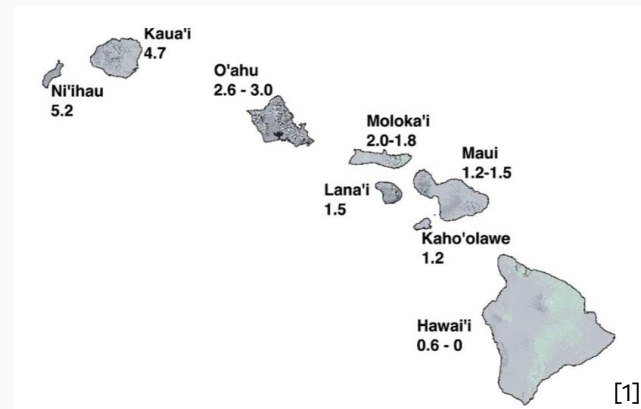


Investigating Scatella Migration Patterns as a Potential Falsification of the Progression Rule in Phylogeography

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

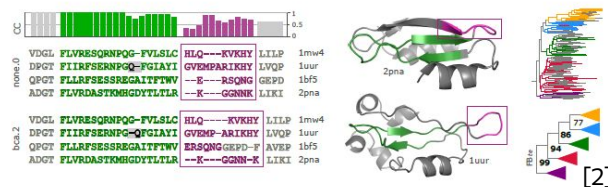
Background Information

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



Muscle5

Top benchmark scores. Scalable. Replicate alignments.



[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).

```

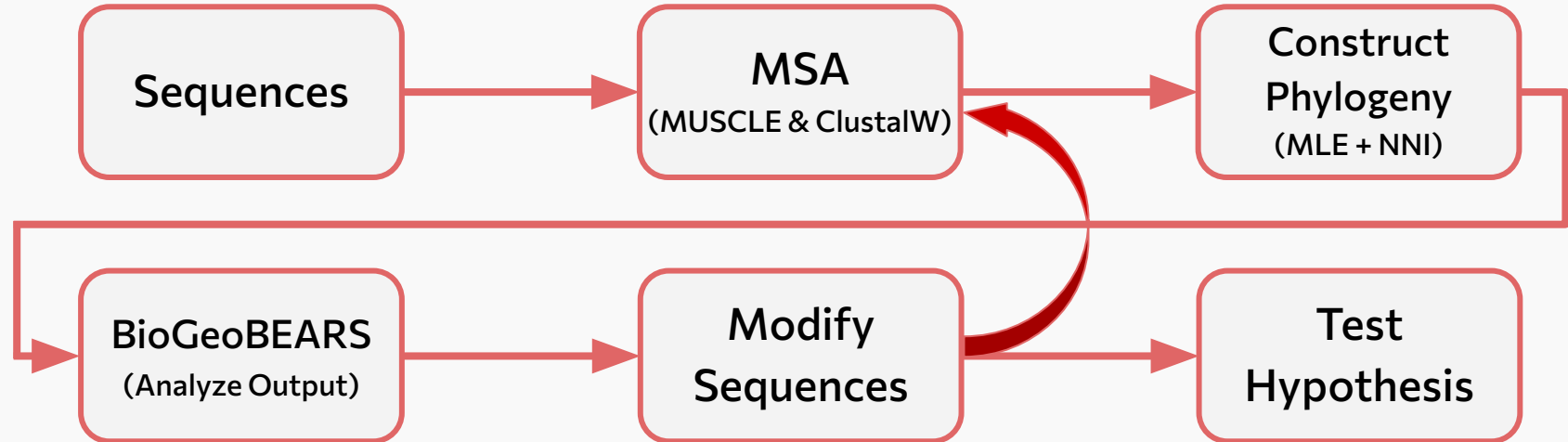
1998 TTCCGCTGCAACCCAGCACTCTCCAGCTGAATGGAGGCGCTCTTCAGCAACGCCAAGCATCAGACAAATGAGGCT— ATATTGAAA
1999
2000 GTCCGAGGCGCTACTTTTGGTACAAATTCATCCCGAGCACACAGGGGCTCCGAGGATTAGAGCGCTGATTTCGAGCTCT
2001 TCCTTGAGAGCGTAAATGGAGAGC— GCTCCGAGTACAGATTTGGTGGAGAGCTTCAGAGAGCTTGAAATACAG
2002 CCGAGGCGGATCAATCCCTGAAAAGGTCACGCGAGGTTGGATTCTTGGTTGGGTGCTATCTATCGAGAGG
2003 TGGCGAATTCGATTTCTGCTGCGAGGATTAAGAGGATGAAGAGAGAGAGAAATCAACATATTCATCATATCCCA
2004 ATATTCGAGAGTACAGCTGGAAGAGGCTGGCGAAGATGATTTCTGCTGCTTACGCGAGATATCTGAGCAA
2005 GTATTAAAGCTGAGCTCCCAATGGGCTCTTTGACATTTGGTGAACAACTGCTTTGAATTTGGAGTGGAGCTGA
2006 GAGAGCTGGAGTTTCCAAAGTACATTTAGATCTCTGGGACAGCTGAGATCGATATGAGAGCGAGATACGA
2007 AGATATTGGGATTCGGCTCAATGAATATGGTGAAGCTGTGGGCGCTCTGAAGCGCTACTCTGCTGGCGAGGCTCG
2008 GAGGCGCTGAGCACTAGGCTATCGGATGAATGGAGCTGGGCTCTCTCGCTGGGCGCTCTGGGCTTTGGCCAA
2009 TAATGAGGAGGCTGGAGATTTTGGCGATCACTGCTTGGCTG—
2010 —CACTCTGAGGTTATATTATTTACCGGATTCGGAATATTTCTCATATATTATGCTCAGAGAGCTGTGTAAGAG
2011 GAAATTTGGTCTATTAGGATATTTATGCAATACAGAAATGGTTAGATGATCATCTGTGGAGTCATATAT
2012 ATTTACAGTAGATAGATGTAGATACCGGAGCTATTTTACTTCAGCTACTATATATTTGCTGCTCCAACTGAAATTA
2013 AATTTTGGTGTATGCTACATACAGGAGAGCAATAGTATGCTCTGATATATATGAGAGCTGGATTTGTA
2014 TTTTATTCAGTAGGAGGTTTAACTGGAGTGTATAGCAAAATTCATCCCTGGATATATTTACATGACACTATTA
2015 CTAGTGTGCTACTTCATATATTTATCAATAGAGAGTGTATTTGCTATATAGCAAAATTTTATCATTTGATATCCT
2016 TATTACTGATACAAATATATAGATGATTAAGAGTCAATTTGATATATATTTATTTGGGTAATTTACTTCTT
2017 TTCCCCAGAGTTCTTAGGATTAGAGAGATACCTCGAGATATTCGACTATCAGATGCTTACACATCATGAAATGT
2018 AATTCAGCAATGGGCACTATTTCTTACTAGGATTTATTTCTTTTACATATTTGAAAGGTTAGTTCCCA
2019 AAGTCAGTAAATTTACCAATACCACTAATTCATCAATGTGATGTACAAATCTCTTCCAGCTGAACACAGTTAT
2020 GCGAATTCGCTTATACATGA—
2021 —CAAAATAAATGTCACAGATGATCAATTTAGGTTTACAAAATAGTCTTCACTTTAATAGAACATATAC
2022 TTTTTCATGACCAAGCACTAATATTTAGTAAATATCTAGTATAGTATGATCTATATTTATATTTATTTTAT
2023 ATAAATTTAGAAATGATTTTACTTGAGAGCAATTTGAGTAAATGAGCAATGATACAGAGATATTTATCTA
2024 TTATTTGCTTTTACCTTCACTTATATATTTCTCTGATGAATTAATGAGCTCACTCACTATTAATCAATTTGG
2025 TCATCAATGATCTGAGATATGATCTAGATTTATATGATTTATGATTTATGATATATATATCTTCAACATGAT
2026 TAGCATTAGATGATTTGGGCTAGATGTTGATACCGAGTGTTTTACCATATAATACAAATTCGATCTTAGTG
2027 AGAGTGGATGATTTCACTCTGAGCTGCTCTCTAGTGTAAAGTAAAGAGAGCTCTGGAGCACTAATACCA
2028 AACAAATTTTATTAATGCCCGAGGATATTTTATGGACATGTTTCAGAAATTTGGGGCCAAACATAGATTTATAC
2029 CTATTTGATTT—
2030

```

Table 1. DNA sequences and amplification conditions.

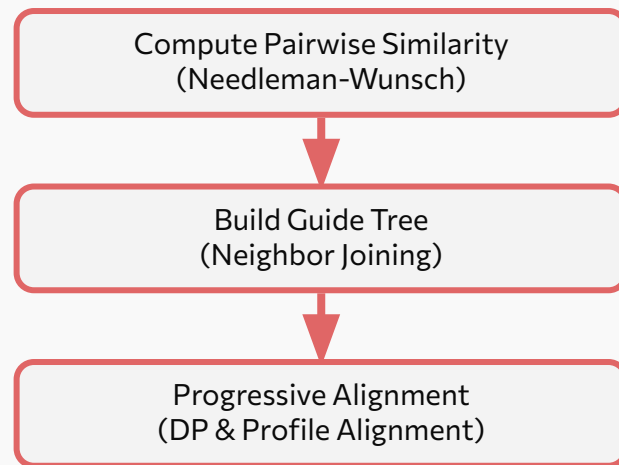
Gene	Primer Sequence	Reference
ND2	192 (AGCTATTGGGTTTCAGACCCC)	1
	732 (GAAGTTTGGTTTAAACCTCC)	
COI	2183 (CAACATTATTTGATTTTTTGG)	1
	3041 (TYCATTGCACTAATCTGCCATATTAG)	
COII	3037 (ATGGCAGATTAGTGCATATGG)	1
	3771 (GTTTAAAGACAGTACTTGG)	
16S	168F (CCGGTTTGAACCTCAGATCAGCT)	2
	168R (CGCCTGTTTAAACAAAACAT)	
CAD	320F (ATHTTYGGNATYTYGYTGGGNCAYCA)	3
	338F (ATGAARTAYGYAATCGTGGHCAYAA)	
	680R (AANGCRTCNCGNACMACYTCRTATYC)	
	843R2 (TCNACCATWCKNARWGCYTYTGRAA)	
wee	weel (GCCCTGGGCCGAGGAYGAYCATG)	4
	weeR (TCACGTGGCCAGGTCNCCDATYTT)	

Pipeline

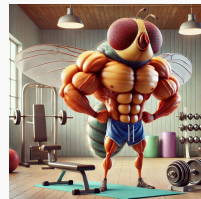


MSA: Progressive Algorithm (ClustalW)

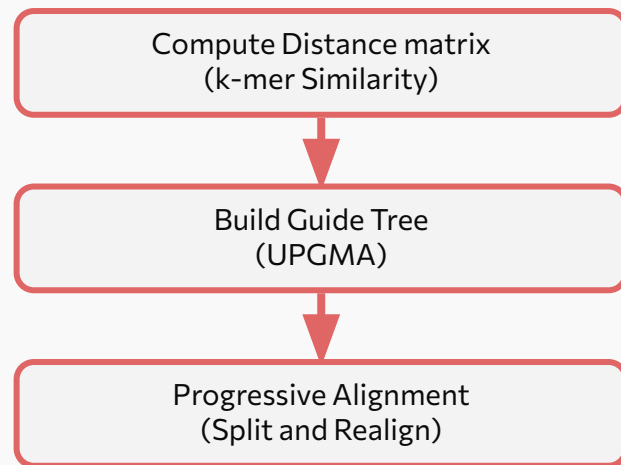
- ❑ Pros
 - ❑ Fast
 - ❑ Effective for close sequences
- ❑ Cons
 - ❑ No convergence guarantee
 - ❑ Cascading errors
 - ❑ “Once a gap, always a gap”



MSA: Iterative Algorithm (MUSCLE)

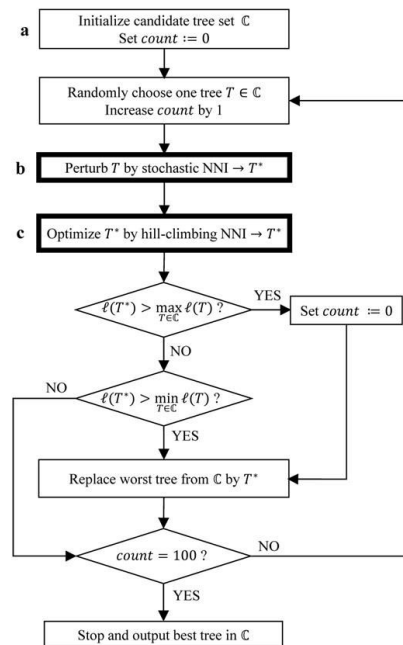


- ❑ Pros
 - ❑ High accuracy and speed
- ❑ Cons
 - ❑ Limitations in handling complex alignments with large insertions/deletions



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)



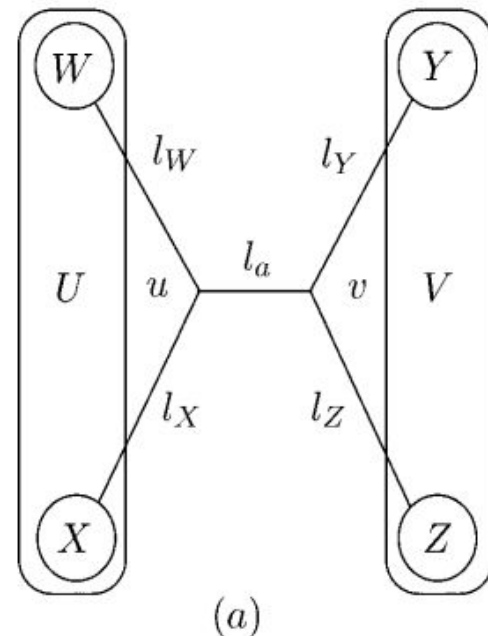
ML Branch Length Optimization

□ Equation 1:

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

□ Equation 2:

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



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)



BioGeoBEARS (Matzke et al., 2013)

- ❑ R Package to test our phylogeny

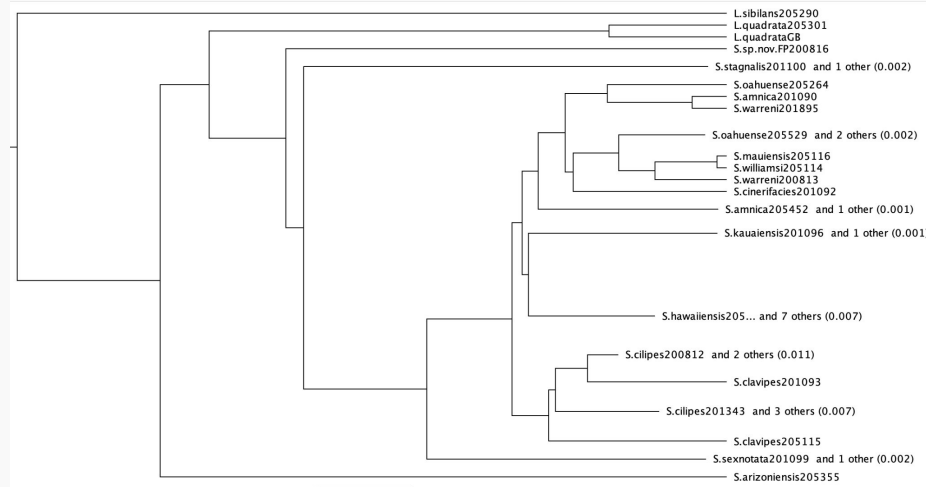
Output Parameters:

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

	Process	Ranges Before After	Character mapping	DIVA	DEC (GeoSSE, LAGRANGE)	BayArea, BBM (RASPL)	Parameter of BioGeoBEARS Supermodel
Anagenetic	Dispersal		✓	✓	✓	✓	d (& x.b)
	Extinction		✓	✓	✓	✓	e (& u.b)
	Range-switching		✓				a (& x.b)
Cladogenetic	Sympatry (narrow)		✓	✓	✓	✓	y (& mx01y)
	Sympatry (widespread)					✓	y (& mx01y)
	Sympatry (subset)				✓		s (& mx01s)
	Vicariance (narrow)		✓	✓			v (& mx01v)
	Vicariance (widespread)		✓				v (& mx01v)
	Founder						j (& x, mx01j)

Preliminary Results

- Constructed MSA using MUSCLE
- Constructed phylogenetic tree using UPGMA



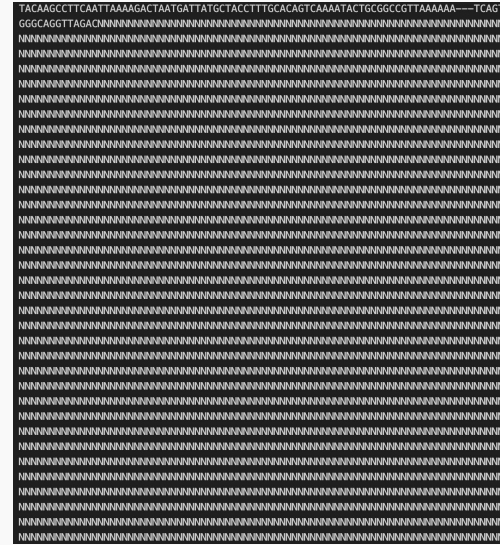
Consensus

```

L.sibilans205290
L.quadrata205301
L.quadrataG8
S.sp.nov.FP200816
S.stagnalis201100
S.stagnalis205494
S.oahuense205264
S.oahuense205529
S.oahuense205536
S.sexnotata201099
S.sexnotata205382
S.amnica201090
S.warreni201895
S.cilipes200812
S.cinerifacies201092
S.mauiensis205116
S.oahuense205529 and 1 other (0.002)
S.mauiensis205116
S.williamsi205114
S.warreni200813
S.oahuense200950
S.cilipes201343
S.cilipes201089
S.cilipes205113
S.cilipes201091
S.cilipes200969
S.kauaiensis201096
S.kauaiensis200814
S.hawaiiensis205537
S.hawaiiensis201621
S.hawaiiensis205258
S.hawaiiensis201344
S.hawaiiensis201094
S.cilipes200812 and 2 others (0.011)
S.hawaiiensis201095
S.hawaiiensis105726
S.clavipes201093
S.hawaiiensis204002
S.cilipes205115
S.clavipes201093
S.arizoniensis205355
S.clavipes205115
S.sexnotata201099 and 1 other (0.002)
S.amnica205452
S.amnica205453
S.arizoniensis205355
S.cilipes201623
  
```

Limitations and Remaining Work

- ❑ Trim sequences to minimize gaps
 - ❑ Too many gaps
- ❑ Continue reimplementations 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

