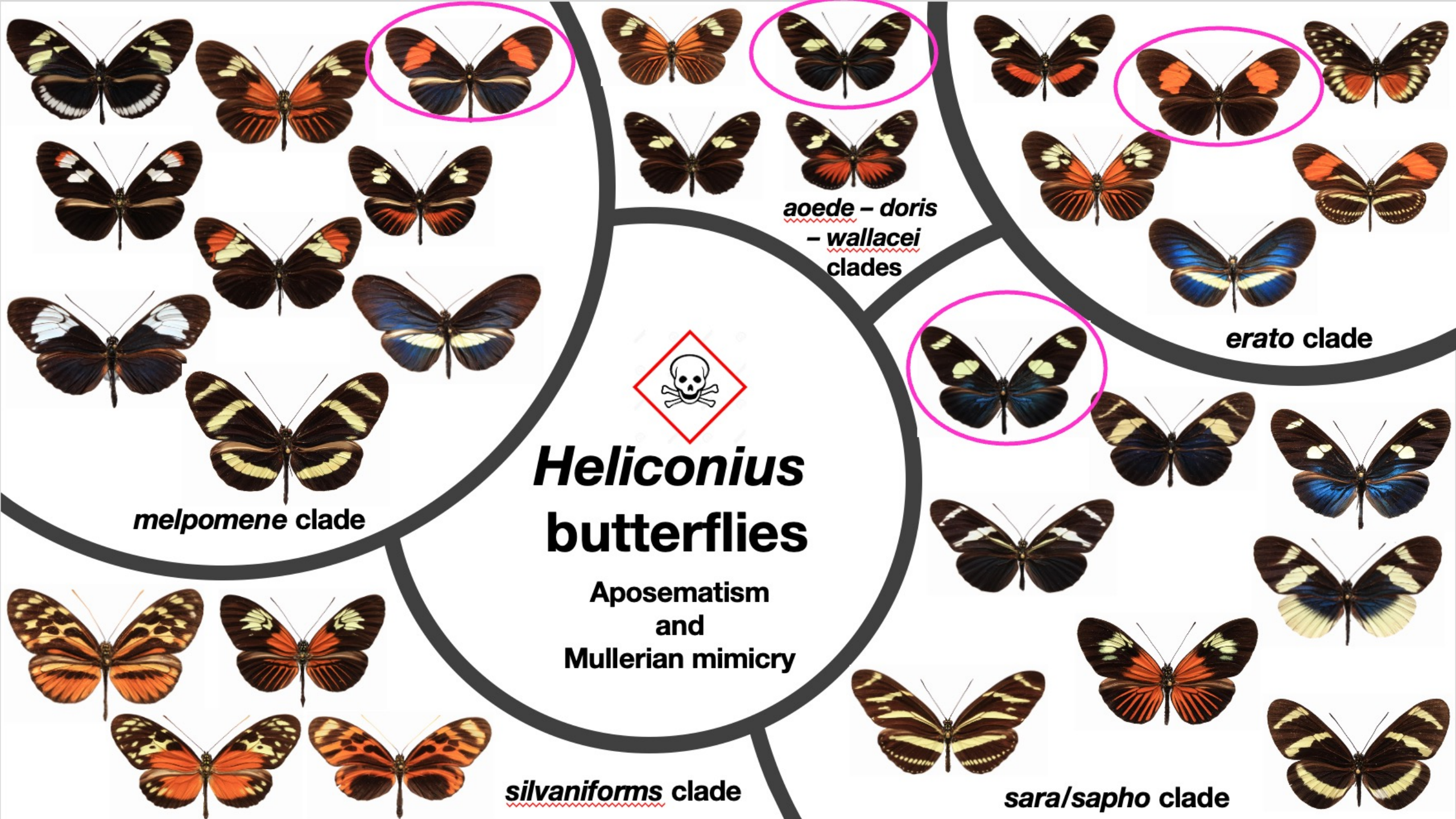
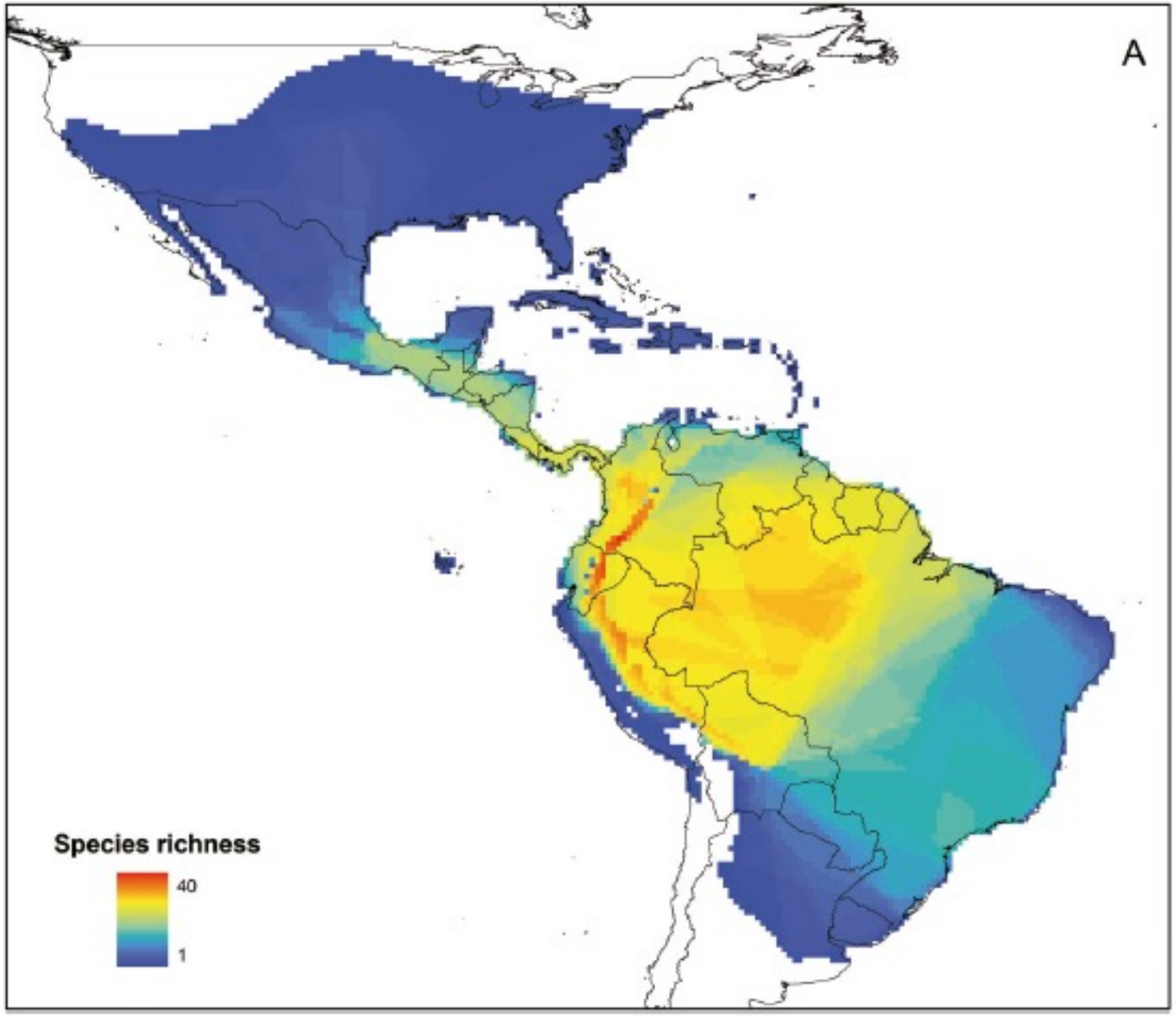
# Introduction to *Heliconius* butterflies

## Biodiversity genomics course

## Tena-Ecuador 2024

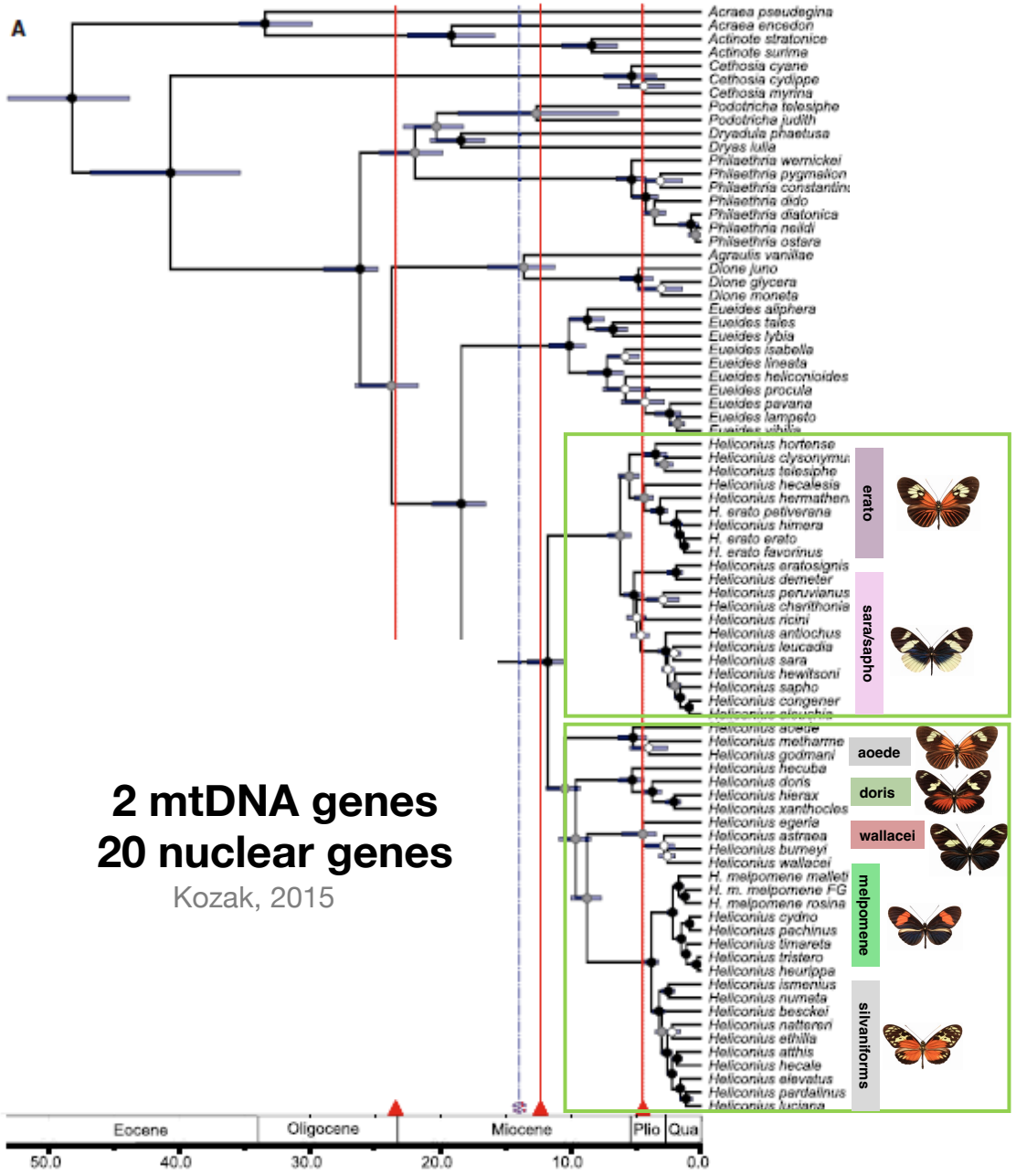

Heliconius butterflies
Aposematism and Mullerian mimicry
melpomene clade
aoede – doris – wallacei clades
erato clade
silvaniforms clade
sara/sapho clade

# Introduction




Rosser et al., 2012

# Introduction




**2 mtDNA genes**
**20 nuclear genes**

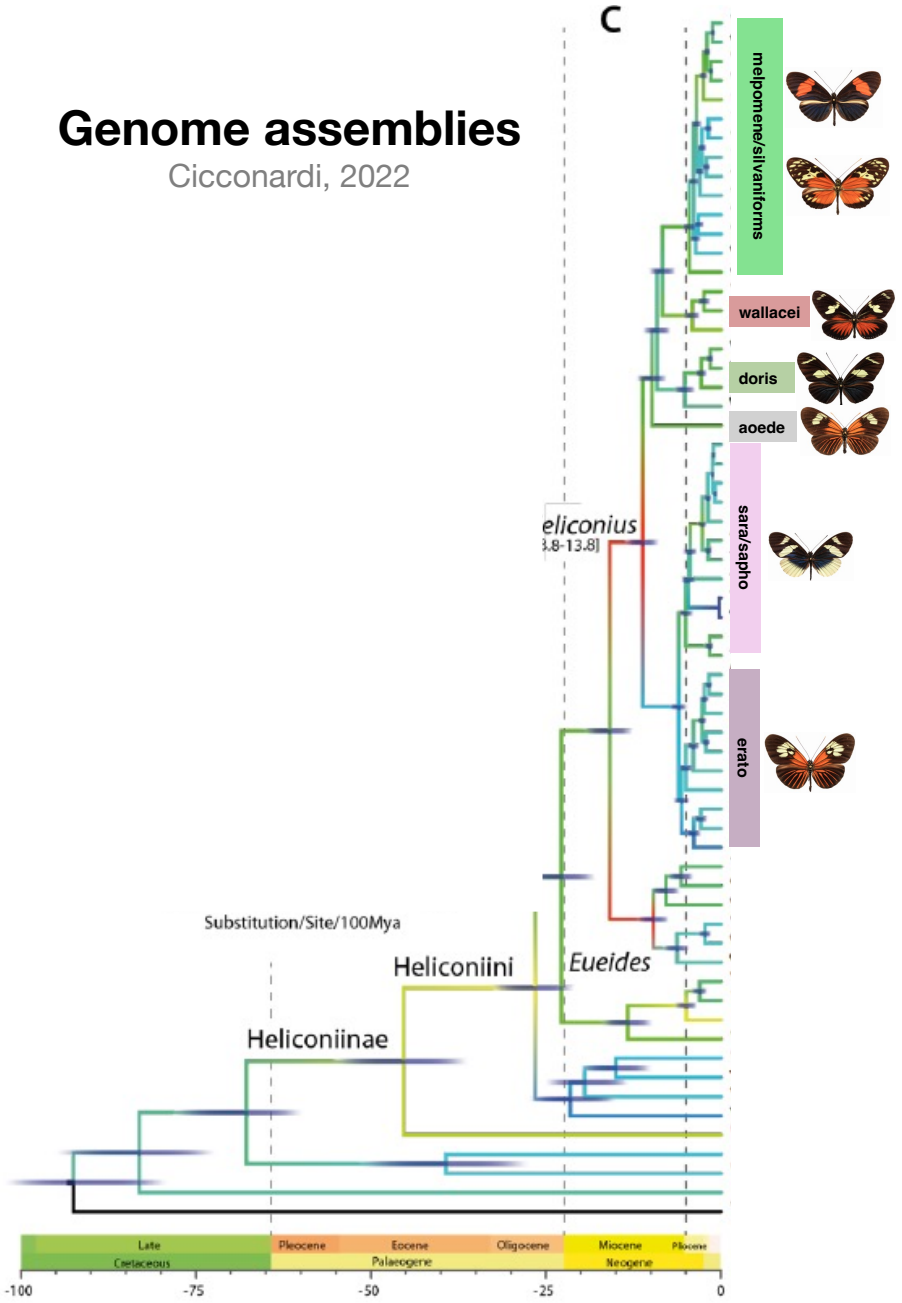
Kozak, 2015

**pupal mating clade**

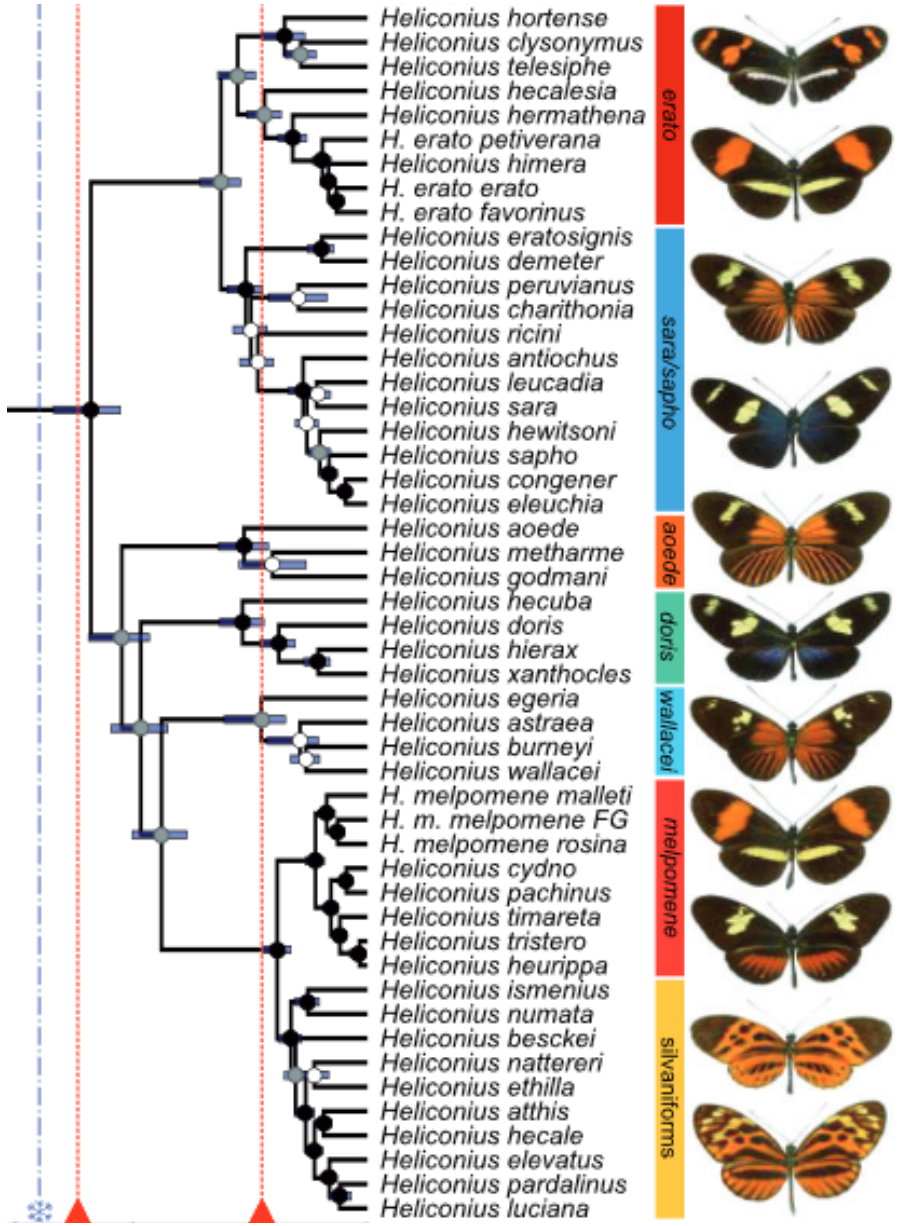
**nonpupal mating clade**

## Genome assemblies

Cicconardi, 2022



# Introduction




Kozak, 2015

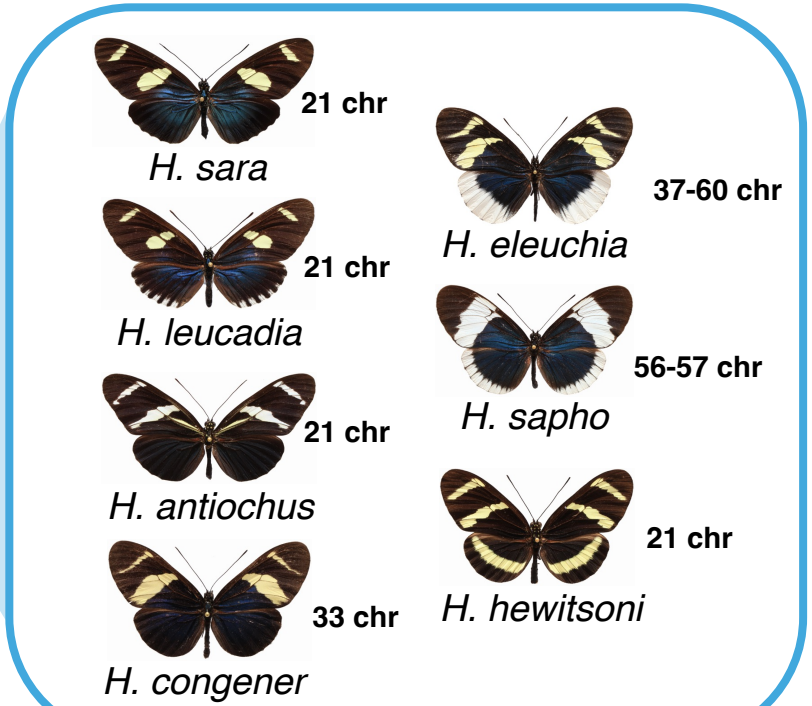
**21 Chromosomes**

**21 - 60 chromosomes**

**20 - 38 chromosomes**

**21 Chromosomes**

**21 Chromosomes**




Brown, 1992

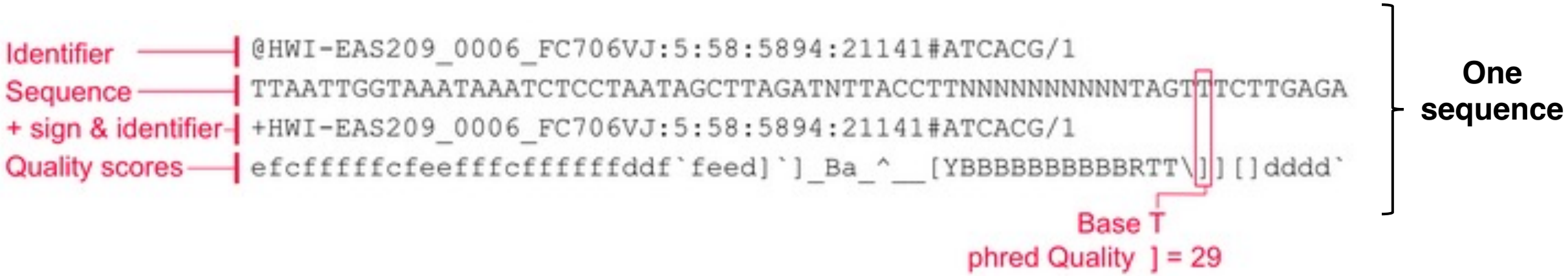
The data (raw reads) we will be working with today and tomorrow are from this clade.

How do raw sequences files look like?

How can we do a quality control of them?

# Fastq Format

This format is designed to handle base quality metrics output from sequencing machines.




**Line 1** begins with the '@' character and is followed by a sequence identifier and an optional description.

**Line 2** is the sequence letters.

**Line 3** begins with a '+' character; it marks the end of the sequence and is optionally followed by the same sequence identifier again in line 1.

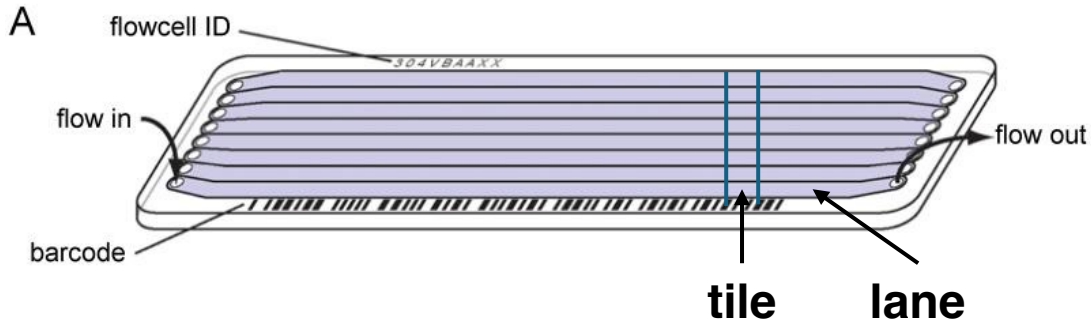
**Line 4** encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

# Read header

Colon

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

| | |
|---|---|
| **EAS139** | the unique instrument name |
| **136** | the run id |
| **FC706VJ** | the flowcell id |
| **2** | flowcell lane |
| **2104** | tile number within the flowcell lane |
| **15343** | 'x'-coordinate of the cluster within the tile |
| **197393** | 'y'-coordinate of the cluster within the tile |
| **1** | the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)* |
| **Y** | Y if the read is filtered, N otherwise |
| **18** | 0 when none of the control bits are on, otherwise it is an even number |
| **ATCACG** | index sequence |



# Quality scores

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
CCGTCAATTCATTAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA:9@::::??@@::FFAAAAACCAA::::BB@@?A?
```

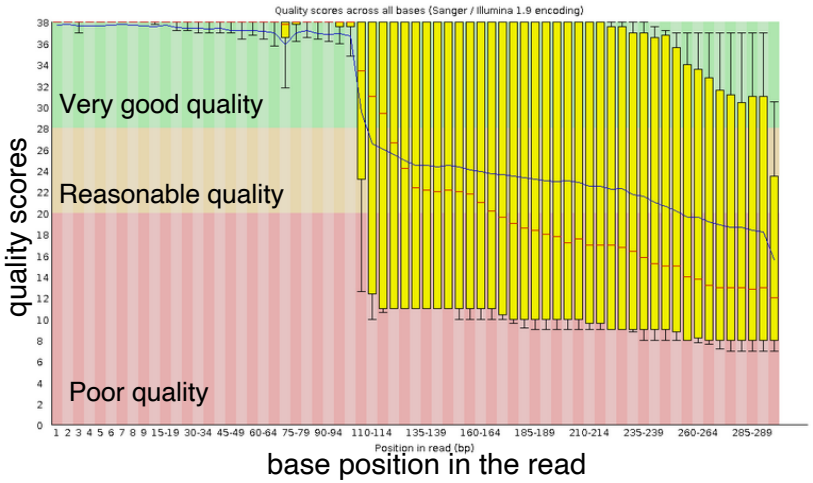
## Quality score: ASCII encoding

| | | |
|---|---|---|
| 40:@ | 90:Z | 141:a |
| 41:A | 91:[ | 142:b |
| 42:B | 92:\ | 143:c |
| 43:C | 93:] | 144:d |
| 44:D | 94:^ | 145:e |
| 45:E | 95:_ | 146:f |
| … :… | … :… | … :… |

| Quality Score | Probability of incorrect base call | Base call accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1000 | 99.9% |
| 40 | 1 in 10000 | 99.99% |

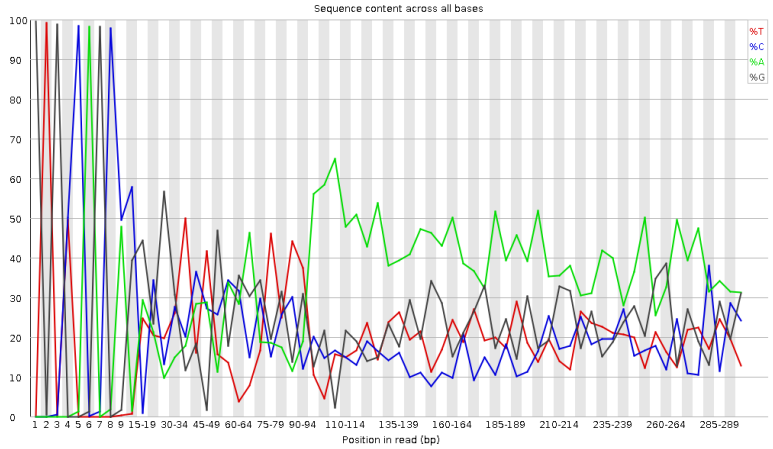
# Assess quality with FastQC → .html

## Per base sequence quality




- The higher the score, the better the base call.
- The quality of reads on most platforms will drop at the end of the read. This is often due to signal decay or phasing during the sequencing run.

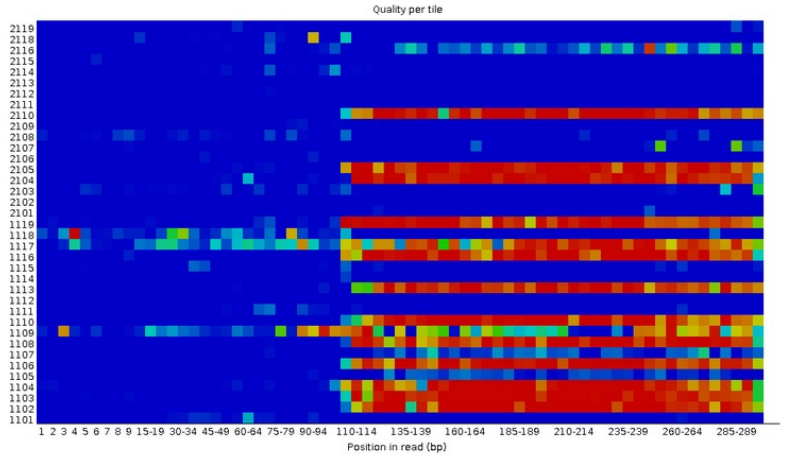
## Per base sequence content




- Percentage of each of the four nucleotides (T, C, A, G) at each position across all reads in the input sequence file
- In a random library, little or no difference is expected between the different bases of a sequence, so the lines in this graph should be parallel to each other.

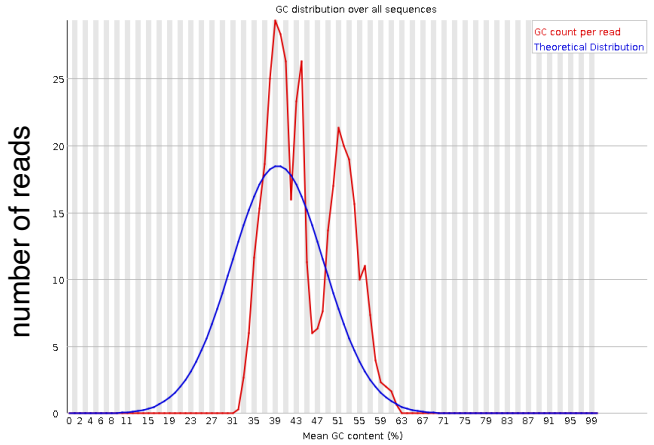
https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#sequence-length-distribution

## Per tile sequence quality




- Quality scores from each tile across all of your bases to see if there was a loss in quality associated with only one part of the flowcell.
- The hotter colours indicate that reads in the given tile have worse qualities for that position than reads in other tiles.

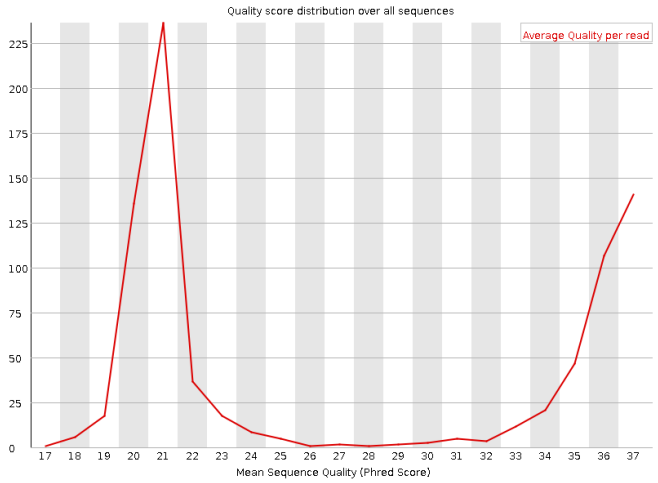
## Per sequence GC content




- This plot displays the number of reads vs. percentage of bases G and C per read.
- An unusually-shaped distribution could indicate a contaminated library or some other kind of biased subset.

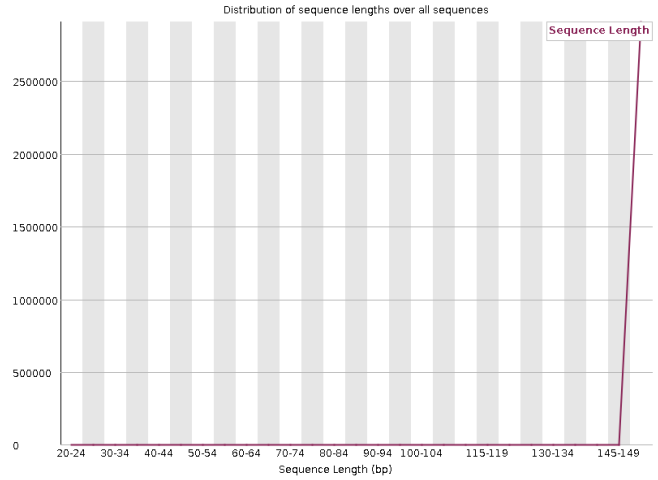
https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#sequence-length-distribution

## Per sequence quality scores




- It plots the average quality score over the full length of all reads on the x-axis and gives the total number of reads with this score on the y-axis
- It can also report if a subset of the sequences have low quality values

## Sequence length distribution




- This plot shows the distribution of fragment sizes in the file which was analysed.
- In many cases, this will produce a simple plot with a peak at one size. But, some sequencers could produce reads of widely varying lengths.

https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#sequence-length-distribution

Let’s have a look at the first few sequences and check the sequencing quality with fastqc