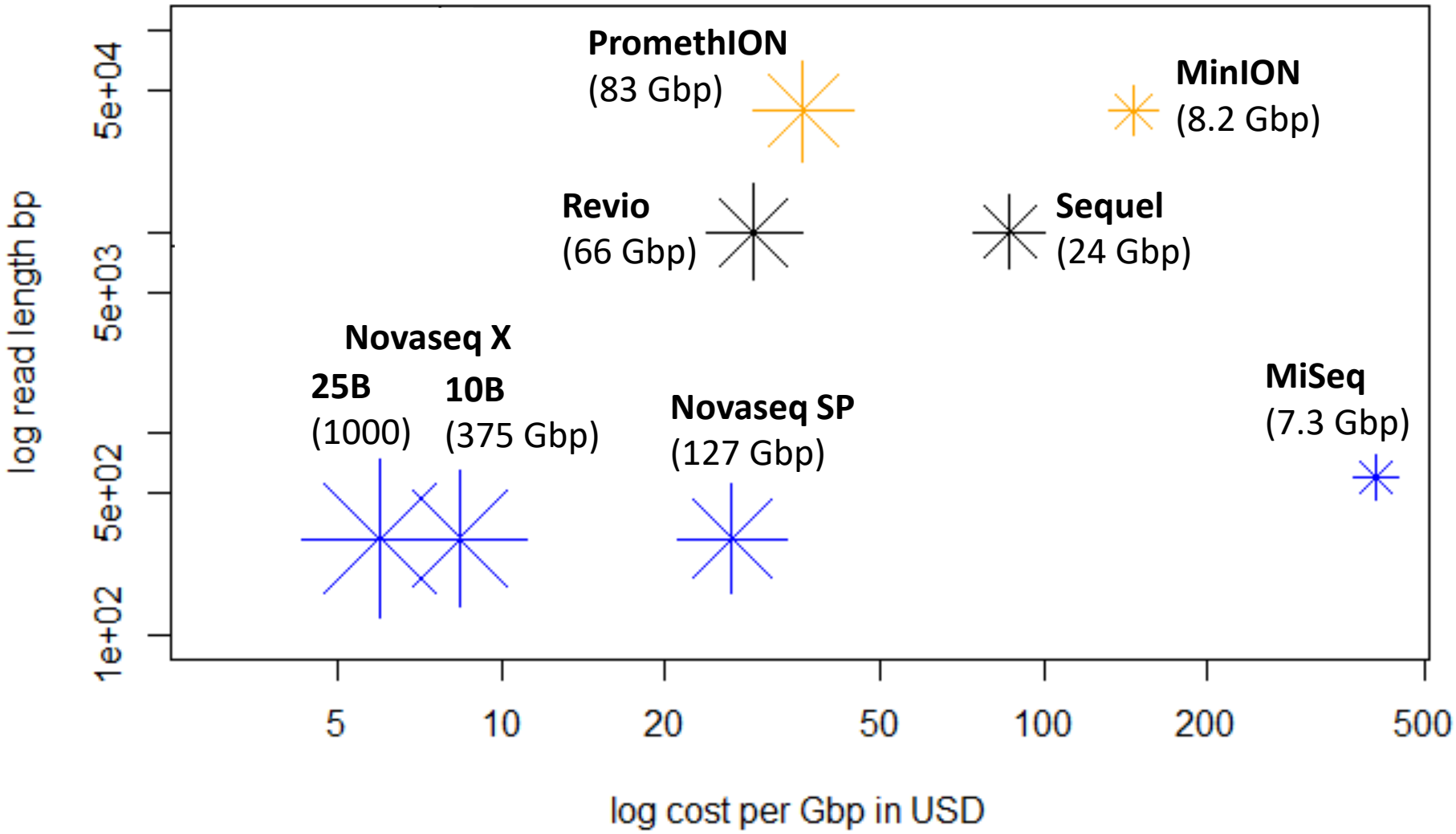
- **Short-read sequencing**
  - Reads are typically 150 bp long
  - Cheaper than long-read sequencing
  - E.g. Illumina, soon probably also Ultima Genomics

- **Long-read sequencing**
  - Reads are typically >10 kb long (PacBio: 15-20 kbp, Nanopore: 10-100 kbp)
  - More expensive than short-read technologies
  - Required for making a reference genome
  - E.g. PacBio or Nanopore

Read length versus per Gbp sequencing costs for different sequencing machines (note the axes are in logarithmic scale)

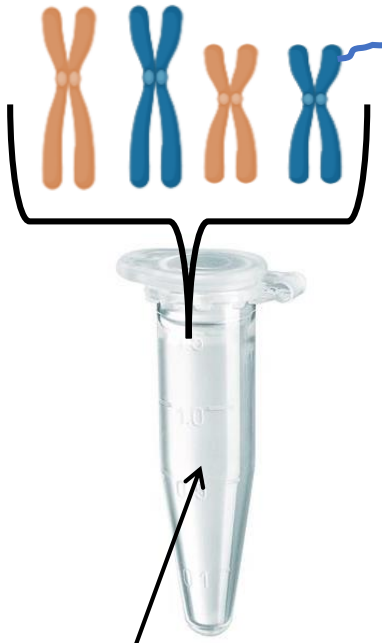
ONT
PacBio
Illumina

Star size shows the total throughput per lane, also given in parentheses ()



Whole-genome sequencing

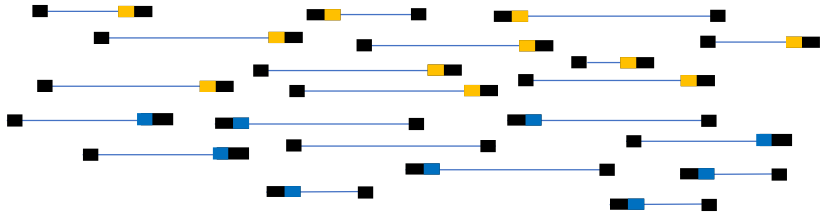
Genome
= complete set of chromosomes



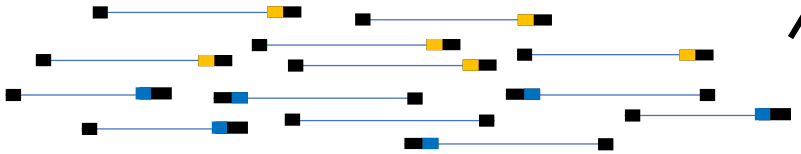
DNA (chromosomes)

Breaking chromosomes into shorter pieces for sequencing

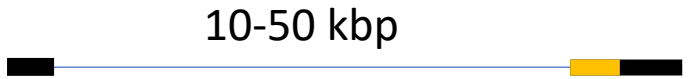
Adding **sequencing adapters** incl. individual index



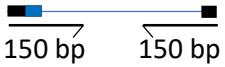
Size selection



Long-read sequencing (PacBio or Nanopore/ONT)
paired-end sequencing

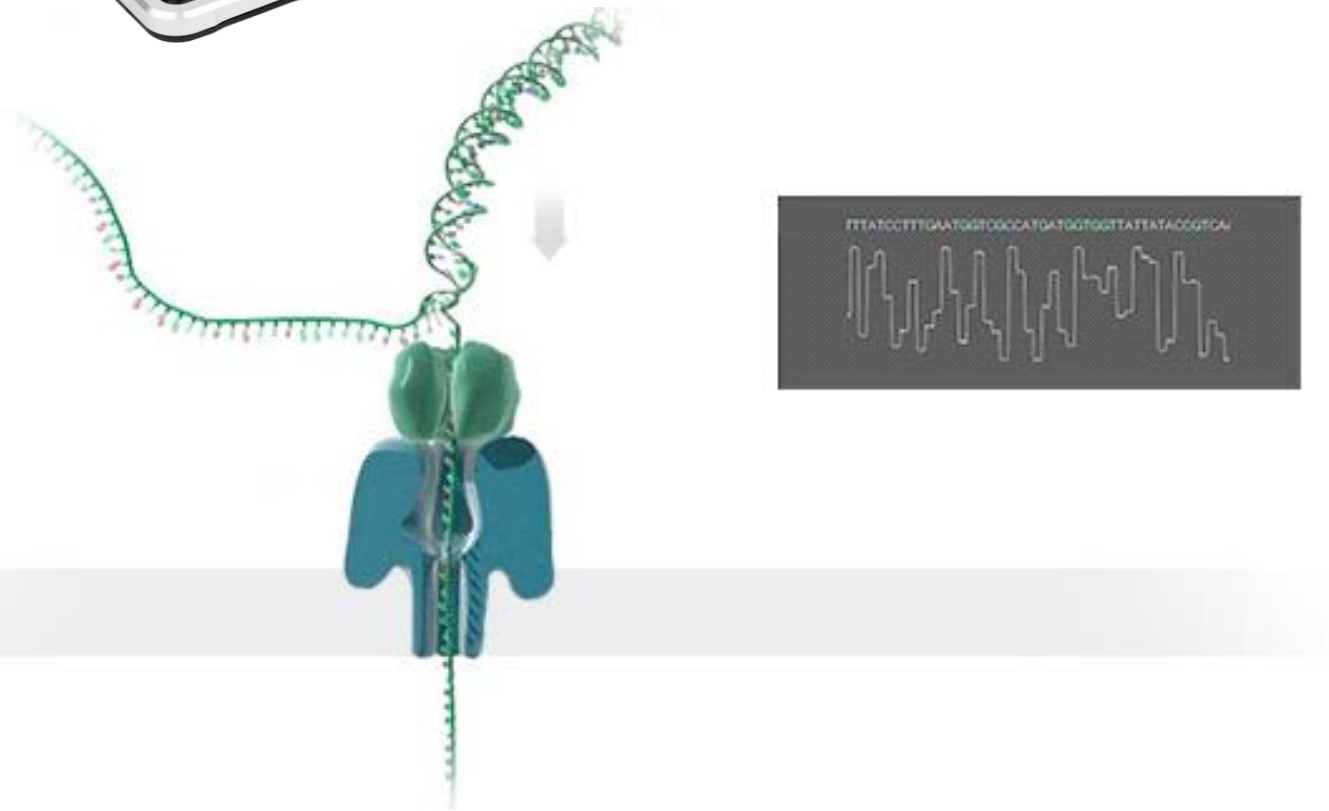


Short-read sequencing (e.g. Illumina)
paired-end sequencing

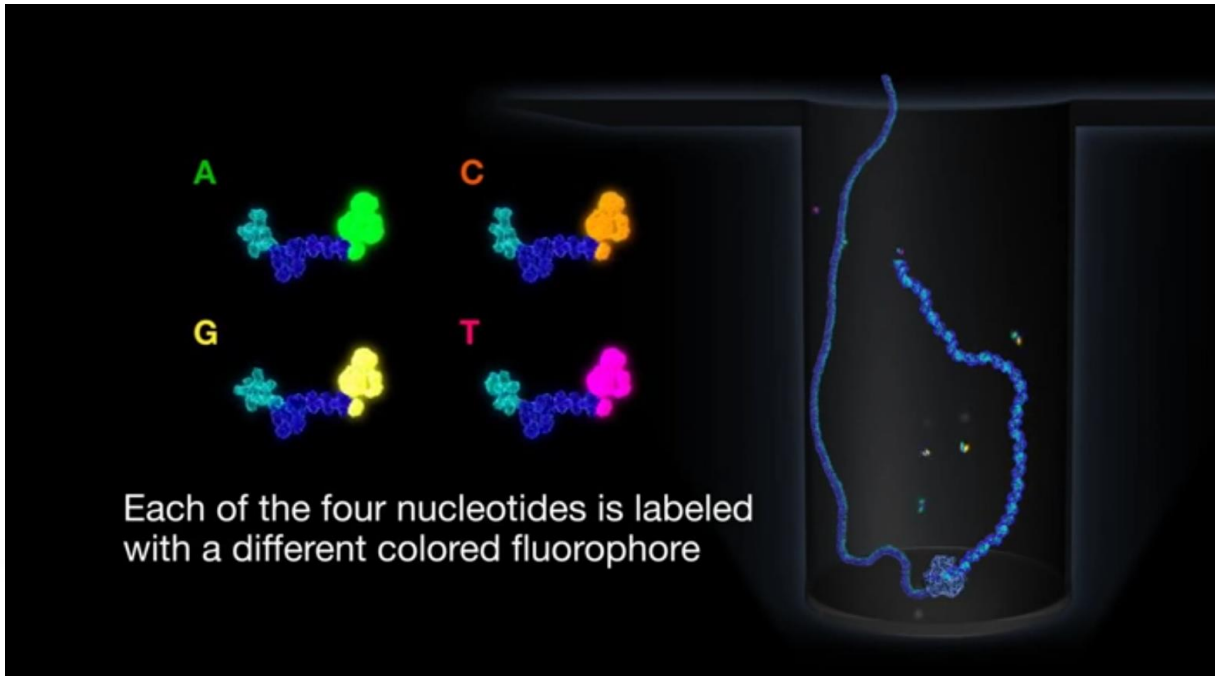


Long read sequencing technologies

Nanopore

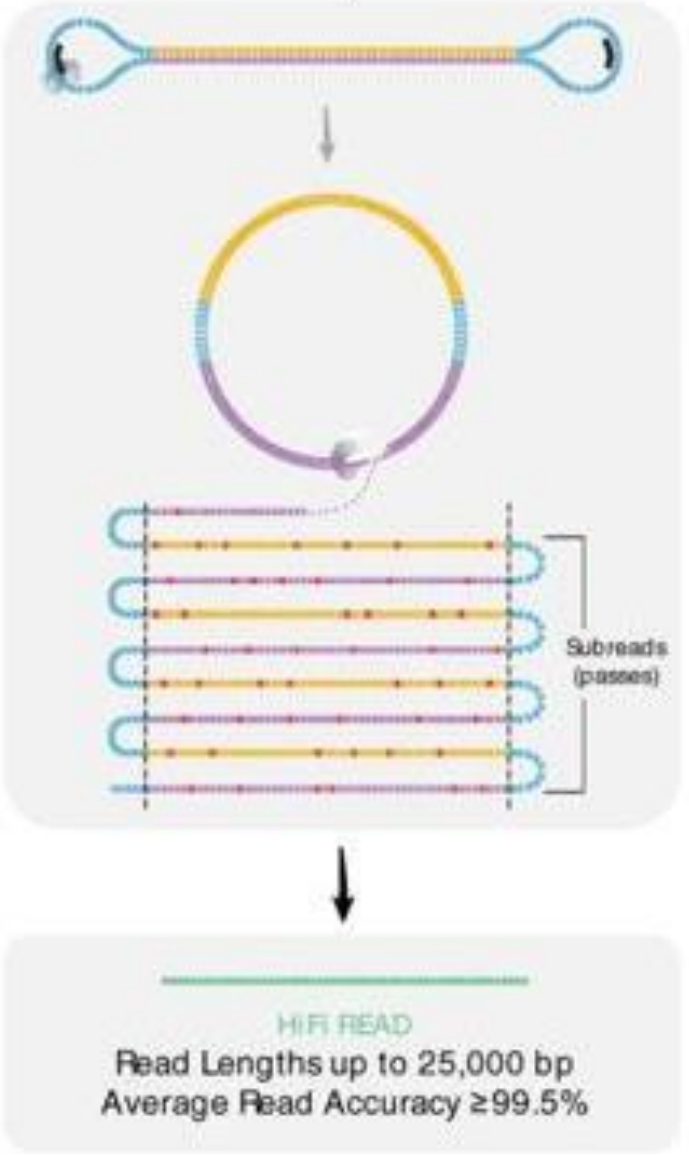


PacBio



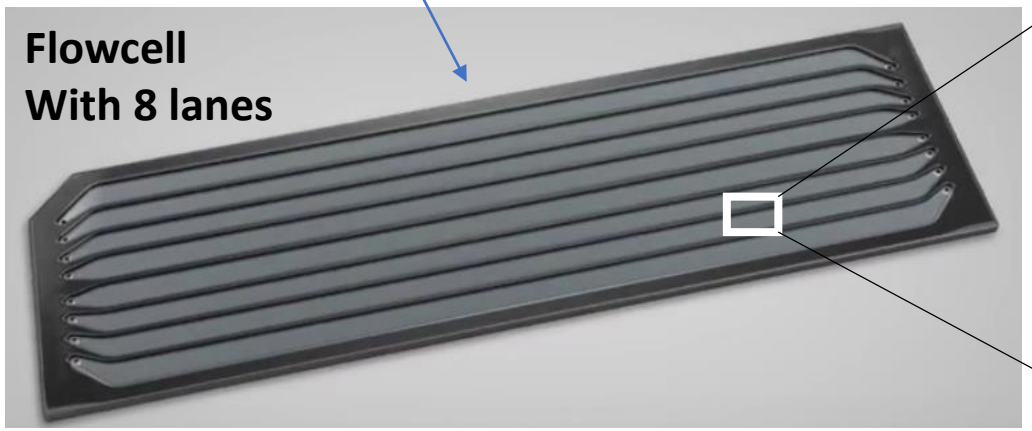
PacBio HiFi reads (99.95% accurary)

Each DNA-fragment is sequenced many times to get a high-quality consensus (=summary) read

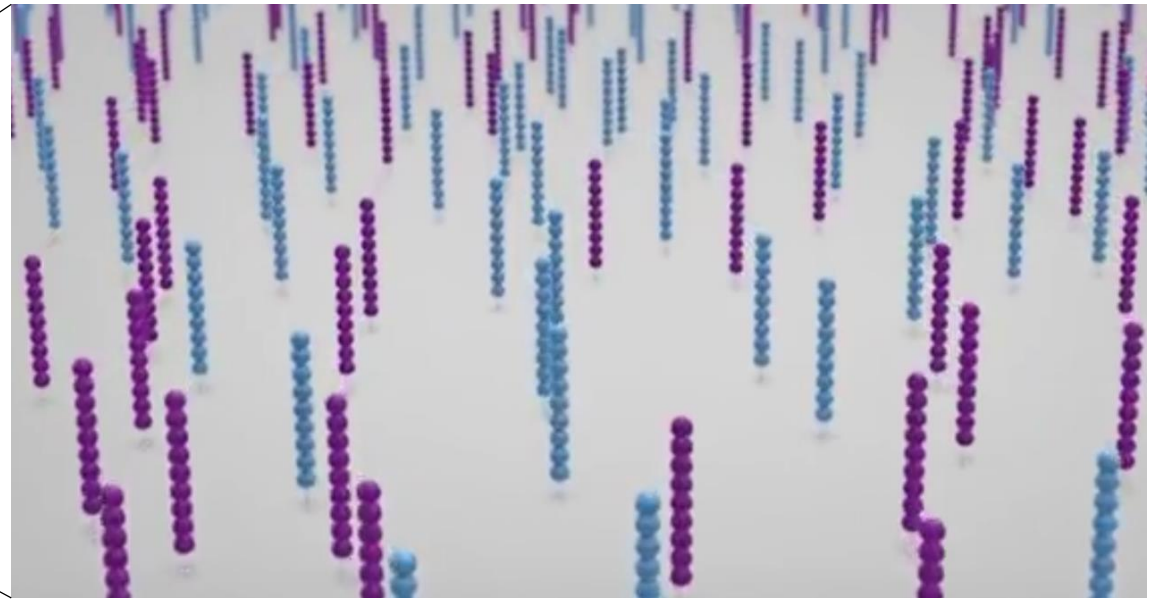


Illumina flowcell

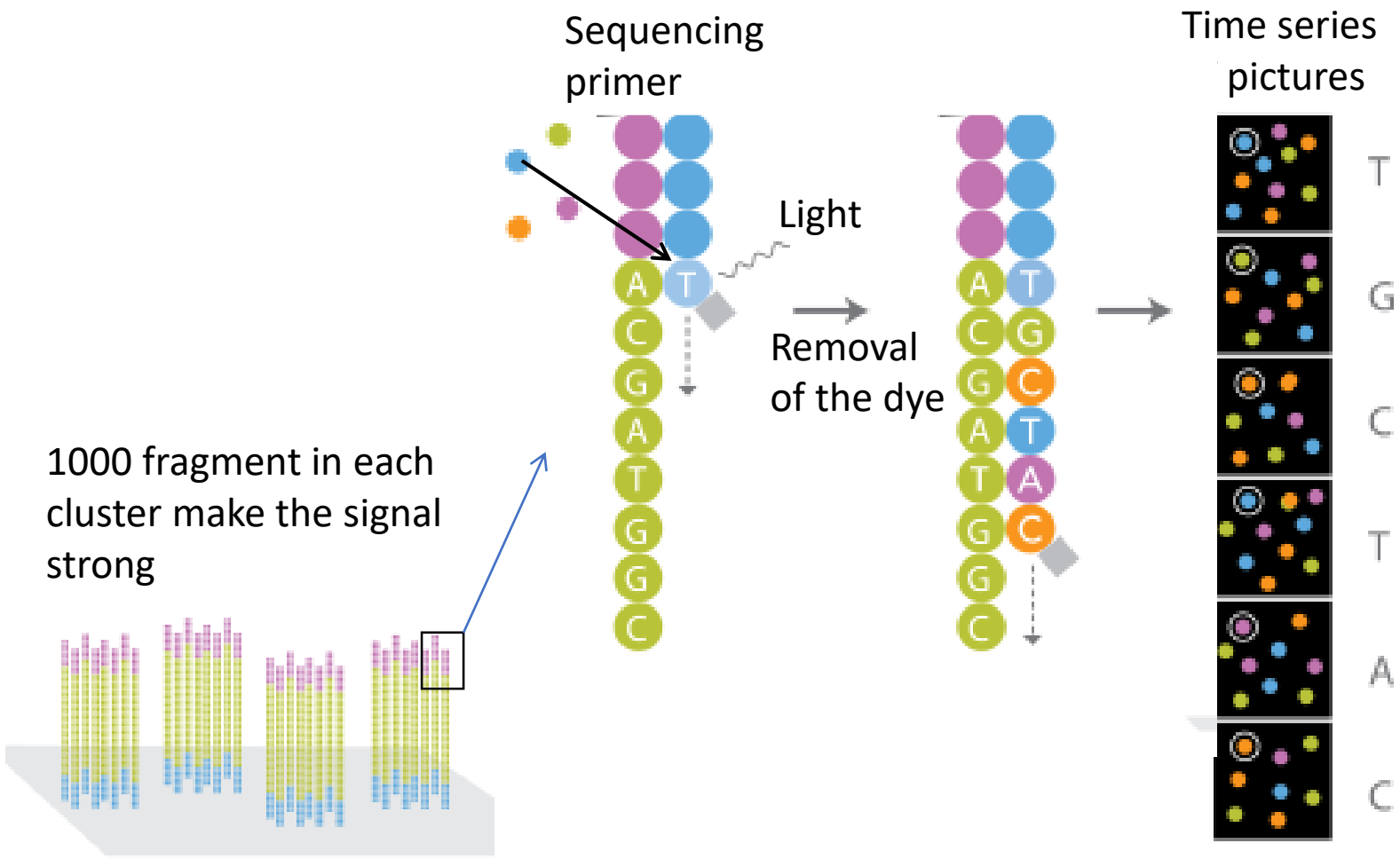
DNA fragments with Illumina adapters

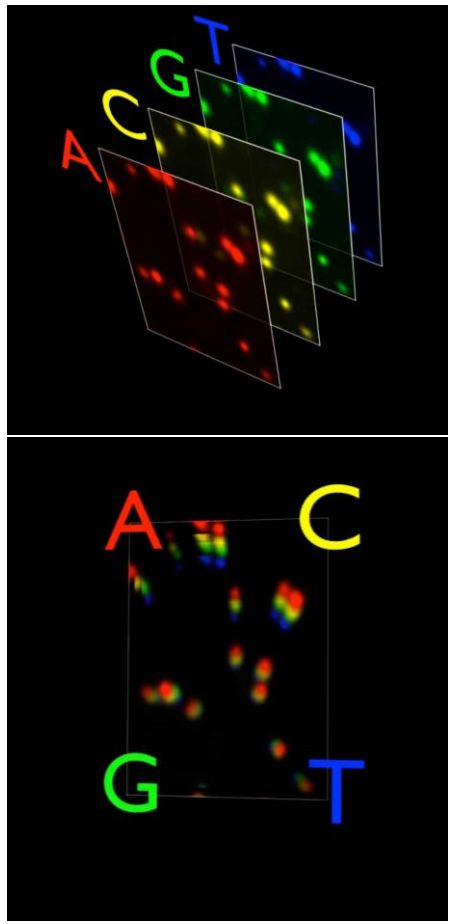


Each lane contains a dense lawn of Illumina primers



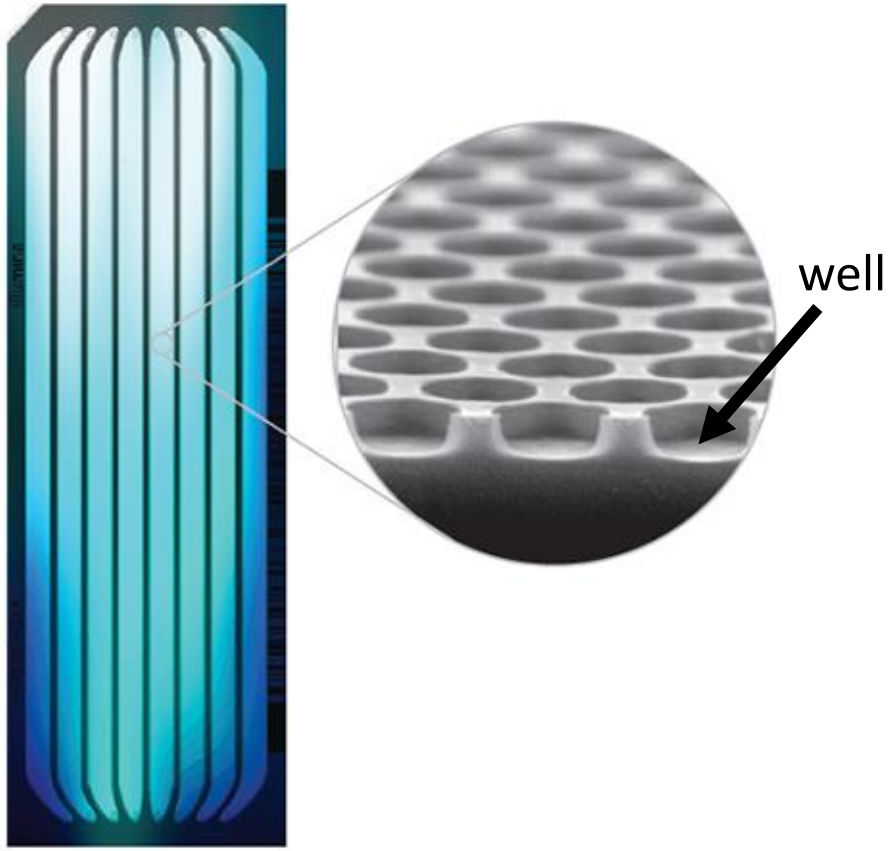
Short-read sequencing with Illumina

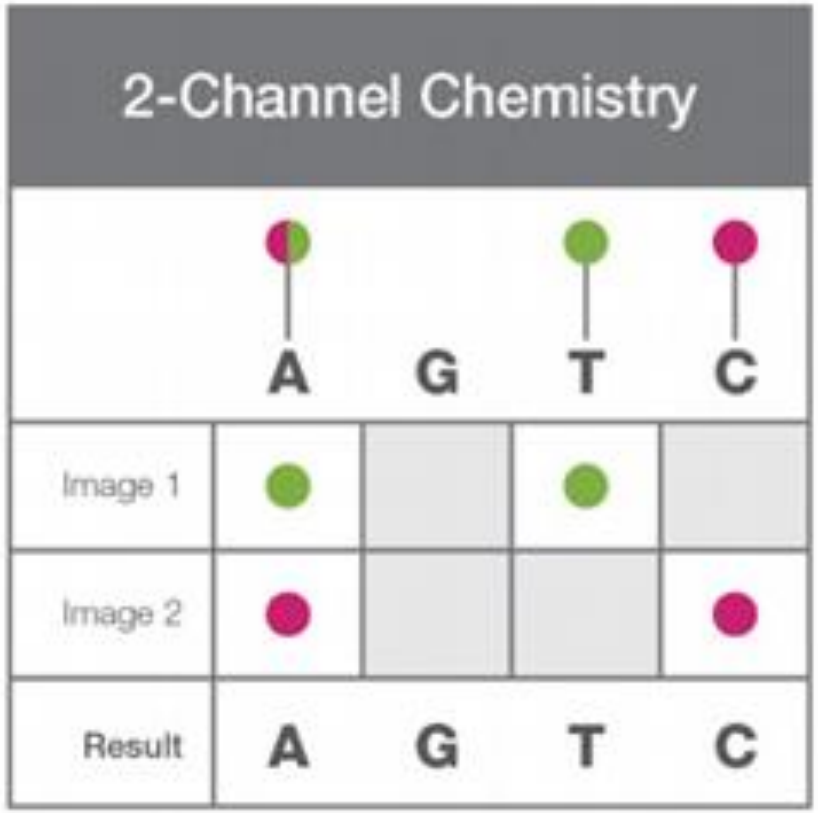




Newer Illumina machines use wells and only 2 colours

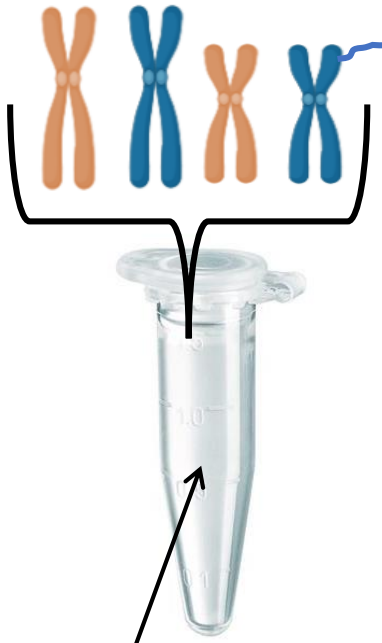
(e.g. Novaseq, Nextseq, MiniSeq. This makes it faster and cheaper)





Whole-genome sequencing

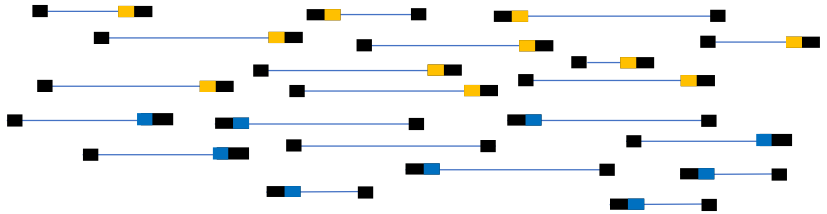
Genome
= complete set of chromosomes



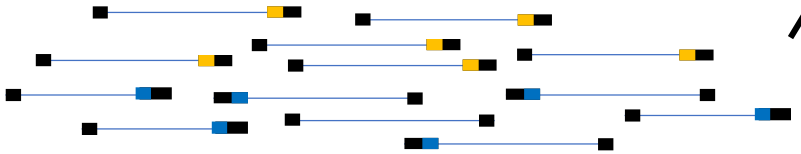
DNA (chromosomes)

Breaking chromosomes into shorter pieces for sequencing

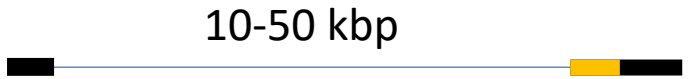
Adding **sequencing adapters** incl. individual index



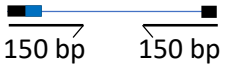
Size selection



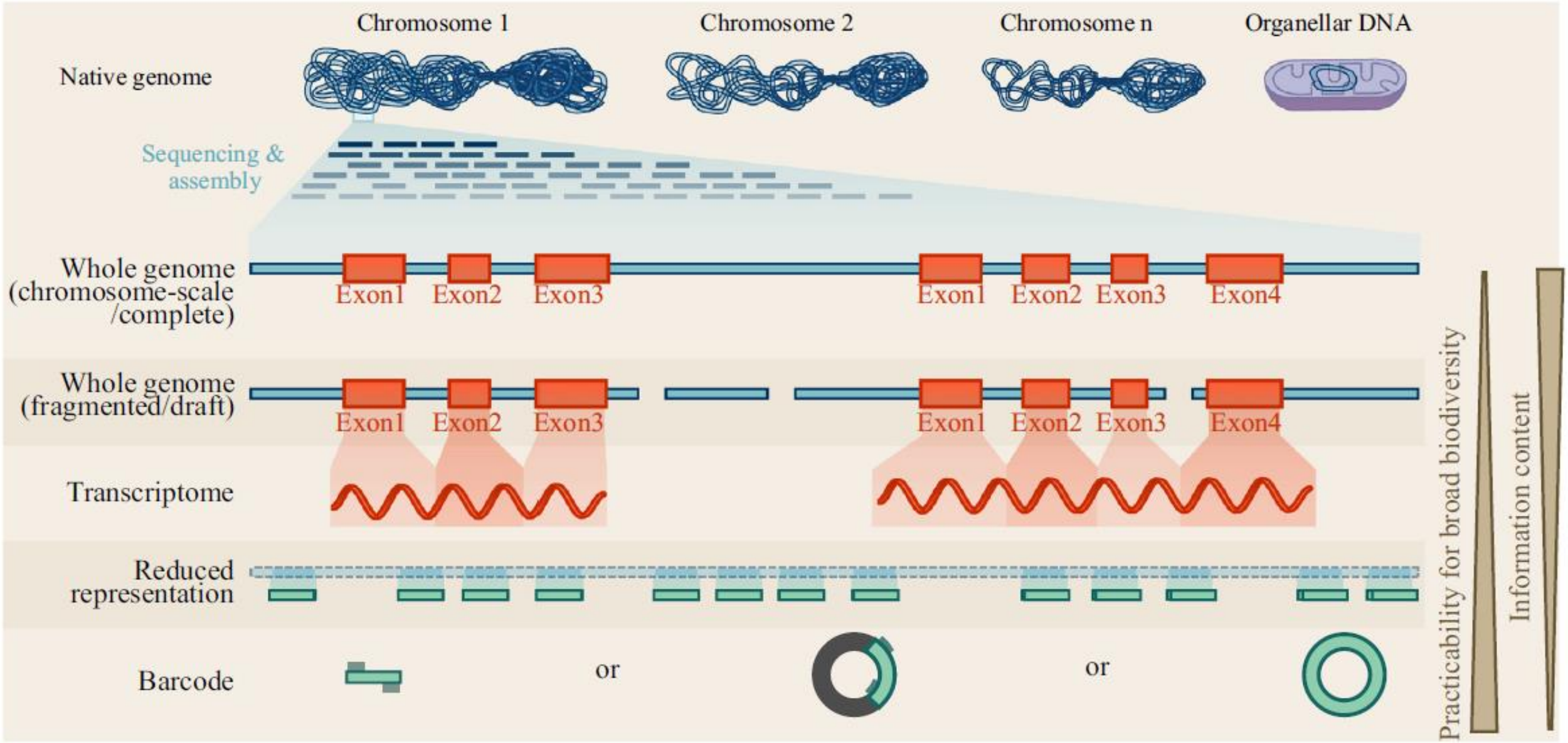
Long-read sequencing (PacBio or Nanopore/ONT)
paired-end sequencing



Short-read sequencing (e.g. Illumina)
paired-end sequencing



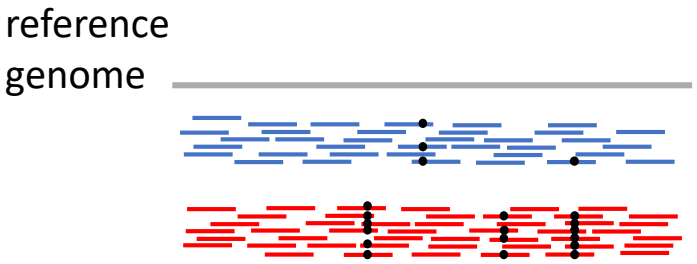
Sequencing approaches for biodiversity genomics



Sequencing approaches for biodiversity genomics

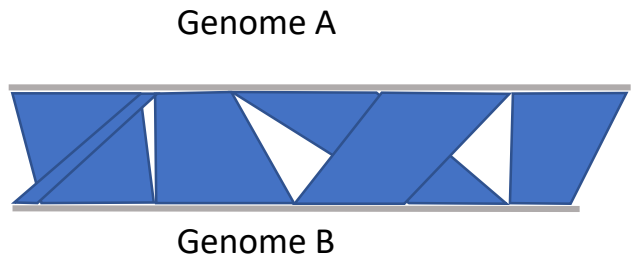
Whole-genome requencing (short-read data)

- Requires a reference genome
- individuals need to be from the same or closely related species
- Complete genome sequenced



Genome assembly comparisons (long-read data)

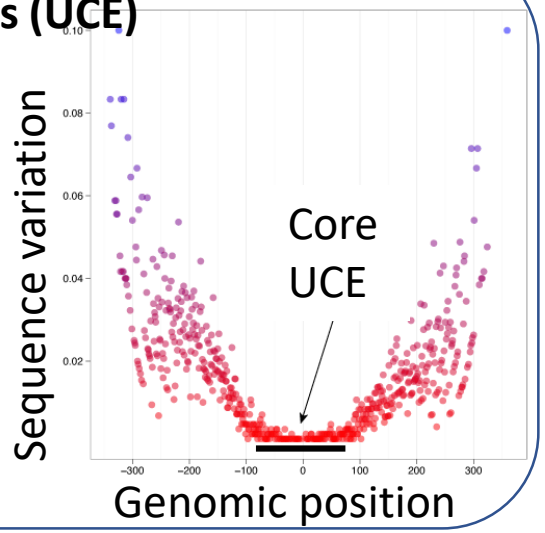
- Comparative genomics – studying structural variation between species, can be distantly related
- Gene expansions, transposable elements etc
- Phylogenomics across deeply divergent species
- Pangenomics – multi-genome assemblies to study within-species variation in structural variants



Reduced-representation techniques
(only parts of the genome sequenced)

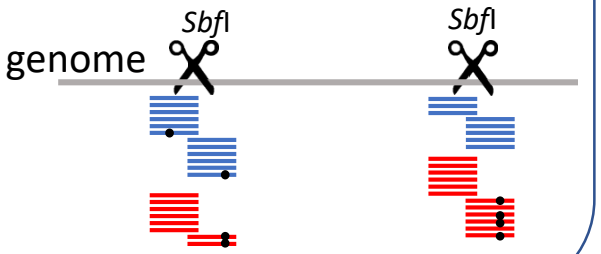
Ultra-conserved elements (UCE)

- Sequence capture with baits based on genomic regions that are conserved across many species
- Works with highly divergent species



Restriction Associated Sequencing (RAD)
(similar methods: GBS, ddRAD)

- does not require primers/baits or reference genome
- individuals need to be from the same or closely related species
- Information from thousands of loci distributed across the genome

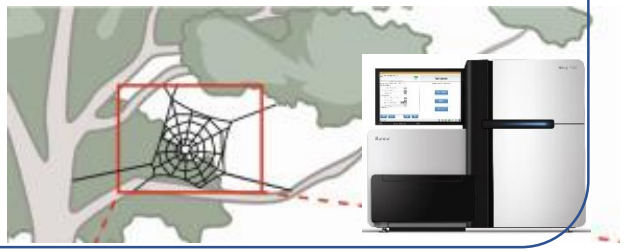


Targeted or amplicon sequencing, e.g. barcoding

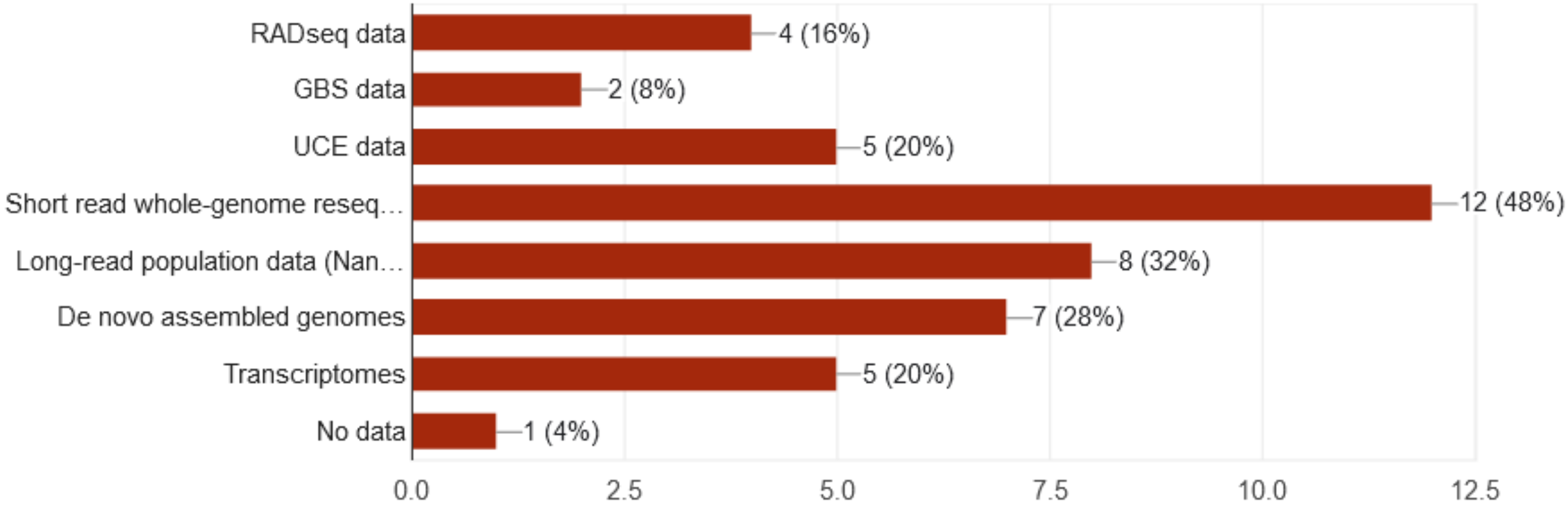
- Sequencing one or few genes
- requires primers
- e.g. CO1 (mitochondrial barcoding region), advantage: large database (BOLD) available to compare to for species identification

Environmental DNA (eDNA)

- Mostly CO1 sequencing from soil, water, air (spider webs)
- Identifying local species
- Studying species richness



Data that course attendees are working on



Trade-offs: Splitting reads (i.e. costs) among:

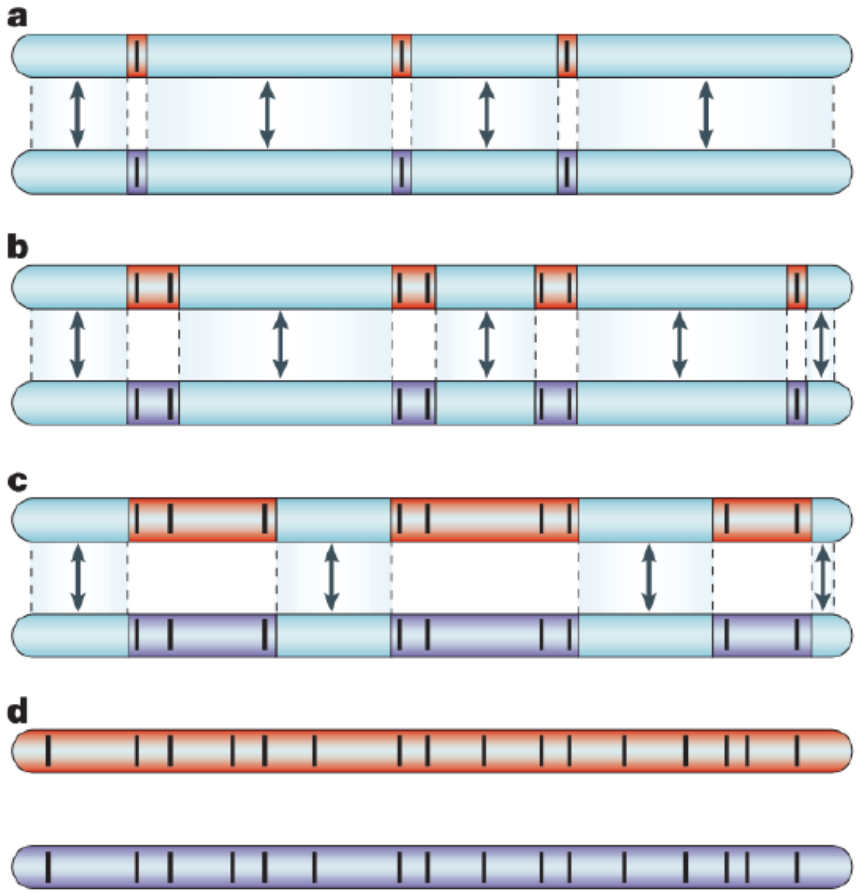
Total data gets divided by:

- Number of sites to sequence
  - Depends on genome size and sequencing stragegy, e.g. RAD versus whole-genome
- Sequencing depth (e.g. sequencing at 10x depth of coverage)
- Number of specimens to sequence

- Example: 1 Novaseq X 10 B lane
  ~2.5 billion paired-end reads of 150 bp each -> 375 Gbp data
  - 100 whole-genomes of a species with 0.375 Gbp genome size at 10x coverage
  - 19 whole-genomes of a species with 1 Gbp genome size at 20x coverage
  - 375 individuals sequenced with a RAD sequencing approach resulting in 50 Mbp at a sequencing depth of 20x

A few applications of these methods and some insights into my work

The genic concept of speciation (Wu, 2001)



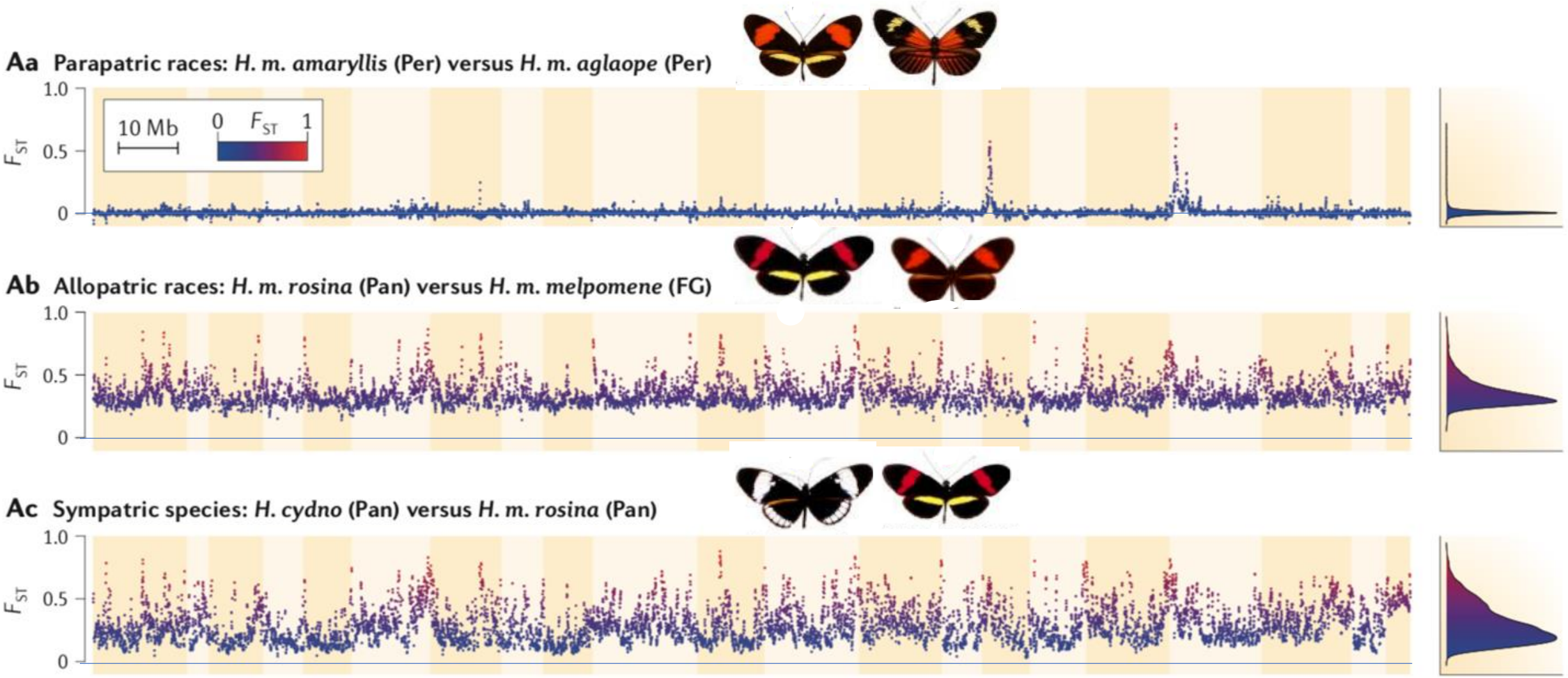
Divergent loci resist gene flow

Gene flow continues but linkage builds and divergent regions grow

Complete reproductive isolation evolves

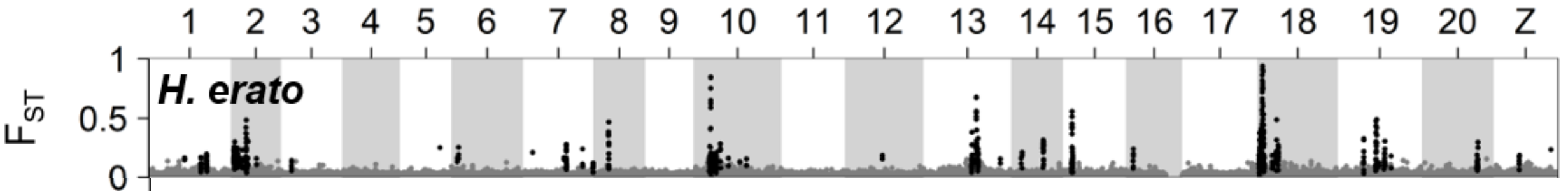
Wu (2001) **J Evo Bio**
Wu & Ting (2001) **Nat Rev Gen**

Genomics evidence for Wu's genic view



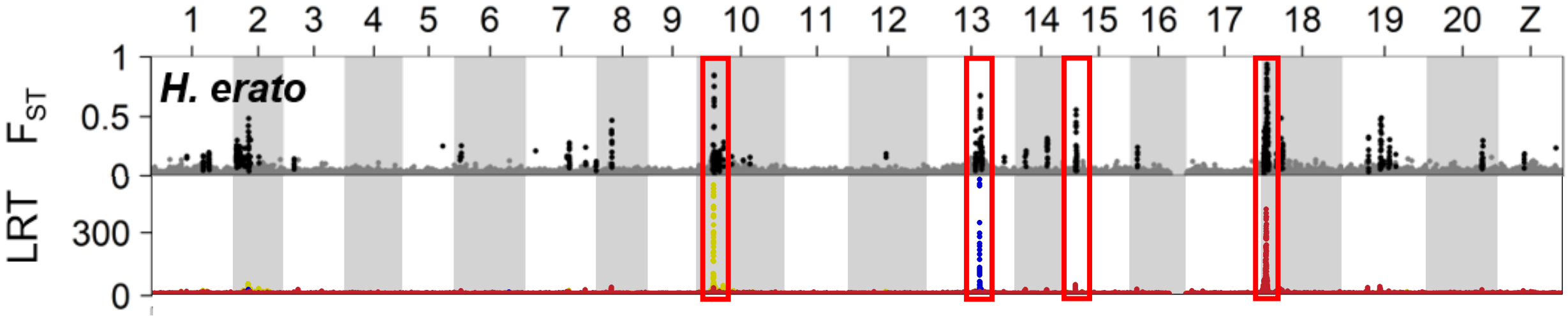
Martin et al. (2013) Genome Research; Seehausen et al (2014) Nat Rev Gen

F_{ST} peaks coincide with GWAS peaks and show signatures of selective sweeps



Meier at al. 2021 PNAS

F_{ST} peaks coincide with GWAS peaks and show signatures of selective sweeps



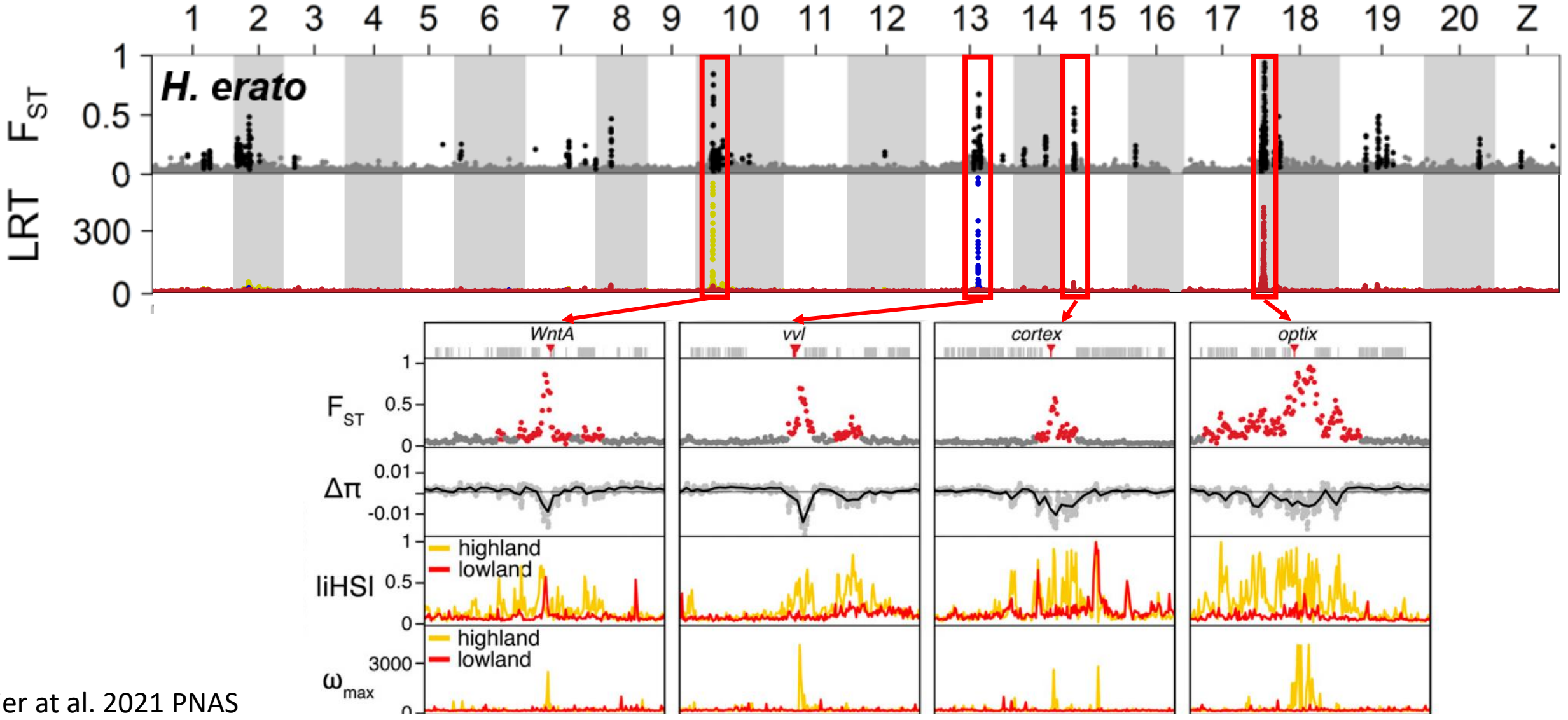
2 or 1 band

band shape

red patterns

Meier at al. 2021 PNAS

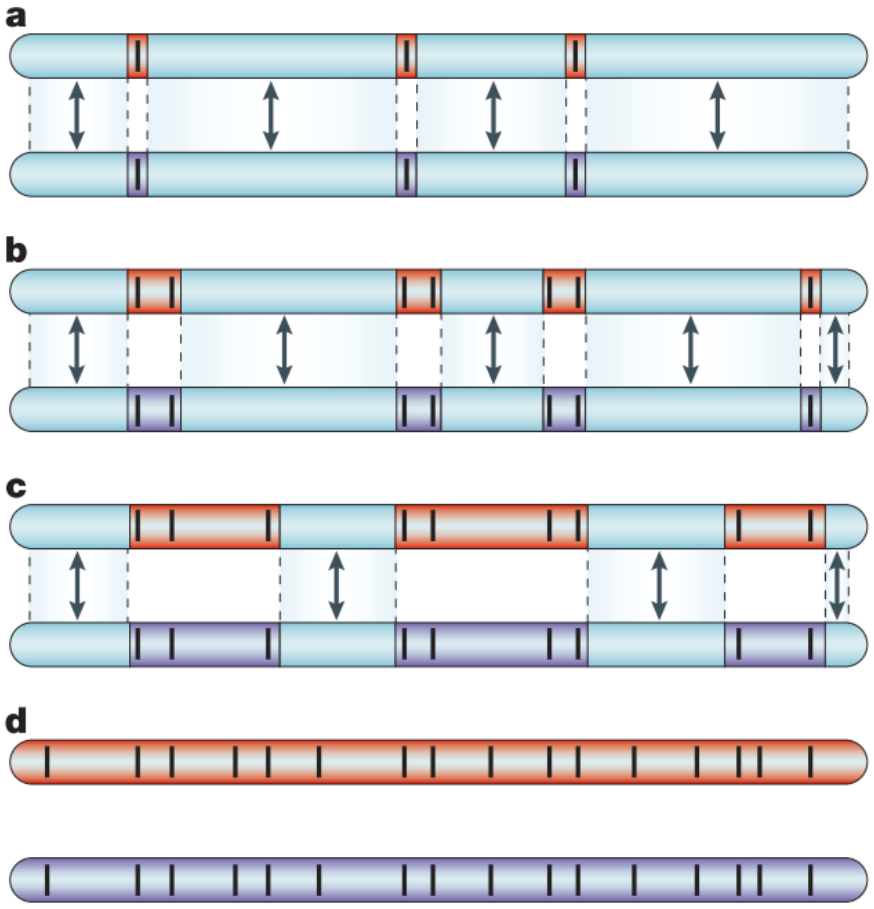
F_{ST} peaks coincide with GWAS peaks and show signatures of selective sweeps



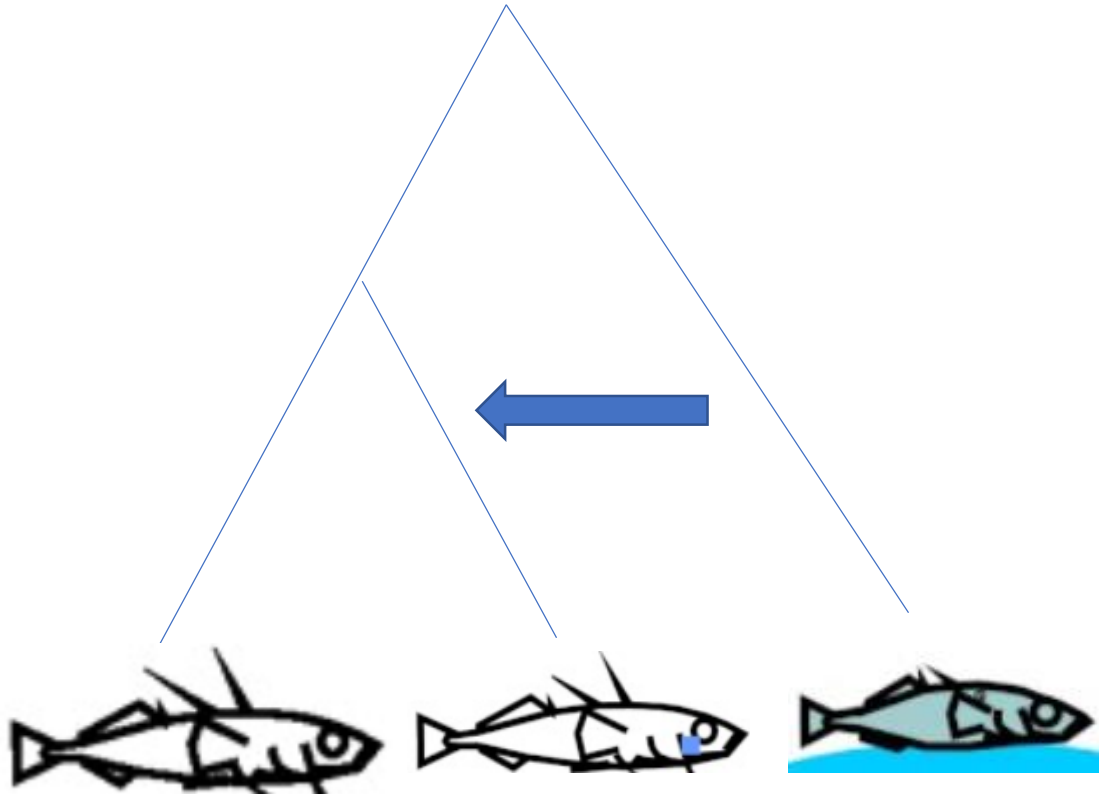
Meier at al. 2021 PNAS

The dual role of hybridisation

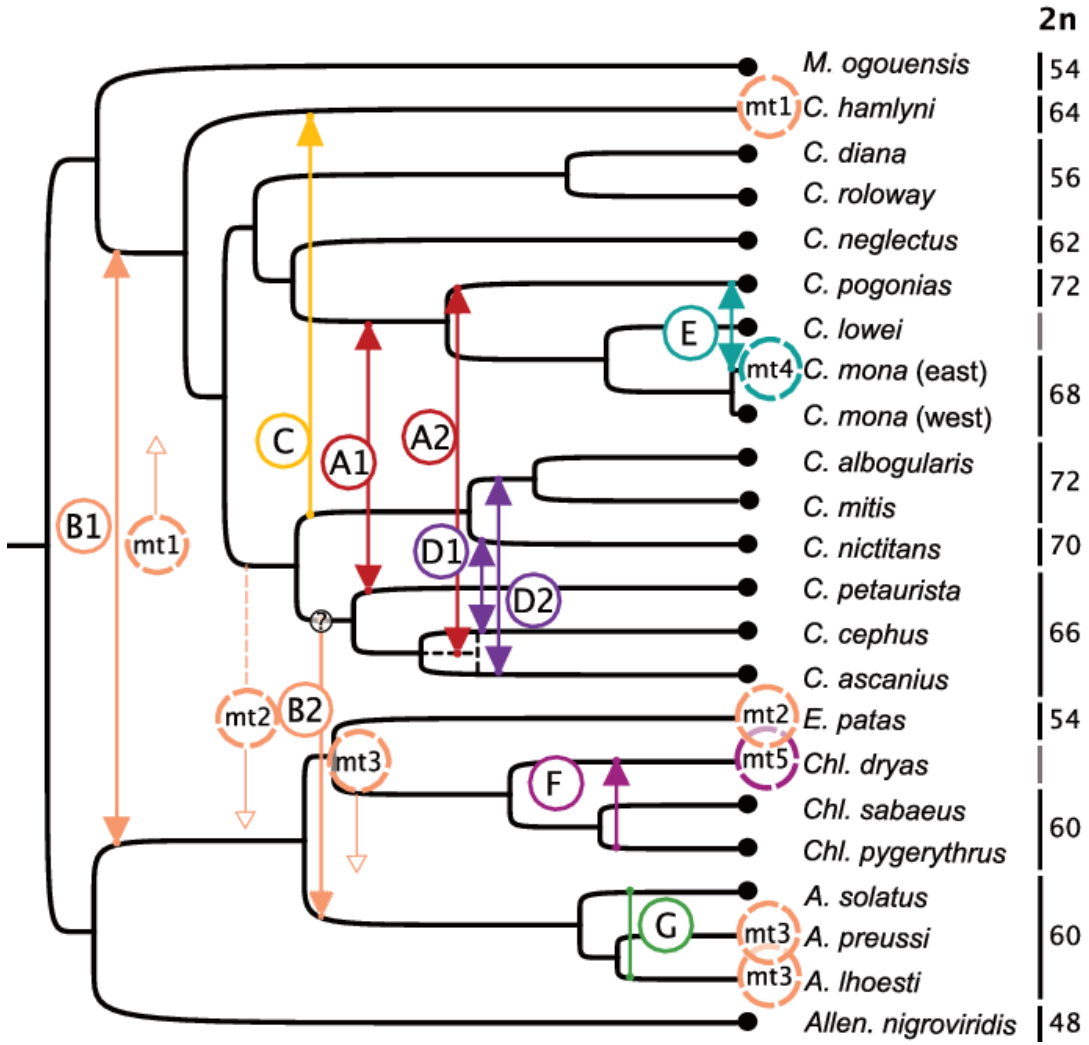
Gene flow between sister taxa counteracts speciation

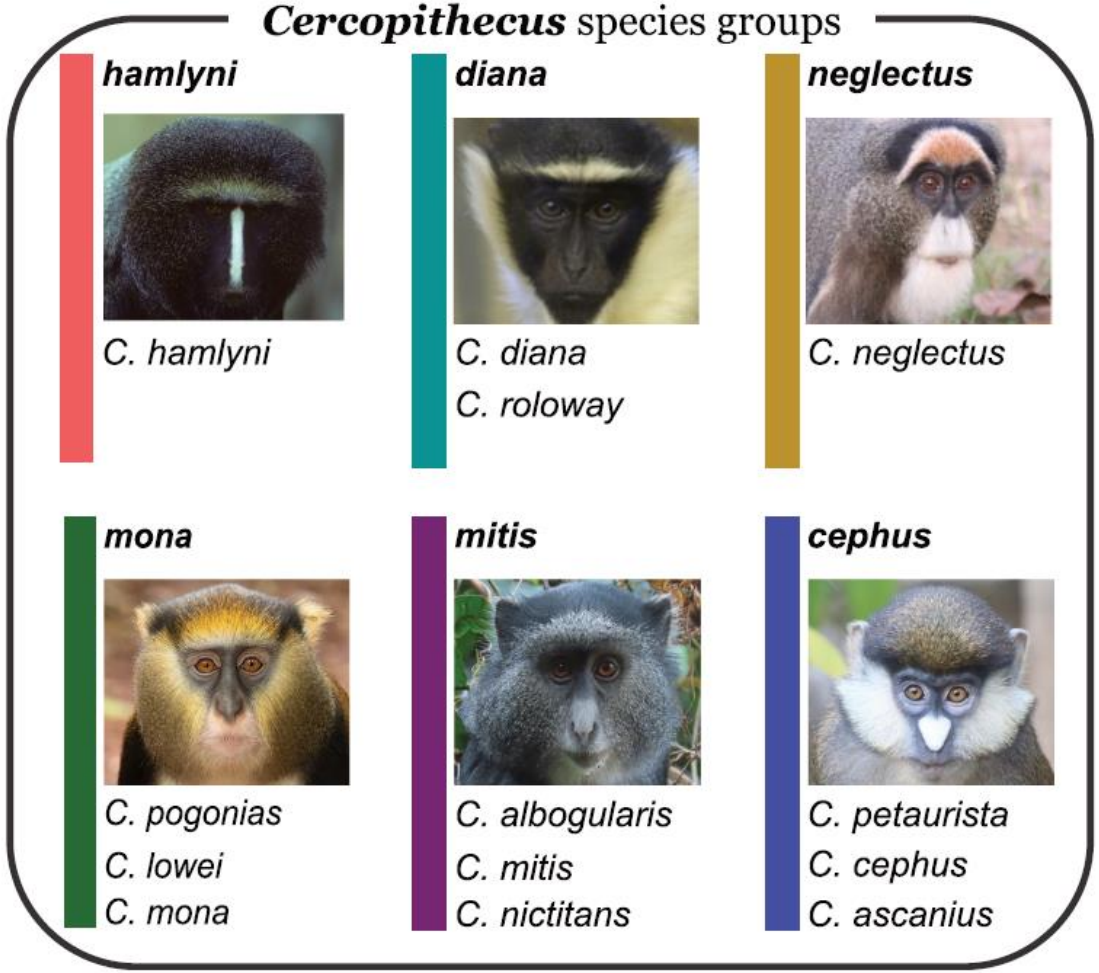


Hybridisation with a third taxon can introduce new genetic variation



Genomics shows that hybridisation is much more widespread than expected (e.g. guenons)



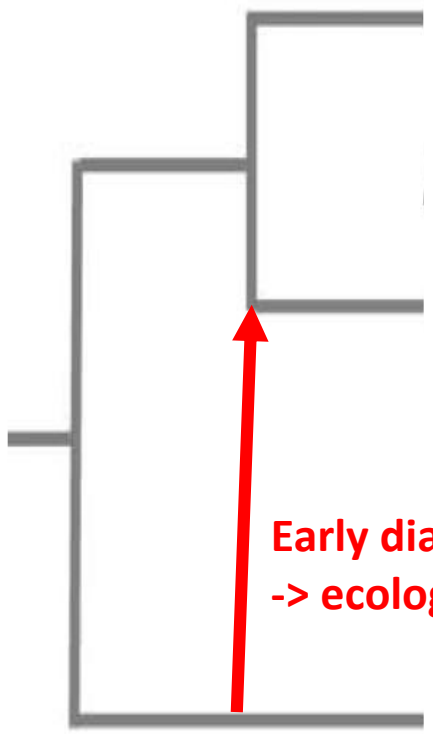


Jensen et al. 2023, MBE

Introgression of key traits contributes to speciation

Hawthorn race

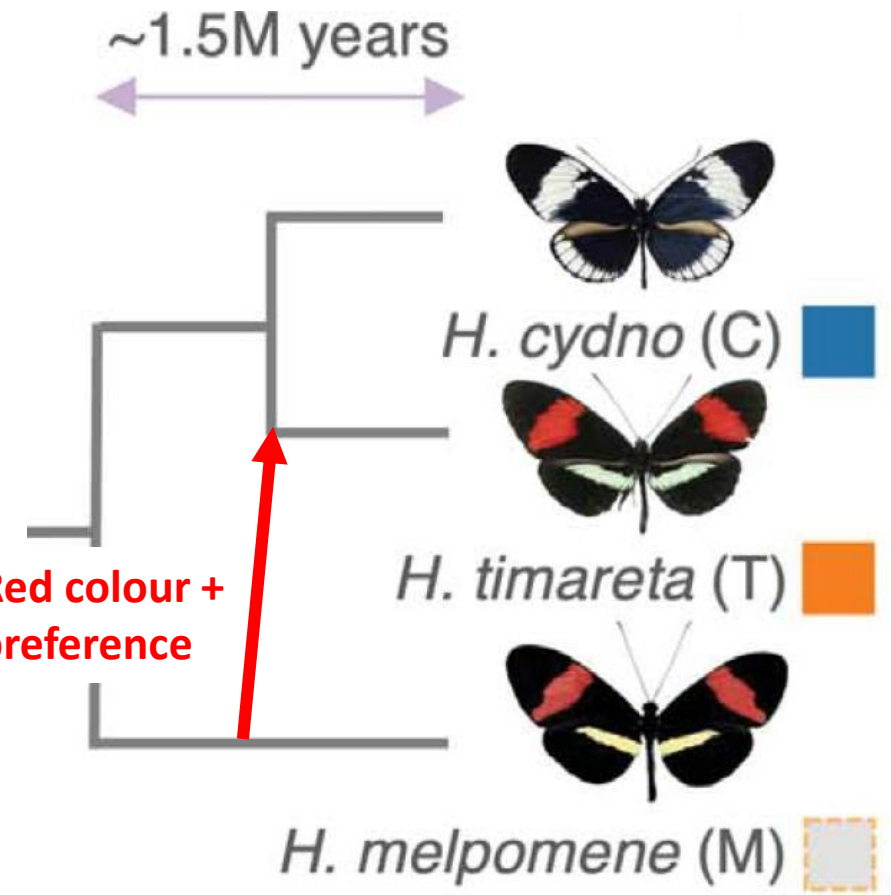
Apple race



Early diapause termination
-> ecological and allochronic isolation

Mexican Rhagoletis flies

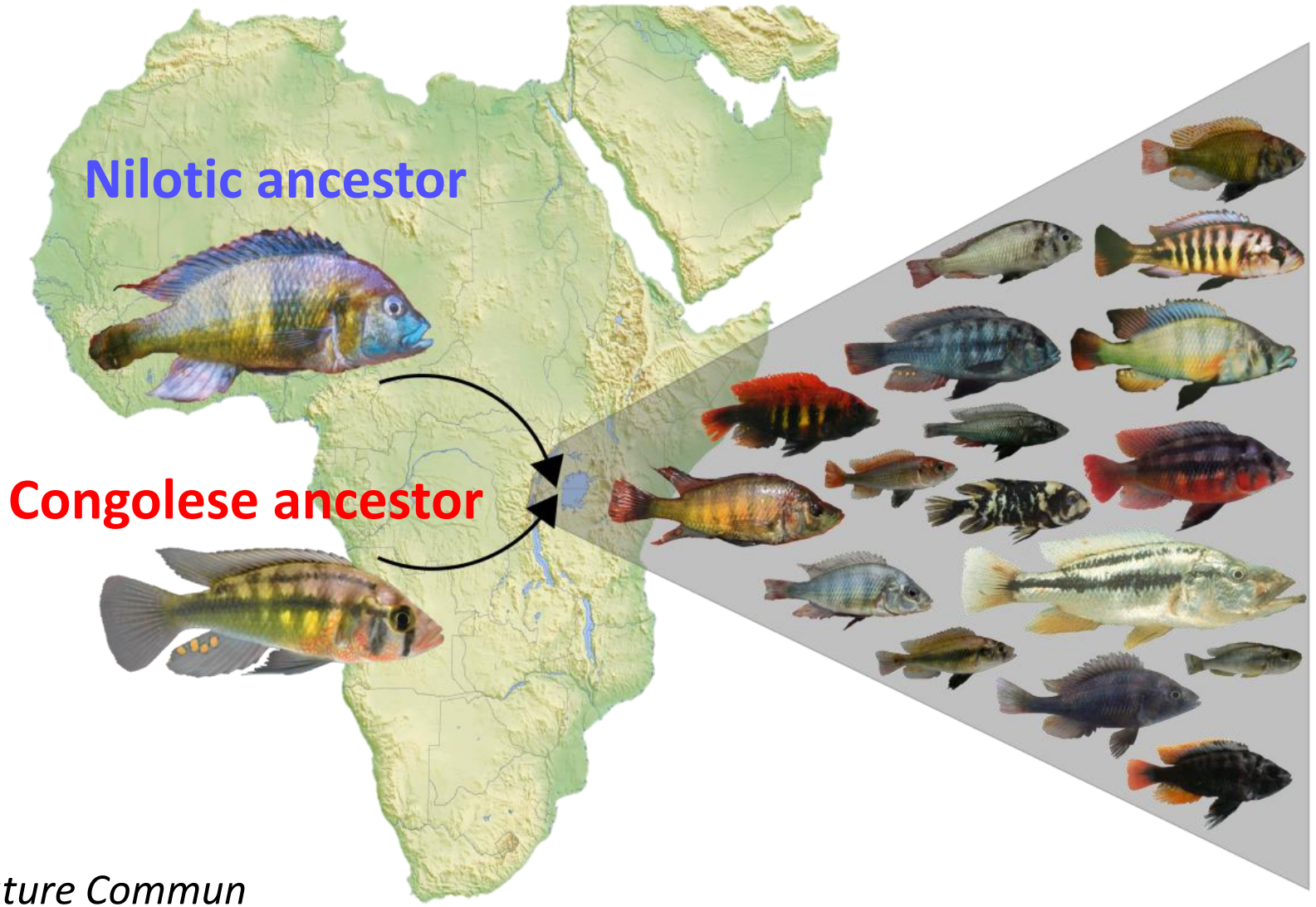
Feder et a. 2003 PNAS; Xie et al 2008 JEB



Rossi et al. 2024 Science

Hybridisation facilitates the origin of adaptive radiations

RAD sequencing without a reference genome



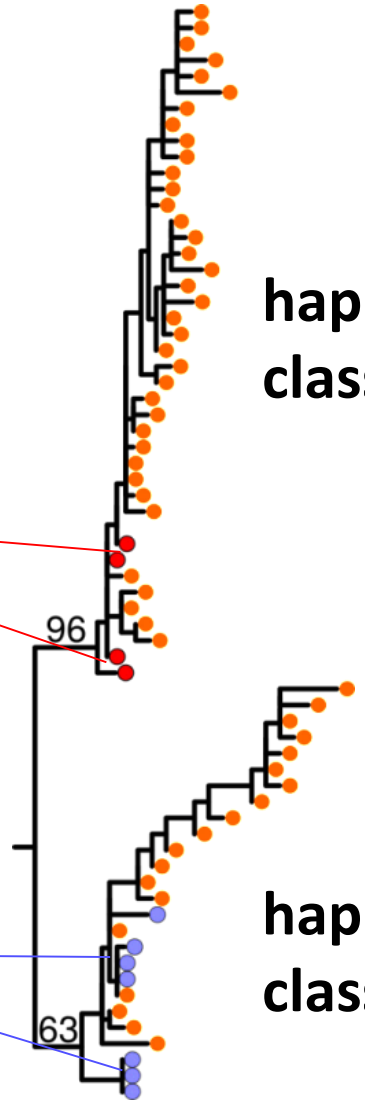
Meier *et al*. (2017) *Nature Commun*
Meier *et al*. (2023) *Science*

The ancestral admixture provided key genetic variation

***LWS* opsin gene**
(red-sensitive)

Congo

Upper Nile



haplotype class I

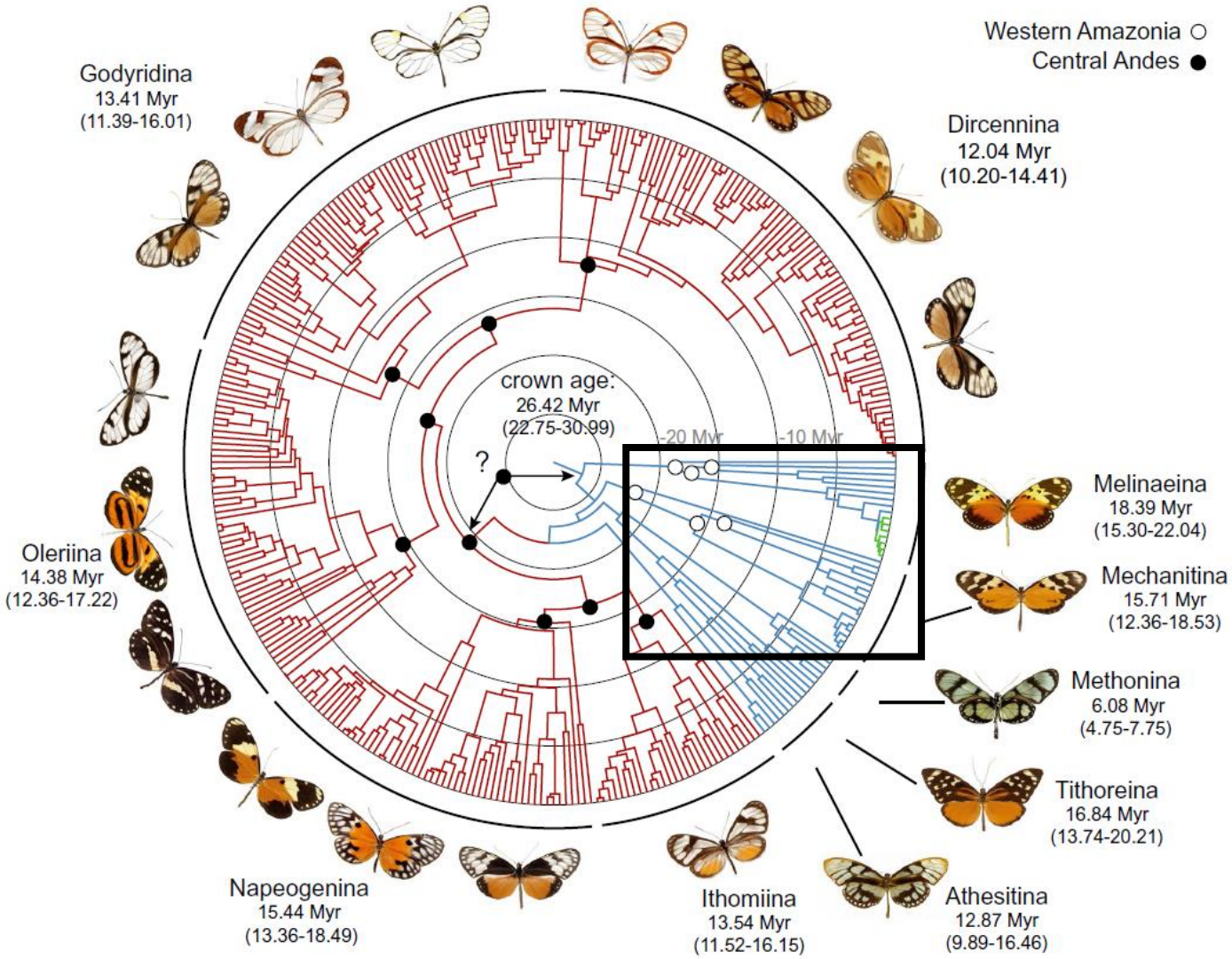
e.g. clear shallow water algivores

haplotype class II

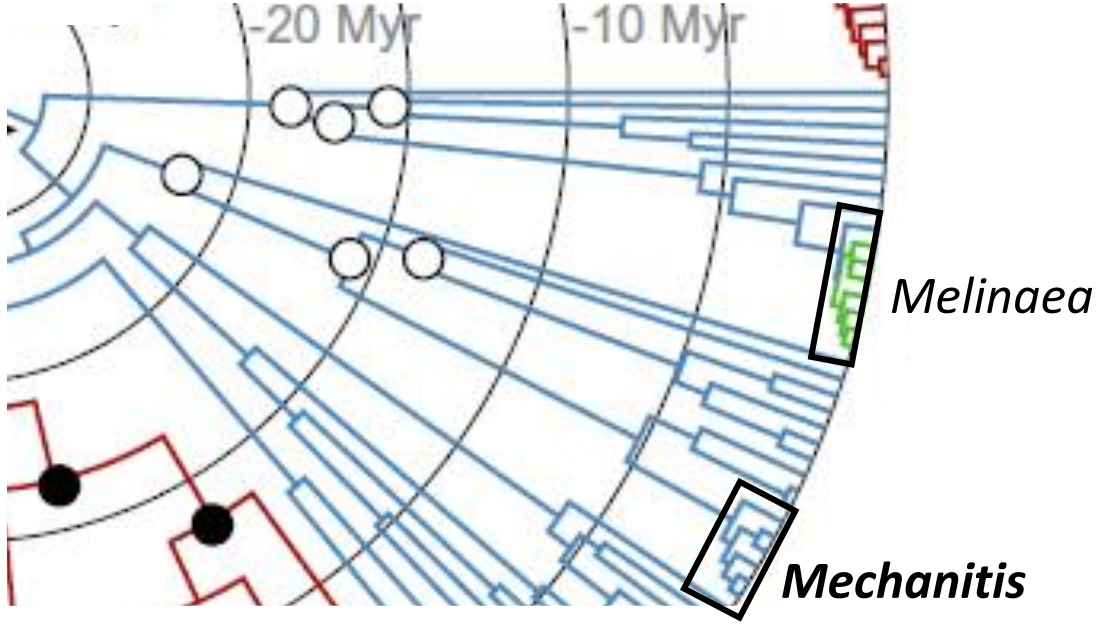
e.g. turbid deep water detritivores

Meier *et al*. (2017) *Nature Commun*

Comparing rapidly and slowly speciating Ithomiini genera



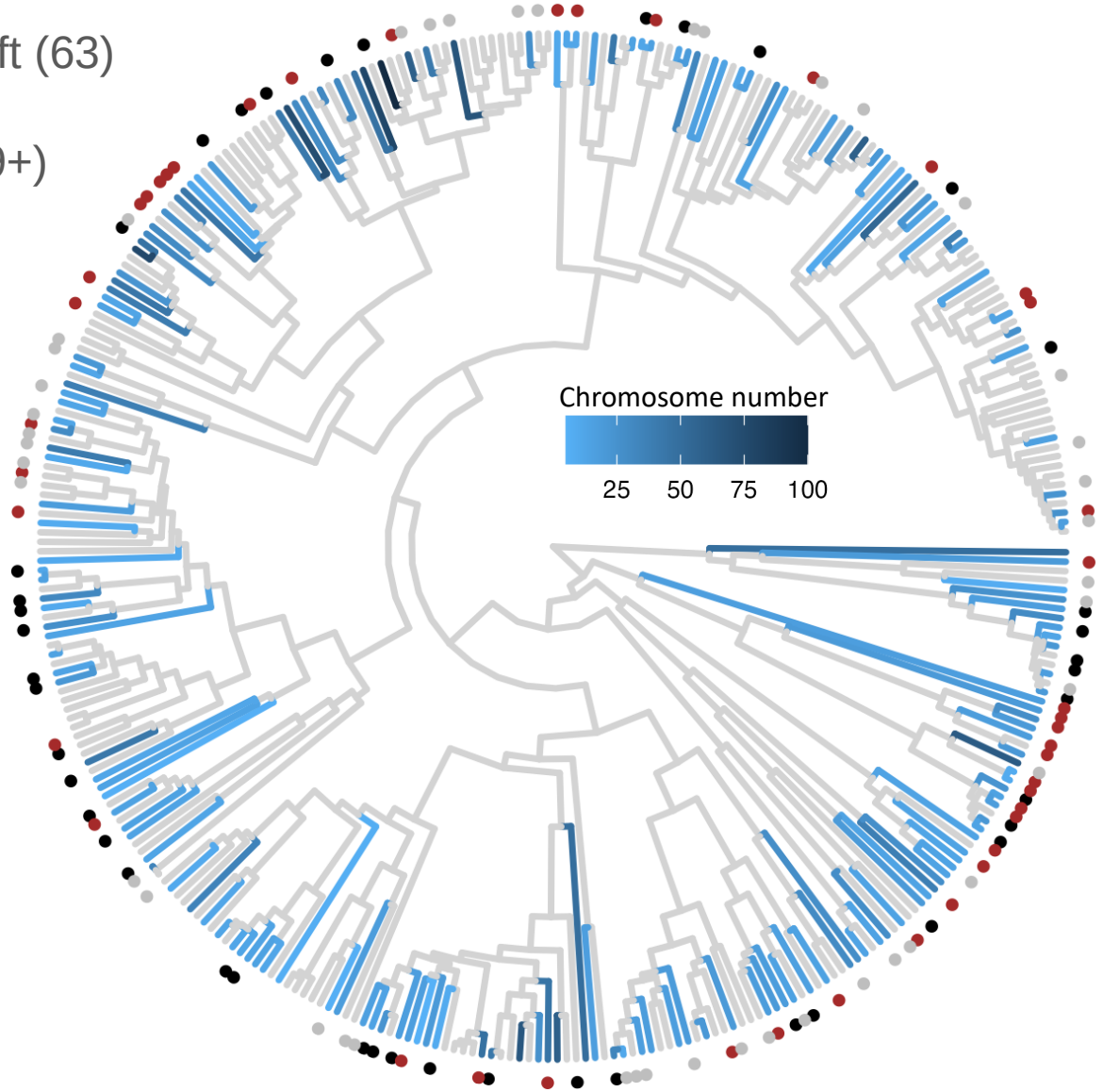
Time since last common ancestor



Chazot et al., 2019

Sequencing 200 ithomiini genomes

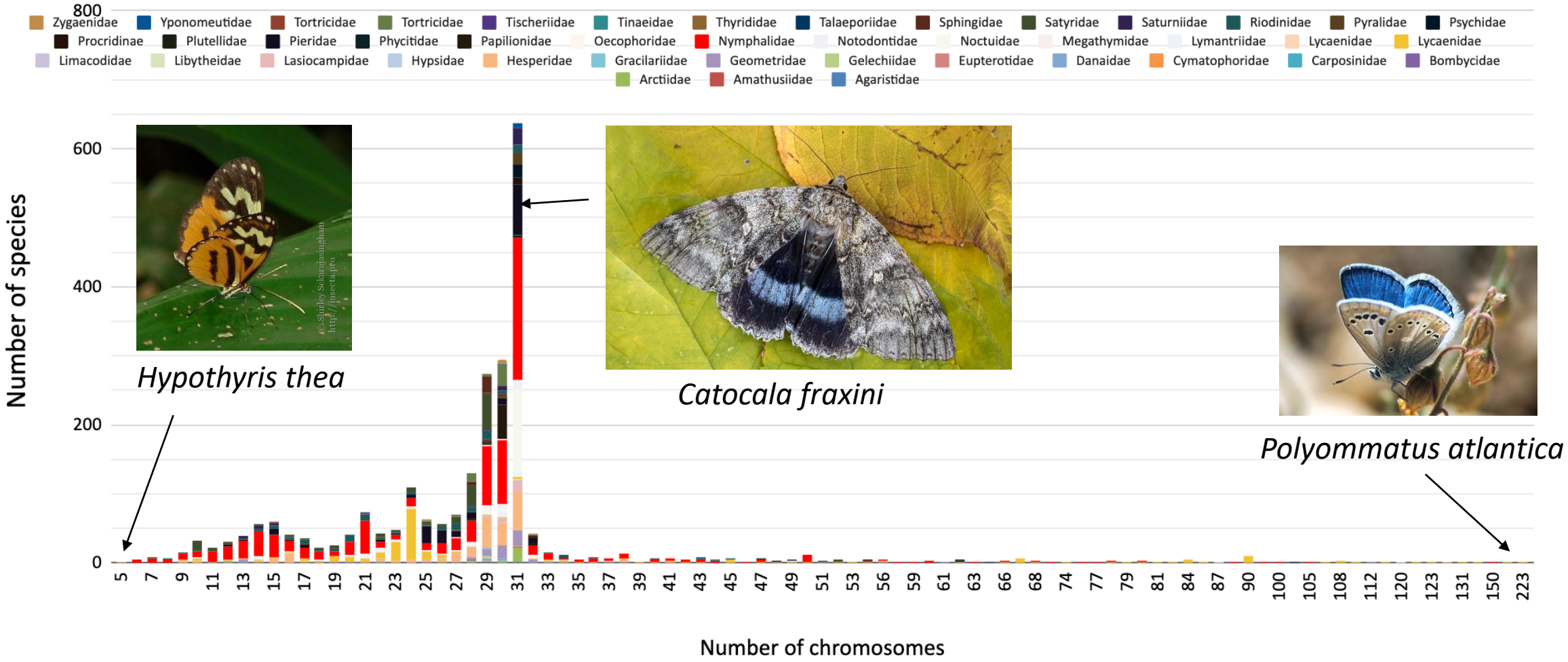
- Curated/Draft (63)
- Lab (47)
- Sampled (49+)



Patricio Salazar

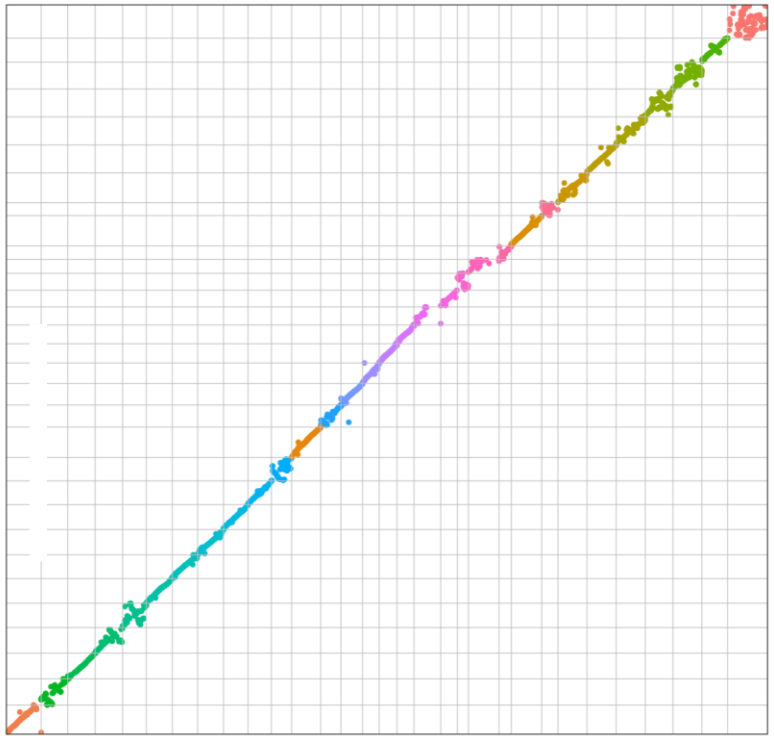
Karin Näsvall

Most butterflies and moths have 31 chromosomes



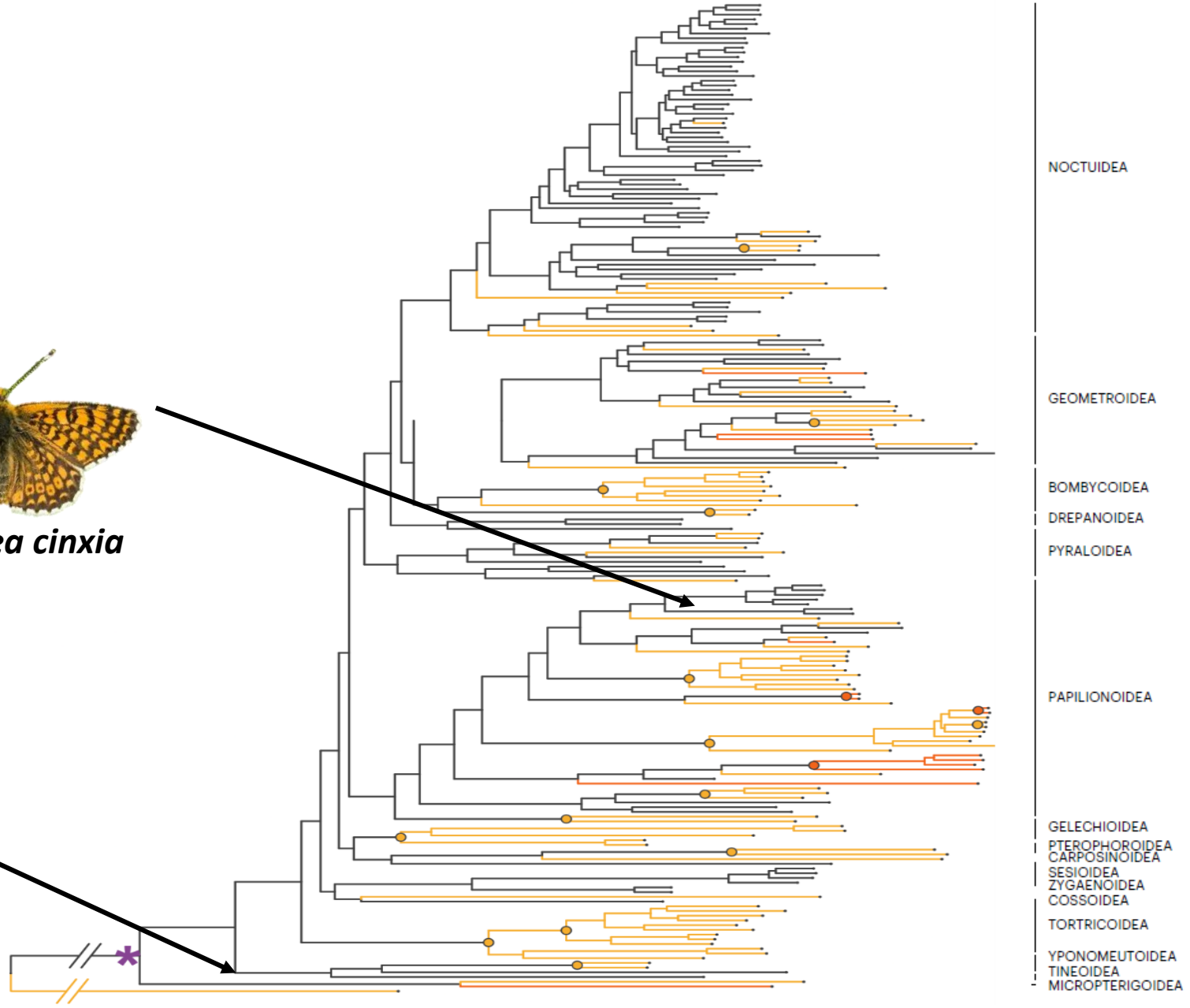
Data from de Vos *et al* 2020

The chromosomes tend to be highly conserved



Melitaea cinxia

Tinea trinotella

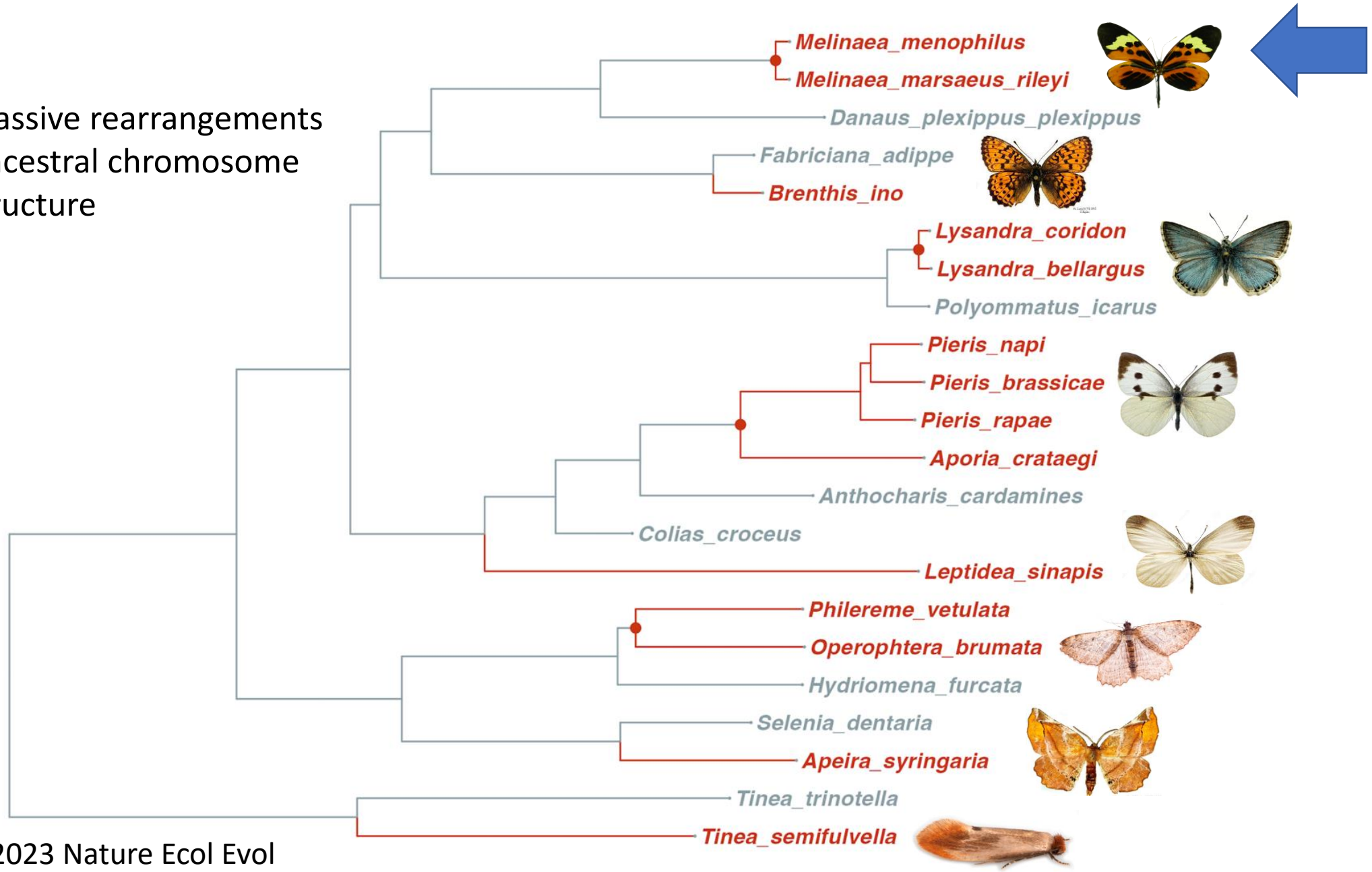


Wright et al. 2024 Nature Ecol Evol

However, there are some exceptions

Massive rearrangements

Ancestral chromosome structure



Wright et al. 2023 Nature Ecol Evol