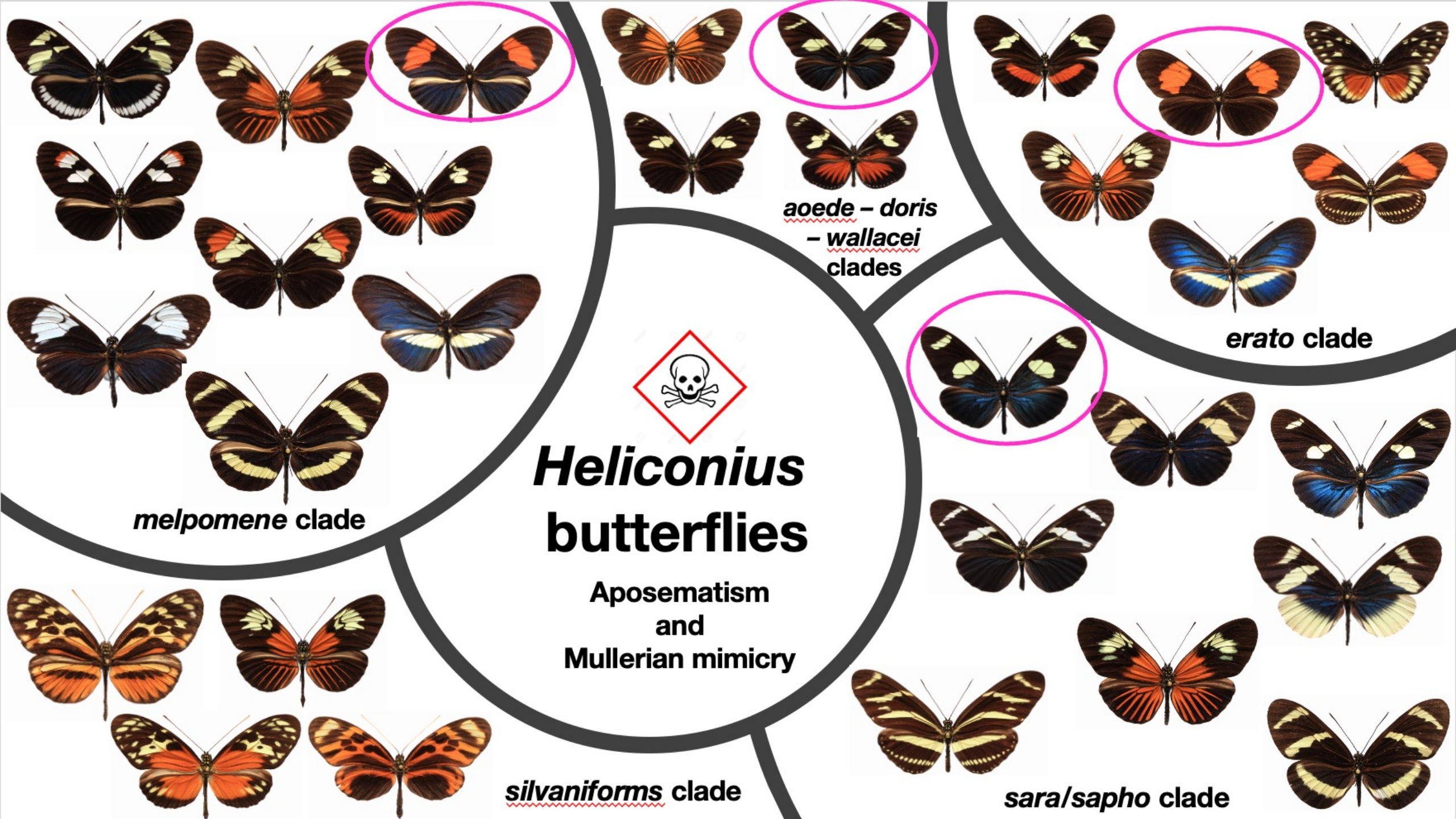
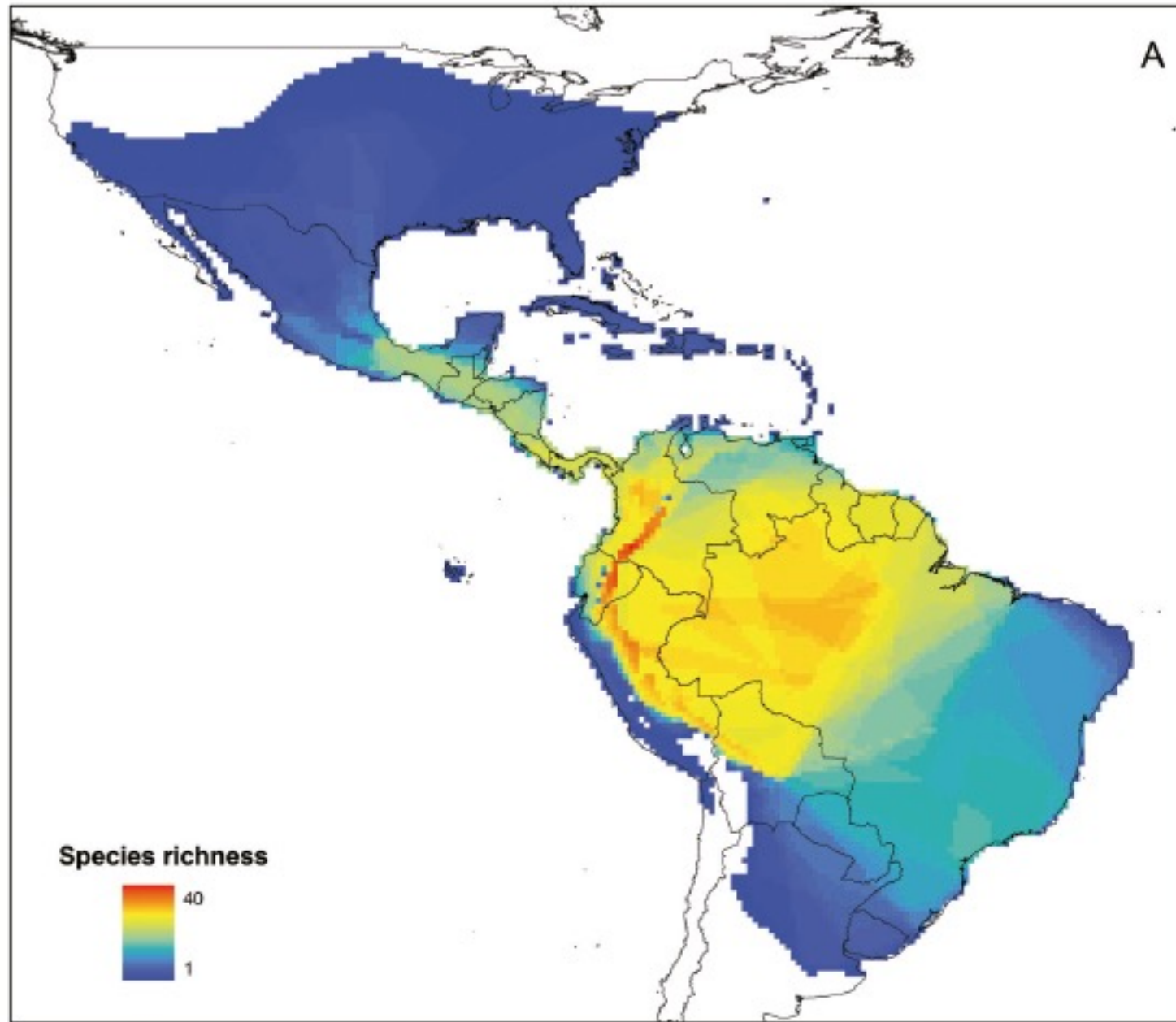


***Heliconius* butterflies**  
**Biodiversity genomics course**  
**Tena-Ecuador 2024**

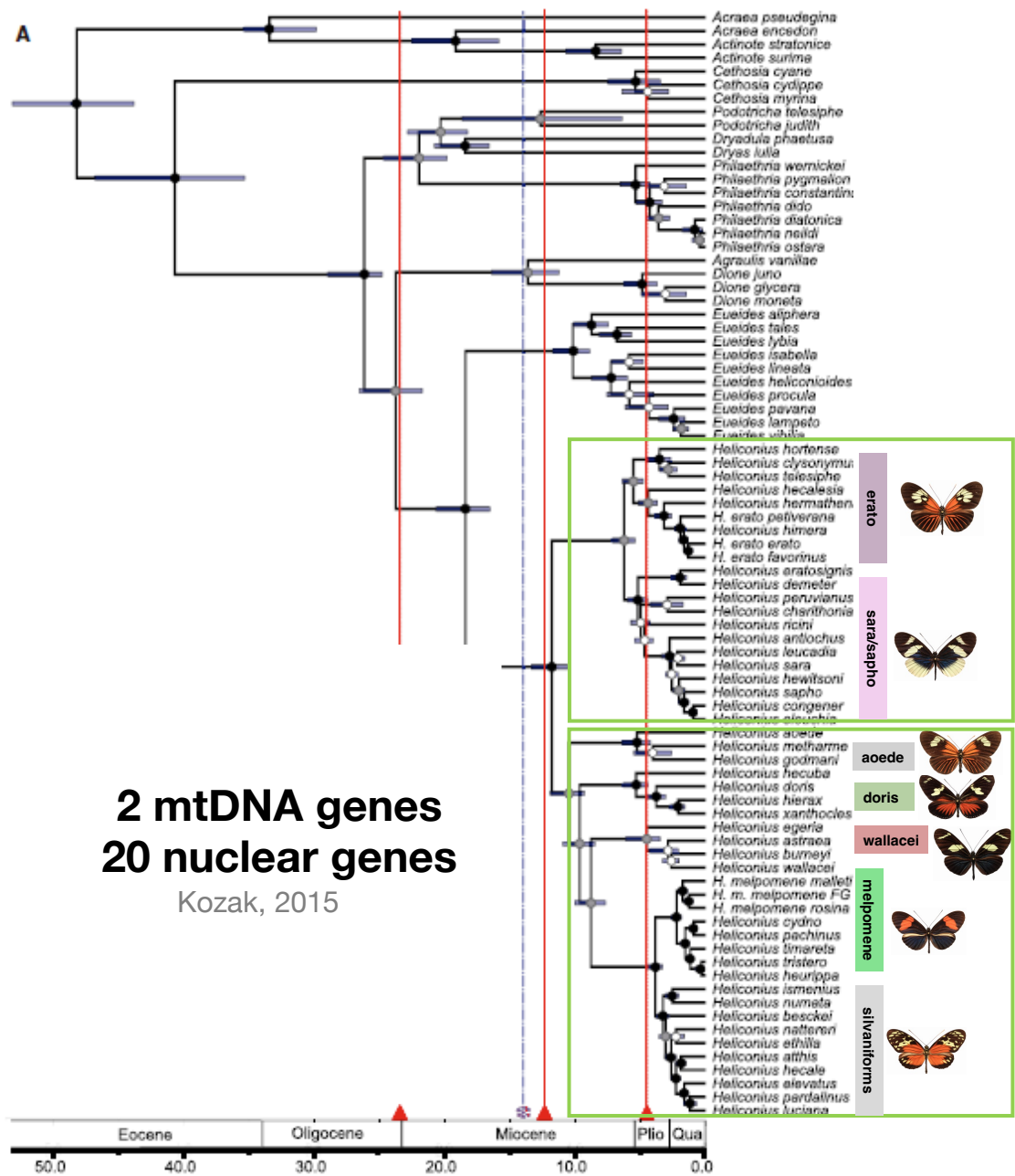


# Introduction



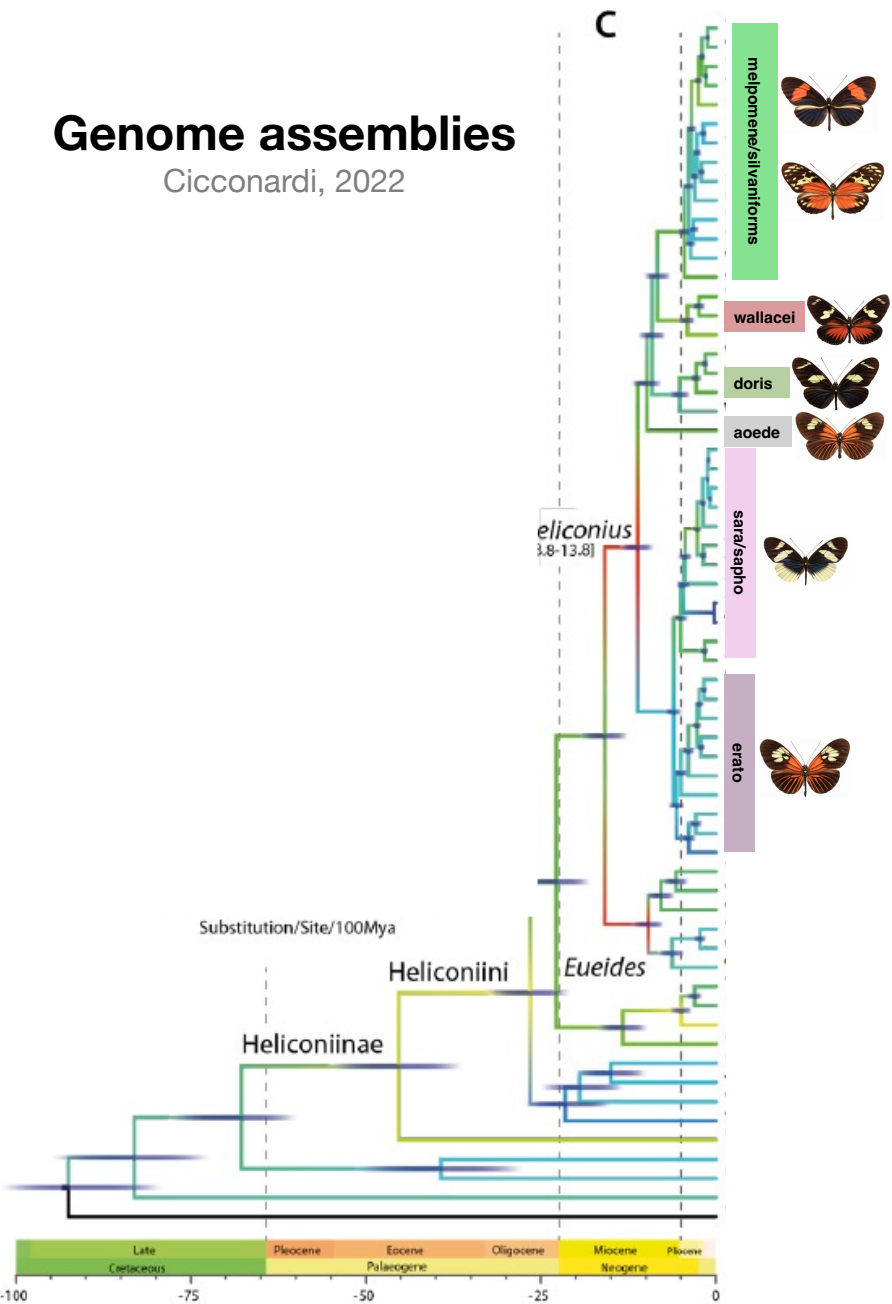


# Introduction

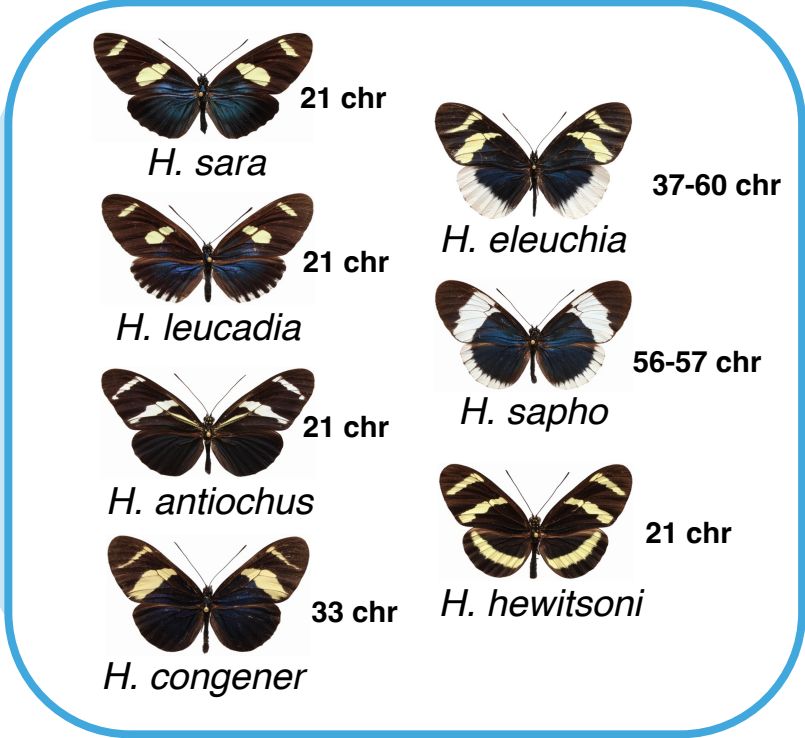
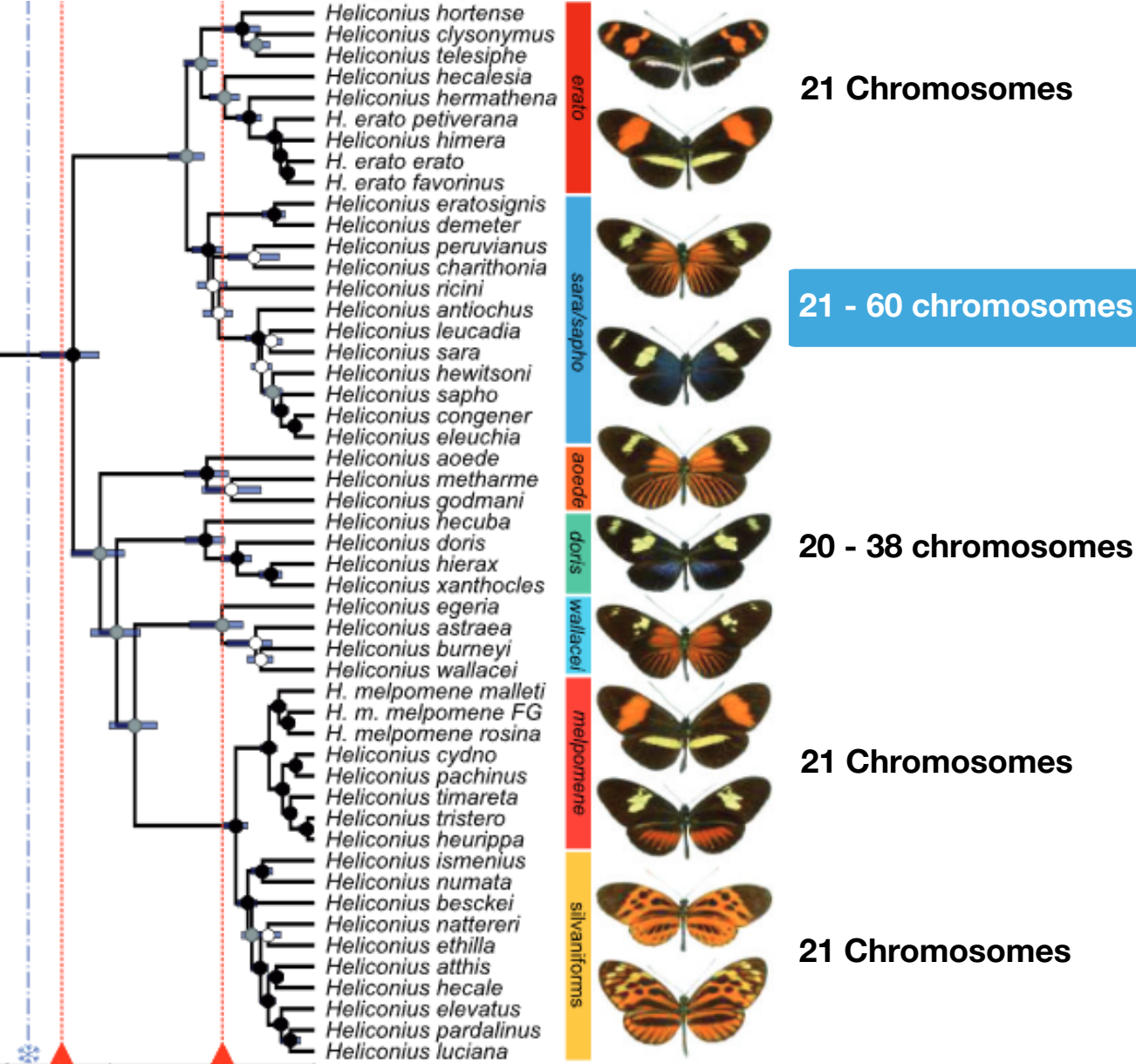


## Genome assemblies

Cicconardi, 2022



# Introduction



Brown, 1992

# Raw sequences files and quality control

# Fastq Format

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

```
Identifier | @HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Sequence  | TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNNTAGTTTCTTGAGA
+ sign & identifier | +HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Quality scores | efcfffffcfeeffffcfffffdddf`feed)`]_Ba_^__[YBBBBBBBBBBRTT\]][] dddd`
```

Base T  
phred Quality ] = 29

One sequence

**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.

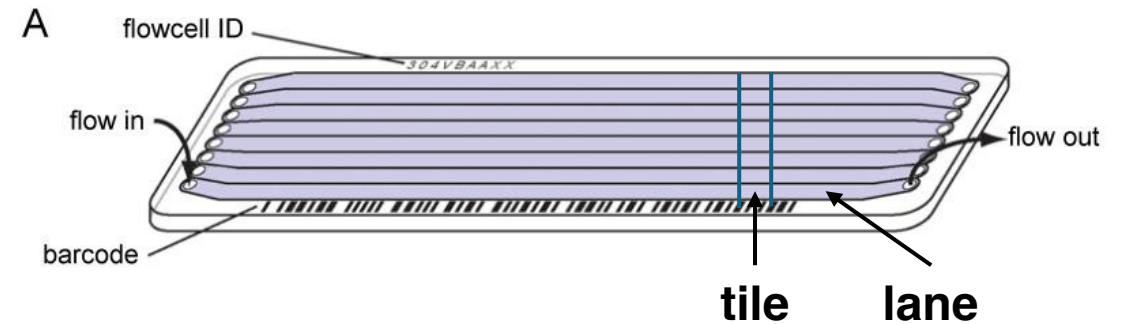
# Read header

Colon



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

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





# Quality scores

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

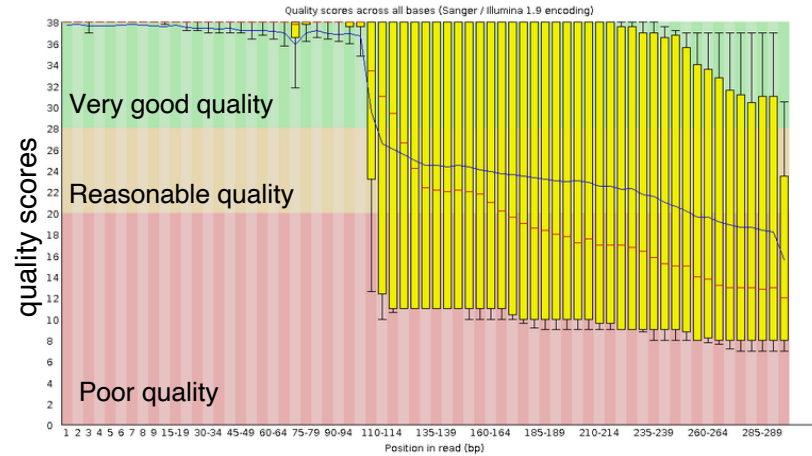
## Quality score: ASCII encoding

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
... :...	... :...	... :...

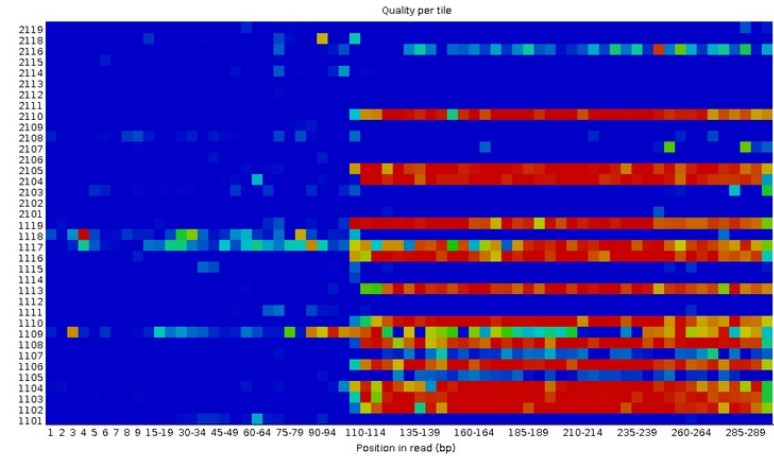
Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

# Assess quality with FastQC → .html ★

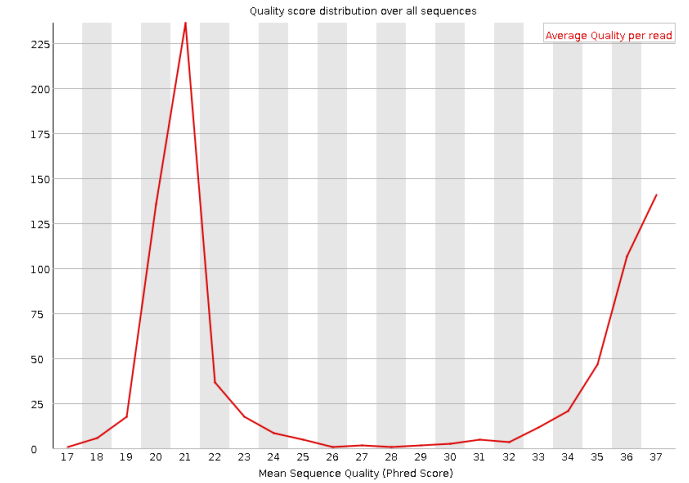
## Per base sequence quality



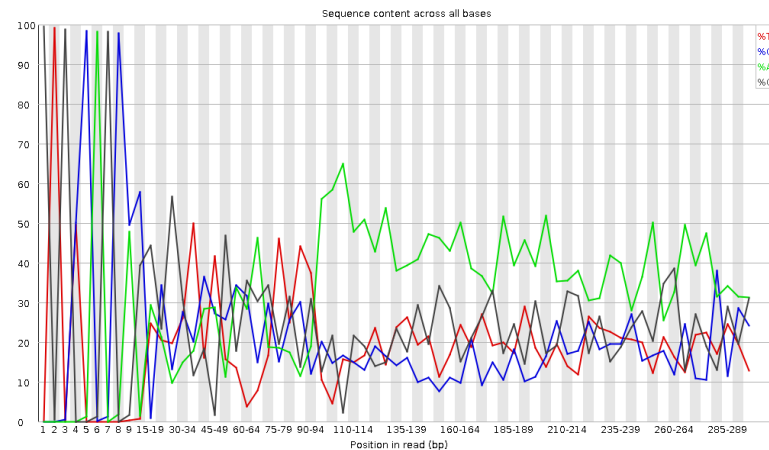
## Per tile sequence quality



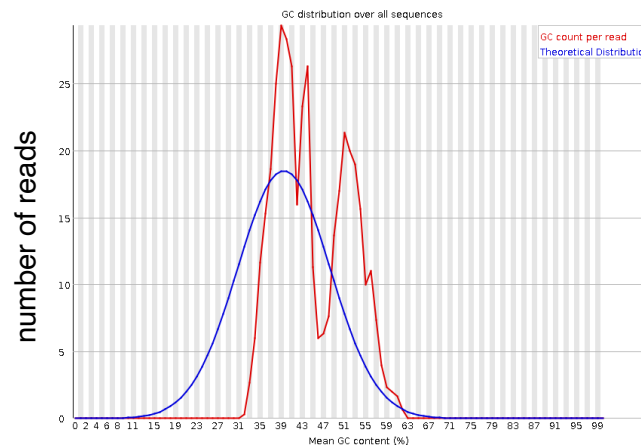
## Per sequence quality scores



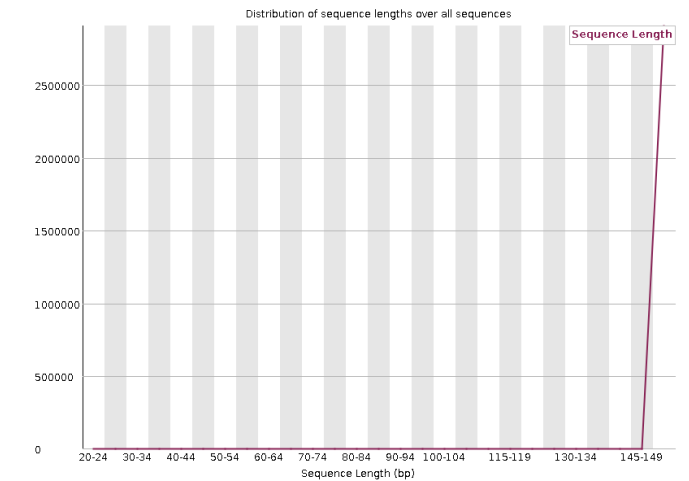
## Per base sequence content



## Per sequence GC content



## Sequence length distribution



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