GENETIC VARIATION AND INTROGRESSION ACROSS A HYBRID ZONE OF DIVERGENT PINE SQUIRRELS (*TAMIASCIURUS*) USING DDRADSEQ DATA

An Undergraduate Research Thesis

Presented to
The Department of Evolution, Ecology, and Organismal Biology
In Partial Fulfillment of the Requirements for Graduation with Research Distinction in Biology
The Ohio State University

By
Sahil Patel
May 2019

Project Advisor: Dr. Andreas S. Chavez, Department of Evolution, Ecology and Organismal Biology

Abstract

The North American tree squirrels *Tamiasciurus hudsonicus* and *Tamiasciurus douglasii* are parapatric species that hybridize over a transitional forest zone in the Northern Cascade mountains. In a previous study, Chavez (2011) found pure parental populations at the ends of the transect and found hybrids in a transitional forest habitat with the use of mtDNA and microsatellite data. The levels of hybridization and strength of barriers to gene flow are still unclear. The objective of this study is to use genome-wide SNP data from ddRADseq approach to provide new insight into the dynamics of gene flow and introgression between *Tamiasciurus hudsonicus* and *Tamiasciurus douglasii*. Population assignment analyses of SNP DNA generated using ddRADseq allowed the genotyping of individuals in populations, and we found different distributions of hybrids across the transect and different levels of hybridization among hybrids and confirmed the presence of first generation hybrids. The geographical cline shows differential introgression asymmetry of the hybrid zone toward *T. douglasii* range. Perhaps the Northern Cascade mountains is not providing a significant physical barrier to gene flow and there is movement of *Tamiasciurus hudsonicus* into the *Tamiasciurus douglasii* range.

INTRODUCTION

Next-generation sequencing has allowed for the relatively cheap collection of genome-wide for evolutionary genetic studies. As opposed to earlier studies using DNA sequence data from just a few select loci, these genome-wide datasets allow researchers a more holistic and precise investigation of species' demography and evolutionary history. These new methodological approaches are valuable in situations where species are naturally hybridizing and forming new genomic combinations from divergent taxa.

Hybrid zones are regions where divergent species meet and interbreed in secondary contact and produce hybrids (Gompert & Berkle 2010; Barton & Hewitt 1985; Harrison 1993). Hybrid zones act as models for early stages of speciation and allow for the study of reproductive barriers because of the potential ability to discern important phenotypic traits and genomic regions involved in natural selection against hybrids or sexual selection for increased assortative mating. Many hybrid zones form tension zones, where dispersal (potential gene flow) of parental individuals into the hybrid zone is balanced by selection against hybrids and hybrid phenotypes (Barton& Hewitt 1985; Harrison 1990; Barton and Bengtsson 1986; Payseur 2010).

Pine squirrels (genus: *Tamiasciurus*) are found throughout coniferous forests in North America because of their reliance on conifer seeds as a major food resource. The hybrid zone between Douglas squirrels, *T. douglasii*, and red squirrels, *T. hudsonicus,* occurs in a 25-km transitional forest zone near the crest of the North Cascades in Washington and British Columbia (Smith 1968; Chavez et al. 2011). Pure *T. douglasii* are found on the western side of the Cascade Mountains in forests

dominated by Douglas fir (*Pseudotsuga menziesii*), western hemlock (*Tsuga heterophylla*), and subalpine fir (*Abies amabilis*), (Smith 1968). Pure red squirrels are found on the eastern side of the Cascade Mountains and east of the transitional forest zone. These squirrels differ in multiple phenotypic traits including vocalization, fur color, and skull morphology that are likely adaptations to the different forest environments that they inhabit (Smith 1981). It is possible that strong divergent selection against admixed individuals with hybrid phenotypes may be contributing to the continued divergence between these hybridizing species that are experiencing some gene flow (Chavez et al. 2011).

Chavez et al. (2011) used microsatellite data from nine loci to characterize hybridization and clinal patterns of genetic variation across the pine squirrel hybrid zone in the North Cascades. They found second generation hybrids, which demonstrated that hybrids appear to be reproductively viable. However, no first generation hybrids were detected in their sampling. One limitation from this previous work was whether the use of relatively few loci accurately characterized levels of hybridization and the geographic position of the hybrid zone.

We used double-digest restriction associated DNA sequencing (ddRADseq) to attempt to better characterize patterns of hybridization and the geographic position of the hybrid zone. This genome-wide DNA sequencing approach affords better coverage of loci for Bayesian analyses and give more precise population structures and geographic clines.

METHODS

Sampling, DNA Extraction, ddRAD seq Library Preparation, and Sequencing.

We collected 160 *T. douglasii*, *T. hudsconicus*, and hybrid individuals from 28 localities along an 130 kilometer transect through the North Cascades hybrid zone in Washington. Animals were collected between the years 2008-2018.

DNA was extracted from frozen liver tissues using a Qiagen DNeasy kit and quantified using a High Sensitivity DNA assay on a Qubit Fluorometer. DNA was digested and size selected according to the ddRADseq protocol described by Peterson et al. (2012), with some modifications. Restriction enzymes *MspI* and *SbfI* were used for the double digest, with *MspI* being the common cutter and *SbfI* the rare cutter. DNA fragments were ligated with barcoded Illumina adapters. The samples were then pooled by eight samples, size-selected for fragments in the range of 300‑400 base pairs (bp) using the BluePippin size selector, ligated with Illumina multiplexing indices, and final pools were assessed for quality using Agilent TapeStation electrophoresis. All pools were combined and sequenced for 50bp single-end reads on a single lane on an Illumina HiSeq4000 at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant.

Bioinformatics and Data Analysis

Raw Illumina data were first demultiplexed using the unique combination of barcodes and indices, and filtered using iPyrad 0.7.8 (Eaton 2014), according to the

default filters with several modifications. Alterations were made in the params file at lines 9, 10, 14, 21, 22, and 23. At line 9, a maximum of 4 low-quality base pairs was allotted per read to minimize sequence uncertainty. Lines 11 and 12 were set to a minimum depth of 10, to set the lower bound of the read depth at which statistical and majority base calls were made, respectively. Line 14 was increased to a value of 0.95 from the default of 0.90 in order to increase the percent sequence similarity required for sequences to be collapsed into a single cluster; this decreases the chances of falsely calling similar sequences the same sequence. The minimum number of samples that must have data at a given locus for it to be retained in the final data set was set at 98% (156 out of 160 samples: line 21 in params file). Lines 22 and 23 were both assigned values 4 to control the maximum number of SNP's and indels allowed in a locus, respectively.

Admixture Analysis

The number of genetic population groupings was estimated with a 493 SNP dataset from iPyrad using a Bayesian clustering method in STRUCTURE 2.3.4 (Pritchard et al. 2000). The population assignment was analyzed using K =2 and with a 10,000 burnin period followed by 100,000 reps. Following Vaha & Primmer (2006), we categorized individuals into three clusters, using a range of q-values between 0 and 0.10 as pure *T. hudsonicus*, 0.90–1.0 as pure *T. douglasii* and 0.11–0.89 for admixed individuals.

We also used the program NEWHYBRIDS 1.1 to identify hybrid individuals (Anderson & Thompson 2002). This program is a more specific Bayesian method for

identifying hybrids and can be used to identify individual assignment to six genotype frequency classes (pure parental 1, pure parental 2, F1, F2, backcross to pure parental 1 and backcross to pure parental 2). NEWHYBRIDS uses a MCMC sampling approach to acquire estimates from the posterior distribution that reflect the level of certainty that an individual belongs to a certain hybrid class We performed 1 X 10^6 MCMC sweeps with a burn-in of 1 X 10^4. We ran the analyses using the Uniform prior.

Geographical Cline

We used Q-values from our STRUCTURE results from each individual in each population to fit a null model cline and estimate cline center and cline width using the R package HZAR (Derryberry et al.2014). The geographical cline allows us to analyze the correlation between the hybrid zone and allele frequency. Cline center and width allow us to see the center of the hybrid zone, range of the hybrid zone, and determine introgression patterns across the transect (Endler 1977).

RESULTS

iPyrad

We obtained a total of 66.52 million sequence reads and an average of 1.96 (+ 1.24 SD) million reads per individual. All individuals were retained in the final analyses with a range of 1.57–3.13 million reads. On average, the total number of retained loci per individual was 823 (+61 SD).

Bayesian Analyses

Thirteen populations were generated using STRUCTURE (Figure 2 and Table 1). Four of the populations that lie farther than ninety-three kilometers on the transect lie east of the transitional forest zone and contain mostly *T.hudsonicus*, and corresponds with the pure zone of *T.hudsonicus*. Four populations located between seventy-seven kilometers and seventy-six kilometers, centrally located in the transitional forest region, contained parental *T.hudsonicus*, parental *T.douglasii*, and admixed individuals. The five remaining populations span the transitional forest region and the western Pacific coastal zone previously determined to be the *T. douglasii* pure zone (Chavez et a. 2011). Pure *T.douglasii* and admixed individuals are seen from the transitional forest zone to the *T.douglasii* pure zone. The presence of admixed individuals in the *T. douglasii* pure zone contradicts the tension zone observed in Chavez (2011).

The NEWHYBRIDS results is consistent with STRUCTURE results and found varying levels of introgression represented by the different hybrid classes seen in populations from the transitional forest zone through to the *T. douglasii* parental zone. The *T. hudsonicus* range consisted of pure *T. hudsonicus* individuals. in eastern populations near the transitional forest region F1 hybrids and back crossed with *T.hudsonicus* made up a percentage of the populations.

Geographical Cline

The no tail fitted geographical cline model was generated from the R package HZAR. It determined the center of the cline to be 76.5 km on the transect and the width of the hybrid zone to be ±3.15 km (Figure 2 , Table 2). This corresponds with the location of the transitional forest zone region determined to be the tension zone (Chavez et al. 2011).

DISCUSSION

Our results using high-throughput genotyping support previous findings (Chavez et al. 2011) that the hybrid zone between *Tamiasciurus douglasii* and *T. hudsonicus* occurs within the transitional forest zone in the North Cascade Mountains in Washington. Furthermore, these findings show similar clinal patterns with sharp clinal transition between Douglas and red squirrels in this forest zone. However, our results uncover both more levels and a broader geographic region of hybridization than what was previously understood.

Levels of hybridization.

Our findings demonstrate that genome-wide SNP data has provided a greater resolution of hybridization in this system. We found that the proportion of admixed individuals was roughly equal to proportions of each parental species within the hybrid zone. These findings provide strong evidence that assortative mating

between the parental species is relatively weak. Unlike the previous genetic study of this hybrid zone (Chavez et al. 2011), we also discovered the presence of F1 hybrids and later generation backcrossing, which confirms that interbreeding between pure individuals of each species is occurring, as well as between hybrids and pure parentals. The presence of later generation hybrids and backcrosses with both parental species gives strong evidence that heterogametic reproductive sterility is not a barrier to gene flow between these squirrel species.

Geography of the Hybrid zone

Consistent with previous genetic work (Chavez et al. 2011), our geographical cline result based on genome-wide SNP data shows the genetic transition between these hybridizing species occurs within the transitional forest zone near the crest of the Cascade Mountains. However, despite the cline center occurring within this transitional forest, we also detected F1 hybrids within the western coastal forests that is typically inhabited by *T.douglasii*. This implies that some pure *T. hudsonicus* individuals must also occur in this region, which has never been documented, or that long distance dispersal of F1 individuals has occurred from the hybrid zone. If the latter is true, it is curious that F1 individuals have not been detected in the eastern interior forests within the geographic range of pure *T. hudsonicus*. Thus, it is possible that the hybrid-zone is moving further into the geographical range of *T.douglasii*.

CONCLUSION

Our analyses of the hybrid zone between *T. douglasii* and *T. hudsonicus* using genome-wide data provide new insights into the dynamics of gene flow and introgression among divergent, but interbreeding parapatric species. This hybrid zone shows some evidence of asymmetry where hybridization is occurring, which suggests that there may be movement of the hybrid zone. Several hybrid zones have differential introgression asymmetry, like the Townsend and Hermit Warbler hybrid zone (Krosby et al. 2008). Surprisingly, the possible hybrid-zone movement in these divergently adapted pine squirrels suggests that the Northern Cascade mountains is not providing a significant physical barrier to gene flow and the movement of one species into the other species' range. This has important consequences for the viability of *T. douglasii* and whether they are being outcompeted for resources and mating opportunities by *T. hudsonicus*.

Acknowledgements

I would like to thank my project advisor, Dr. Andreas Chavez and department representative, Dr. Ryan Norris, for their valuable support, encouragement and guidance throughout this process. I would like to thank Zachery Hanf for lab technical assistance and insightful discussions throughout my time in the lab. I would like to thank Brooke Rawson for work on the first ddRADseq library and for discussions and lab technical assistance during this process.

References

Anderson, E. C., & Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, *160*(3), 1217–1229..

Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridizing populations. Heredity 57:357–376

Barton, N., & Hewitt, G. (1985). Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics, 16*, 113-148.

Chavez, A. S., Saltzberg, C. J., & Kenagy, G. J. (2011). Genetic and phenotypic variation across a hybrid zone between ecologically divergent tree squirrels *(Tamiasciurus)*. *Molecular Ecology, 20*: 3350-3366.

Derryberry, E. P., Derryberry, G. E., Maley, J. M. and Brumfield, R. T. (2014), hzar: hybrid zone analysis using an R software package. Mol Ecol Resour, 14: 652-663.

Eaton, D.A.R. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* 30:844– 1849.

Endler J (1977) Geographical Variation, Speciation, and Clines, Princeton University Press, Princeton, New Jersey. Evanno.

Harrison RG (1990) Hybrid zones: windows on evolutionary process. Oxford surveys in evolutionary biology 7:69–128.

Krosby M, Rowher S (2009) A 2000 km genetic wake yields evidence for northern glacial refugia and hybrid zone movement in a pair of songbirds. Proceedings of the Royal Society B: Biological Sciences, 276, 615–621.

Payseur BA (2010) Using differential introgression in hybrid zones to identify genomic regions involved in speciation. Molecular Ecology Resource 10:806–820.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS One* 7(5):1-11.

Vähä, J. and Primmer, C. R. (2006), Efficiency of model‐based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. Molecular Ecology, 15: 63-72.

Smith, C. C. (1968), The Adaptive Nature of Social Organization in the Genus of Three Squirrels Tamiasciurus. Ecological Monographs, 38: 31-64.

Smith, C. C. (1981), The Indivisible Niche of Tamiasciurus: An Example of Nonpartitioning of Resources. Ecological Monographs, 51: 343-363.

Figures

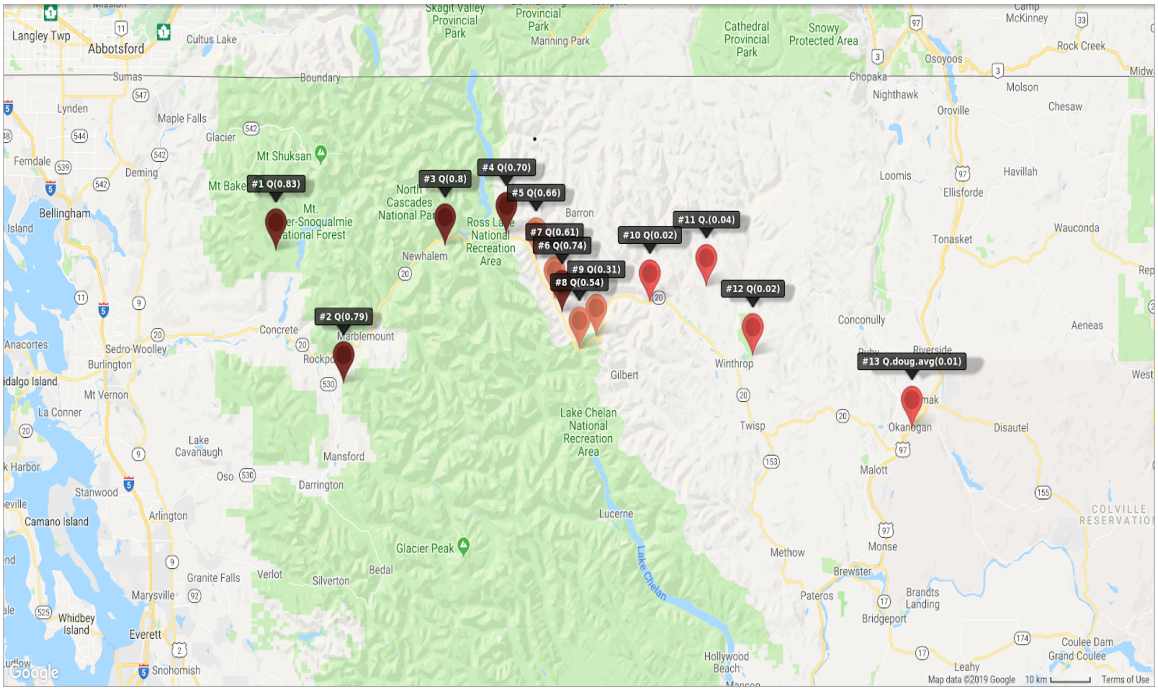

Figure 1 Sample distribution along transect. Colors correspond to STRUCTURE populations: Red represents the eastern *T. hudsonicus*, brown represents western *T. douglasii*, orange represents the hybrid *T. hudsonicus/douglasii* group.

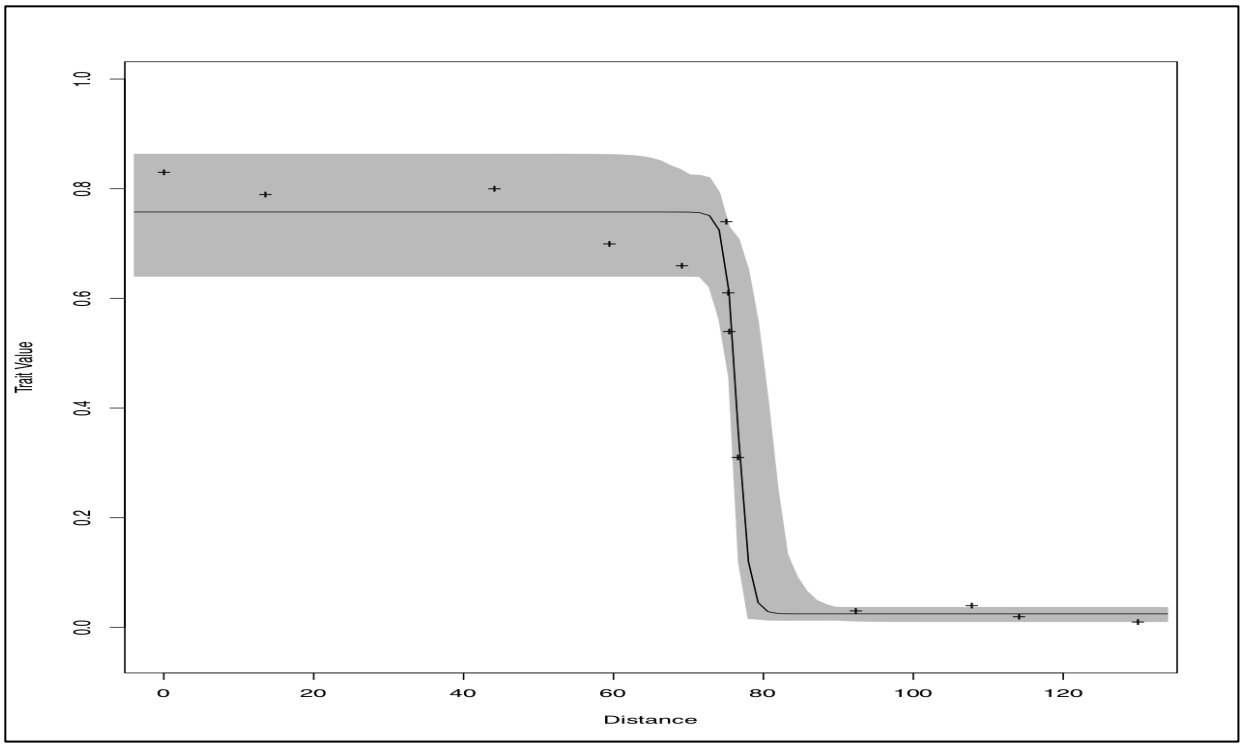

Figure 2. Geographic cline none fitted model constructed from the R package HZAR using the average Q-values of the sampling locations returned from STRUCTURE. The Distance represents the one-dimensional transect distance in kilometers. The trait value is the average Q-value of each population. 1.0 represents pure *T. douglasii*, populations, and 0.0 is pure *T.hudsonicus* populations. The average Q-value ranges from 0.01-0.83, which indicates that the western sampling populations contained hybrids within the parental *T. douglasii* zone..

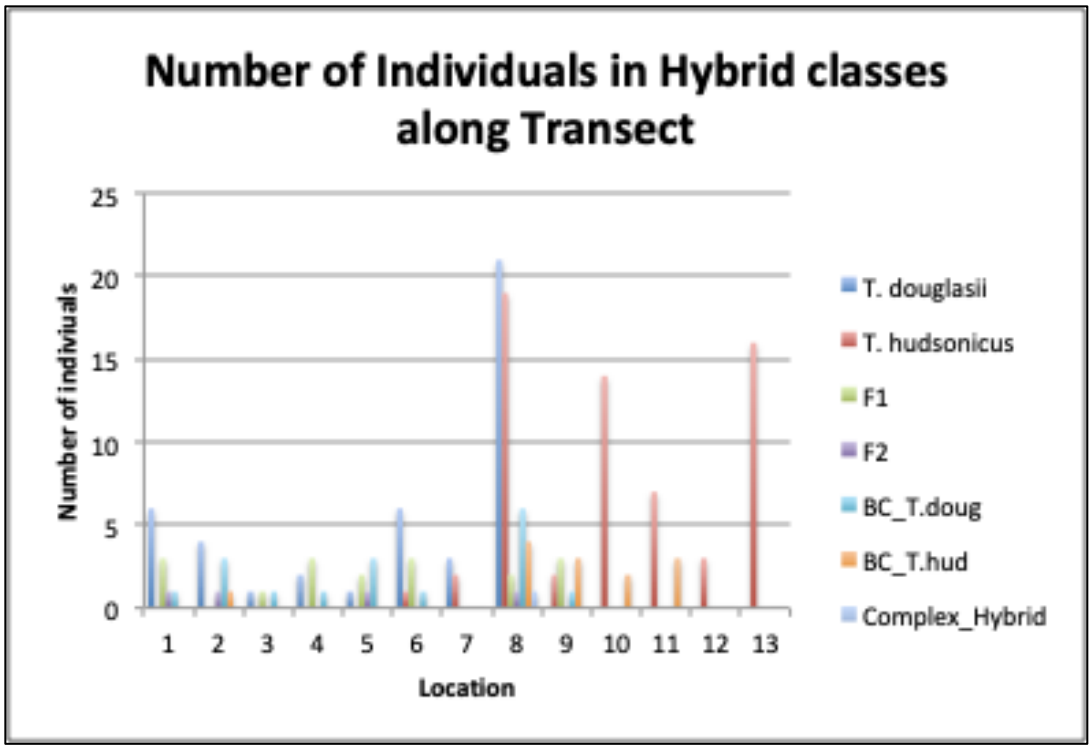

Figure 3: NEWHYBRIDS results classify genotype frequencies into classes (*T.douglasii* parental, *T. hudsonicus* parental, F1, F2, backcross to *T.douglasii* parental and backcross to *T. hudsonicus* parental). More introgression and a higher number of hybrid classes are seen in populations 1 through 9. While populations 10-13 remain mostly pure *T.hudsonicus* parental.

Table 1: STRUCTURE results. Avg Q. *douglasii*/Avg Q. *hudsconicus* represents the proportion of individuals from each species at each population. Variance is the variance of Q values among individuals in the population.

Locality	Transect Distance	samples	Avg.Q. *douglasii*	Avg.Q. *hudsonicus*	Variance
near Baker Lk.	0	11	0.83	0.17	0.0437271
near Rockport	13.6	9	0.79	0.21	0.0623484
Newhalem	44.1	3	0.8	0.2	0.061697
HWY 20	59.5	6	0.7	0.3	0.1077854
Hwy 20, 12.6mi W of Rainy Pass	69.2	7	0.66	0.34	0.0299486
Rainy Pass	75.12	11	0.74	0.26	0.1571753
Easy Pass	75.37	5	0.61	0.39	0.1558687
Bridge Creek Trailhead	75.46	54	0.54	0.46	0.1881784
Hwy 20 near Washington Pass	76.6	9	0.31	0.69	0.504032
Mazama	92.4	16	0.03	0.97	0.002072
Methow Valley	107.8	10	0.04	0.96	0.0017507
Winthrop	114.2	3	0.02	0.98	0.0001773
Twisp	130	16	0.01	0.99	0.0003858

Table 2: HZAR None fitted trait model and geographic cline. *Center* represents estimated center of the hybrid zone *of T.douglasii* and *T. hudsoncius* . *2-log High/Low Center value* represents the maximum and minimum log likelihood for hybrid zone center. *Width* represents estimated center of the hybrid zone of *T.douglasii* and *T. hudsoncius*. *2-log High/Low Width value* represents the maximum and minimum log likelihood for hybrid zone width.

Geographic Trait Model	Center	2- log High Center Value	2-log Low Center Value	Width	2-log High Width Value	2-log Low Width Value
None-fitted model	76.50303	79.61808	76.10919	3.150201	10.55019	1.468564

Table 3: NEWHYBRID results show the number of individuals in each hybrid class at the populations along the transect.

Local	Transect Distance	samples	T doug	T. hud	F1	F2	BC_T. doug.	BC_T .hud	Complex _Hyrbid
near Baker Lk.	0	11	6	0	3	1	1	0	0
near Rockport	13.6	9	4	0	0	1	3	1	0
Newhalem	44.1	3	1	0	1	0	1	0	0
HWY 20	59.5	6	2	0	3	0	1	0	0
Hwy 20, 12.6mi W of Rainy	69.2	7	1	0	2	1	3	0	0

Pass									
Rainy Pass	75.12	11	6	1	3	0	1	0	0
Easy Pass	75.37	5	3	2	0	0	0	0	0
Bridge Creek Trailhead	75.46	54	21	19	2	1	6	4	1
Hwy 20 near Washington Pass	76.6	9	0	2	3	0	1	3	0
Mazama	92.4	16	0	14	0	0	0	2	0
Methow Valley	107.8	10	0	7	0	0	0	3	0
Winthrop	114.2	3	0	3	0	0	0	0	0

Twisp	130	16	0	16	0	0	0	0	0