

Pipeline filogenetyczny

Badanie filogenezy *Drosophila*

Anna Szymik, 22.01.2025

Plan prezentacji

- 1 Charakterystyka analizowanych organizmów
- 2 Pipeline
- 3 Wyniki

Wprowadzenie



- Rodzaj *Drosophila* obejmuje ponad 1,500 gatunków - jeden z najbardziej zróżnicowanych taksonów wśród owadów.
- Genomy stosunkowo niewielkie (np. *Drosophila melanogaster* ~ 175 milionów par zasad).
- Liczba genów kodujących białka wynosi około 12-13,000.
- Rodzaj charakteryzuje się wysokim tempem specjacji.
- Gatunki są szeroko rozpowszechnione i zajmują różnorodne nisze ekologiczne – doskonałe obiekty badań ewolucyjnych.

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE

 Open Access



Phylogenomic analyses of the genus *Drosophila* reveals genomic signals of climate adaptation

Fang Li, Rahul V. Rane, Victor Luria, Zijun Xiong, Jiawei Chen, Zimai Li, Renee A. Catullo, Philippa C. Griffin, Michele Schiffer, Stephen Pearce, Siu Fai Lee, Kerensa McElroy, Ann Stocker ... [See all authors](#) 

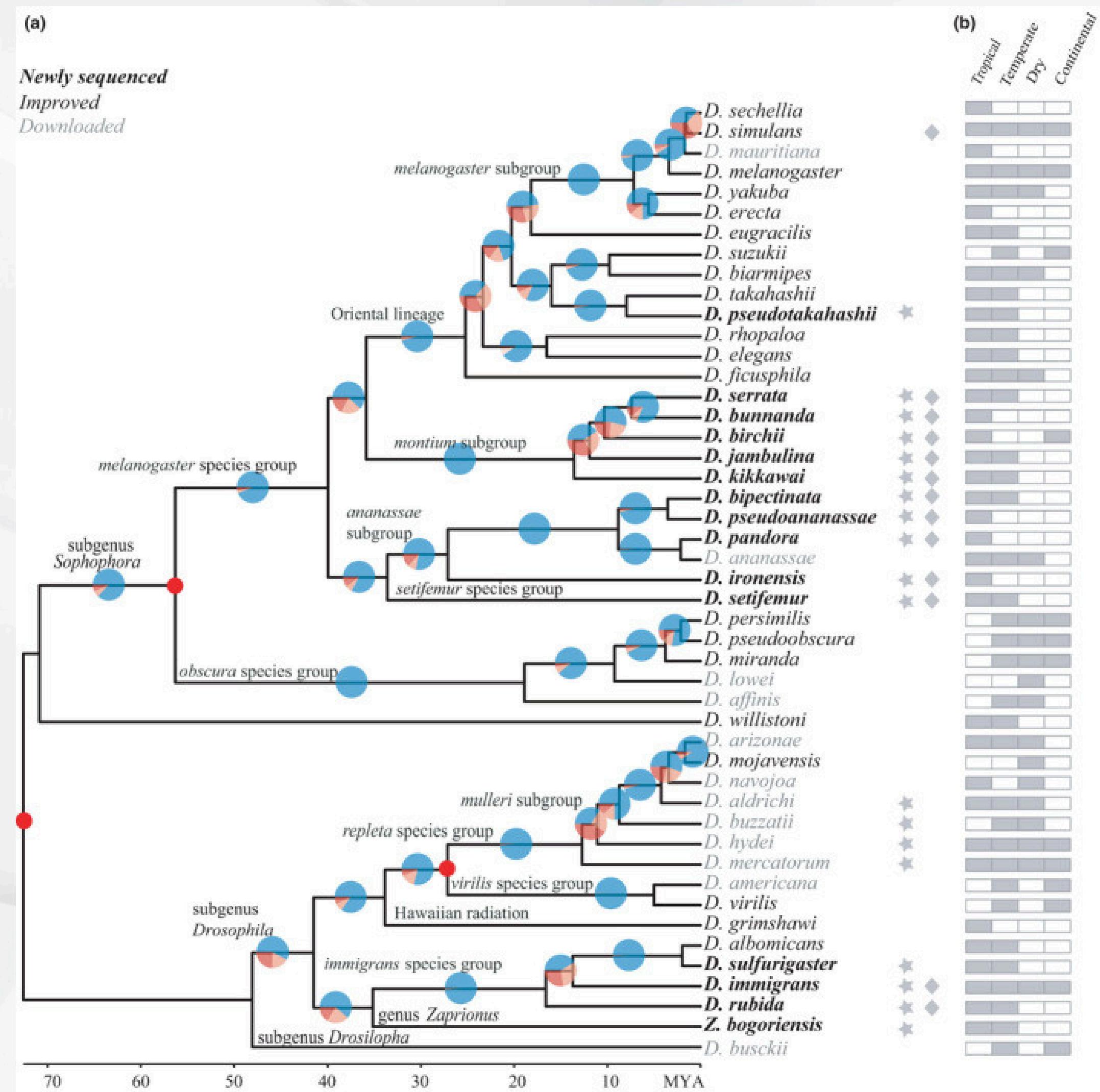
First published: 27 November 2021 | <https://doi.org/10.1111/1755-0998.13561> | Citations: 5

- Autorzy przeprowadzili analizy filogenomiczne, które rozwiązały kilka wcześniej niejasnych relacji filogenetycznych.
- Skupili się na różnicach związanych z różnymi niszami klimatycznymi.
- Zsekwencjonowali na nowo 15 genomów, a kilkadziesiąt innych poprawiono poprzez dodatkowe sekwencjonowanie lub dodano adnotacje.

Gatunki

15 genomów zsekwencjonowanych w cytowanej publikacji:

1. *Drosophila bipectinata*
2. *Drosophila ironensis*
3. *Drosophila pseudoananassae*
4. *Drosophila setifemur*
5. *Drosophila birchii*
6. *Drosophila jambulina*
7. *Drosophila pseudotakahashii*
8. *Drosophila sulfurigaster*
9. *Drosophila bunnanda*
10. *Drosophila kikkawai*
11. *Drosophila rubida*
12. *Zaprionus bogoriensis*
13. *Drosophila immigrans*
14. *Drosophila pandora*
15. *Drosophila serrata*



Pipeline

- 1 Pobranie sekwencji z bazy danych
- 2 Klastrowanie
- 3 Poprawa klastrowania (z zachowaniem paralogów bądź bez)
- 4 MSA
- 5 Inferencja drzew metodą NJ (z bootstrapem bądź bez)
- 6 Obliczanie drzewa konsensusowego i superdrzewa

Pobranie sekwencji

- Wszystkie genomy, wraz z adnotacjami, zdeponowane w bazie danych NCBI pod wspólnym numerem projektu PRJNA736147.
- Pobranie z użyciem modułu Entrez z pakietu BioPython.

BIOPROJECT

[Phylogenomic analyses of the genus Drosophila reveals genomic signals of climate adaptation](#)

We newly sequenced and assembled 15 drosophila species, and 19 transcriptome to assist the annotation. As well, we re-sequenced about 20 individuals for each of 13 species.

[PRJNA736147](#)

[Genomes](#) [BioSample](#) [PubMed](#)

Klastrowanie i MSA

- Narzędzie: MMSeqs2 (Many-against-Many sequence searching)
- Tryb: easy-linclus - zoptymalizowany pod kątem szybkości i pamięci z progiem identyczności: 35%
- 37 022 surowych klastrów, w tym 3 829 zawierających sekwencje wszystkich organizmów
- Usuwanie paralogów: wybór reprezentanta o najwyższym sumarycznym bit-score w porównaniach BLASTP z pozostałymi elementami klastra
- Liczba klastrów do dalszej analizy: 3 829
- Następnie MSA rodzin przy użyciu programu MAFFT

Inferencja drzew

- Metoda: Neighbor-Joining (NJ), zaimplementowana w pakiecie phangorn
- Model substytucji do wyliczenia macierzy odległości: JTT (Jones-Taylor-Thornton)
- Analiza bootstrap: 100 powtórzeń dla każdego drzewa i usunięcie drzew z kladami o wsparciu <60
- Liczba drzew do analizy: 3 829 bez analizy bootstrap i jedynie 135 z analizą bootstrap.

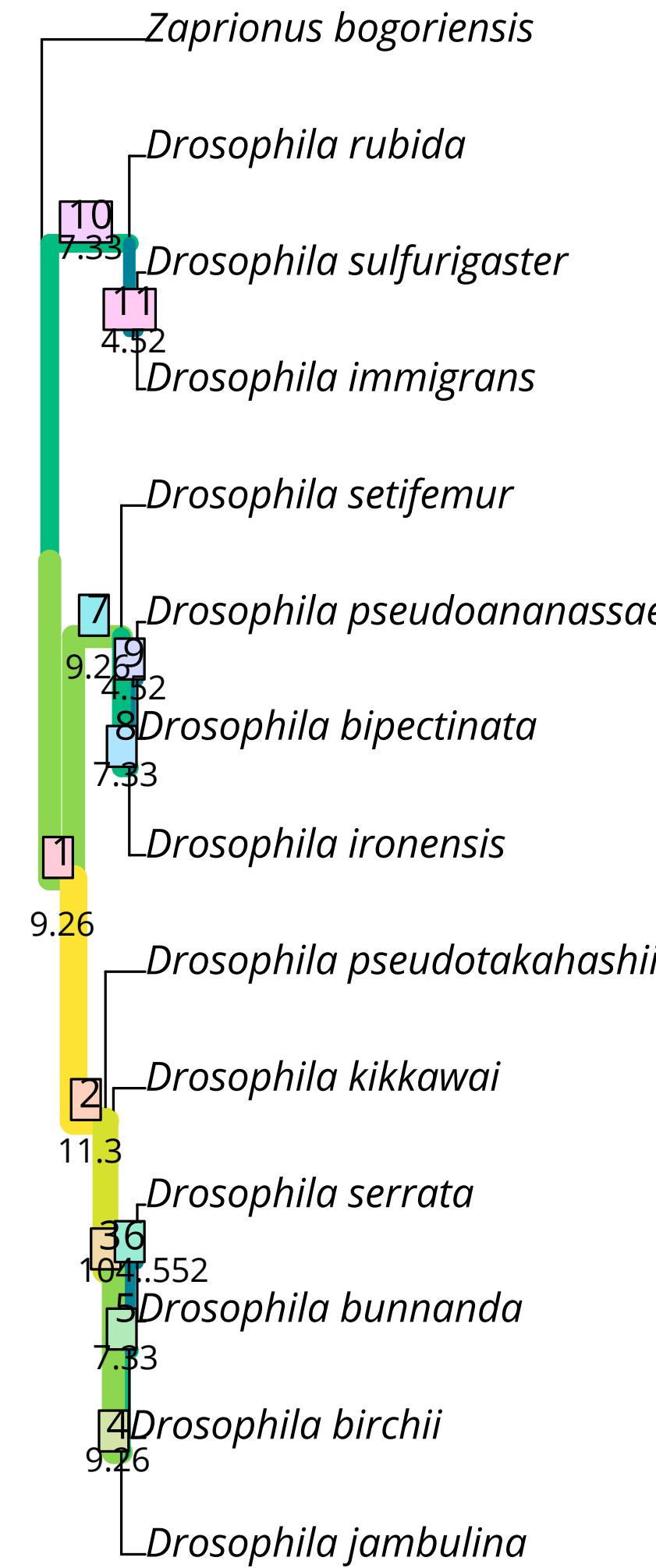
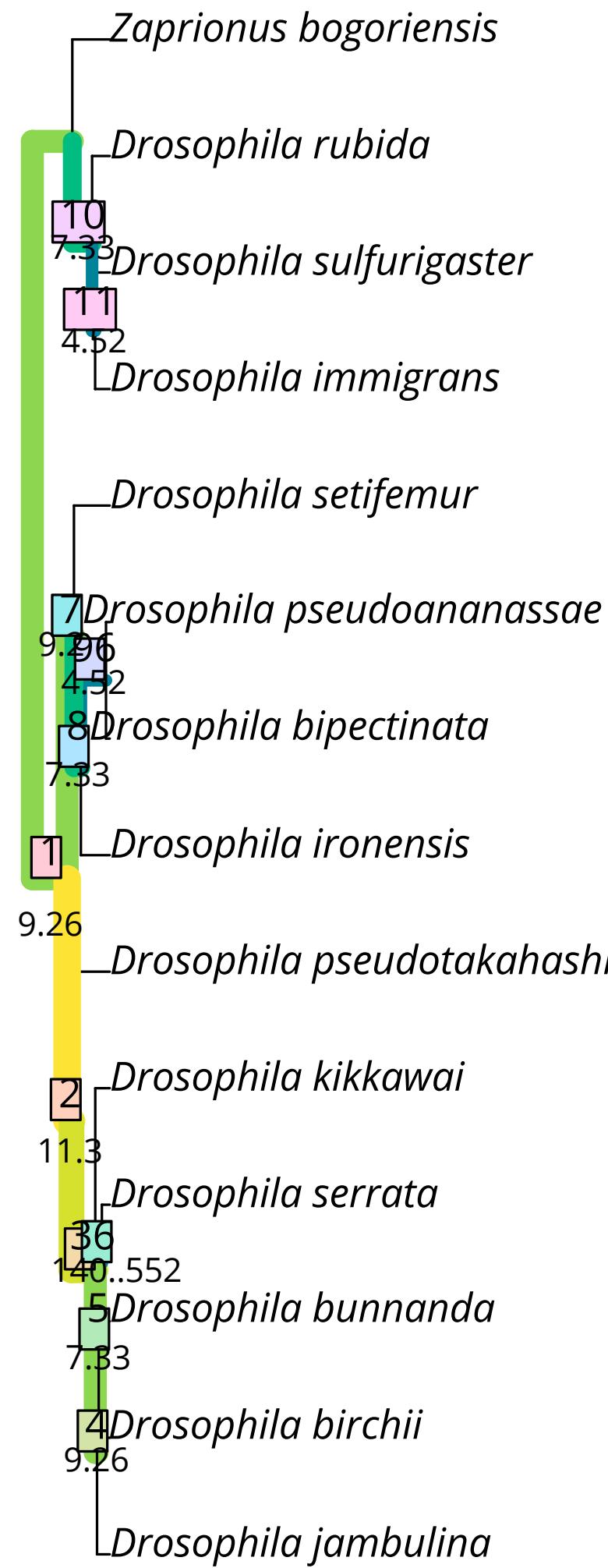
Drzewo genomów

- Konsensus większościowy i zachłanny, wyliczony przy pomocy programu IQ-TREE z klastrów ortologicznych – dla drzew z i bez analizy bootstrap.
- Superdrzewo obliczone przy pomocy programu Fasturec dla klastrów ortologicznych i zawierających paralogi, również dla drzew z i bez analizy bootstrap.
- Ukorzenienie drzewa: *Zaprionus bogoriensis* jako klad zewnętrzny.

Wyniki

Wspólna informacja filogenetyczna

Metoda	Wspólna informacja filogenetyczna
Konsensus większościowy	0.643335
Konsensus zachłanny	0.9468724
Superdrzewo	0.9468724
Superdrzewo z zachowaniem paralogów	0.9468724
Konsensus większościowy z bootstrap	0.7520542
Konsensus zachłanny z bootstrap	1.0
Superdrzewo z bootstrap	1.0



Czas działania etapów

Etap	Czas działania
Pobranie proteomów z bazy danych	1.89 min
Klastrowanie sekwencji	1.49 min
Poprawa klastrowania	58.72 min
Multiuliniowienia rodzin genów	18.20 min
Budowa drzew rodzin metodą NJ (bez analizy bootstrap)	1.52 min
Konstrukcja drzewa genomów metodą konsensusową	1.18 s
Konstrukcja drzewa genomów metodą superdrzewową	5.66 s

Wnioski

- Pipeline w pełni zautomatyzowany (Snakemake) i czasowo efektywny.
- Możliwość optymalizacji skryptu do usuwania paralogów - obecnie nieefektywne operacje na plikach.
- Inferencja drzew metodą NJ: szybka, ale może dawać błędne długości gałęzi i niską rozdzielcość kladów - przy większych zasobach obliczeniowych możliwość zastosowania ML lub metod bayesowskich.
- Wysoka zgodność drzew konsensusu zachłannego i superdrzewa z drzewem referencyjnym, szczególnie w przypadku drzew o wysokim wsparciu bootstrap.

Dziękuję za uwagę!