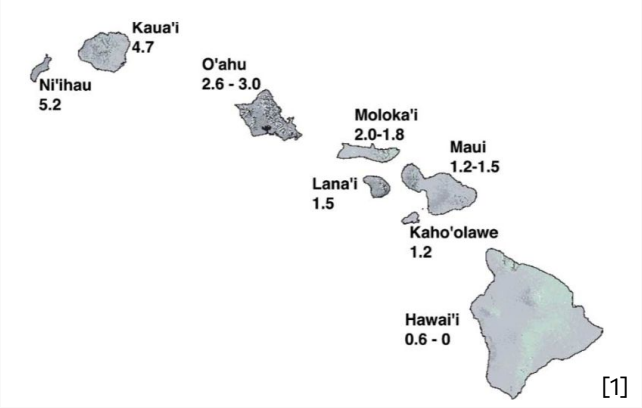
Investigating Scatella Migration Patterns as a Potential Falsification of the Progression Rule in Phylogeography

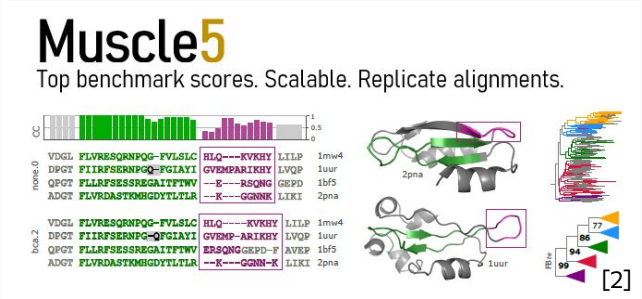
Aaron Nagler (atn45), Aaron Song (ams799), Parker Rho (pkr47), Tony Oh (do256)

Background Information

- Scatella in Hawaii
- Phylogeography: Progression Rule
  - Are Scatella a potential counterexample?
- MSA & MLE to determine ancestry



[1]



[2]

[1] https://www.researchgate.net/publication/7582299_Phylogeny_and_age_of_diversification_of_the_planitibia_species_group_of_the_Hawaiian_Drosophila
[2] https://drive5.com/images/muscle5_header.jpg

Dataset

- Obtained from Professor Patrick O'Grady
- List of DNA sequences containing 6 concatenated sequences from the CO1 gene per species of Scatella (40 different species).

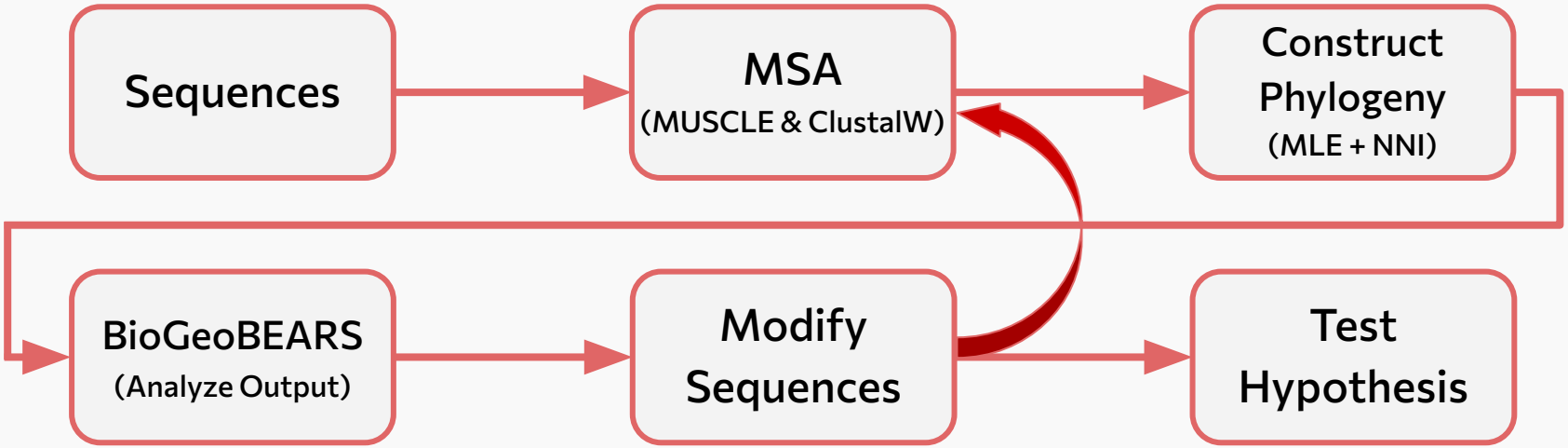
Table 1. DNA sequences and amplification conditions.

Gene	Primer Sequence	Reference
ND2	192 (AGCTATTGGGTTCAGACCCC) 732 (GAAGTTTGGTTTAAACCTCC)	1
COI	2183 (CAACATTTATTTTGATTTTTTGG) 3041 (TYCATTGCACTAATCTGCCATATTAG)	1
COII	3037 (ATGGCAGATTAGTGCAATGG) 3771 (GTTTAAGAGACCAGTACTTG)	1
16S	16sF (CCGGTTTGAACTCAGATCACGT) 16sR (CGCCTGTTTAACAAAAACAT)	2
CAD	320F (ATHTTYGGNATYTGYYTGGGNCAYCA) 338F (ATGAARTAYGGYAATCGTGGHCAYAA) 680R (AANGCRTCNCGNACMACYTCRTAYTC) 843R2 (TCNACCATWCKNARWGCYTTYTGRAA)	3
wee	weeL (GCCTGGGCCGAGGAYGAYCAYATG) weeR (TCACGTGGCCCAGGTCNCCDATYTT)	4

[1] https://www.drosophilaevolution.com/uploads/6/6/8/3/6683664/0130-ogrady_orig.jpg

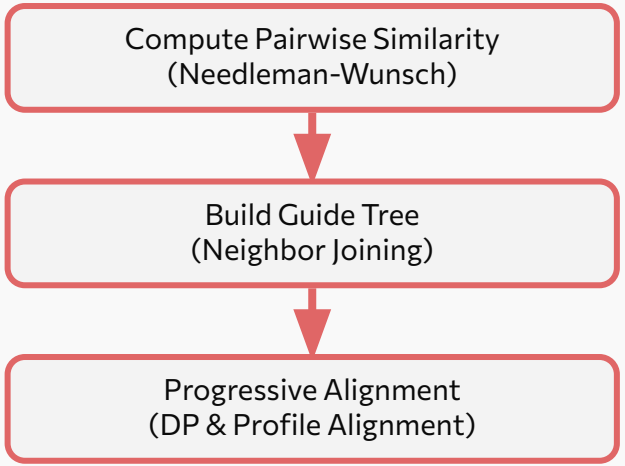
O'Grady et al., 2019

Pipeline


Sequences
MSA
(MUSCLE & ClustalW)
Construct
Phylogeny
(MLE + NNI)
BioGeoBEARS
(Analyze Output)
Modify
Sequences
Test
Hypothesis

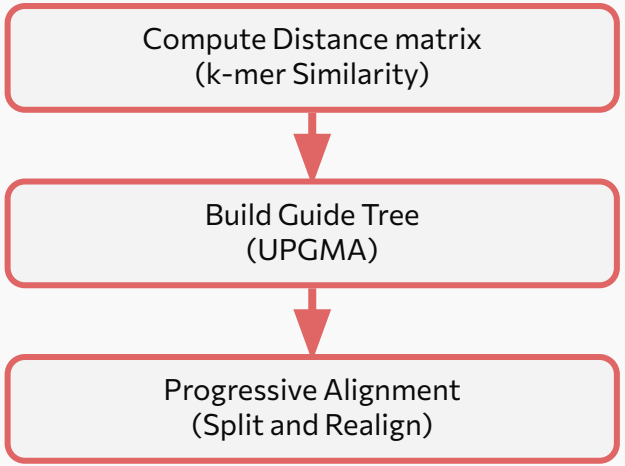
MSA: Progressive Algorithm (ClustalW)

- Pros
  - Fast
  - Effective for close sequences
- Cons
  - No convergence guarantee
  - Cascading errors
  - “Once a gap, always a gap”



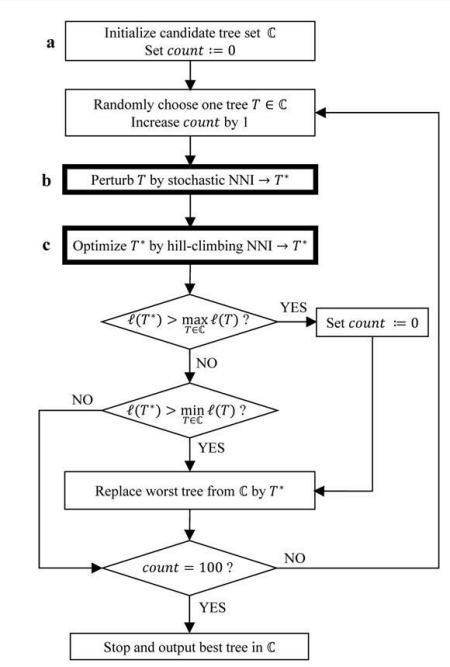
MSA: Iterative Algorithm (MUSCLE)

- Pros
  - High accuracy and speed
- Cons
  - Limitations in handling complex alignments with large insertions/deletions



Building Trees: MLE + NNI (IQ-TREE)

- Make distance matrix, $O(n^2s)$
- Initialize an unrooted tree with NJ, $O(n^3)$
- Maximum Likelihood Branch Length Optimization, O(ns)
- Hill Climbing NNI, O(pns) where p is the number of refinement stages (shown to be asymptotically less than n)



Nguyen et al., 2015

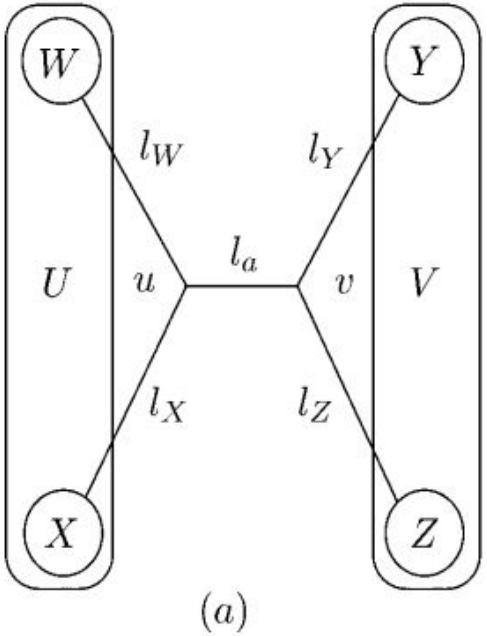
ML Branch Length Optimization

- Equation 1:

$$L = \prod_{i} \sum_{h,h' \in \{A,C,G,T\}} \pi_h L(i = h \mid U) L(i = h' \mid V) P_{hh'}(l).$$

- Equation 2:

$$L(i = h \mid U) = \left(\sum_{g \in \{A,C,G,T\}} L(i = g \mid W) P_{hg}(l_W) \right) \times \left(\sum_{g \in \{A,C,G,T\}} L(i = g \mid Y) P_{hg}(l_Y) \right),$$



Guindon et al., 2003

Hill-Climbing NNI

- NNI: Nearest Neighbor Interchange
- Hill-Climbing:
  1. Create a NNI-neighborhood of trees that strictly increase the computed ML branch length optimization
  2. Remove NNIs with “conflicting branches”
  3. Apply all NNIs from the neighborhood to the current tree and choose the tree with the best likelihood
  4. Repeat to get the best tree ever :D (the locally optimal tree)

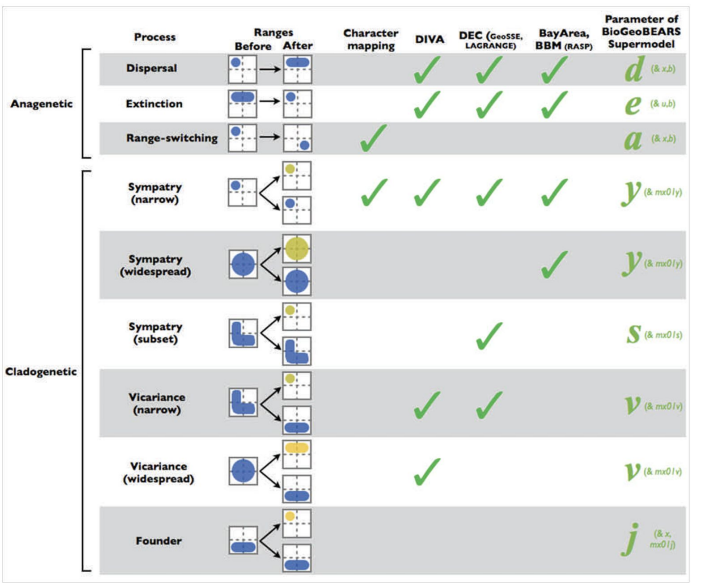
Nguyen et al., 2015

BioGeoBEARS (Matzke et al., 2013)

- R Package to test our phylogeny

Output Parameters:

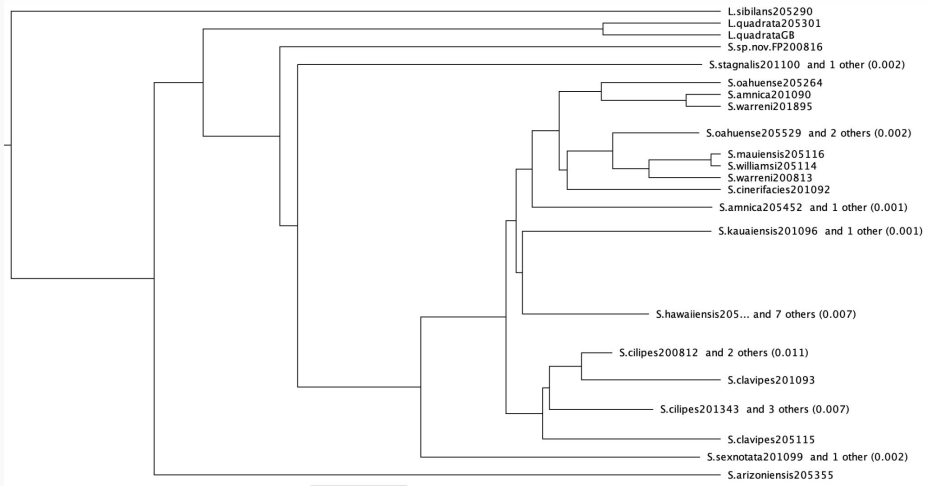
- Dispersal
- Extinction
- Range-switching
- Various Indications of Sympatry and Vicariance

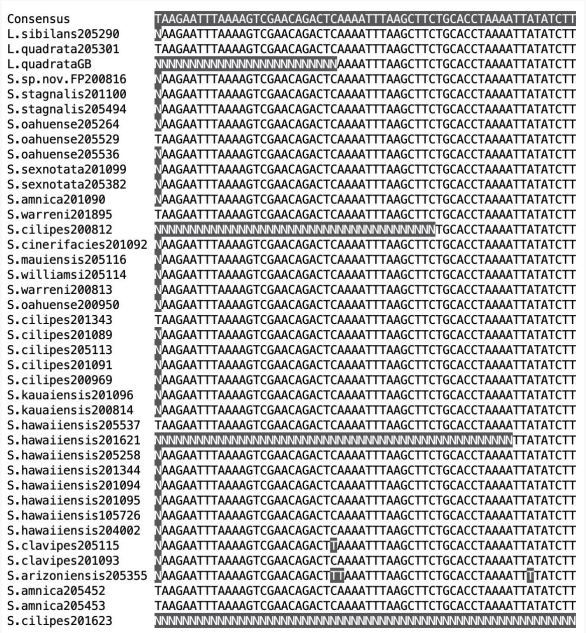


Matzke et al., 2013

Preliminary Results

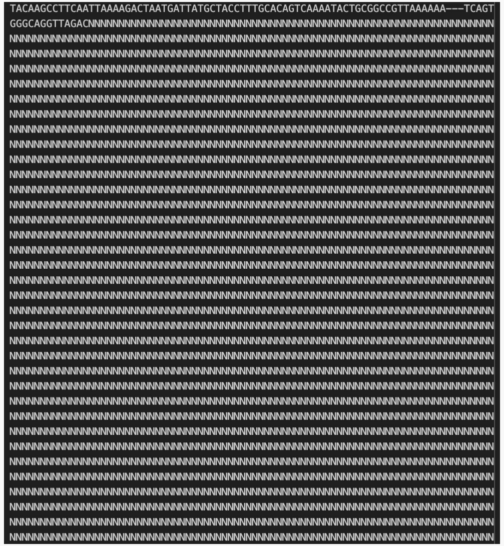
- Constructed MSA using MUSCLE
- Constructed phylogenetic tree using UPGMA





Limitations and Remaining Work

- ❑ Trim sequences to minimize gaps
  - ❑ Too many gaps
- ❑ Continue reimplementation of MUSCLE and MLE + NNI
- ❑ Analyze our phylogenies using BioGeoBEARS



Resources

ML Branch Length Optimization and Hill-Climbing NNI: Stéphane Guindon, Olivier Gascuel, A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood, Systematic Biology, Volume 52, Issue 5, 1 October 2003, Pages 696–704, https://doi.org/10.1080/10635150390235520

BioGeoBears: Matzke, N. J. (2013). Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. Frontiers of Biogeography, 5(4). http://dx.doi.org/10.21425/F5FBG19694 Retrieved from https://escholarship.org/uc/item/44j7n141

IQ-Tree: Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, Bui Quang Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies, Molecular Biology and Evolution, Volume 32, Issue 1, January 2015, Pages 268–274, https://doi.org/10.1093/molbev/msu300

Supplementary images of supporting Scatella were created with the assistance of DALL-E 2

Thank You :3

[1]
https://tasmanianinsectfieldguide.com/hexapoda/insectsoftasmaniadiptera/suborder-brachycera/infraorder-cyclorrhapha/ephdridae-marsh-shore-flies/genus-scatella/scatella-scatella-nitidithorax/