**Long-distance dispersal, ice sheet dynamics, and mountaintop isolation underlie the genetic structure of glacier ice worms**

Scott Hotaling ([scott.hotaling@wsu.edu](mailto:scott.hotaling@wsu.edu)), Daniel H. Shain, Shirley A. Lang, Robin K. Bagley, Lusha M. Tronstad, David W. Weisrock, and Joanna L. Kelley

*Proceedings of the Royal Society B (DOI: 10.1098/rpsb.2019.0983)*

**Supplementary Materials:**

**Methods:**

*Sample collection, library preparation, and SNP calling*

During the summer of 2009, ice worms were collected from nine glaciers across most of their geographicrange (Figs. 1, S1; Table 1). Samples were stored in > 80% EtOH until DNA was extracted from 59 ice worms using a Qiagen DNEasy Blood and Tissue Kit. Double-digest restriction-site associated DNA (ddRAD) sequencing libraries were prepared following Peterson *et al.* [1] with restriction enzymes EcoRI and NlaIII. During library preparation, samples were divided into two groups and each sample was assigned a unique, variable-length barcode [2] which was incorporated during adapter ligation. Size selection for a 350 bp ± 35 bp window was performed with a Pippen Prep (Sage Science), and both sample groups were subsequently amplified using PCR primers containing a group-specific barcode. The 59-sample library was sequenced on one lane of an Illumina HiSeq4000 at the University of Illinois High-Throughput Sequencing and Genotyping Unit with single-end, 100 bp chemistry.

Raw reads were demultiplexed, quality-filtered, and ddRAD loci were assembled *de novo* using the *process\_radtags* and *denovo\_map* functions of the Stacks v1.46 pipeline [3]. We allowed a maximum distance between stacks of 2 and a minimum read depth of 10. Next, we applied a stringent filtering scheme to identify high-confidence SNPs that were shared among many individuals. We only included SNPs if they were present in ≥ 5 populations, genotyped in ≥ 50% of individuals per population, and were in Hardy-Weinberg equilibrium with a minor allele frequency of ≥ 0.025 overall. We further restricted analyses to one random SNP per locus for all analyses except fineRADstructure (see below). All post-Stacks filtering steps were performed in PLINK v1.07 [4] and the commands used in this study are provided on GitHub (<https://github.com/scotthotaling/ice_worm_ddRAD>).

*Population genetic and phylogenetic analyses*

Population structure was inferred in two ways: a maximum likelihood-based method using ADMIXTURE 1.3.0 [5] and a discriminant analysis of principal components (DAPC) with the R package *adegenet* [6]. ADMIXTURE analyses were performed with default settings, a range of clusters (*K*) from 1-12, and 25 replicates per *K* with the current time as the random seed. The cross-validation (CV) error for each *K* was plotted to identify the best-fit *K* (minimized CV across the mean of all replicates for each *K*). After identifying the best-fit *K*, we considered the replicate that minimized CV across all 25 replicates for all *K*’sto be the best-fit solution overall. However, because all runs did not converge on the same result, we also inspected best-fit solutions for other replicates of *K* = 7 (the best-fit *K* overall)to clarify the distribution of best-fit solutions. For DAPC, we first used the *find.clusters* function to identify the optimal *K* [i.e., the *K* with the lowest Bayesian Information Criterion (BIC)]*.* Next, to avoid over-fitting of the model, we retained the appropriate number of principal components (PCs) according to the α-score [PCs retained = 6, Fig. S2]. We performed a final DAPC analysis using the best-fit *K* and optimal number of PCs identified in the previous two steps.

We extended our population structure analyses to infer both shared ancestry and phylogenetic relationships in two ways: a nearest neighbor haplotype approach to infer coancestry with fineRADstructure [7] and phylogenetic relationships inferred from singular value decomposition estimates for quartets of tips using SVDQuartets [8] as implemented in PAUP\* v4.0a159 [9]. For fineRADstructure, we used 100,000 burn-in iterations followed by 100,000 iterations sampled every 1,000 steps for the Markov chain Monte Carlo clustering algorithm. Next, we used 10,000 iterations of the tree-building algorithm to assess genetic relationships among clusters. Since fineRADstructure is a haplotype-based approach, analyses were performed using all variable sites for a given ddRAD locus (i.e., a haplotype) rather than randomly selected single SNPs per locus. For SVDQuartets, we performed exhaustive sampling of all possible quartets (every combination of four tips) and branch support was estimated with 100 nonparametric bootstrap replicates.

*Demographic modeling*

To explicitly assess the demographic history and timing of divergence for the major groups identified in our population genetic and phylogenetic analyses (I and II; see Results), we performed demographic modeling in fastsimcoal2 v2.603 [10], a coalescent-based program which estimates demography from the site frequency spectrum (SFS). We developed four two-lineage models (Fig. 2A) which included no gene flow (M1), unidirectional gene flow from group I into II (M2), unidirectional gene flow from group II into group I (M3), and bidirectional gene flow (M4). Model definition files (\*.est and \*.tpl) are provided in Appendix A. To ensure that our models accurately reflected the evolutionary scenario we sought to model, we visualized models in R with the script ParFileInterpreter-v6.3.1.r which is provided with the fastsimcoal2 documentation. Our demographic analyses included an initial set of model selection runs, comparisons of maximum observed and expected likelihoods to select the best-fit model, then parameter estimation for the best-fit model through parametric bootstrapping.

Because demographic inference from the SFS is particularly dependent on the presence of rare alleles and can be biased by missing data [11], we maximized the number of shared SNPs between our focal groups by selecting four individuals from the same population in each group (group I = Davidson; group II = Treaty). We selected Davidson and Treaty because they are the most geographically proximate populations that belong to groups I and II and both populations were robustly sampled for this study. For each population, we selected the four individuals with the least missing data across the same post-filtering data set used in other analyses. Only loci with no missing data were retained for demographic analyses. This yielded 2,714 SNPs across our eight focal individuals. We converted PLINK-formatted allele counts (as output from Stacks) into folded SFS with a modified version of the fs\_from\_data.py script included in δaδi [11]. After constructing observed SFS files for variable sites, we adjusted the monomorphic counts for all \*.obs files to the total number of monomorphic sites for our RAD loci that were 94 bp long (249,064).

For each model, we performed 50 replicate runs with 100,000 simulations and 100 cycles of a conditional maximization algorithm per run. We specified the nuclear mutation rate at 3.5E-9 per site per generation which was previously estimated for *Drosophila melanogaster* [12]. To identify the best-fit model, the maximum expected likelihood (MEL) was compared to the maximum observed likelihood (MOL) for each model replicate. The best-fit run minimized the difference between MEL and MOL for each model’s set of fifty replicates. Using these best-fit runs, we calculated an Akaike Information Criterion (AIC) score for each model using the formula: AIC = [(2*k*) ­– (2 x *ln*(10) x MOL)], where *k* is the number of parameters in the model. We identified the best-fit model as the one with the lowest AIC. We calculated the difference between each model AIC and the best-fit model to rank models according to ΔAIC. Use of the AIC allowed for model comparison despite varying numbers of parameters.

To generate 95% confidence intervals (CIs) of parameter estimates for our best-fit model, we used a combination of the best-fit run of the initial model runs and parametric bootstrapping. We first simulated 50 replicates of the SFS from the \*.maxL.par file (i.e., the parameter estimates that produced the maximum likelihood) for the best-fit run (minimized difference between MEL and MOL) of the best-fit model (minimized AIC). Next, we performed the same 50 replicate analyses described above for each of the 50 newly simulated SFS files. Finally, we calculated mean parameter estimates and 95% CIs from the 50 best-fit bootstrapping replicates (i.e., the runs with least difference between MEL and MOL for each of the 50 simulated SFSs).

**Discussion:**

*Using previous studies to inform the effect of putative ice ridge formation on ice worm evolution*

Our study is the fourth to investigate ice worm biogeography and population genetic structure using molecular data. The first three [4-6] leveraged either 28S or COI data to draw conclusions. In total, 37 ice worm populations (perhaps a few more depending on how disjunct glaciers are grouped) have been sampled for population genetic analyses (Figure S4). Of those, only ~7-10 are within ~100 km of the putative ice ridge and the bulk (26/37 = ~70%) are far to the northwest or southeast of the key area and well within the area of Groups I and II (or previously, the “Northern” and “Southern” clades). Thus, we are unable to further assess our population genomic conclusions with regards to the effects of the putative ice ridge on ice worm structure in light of past genetic data.

**Supplementary Tables:**

**Table S1.** Results of demographic model testing.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Description | *k* | AIC | ΔAIC | Model choice |
| M1 | No gene flow | 4 | 39,278.96 | -- | 1 |
| M2 | Unidirectional gene flow; I → II | 5 | 39,288.06 | 9.10 | 1.E-02 |
| M3 | Unidirectional gene flow; I ← II | 5 | 39,397.36 | 118.40 | 2.E-26 |
| M4 | Bidirectional gene flow; I ↔ II | 6 | 39,397.59 | 118.63 | 2.E-26 |

**Supplementary Figures:**

****

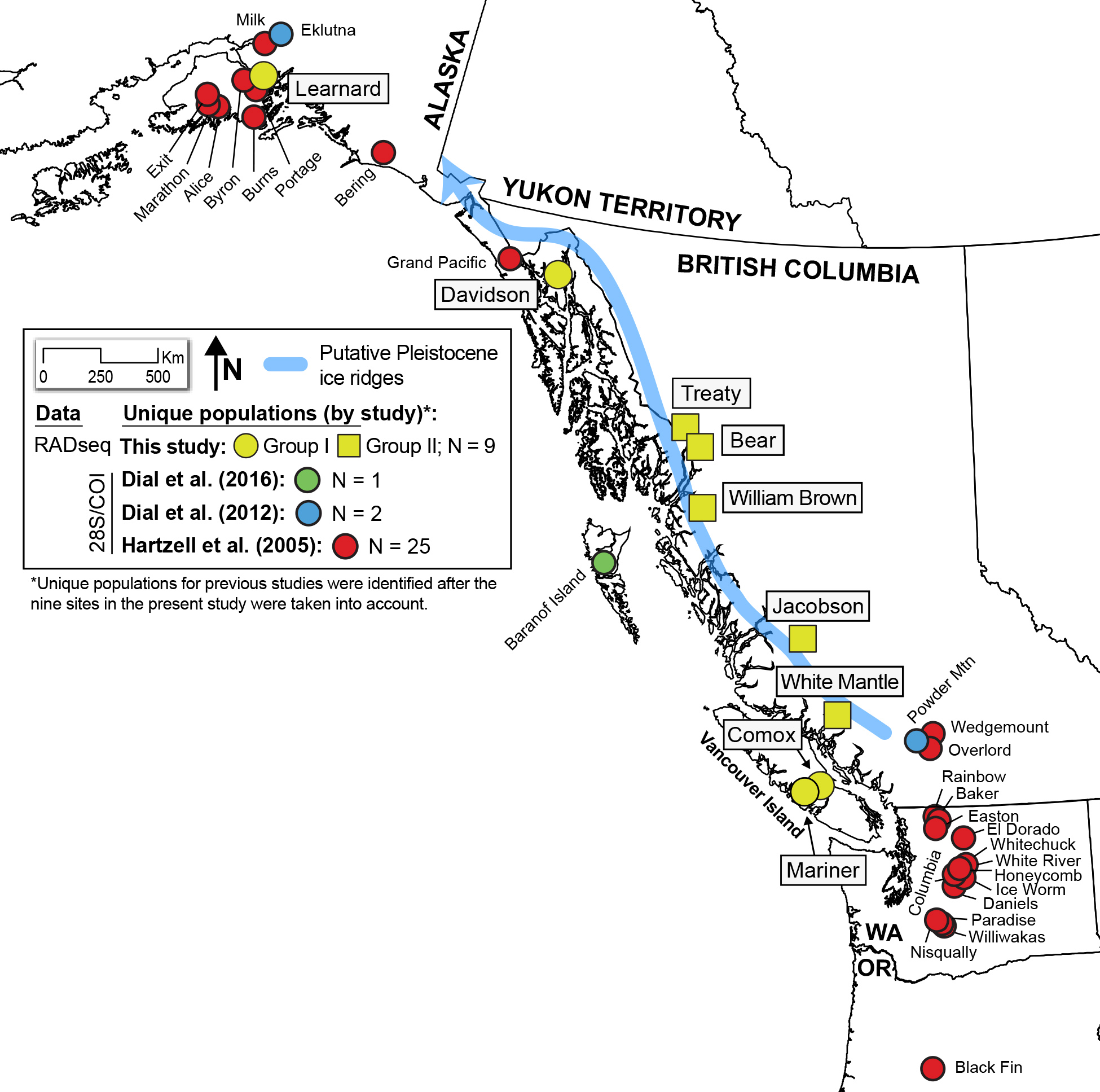
**Figure S1.** The known distribution of ice worms (*Mesenchytraeus solifugus*). Presence and absence information stems from our own surveys, personal communication with Roman Dial, and previous published manuscripts [13-15].

****

**Figure S2.** An α-score plot for identifying the optimal number of principal components (PCs) to retain in DAPC analyses. The optimal number of PCs to retain (6) is highlighted by a red circle.

****

**Figure S3.** Mean coverage and heterozygosity (variant positions only) for all ice worms included in this study. Red bars highlight a single individual (MS5) which bore a signature of substantial admixture between southern Alaska (Learnard Glacier population) and Vancouver Island (Mariner Glacier population).

****

**Figure S4.** All of the sites where population genetic data has been collected for ice worms (*Mesenchytraeus solifugus*). These data include sites from the present study and three previous studies [13-15].

**References:**

1. 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), e37135. (doi:10.1371/journal.pone.0037135).

2. Burford Reiskind M., Coyle K., Daniels H., Labadie P., Reiskind M., Roberts N., Roberts R., Schaff J., Vargo E. 2016 Development of a universal double‐digest RAD sequencing approach for a group of nonmodel, ecologically and economically important insect and fish taxa. *Molecular ecology resources* **16**(6), 1303-1314.

3. Catchen J., Hohenlohe P.A., Bassham S., Amores A., Cresko W.A. 2013 Stacks: an analysis tool set for population genomics. *Molecular ecology* **22**(11), 3124-3140. (doi:10.1111/mec.12354).

4. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., De Bakker P.I., Daly M.J.J.T.A.j.o.h.g. 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. **81**(3), 559-575.

5. Alexander D.H., Novembre J., Lange K. 2009 Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**(9), 1655-1664. (doi:10.1101/gr.094052.109).

6. Jombart T., Devillard S., Balloux F. 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics* **11**(1), 94.

7. Malinsky M., Trucchi E., Lawson D.J., Falush D. 2018 RADpainter and fineRADstructure: population inference from RADseq data. *Molecular biology and evolution* **35**(5), 1284-1290.

8. Chifman J., Kubatko L. 2014 Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30**(23), 3317-3324.

9. Swofford D.L. 1998 Phylogenetic analysis using parsimony.

10. Excoffier L., Dupanloup I., Huerta-Sánchez E., Sousa V.C., Foll M.J.P.g. 2013 Robust demographic inference from genomic and SNP data. **9**(10), e1003905.

11. Gutenkunst R.N., Hernandez R.D., Williamson S.H., Bustamante C.D.J.P.g. 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. **5**(10), e1000695.

12. Keightley P.D., Trivedi U., Thomson M., Oliver F., Kumar S., Blaxter M.L.J.G.r. 2009 Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. **19**(7), 1195-1201.

13. Dial R.J., Becker M., Hope A.G., Dial C.R., Thomas J., Slobodenko K.A., Golden T.S., Shain D.H. 2016 The role of temperature in the distribution of the glacier ice worm, Mesenchytraeus solifugus (Annelida: Oligochaeta: Enchytraeidae). *Arctic, Antarctic, and Alpine Research* **48**(1), 199-211.

14. Dial C.R., Dial R.J., Saunders R., Lang S.A., Lee B., Wimberger P., Dinapoli M.S., Egiazarov A.S., Gipple S.L., Maghirang M.R., et al. 2012 Historical biogeography of the North American glacier ice worm, Mesenchytraeus solifugus (Annelida: Oligochaeta: Enchytraeidae). *Mol Phylogenet Evol* **63**(3), 577-584.

15. Hartzell P.L., Nghiem J.V., Richio K.J., Shain D.H. 2005 Distribution and phylogeny of glacier ice worms (Mesenchytraeus solifugus and Mesenchytraeus solifugus rainierensis). *Canadian Journal of Zoology* **83**(9), 1206-1213.

**Appendix A:** fastsimcoal2 model definition files

**M1. est**

// Priors and rules file

// \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[PARAMETERS]

//#isInt? #name #dist.#min #max

//all Ns are in number of haploid individuals

1 ANCSIZE unif 10 1e6 output

1 NPOP1 unif 10 1e6 output

1 NPOP2 unif 10 1e6 output

1 TDIV unif 10 1e7 output

[RULES]

[COMPLEX PARAMETERS]

0 RESIZE = ANCSIZE/NPOP1 hide

**M1.tpl**

//Number of population samples (demes)

2 populations to simulate

//Population effective sizes (number of genes)

NPOP1

NPOP2

//Sample sizes

8

8

//Growth rates

0

0

//Number of migration matrices : 0 implies no migration between demes

2

//Migration matrix 0

0 0

0 0

//Migration matrix 1

0 0

0 0

//historical event: time, source, sink, migrants, new deme size, growth rate, migr mat index

1 historical event

TDIV 1 0 1 RESIZE 0 1

//Number of independent loci [chromosomes]

1 0

//Per chromosome: Number of linkage blocks

1

//per block: Datatype, numm loci, rec rate and mut rate + optional parameters

FREQ 1 0 2.8e-9 OUTEXP

**M2. est**

// Priors and rules file

// \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[PARAMETERS]

//#isInt? #name #dist.#min #max

//all Ns are in number of haploid individuals

1 ANCSIZE unif 10 1e6 output

1 NPOP1 unif 10 1e6 output

1 NPOP2 unif 10 1e6 output

0 N1M21 logunif 1e-2 20 hide

1 TDIV unif 10 1e7 output

[RULES]

[COMPLEX PARAMETERS]

0 RESIZE = ANCSIZE/NPOP1 hide

0 MIG21 = N1M21/NPOP1 output

**M2.tpl**

//Number of population samples (demes)

2 populations to simulate

//Population effective sizes (number of genes)

NPOP1

NPOP2

//Sample sizes

8

8

//Growth rates

0

0

//Number of migration matrices : 0 implies no migration between demes

2

//Migration matrix 0

0 MIG21

0 0

//Migration matrix 1

0 0

0 0

//historical event: time, source, sink, migrants, new deme size, growth rate, migr mat index

1 historical event

TDIV 1 0 1 RESIZE 0 1

//Number of independent loci [chromosomes]

1 0

//Per chromosome: Number of linkage blocks

1

//per block: Datatype, numm loci, rec rate and mut rate + optional parameters

FREQ 1 0 2.8e-9 OUTEXP

**M3. est**

// Priors and rules file

// \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[PARAMETERS]

//#isInt? #name #dist.#min #max

//all Ns are in number of haploid individuals

1 ANCSIZE unif 10 1e6 output

1 NPOP1 unif 10 1e6 output

1 NPOP2 unif 10 1e6 output

0 N2M12 logunif 1e-2 20 hide

1 TDIV unif 10 1e7 output

[RULES]

[COMPLEX PARAMETERS]

0 RESIZE = ANCSIZE/NPOP1 hide

0 MIG12 = N2M12/NPOP2 output

**M3.tpl**

//Number of population samples (demes)

2 populations to simulate

//Population effective sizes (number of genes)

NPOP1

NPOP2

//Sample sizes

8

8

//Growth rates

0

0

//Number of migration matrices : 0 implies no migration between demes

2

//Migration matrix 0

0 0

MIG12 0

//Migration matrix 1

0 0

0 0

//historical event: time, source, sink, migrants, new deme size, growth rate, migr mat index

1 historical event

TDIV 1 0 1 RESIZE 0 1

//Number of independent loci [chromosomes]

1 0

//Per chromosome: Number of linkage blocks

1

//per block: Datatype, numm loci, rec rate and mut rate + optional parameters

FREQ 1 0 2.8e-9 OUTEXP

**M4. est**

// Priors and rules file

// \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[PARAMETERS]

//#isInt? #name #dist.#min #max

//all Ns are in number of haploid individuals

1 ANCSIZE unif 10 1e6 output

1 NPOP1 unif 10 1e6 output

1 NPOP2 unif 10 1e6 output

0 N1M21 logunif 1e-2 20 hide

0 N2M12 logunif 1e-2 20 hide

1 TDIV unif 10 1e7 output

[RULES]

[COMPLEX PARAMETERS]

0 RESIZE = ANCSIZE/NPOP1 hide

0 MIG21 = N1M21/NPOP1 output

0 MIG12 = N2M12/NPOP2 output

**M4.tpl**

//Number of population samples (demes)

2 populations to simulate

//Population effective sizes (number of genes)

NPOP1

NPOP2

//Sample sizes

8

8

//Growth rates

0

0

//Number of migration matrices : 0 implies no migration between demes

2

//Migration matrix 0

0 MIG21

MIG12 0

//Migration matrix 1

0 0

0 0

//historical event: time, source, sink, migrants, new deme size, growth rate, migr mat index

1 historical event

TDIV 1 0 1 RESIZE 0 1

//Number of independent loci [chromosomes]

1 0

//Per chromosome: Number of linkage blocks

1

//per block: Datatype, numm loci, rec rate and mut rate + optional parameters

FREQ 1 0 2.8e-9 OUTEXP