Demographic history, linked selection, and recombination shape the genomic landscape of a broadly distributed Pacific salmon.

Quentin Rougemont[1*], Jean-Sébastien Moore[1], Thibault Leroy[2], Eric Normandeau[1], Eric B. Rondeau[3,4], Ruth E. Withler[5], Donald M. Van Doornik[6], Penelope A. Crane[7], Kerry A. Naish[8], John Carlos Garza[9] Terry D. Beacham[5], Ben F. Koop[3,4], Louis Bernatchez[1]

[1] Département de Biologie, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, Québec, Canada QC G1V OA6

[2] ISEM, Univ. Montpellier. CNRS, EPHE, IRD, Montpellier, France

[3] Centre for Biomedical Research, University of Victoria, Victoria, BC, Canada V8P 5C2

[4] Department of Biology, University of Victoria, Victoria, BC, Canada V8P 5C2

[5] Department of Fisheries and Ocean, Pacific Biological Station, 3190 Hammond Bay Road, Nanaimo, British Columbia, Canada V9R 5K6

[6] National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Northwest Fisheries Science Center, Manchester Research Station, 7305 Beach Drive East, Port Orchard, Washington 98366, USA

[7] Conservation Genetics Laboratory, U.S. Fish and Wildlife Service, 1011 E. Tudor Road, Anchorage, Alaska 99503

[8] School of Aquatic and Fishery Sciences, University of Washington, 1122 NE Boat Street, Box 355020, Seattle, WA 98195 5020, USA

[9] Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service and Institute of Marine Sciences, University of California–Santa Cruz, 110 McAllister Way, Santa Cruz, California 95060

*Corresponding author e-mail: quentinrougemont@orange.fr (QR)

Abstract

Understanding the impacts of current human activities on within-species genetic variation requires a thorough description of the historical factors that have shaped the genomic and geographical distribution of nucleotide diversity. Past and current conditions influencing effective population size have important evolutionary implications for the efficacy of selection, increased accumulation of deleterious mutations, and loss of adaptive potential under the nearly neutral theory. Here, we gather extensive genome-wide data that represent the extant diversity of the Coho salmon (*Oncoryhnchus kisutch*) to address three issues. First, we demonstrate that a single glacial refugium is the source of the majority of present-day genetic diversity, with minor but detectable inputs from secondary micro-refugia. We propose a scenario whereby several ancestral populations located south of the ice sheets expanded in postglacial time, swamping out most of the diversity from other putative micro-refugia. Following this expansion, we identify particular populations having undergone continuous declines in population size (*Ne*). Second, we combine multiple evidence from demographic modelling, analysis of recombination landscape, and genome-wide landscape of diversity to demonstrate that selection at linked sites and Hill-Robertson interference played a major role in shaping genetic diversity across the Coho salmon genome. Third, we demonstrate that this demographic history generated subtle differences in the load of deleterious mutations among populations, a finding that mirrors recent results from human populations. Taken together, we found considerable support for the joint contributions of demographic history and linked selection in the load of deleterious mutations. We suggest that these inferences should be better integrated in conservation genetics of managed fish species which currently focuses largely on within-population adaptation.

Author Summary

Reconstruction of a species' past demographic history from genome-wide data allows understanding how historical factors interact with intrinsic genomic properties to shape the distribution of genetic diversity along its genome and its geographic range. Here, we combine genotyping-by-sequencing and whole genome sequence data with demographic modelling to address these issues in the Coho salmon, a Pacific salmon species with rapidly declining census size in some parts of its range, notably in the south. Our demographic reconstructions indicate a linear decrease in genetic diversity towards the north of the species range, supporting the hypothesis of a major southern refugia for the Coho salmon and a northern route of postglacial recolonization. Accordingly, the number of candidate deleterious homozygous derived mutations was higher in northern populations. Demographic modelling also suggested the existence of cryptic refugia that may have been missed with the use of simpler summary statistics. We further showed that the species' genome was shaped by linked selection and biased gene conversion. In particular, local variation in recombination rates have modulated the efficacy of natural selection. These processes, together with a complex demographic history, can contribute to the load of deleterious mutations – an effect we argue should be taken into account more routinely in conservation genetics studies.

Introduction

Both plant and animal biodiversity are currently declining or disappearing at unprecedented rates due to human activity [1]. This leads to population size reduction, reduced genetic diversity, and to the sixth mass extinction [2]. Before humans became major drivers of changes in species distributions and abundance, long-term climate change had a major influence [3]. The Pleistocene glaciations resulted in major contractions in the geographical distributions of many species into refugia that persisted in unglaciated areas [4]. Postglacial range expansions often led to contacts between ancestral populations previously segregated in different refugia [4,5]. The effects of long-term climate change combined with recent human-induced population declines can foster genetic changes including a loss of genetic diversity, increased inbreeding, increased load of deleterious mutations, and a loss of local adaptation [6].

In this context, it becomes important to understand how demographic history interacts with past and ongoing selection and recombination to shape genetic variation. By disentangling past and current drivers of range-wide genomic diversity, this information can inform management and conservation decisions [7]. Beyond conservation implications, such context provides a unique opportunity to address outstanding questions in evolutionary biology. In particular, what is the role of gene flow in shaping heterogeneous differentiation landscape during population divergence [8,9]? What are the demographic conditions required to generate substantial differences in deleterious load among populations [10]?

A major challenge to understanding drivers of genome-wide patterns of diversity is that different demographic processes can lead to similar contemporary genomic footprints [11]. As populations diverge, they accumulate genetic incompatibilities forming barriers to gene flow [12], while the rest of the genome may continue to be freely exchanged. As a consequence, the genomic landscape of divergence is expected to vary, with greater differences between populations at genomic barriers as compared to genomic regions exhibiting ongoing gene flow. However, similar patterns of heterogeneous genome-wide divergence can be due to genetic hitchhiking of neutral alleles linked to selective sweeps [13] or to background selection (BGS; [14]). These combined effects, refered to as linked selection reduces polymorphism at sites closely linked to advantageous or deleterious variants, and therefore reduces local effective population size (*Ne*) along the genome. The intensity of selection on linked loci will be mostly modulated by variation in local recombination rate and by gene density [15,16]. Under linked selection, diversity (π, D_{XY}) and differentiation (F_{ST}) metrics are expected to be positively and negatively correlated with genome-wide variation in recombination rate (ρ) respectively, and with the density of targets (e.g., genes, regulatory regions) that are subject to selection [14–18]. It is now increasingly recognized that neglecting BGS can bias demographic inferences [19,20] or lead to false adaptive interpretations [21].

An understanding of historical demography is also essential for a sound interpretation of patterns of deleterious mutation load observed among contemporary populations [10,22]. Population bottlenecks are predicted to reduce potential for local adaptation, but also to reduce standing genetic variation and the efficacy of selection [23,24]. In turn, a reduced efficacy of purifying selection leads to an increase in the number of deleterious variants segregating in a population. Moreover, intrinsic genome properties, in particular local variation in recombination rate or background selection, can favour the accumulation of

deleterious mutations [25]. From a conservation standpoint, populations harboring an elevated number of deleterious variants might need to be monitored more closely.

Combining population genomics data with demographic modelling represents a powerful strategy to test alternative hypotheses about historical drivers of existing genomic diversity. Previous studies employing a similar approach have focussed mostly on species with a narrow geographic range, such as small islands [26,27], which are on the verge of extinction [28–31] are strongly bottlenecked [32], sheddig light on the evolutionary consequences of small population size. Few studies, however, have investigated how historical processes have shaped the geographical patterns in the distribution of genomic diversity in more broadly distributed species, e.g., at the scale of a whole continent. An exception to this observation is the vast literature on demographic reconstructions of human populations. Long-lasting debates in this literature regarding the role of demography in generating mutation load differences among populations [22,33,34] could benefit from studies of species displaying similarly complex demographic histories and broad geographic distributions.

Salmonid fishes are economically important species that have suffered recent demographic declines [35,36]. This is particularly the case for Coho salmon (*Oncorynchus kisutch*), one of the five anadromous species of Pacific salmon that supports important recreational and indigenous subsistence fisheries, which has suffered dramatic population declines (> 90%) over the last three decades in parts of its range [36,37]. A previous study investigated the range-wide population structure and demographic history of the species and found a cline of decreasing diversity from south to north, as well as some endemic diversity in small putative refugia [38] (see also [39]). This study indicated that Coho salmon may have survived the last glacial maximum (LGM, i.e. the Fraser Glaciation in British Columbia, and the McConnell/McCauley Glaciation in Yukon and Alaska; 23 to 18 Ky ago) in unglaciated areas of Haida Gwaii and Beringia in addition to areas south of the ice sheets. This study, however, predates the genomic era and could not eliminate alternative hypotheses regarding the origin and number of glacial refugia during the LGM. Most importantly, the impacts of confounding factors such as background selection, recombination rate variation, and how these factors may facilitate the accumulation of deleterious mutations could not be studied with the limited number of genetic markers available at the time. In North America, the species is currently distributed from California to Alaska [40]. Unglaciated areas that could potentially serve as glacial refugia persisted both north (e.g. the Beringian refugium in Alaska, the Yukon Territory of Canada and areas of Asia and the Bering Land Bridge) and south (e.g. all of the deglaciated area south of British Columbia, Canada) of the ice sheet [40–42]. Other unglaciated areas (e.g. Haida Gwaii in British Columbia) could also have been micro-refugia [43,44]. In this context, distinct demographic scenarios can be tested. Under a first scenario whereby populations expanded north from a single southern refugium, we predict: *i*) a latitudinal decrease in genetic diversity from south to north along with a pattern of IBD, and *ii*) ancestral populations located in areas south of the ice sheets. Under a second scenario, populations expanded south from a single northern refugium, and we predict the opposite geographic pattern. The third scenario corresponds to the survival of populations in different refugia where we predict: *i)* the existence of clearly distinct genetic clusters, and *ii)* postglacial gene

flow with signatures of secondary contacts, with contact zones displaying higher genetic diversity through postglacial admixture between different genomic backgrounds.

In order to test these alternative scenarios, we generated genome-wide data from nearly 2,000 Coho salmon from California to Alaska, one of the most extensive genomic datasets for a non-model vertebrate species to date. First, to resolve the species demographic history, we used a modelling approach that accounts both for barriers to gene flow affecting migration locally, and for linked selection affecting the rate of drift. Next, we tested the above predictions related to linked selection. Finally, we hypothesized that demographic history and background selection shaped the pattern of deleterious mutation load, both within and among populations. In particular, we hypothesized that postglacial re-colonisation influenced levels of standing genetic variation and favoured the accumulation of deleterious mutations at the expansion front. In these conditions, we predicted that genetic diversity should decrease as a function of the distance from the ancestral populations, while the accumulation of putatively deleterious mutations should increase as a function of the distance to the ancestral populations.

Results

Overall genetic diversity and population structure

A total of 1,957 individuals was sampled from California to Alaska representing 58 sampling locations (mean n = 34 fish per location, Fig 1a, S1 Table) and genotyped using a genotype by sequencing (GBS) method that generated 82,772 high quality filtered single nucleotide polymorphisms (SNPs). Another set of 55 individuals representing 11 sampling locations from the same range (Fig 1a), were whole genome sequenced (WGS) to ~30X coverage, and used in specific analyses (S2 Table).

Levels of genetic diversity (observed and exepected heterozygosity, π_{SNP}) were highest in formerly deglaciated areas in the south (California, Cascadia, Fig 2a, Fig 1b) and decreased as a function of distance from the southernmost site up to Alaska ($r = 0.64$, $p < 0.0001$, Fig 2a, S1 Fig). The Thompson River watershed (Thompson R. hereafter) in southern British Columbia was an exception to this latitudinal pattern and displayed the lowest average level of regional genetic diversity of all sampling locations which we hypothesized to results from bottlenecks in this area. The remaining samples from British Columbia were intermediate in genetic diversity.

The distribution of singletons provided further information regarding the most ancestral populations, with older populations expected to have accumulated a higher density of singletons [45]. Counting the number of singletons by sampling site and averaging by regional groups revealed the following differences: Cascadian samples contained the highest number of singletons, with a mean of 1,263 singletons per site. Californian samples had the fewest number of singletons ($n_{MEAN} = 55$) while Alaska harbored intermediate density ($n_{MEAN} = 966$). Consistently, WGS data revealed 2.7 times more singletons in Cascadia (Tsoo-Yess River) as compared to Alaska (Kwethluk River), whereas Thompson R. samples contained 6.7 times less singletons, supporting the hypothesis of a pronounced bottleneck in the these populations (S3 Table). Similarly, the occurrence of private alleles was more prevalent among southern than northern populations, being nearly twice as high in Cascadian (n = 10,097) than in Alaskan populations (n = 5,270). Again,

populations from the Thompson R. were an exception to this pattern with the lowest level of private polymorphism (n = 1,479, S2 Fig).

Comparing the decay of linkage disequilibrium (LD) across populations based on the WGS data indicated rapid LD decay across all samples (S3 Fig). The Thompson R. populations again departed from the general pattern with a much reduced LD decay. Indeed, r^2 values decreased to 0.4 at approximately 342 kb for Thompson R. whereas this r^2 value was attained between 13 and 34 kb for all other populations.

Next we used the β_{ST} coefficient to identify ancestral population [46]. Unlike F_{ST} estimates [47], this index can account for the non-independence among populations and negative values are indicative of ancestral populations [46]. Here, β_{ST} indicated that ancestral populations were located in previously unglaciated areas corresponding to Cascadia (n= 5 localities), California (n = 3 localities) as well as one site from southern British Columbia (Fig 2b, S4 Table). A linear decrease in β_{ST} as a function of distance from the southernmost site was observed (r = 0.60, slope = 1.03e-04, p <0.0001) as expected under IBD. Support for this IBD pattern was also observed using F_{ST} (S4 Fig, r = 0.66, slope= 4e-05, p < 0.0001) as well as Mantel tests (r = 0.64; p < 0.0001, r = 0.72; p < 0.0001 when removing Thompson R. populations). Average pairwise F_{ST} across all populations was 0.095 and varied from 0.002 to 0.334 (S5 Fig), indicating moderate population structure, typical of anadromous species connected by gene flow [48].

Model-based analysis of population structure failed to reveal a clear number of distinct populations (K value). Instead, K values ranging from 30 to 60 all fit the data well (S6 Fig), due to the counfonding effect of IBD. The first axis of a Principal Components Analysis (PCA, Fig 3a) revealed a separation of the sample from South to North with the most divergent samples found in California. The second axis revealed the divergence from East to West but with a strong separation of the Thompson R. populations. Along these axes populations followed an IBD pattern. These results were also supported by an MDS analysis (S7 Fig). The third and fourth axis did not yield further information (S8 Fig).

We then used Treemix [49] to infer population splits and gene flow (Fig 3b). A first tree assuming no migration (i.e. drift only) explained 98.1% of the variance observed. Adding up to four significant migration events (p <0.0001) explained over 99.1% of variance, then the proportion of explained variance plateaued (S9 Fig). Populations from Cascadia occupied basal positions. California populations displayed pronounced genetic drift, corroborating the high divergence observed in the PCA. Populations from Alaska (MSL River) and Thompson R. also displayed higher genetic drift, in line with evidence based on analyses of genetic diversity and genetic structure. The two most supported migration events occurred from Cascadia south to California and north to the Thompson R. We note that populations followed the south to north arrangement, with the samples from Cascadia displaying less drift than those further north.

Demographic history

In order to assess more formally the occurrence of one or more refugial origins for contemporary populations, we performed the following explicit model-based inferences of population divergence scenarios using ∂a∂i [50]. Our models account for the confounding effects of selection at linked sites and that of the accumulation of local barriers to gene flow in the genome [19,51]. Four major demographic models were

statistically compared using groups identified in the PCA and with a focus on previously hypothesized refugia (i.e., Cascadia, California, Haida Gwaii, Alaska). The following models were tested: strict isolation (SI model), divergence with ancient migration (AM model), divergence with continuous gene flow (IM model) and secondary contact (SC model; S10 Fig). A total of 69 pairwise comparisons was performed with ∂a∂i. In each pairwise comparison one single representative population from a putative refugium was compared against one population from another refugium. Due to the high local structure within groups, we avoided pooling samples into higher order groups (e.g., regional groups) as this would unavoidably bias our results. Models incorporating linked selection and restricted introgression along the genome always received highest support (S5 Table). ΔAIC (minimum value > 5) confidently discriminated models in 87% of the ∂a∂i comparisons (S5 Table). The SC model received the highest support in 46% of the comparisons, the AM model in 30%, and the IM model in 14%, with 13% remaining unclassified. Scenarios including periods of gene flow clearly outperform scenarios assuming no gene flow. The fact that secondary contact was the best-supported model suggests that more than one glacial refugia contributed to the recolonization of the contemporary range occupied by Coho salmon.

Assuming a generation time of 3.5 years [52] and mutation rate of $8e^{-9}$ bp/generation revealed similar divergence time estimates among models (i.e., 57 Kya under SC, 40 Kya under IM, and 42 Kya under AM, S6 Table). The median time of secondary contact (SC) was 10 Kya [min = 4,800 – max = 32,900], corresponding roughly to the onset of the last glacial retreat (Fig 4A). Parameter estimates under the AM model supported a very recent reduction in gene flow (mean < 1Kya), which therefore represents a demographic model similar to an IM model. Pronounced variation in effective population size (*Ne)* was observed with the highest values in Cascadia populations (mean 12,800; range [3,500 – 16,500]). On average, the smallest *Ne* were observed in the Thompson R. populations (mean 3,000; range [1,300 – 8,300], Fig 4B), again consistent with a population bottleneck. However, smaller *Ne* values were also observed in the isolated Navfac Creek population in Alaska (*Ne* = 500) and the Scott Creek population in California (*Ne* = 1,800), which are the most divergent populations identified in the PCA. The McGarvey (California) and Clackamas (Cascadia) populations identified as “ancestral” according to the β_{ST} displayed the highest *Ne* (23,000 and 16,500). We also note that incorporating linked selection and barriers to gene flow further reduced *Ne*, regardless of the model (S11 Fig). Intrinsic barriers to gene flow reduced the estimated migration rate by half of its value and affected 20 to 40% of the genome (Fig 4C, S6 Table). Similarly, the Hill-Robertson factor suggested that *Ne* was reduced to 37%-50% of its initial estimate and that approximately half of the genome was affected (S6 Table).

We then investigated historical change in effective population size using the Sequentially Markovian Coalescent in SMC++ [54] and the 55 whole genome sequences, which are representative of 11 populations from California to Alaska (Fig 1, S2 Table). For ease of visualization, results for a subset of the populations are presented in Fig 4D and details for all populations are displayed in S12 Fig. This analysis revealed: i) an expansion of most populations approximately 12-20 KyA, concomitant with postglacial re-colonisation, ii) a slow and steady decline in the Thompson R. (Fig 4D, S12 Fig), and iii) a split time between all pairwise combinations of populations (median = 16,3 KyA, range = 6,7KyA - 61KyA, S13 Fig) compatible with the

onset of postglacial population expansion (Fig 4D) which was accompanied by an increase in *Ne* across samples (Fig 4D). We note that using a different mutation rate (1,25e-8 mutations/bp/generations) yielded estimates of split time that were more in line with the estimates of postglacial expansion (median = 9,6 KyA, S13 Fig) and with less variance among compared population (min = 5 KyA – max = 18 KyA). Overall, SMC++ results are consistent with previous analyses indicating that until recently, all populations shared a similar demographic history. We caution against a strict interpretation of the most recent time. Indeed, reliable estimates for present time require a high number of samples, whereas we only had five samples per populations [54]. Similarly, exact *Ne* estimates should be interpreted cautiously and only the trend should be considered rather than the exact values.

Linked selection shape the Coho salmon genomic landscape

We first described patterns of recombination and then tested correlations among genetic diversity (π), divergence (D_{XY}), and differentiation (F_{ST}) parameters measured across the whole genome using 500 kb sliding windows (Fig 5 A-C) and the population-scaled landscape of recombination (ρ) (Fig 5D) and gene density (Fig 5E) to test our predictions using whole genome sequences.

This analysis first revealed a heterogeneous recombination landscape that varied both within and among chromosomes. In particular, recombination rate was higher towards the ends of chromosomes (Fig 5D), as observed across many species [55] Second, we observed a negative correlation between recombination rate and chromosome length (R^2 = 0.481, p <0.0001, S14 Fig). To investigate the effects of selection at linked sites, we summarised π (across all 11 WGS populations), D_{XY} and F_{ST} statistics (55 pairwise comparisons), using the first axis of a PCA (PC1) to obtain the common variation in these metrics and reduce the dimensionality to single summary statistic (Fig 6, S15 Fig, S7 Table). PC1 captured 99% of variance in π and in D_{XY}, indicating that the PC1 was effective at summarizing information about diversity and divergence. All linear correlations among the tested variables were significant (p < 0.0001, S8A Table).
We found a positive correlation between PC1-⊓ and recombination (r = 0.54, S8A Table) and a negative correlation between PC1-⊓ and gene density (r = -0.20) (Fig 6A-D). We also found a positive correlation between PC1-D_{XY} and recombination (r = 0.55) and a negative correlation between PC1- D_{XY} and gene density (r = -0.21). PC1-F_{ST} was negatively, albeit weakly correlated with both recombination (r = -0.05) and gene density (r = -0.04, Fig 6B-E and C-F). We observed a correlation of 0.999 between π and D_{XY}, as expected here because of a very recently shared common ancestor (Fig 6G). We also found a negative relationship between PC1- π and PC1- F_{ST} (r = -0.44, Fig 6H) as well as between PC1- D_{XY} and PC1-F_{ST} (r = -0.43, Fig 6I). The smaller correlations between statistics involving F_{ST} can be explained by the modest variance in F_{ST} explained by PC1 (37%).

In order to investigate further the effects of linked selection, we used mixed linear models that integrate interactions among tested variables (i.e. recombination, gene density). These revealed a significant effect of both recombination (t = -41.17, p <0.0001, S8B Table) and gene density (t = 5.72, p<0.0001) on PC1-D_{XY} (R^2 = 0.348). The same was true when considering the correlation of PC1-π with recombination (R^2 = 0.55, Fig 6A-D, t = -40, p<0.0001) and gene density (t = 5.22, p<0.0001). Significant effects were also

found between PC1-F_{ST} and these variables (recombination: t = 6.91, p<0.0001, gene density t =6.27, p p<0.0001, interaction: t =-4.073, p<0.0001, R^2 = 0.025, S8B Table). Similar correlations were found when analysing populations separately (S9 Table). Therefore, our results suggest a role of linked selection in shaping the patterns of diversity along the Coho salmon genome.

We then tested whether the efficacy of selection also played a role in shaping patterns of diversity along the Coho salmon genome using the whole genome sequences. We measured ratios of non-synonymous to synonymous polymorphisms (π_N/π_S) as a measure of the efficacy of natural selection and GC content at the third codon position (GC3) as a proxy of local recombination rates. This metric was preferred as an indicator of localized recombination rate variation at the gene scale, as opposed to the large population scale estimates of recombination (ρ) over 250kb windows used above. Indeed, biased gene conversion (gBGC) occurs because base mismatch during homologous recombination is known to be biased toward G and C (conversion preferentially favours G+C over A+T bases) [56,57]. Here, we observed that the correlation between GC3 and ρ ranged from r= 0.3 to 0.72, depending on population. These modest correlations likely reflects the loss of information about fine scale recombination rate variation when using 250kb sliding windows, which also provides added justification to use the GC3 as a proxy. We found a strong correlation between π_N/π_S and GC3 for all populations (Fig 7, linear models all R^2 > 0.9, p < 0.0001, S10 Table). Next, we observed a negative correlation (linear model, R^2 = 0.28, p = 0.051) between historical N_e measured before the onset of population expansion ~13,000 years ago and πN/πS ratio (S16 Fig), in agreement with the nearly neutral theory of molecular evolution and indicating that our N_e estimates were good proxies of genetic diversity. Second, among-population differences in πN/πS across all genes were modest (mean = 0.252), with a minimum value of 0.248 observed in Tsoo-Yess R. (Cascadia) and Robertson Creek (BC) and maximum values of 0.257 and 0.256 observed in the Thompson R. and Inch Creek (BC) respectively. The π_N/π_S differences among populations were more contrasted in areas of low recombination of the genome (Fig 7) with lowest π_N/π_S values observed in the Kwethluk (Alaska, π_N/π_S =0.442), Quilcene (BC, π_N/π_S = 0.449), and Deschutes (Cascadia, π_N/π_S =0.451) river populations. The highest values were observed in the Thompson (π_N/π_S =0.551), Klamath (California; π_N/π_S =0.543) and Capilano (BC; π_N/π_S =0.495) river populations.

Spatial patterns of variation in deleterious mutations

Given the complex demographic history of population size changes and spatial expansions leading to secondary contacts in Coho salmon and the inference of linked selection, we predicted that deleterious mutations should segregate at higher frequencies at the edge of the norther expansion front in Alaska, but also in California likely representing a southern expansion from the Cascadian refugium, as well as in the bottlenecked Thompson R. populations.

To do so, we tested for an increase in the derived allele frequency (DAF), and homozygosity of predicted deleterious mutations at non-synonymous sites. We estimated derived alleles using whole genome information from three outgroups: 1) Chinook salmon (*Oncorhynchus tshawytscha)* [58], 2) Rainbow trout (*Oncorhynchus mykiss)* [59] and 3) Atlantic salmon (*Salmo salar*) [60]. Out of 4,427 non-synonymous mutations identified with the GBS dataset from all 58 populations, PROVEAN (63) predicted a total of 1,297

deleterious mutations in at least one of these populations and for which we were able to identify the derived allele. Deleterious mutations were maintained at lower DAF than synonymous variants (Wilcoxon-tests, all p <0.001). The DAF spectrum (Fig 8A) were computed for each major region (using a sample of size n = 100) and for each population separately (S17 Fig). DAF were significantly different, among region (Kruskal-Wallis chi-squared = 100.57, df = 5, p-value < 2.2e-16) as well as among population (Kruskal-Wallis chi-squared = 638.59, df = 57, p-value < 2.2e-16, S11 Table).They showed that Alaska, Thompson R. and California populations displayed more fixed deleterious mutations (n = 18, 12 and 11 respectively) than those from the Cascadia region (n = 3). Haida Gwaii and BC populations were intermediate with 7 and 6 fixed deleterious mutations respectively (Fig 8A). The three former regions also displayed higher frequencies of polymorphic derived deleterious mutation (S18 Fig) compared to British Columbia or Cascadia (Wilcoxon test, $p < 0.001$).

Next, we examined the count of homozygous (proportional to the load under a recessive model [33,34]), heterozygous and of total derived deleterious mutations (proportional to the load under an additive model [33,34]). In particular, we expected that mutations in a heterozygous state should be more frequent than in a homozygous state, especially in populations with higher effective sizes, where selection should be more effective at purging these mutations [23]. We found that 77% of deleterious mutations were maintained in heterozygous states across all samples. Also, fish from Alaska, Haida Gwaii and California populations harbored a significantly higher number of deleterious mutations in a homozygous state when compared to Cascadia or British-Columbia (Wilcoxon-test, $p < 0.01$) (Fig 8B, S12 Table). When considering the total load of derived deleterious mutations, we found that, on average, there were significantly more putatively deleterious variants per individual in California, Cascadia, and Haida Gwaii populations than in fish from Alaska, British-Columbia or the Thompson R. watershed (S12 and S13 Table, Fig 8C, Wilcoxon-test, $p < 0.01$), although these differences were modest. Finally, we tested the prediction that distance from the likely ancestral source predicts the deleterious load, as observed in human populations [33]. We found a nearly linear relationship between the distance from the putative origin of ancestral populations (the site with the lowest β_{ST} in Cascadia) and the number of derived, homozygous, putatively deleterious mutations (HDD; linear models, $p<0.0001$; $R^2= 0.13$, Fig 8D). Under the hypothesis that higher recombination rate leads to more efficient purging of deleterious mutations we also expected deleterious mutations to be preferentially located in areas of low recombination (S19 Fig). As expected, the GLM revealed that the occurrence of deleterious mutations decreased as recombination rate increased ($\chi^2 = 4.90$, DF = 1, $p = 0.027$). This effect was stronger when considering non-synonymous variants instead of putatively deleterious variants ($\chi^2 = 10.07$, DF=1, $p = 0.0015$). Given the negative correlation of recombination with chromsome length, we also found a positive correlation ($r = 0.76$, $p < 0.0001$) between the chromosome length and the number of deleterious variants by chromosomes, as expected when purging is more efficienct in areas of higher recombination (S20 Fig). To further examine this, we classified regions by recombination rate into high (rho > 4.516, see methods), low (rho < 1.411) and intermediate classes and found that there was a significant excess of candidate deleterious mutations in areas of low recombination and depletion of candidate deleterious mutations in areas of high recombination ($\chi^2 = 13.33$, DF= 2, $p = 0.0012$).

Discussion

Coho salmon is an emblematic fish species that has undergone population declines in recent decades throughout its North American range. We generated one of the largest collections of GBS and WGS data to date for a non-model vertebrate species, which revealed: *i)* a complex demographic history involving population splits, gene flow, and secondary contacts; *ii)* linked selection and Hill-Robertson interference shaped genetic diversity along the genome, and *iii)* this demographic history has resulted in modest yet detectable differences in the frequency of deleterious variants across regions and populations, which also varied as a function of recombination rate along the genome. Together, these observations help illuminating the drivers of variation in genetic diversity throughout the genome of a broadly distributed species of economic and cultural importance.

Expansion from a major southern refugium and secondary contact with micro-refugia

Our results revealed the existence of a major ancestral refugium located south of the ice sheets in Cascadia (i.e., Washington and Oregon) where contemporary populations contain most of the genomic diversity present over the entire North American range of the species. This conclusion was supported by observations of: *i)* a pronounced south-north gradual decrease of genetic diversity, singleton density, and private polymorphism, *ii)* a pattern of IBD, and *iii)* the occurrence of ancestral populations south of formerly glaciated areas. Therefore, although the so-called Beringian refugium that persisted north of the ice sheet (mainly comprising Alaska and the Yukon Territory) was important for many temperate species (6), this was not the case for Coho salmon. The gradual northward decrease in diversity also suggests that populations subsequently expanded from Cascadia to British-Columbia and Alaska [4,5] and is indicative of a serial founder effects due to small proportions of individuals issued from ancestral populations colonising new locations, a process amply documented in humans (e.g., [61]). This expansion was likely postglacial, although GBS-based parameter estimates of divergence time under AM, IM, and SC (~ 30 to ~ 45 KyA) were not in full agreement with SMC++ estimates using WGS (~18 KyA). This, however, is expected given the different assumptions made by the various models compared and that gene flow is expected to delay divergence [62]. The last glacial period (Wisconsin Glaciation) which lasted from ~ 120 Ky to ~11 KyA was interrupted by an interglacial period from 55 to 25 KyA ([40]). Therefore, our results indicate that neither Beringia nor Haida Gwaii were refugia before the onset of the Wisconsin Glaciation. However, it is possible that individuals from Cascadia already colonized and subsequently diverged in Alaska (Beringia) 55 to 25 KyA. If this was the case, however, the contribution of Beringia is likely minor as no strong footprint of ancestral diversity and no signal of secondary contact was inferred in this area.

While a single major southern refugium hypothesis is well supported by our data, our analyses also revealed patterns consistent with minor contributions of micro-refugia on extent genomic diversity. Indeed, our demographic modelling supported a model of postglacial (10-20 KyA, S6 Table) secondary contact (SC) detected in several pairwise comparisons among populations between California and Haida Gwaii (n = 10 pairwise comparisons), Cascadia and Haida Gwaii (n = 2), California and Thompson R. (n = 11), and

between Cascadia and Thompson R. (n = 1). These results are counter-intuitive, because contacts between northern sites and California, but bypassing Cascadia, seem unlikely. A possible explanation is that our statistical power to detect SC involving putative micro-refugia was reduced. Here, indeed, the SC period represented a large proportion (22% on average) of the total divergence time across all models. However, models of SC can easily be confounded with models of isolation with migration when SC represents a large period of time (> 10%) relative to the total period of divergence, as observed in our investigations [19,63]. Still, our inference of SC supports the hypothesis that smaller micro-refugia have persisted along the Pacific Coast [43,64]. In particular, the Haida Gwaii archipelago is known for high endemism, and its role as a refugium for mammals, invertebrates, and angiosperms is well established [64–66]. The divergence time inferred between populations from California and Haida Gwaii was relatively recent (39 KyA), in the range of inferred divergence times for bird species from the region [67] and in the time frame of the last interglacial period 55 to 25 KyA. Our results also indicated a pronounced divergence between California and Cascadia, suggesting that these were likely two separated refugia, thus supporting the hypothesis of "refugia within refugia" [43,67]. However, California populations have suffered from strong census size reductions in recent decades [68], possibly increasing genetic structure and lowering effective population size. Given our estimates of split times and population sizes, the possible refugium in Haida Gwaii, and California were demographically small and relatively recent, such that populations in these areas were unlikely to have accumulated significant endemic genetic diversity. The case of the Thompson River, on the other hand, is rather intriguing as this region was entirely covered by ice during the Pleistocene. A possible explanation is that a strong and continuous bottleneck has led to a false signal of SC. The SC was the most parameterized model and was therefore more likely to be falsely supported. A model including more drastic and continuous post-divergence changes in population size may fit the data well but would likely be difficult to statistically separate from a model of secondary contact. British-Columbia populations, located at intermediate latitudes, also displayed intermediate levels of genetic diversity, with some populations displaying similar or higher diversity than those located in Cascadia, further supporting the hypothesis of post-colonization admixture between different refugia [69] Complex history of divergence in different refugia and possible mixing was also infered in the Chinook salmon along the Pacific Coast and was associated to different migratory ecotypes [70]. A potential limitations of our current approach is that we did not incorporate samples from Asia, where the Coho salmon also occurs [41]. Whether this would change our inferences remains an open question. In summary, our modelling approach revealed the presence of multiple cryptic refugia that would otherwise have been missed because their endemic genetic diversity appears to have been largely wiped out by a major demographic expansion out-of-Cascadia, which has clearly been the primary contributor of extent genomic diversity in North American Coho salmon. Knowledge of this non-equilibrium demography will important to help interpret adaptive and deleterious genomics landscape in future studies.

Selection at linked sites shape genome wide variation

We also explored the role of gene flow and linked selection in generating a heterogeneous differentiation landscape along the genome, a hotly debated topic in evolutionary genomics [8,9,71,72]. Disentangling the

two processes is of fundamental importance for correctly interpreting the origin of genomic islands of divergence and genome scan results [73,74]. Although a positive relationship between absolute and relative divergence may reveal the presence of barriers to gene flow [9], this is not relevant for early stages of divergence [75], as in the present study. Instead, our modelling approach best supported a role of both gene flow and linked selection. In particular, the role of linked selection is supported by the positive correlations between π or D_{XY} and recombination rate, while F_{ST} was negatively correlated with recombination rate [73,75]. The negative correlations between π or D_{XY} and gene density provided further evidence for the role of linked selection [17]. Here, heterogeneous genomic divergence did not arose after speciation, as suggested by Cruickshank & Hahn [9]. Instead, our within-lineage study indicated that linked selection arose within structured populations, in line with recent findings in birds [73]. Similar effects of linked selection were suggested between the more diverged ancestral lineages within Atlantic salmon (>1 MyA) [76]. Given that correlated genomic landscapes have been reported among species diverged for over tens of millions of years [77,78], it would be interesting to investigate how linked selection has also shaped genomic landscapes within the entire radiation of salmonid fishes. Indeed, salmonid species have undergone a whole genome duplication approximately 90 MyA [79] affecting recombination throughout their genome [57,80,81]. Precise mapping of the genomic locations of duplicated regions will allow an improved understanding of how recombination rate varies in these regions althougth we do not expect that will affect our main conclusion related to linked selection. Salmonid genomes are also characterised by pronounced male heterochiasmy [82,83] and here, we observed a significant correlation between recombination and chromosome length, indicative of crossover interference [84]. Both heterochiasmy and crossover interference contribute to heterogeneity in recombination rates and likely favour the effect of linked selection. Given that most non-neutral mutations are deleterious [85], and given the impact of recombination across the genomes of fishes, models of background selection could be the null model against which to test adaptive hypotheses [21]. Our data revealed that the GC3 was a good proxy of recombination and that it was strikingly well correlated with our proxy for the efficacy of natural selection (π_N/π_S). As expected, this metric revealed differences in the most bottlenecked populations from the Thompson River drainage, where the severity of the bottleneck mimics the effect of domestication bottleneck (e.g. [86]). Indeed, the Thompson R. populations displayed a higher burden of non-synonymous mutations, and we suggest that the strong bottlenecks in these populations had a dramatic impact on the efficacy of natural selection. When considering only genomic regions of low recombination, however, we also found that some California populations also displayed an increased π_N/π_S ratio. Given that these populations have recently undergone large declines in abundance [68], we expected a lower efficacy of purifying selection in regions of low recombination where Hill-Robertson interference should be high [82]. Furthermore, we suggest that the severe and recent demographic declines documented in California populations compared to more long term and putatively less severe declines in the Thompson R. may explain why differences were detected only in genomic regions of low recombination in the former but genome-wide in the latter. Regardless, the non-equilibrium demography, and role of linked selection, in particular background selection, are predicted to favour the accumulation of deleterious mutations, an important finding to manage declining species.

Accumulation of deleterious mutations under complex demography

In the literature, contradictory results have been reported regarding the role of demographic history on the load of deleterious mutations [10,33]. On the one hand, processes such as strong and repeated bottlenecks [24,32], including domestication [86,88], large expansions [89,90] or postglacial colonization [91] can all increase the deleterious load. On the other hand, empirical studies in human indicate that recent demography should not have a strong impact on the load of deleterious variants over the long term [10]. However, other studies indicate that strongly deleterious variants affecting fitness can still display increased in frequencydue to demographic growth [10]. Here, the inferred demographic history, together with support for linked selection and gBGC, provides ideal conditions to favour the accumulation of deleterious mutations. In particular, we found that populations displaying smaller effective population size (e.g. California), and those at the extreme of the expansion front (Alaska), displayed deleterious mutations at higher derived allele frequencies, and more frequently in homozygous state, in line with the nearly neutral theory. Similar findings have been observed in domesticated species [86] and recently in Isle Royale wolves [26] where decreased population sizes and inbreeding increased the frequency of deleterious recessive mutations. Overall, these results are consistent with recent empirical findings in which small populations or populations at the expansion front carry more homozygous derived deleterious mutations [26,27,86]. These deleterious mutations in homozygous state are expected to be purged by purifying selection [23,92]. Populations from Cascadia, with higher effective population sizes, contain a higher number of putatively deleterious variants in heterozygous state, as expected from population genetics theory. We also found a nearly linear relationship between the number of derived deleterious mutations in homozygous states and geographic distance from the putative refugial source in Cascadia. This relationship mirrors the findings pertaining to the 'out-of-Africa' expansion of human populations [33], although the maximum geographic distance in our study is an order of magnitude smaller than in the human studies, and that the presence of multiple refugia (as opposed to the sole African ancestral origin for humans) may have contributed to obscure somewhat the relationship. Moreover, pronounced genetic drift or bottlenecks observed in some populations (e.g. Thompson R. and California) may have contributed to reduce the efficacy of selection [24]. Finally, the lower prevalence of deleterious mutation in fish from British-Columbia might also be explained by post-glacial admixture. In some small populations it is possible that these mutation initially reached fixation but subseqently became masked in heterozygous state due to gene-flow and introgression [93]. Similar effect could be due to artificial supplementation programs, but the origin of source populations and intensity of stocking in our study populations are poorly documented. Finally, we observed that recombination was significantly correlated with the proportion of non-synonymous mutations and with the load of deleterious mutations, with more deleterious mutations being found in genomic regions of low recombination. This is in line with theoretical predictions (e.g. Hill-Robertson Interference discussed above), and also empirical studies in other species [94,95]. Overall, our results suggest that Coho salmon populations with highly reduced population sizes are exposed to higher inbreeding depression – a prediction with major conservation implications and which should be investigated further in future studies.

Conclusion

In this study, we presented a rare combined assessment of the relative role of complex demographic history (investigated both by empirical genomic data and modelling) and intrinsic genomic factors (recombination, linked selection) in shaping drivers of the genomic landscape of a broadly distributed species. Complex demographic processes including population expansion, isolation, and secondary contact were revealed through the use of an extensive modelling framework. Moreover, our results highlighted the necessity of accounting for local variation in recombination rate, a key driver of linked selection. Altogether, these processes influence the efficacy of selection and can favour the accumulation of mutations affecting fitness. Our findings suggest that such approaches offer enormous potential in the field of conservation genomics to disentangle the impacts of historical vs. recent drivers of demographic declines and for assessing the distribution of not only putatively beneficial, but also deleterious variants. We propose that future studies should also integrate in-depth analysis of selective sweeps which will be necessary to investigate into more details how linked selection, through background selection and hitchhiking, acts to maintain deleterious mutations in the genome [96]. Finally, while it has become routine with new genome-wide datasets to focus on the effect of positive selection and documenting patterns of local adaptation [7], this focus ignores the fundamental prediction that most new mutations are likely to be deleterious [85]. An increased focus on deleterious mutations would provide a more nuanced view of genomic evolution in wild species which would benefit both the fields of evolutionary and conservation genetics

Methods

Genotyping By Sequencing

A total of 2,088 individuals was collected from 58 sample sites located along the Pacific coast from California to Alaska (S1 Table and Figure 1). DNA was extracted from all individuals and sequenced using a GBS method (protocol detailed in [97]). Reads were aligned to the Coho salmon reference genome v1 (GCF_002021745.1) using bwa-mem 0.7.13 [102]. Samtools v1.7 was used to keep reads with a mapping quality above 20, remove supplmentary alignment and unmapped read. Variants were then called with Stacks v1.46 [98]. To do so, the module “pstacks” was used with a minimum depth of 5, and up to three mismatches were allowed in catalog assembly. The module “populations” was run to produce a vcf file that was filtered with a custom python script.We performed stringent filtering to remove SNPs that were 1) genotyped in less than 60% of the individuals; 2) at a mean depth of sequencing below 7, and 3) with observed heterozygosity above 0.60, thus resulting in 93,000 SNPs. The pipeline for SNP calling is available on github at https://github.com/enormandeau/stacks_workflow/releases/tag/coho_demography_paper. Next, we removed any individuals with more than 5% missing data and finally only kept SNPs present in at least 95% of the individuals yielding a total of 82,772 filtered SNPs for 1,957 individuals. Remaining filtration was done according to the requirement of each analysis performed below.

Genetic diversity and ancestral populations

For each sampling location we estimated the observed heterozygosity and π using vcftools 0.1.16 [99] and hierfstat [100] The most likely ancestral populations were identified using β_{ST} [46]. A total of 1,000 bootstraps was performed to obtain the 95% confidence intervals around the β_{ST}. Weir and Cockerham's F_{ST} estimator θ [47] was computed in vcftools. We measured the relationship between observed heterozygosity, β_{ST}, F_{ST} and the distance to the southernmost site using linear models. We also verified the relationship between F_{ST} and the distance to the southernmost site using Mantel tests with 10,000 permutations. Vcftools was also used to identify singletons (i.e. variants present in one single individual across the whole dataset). Their distributions were then summed in each locality. We then computed the averaged (min, max and median) number of singletons at the regional level. The scripts are available on github at https://github.com/QuentinRougemont/utility_scripts

Population structure, admixture and gene flow

Levels of ancestry and admixture proportions were inferred with the snmf function implemented in the R package LEA [101]. We allowed less than 5% of missing data. We then kept a set of SNPs in approximate linkage equilibrium by removing pairs of variants with r^2 greater than 0.2 (option --indep-pairwise 50 10 0.2) resulting in 40,632 SNPs. K-values ranging between 1 and 60 were tested and cross-validations were obtained using the cross-entropy criterion with 5% of masked genotypes. The default value for the regularization parameter was used to avoid forcing individuals into groups and hence underestimating admixture. Similar results were obtained from Admixture [102] and are not presented here. Genetic relationship among all salmon was assessed using a PCA with the R package ade4 [103] based on the LD-pruned dataset (40,632 SNPs). We used a 1% minor allele frequency (MAF) threshold and allowed less than 4% missing data. Formal tests of admixture were performed using Treemix [49] using the LD-pruned dataset of 40,632 SNPs and without any MAF threshold. A MDS was also constructed using plink and plotted with the ggplot2 [105] R package. We ran Treemix allowing up to 20 migration events and performed 500 bootstraps replicates of the best model to obtain robustness of the nodes. The "best" model was inferred after inspecting the relevant migration edges by measuring the percentage of variance explained as migration edge were added to the tree as well as by assessing the p-value associated to each migration edge. A total 500 bootstraps replicate was performed under the "best" model and under a model without migration to infer the robustness of the nodes. The scripts are available on github at https://github.com/QuentinRougemont/treemix_workflow

Explicit demographic inferences.

We tested alternative hypotheses of secondary contact between major regional groups (Haida Gwaii, California, Cascadia (Washington-Oregon), Alaska, British Columbia, Thompson R.) of populations by comparing alternative divergence scenarios represent in Fig S12 and initially described in [19,76]. Alternative hypotheses of secondary contacts were tested between major groups by testing the significance of alternative divergence scenarios. The four major models tested included a model of Secondary Contact

(SC), a model of Strict Isolation (SI), a model of Ancient Migration (AM) and a model of Isolation with Migration (IM).

The models shared the following parameters: the ancestral populations of size N*anc*, splits at time T_{split} into two daughter populations of size N_1 and N_2. Under the SI model, no gene flow occurs between the two populations. Under AM, gene flow occurred between T_{split} and T_{am} and is followed by a period of strict isolation. Under IM, gene flow occurs at a constant rate at each generation between the two populations. Gene flow can be asymmetric, so that two independent migration rates m_{12} (from population 2 to 1) and m_{21} (from population 1 to 2) were modeled. Under the SC model, the population evolved in strict isolation between T_{split} and until T_{sc} where a secondary contact occurs continuously up to present time. Gene flow is modeled as $M = 2Nref.m$. In ∂a∂i, heterogeneous introgression was modeled using two categories of loci occurring in proportions P (i.e., loci with a migration rates M_{12} and M_{21}) and $1\text{-}P$ (i.e., loci with a reduced effective migration rates Me_{12} and Me_{21}) across the genome. The same procedure was used to account for linked selection by defining two categories of loci with varying effective population sizes (proportion Q of loci with a "neutral N_e" and a proportion 1-Q of loci with a reduced effective population size due to either linked or background selection). To quantify how linked selection affects reduced *Ne*, we used a Hill-Robertson scaling factor (Hrf) to relate the effective population size of loci influenced by selection ($Nr = Hrf * Ne$) to that of neutral loci (N_e).

Models were fitted using the diffusion theory implemented in ∂a∂i [50] and also includes the effect of linked selection and barriers to gene flow as detailed in [19,104]. ∂a∂i uses the SFS as a summary of the data. For a given demographic model, the SFS is computed using diffusion approximation and compared to the empirical SFS using AIC. Here, we started from the whole file containing 200,000 SNPs and used one single SNP per GBS locus, filtered the data to minimize missing data. No MAF was used and singletons were kept to obtain ascertainment-free estimates of demographic parameters. Ideally, no Hardy-Weinberg Equilibrium (HWE) filter should be used for demographic inferences, as this also biases the distribution of allele frequencies. However, to remove paralogs present in the Coho salmon genome, a permissive HWE filter based on a p-value of 0.0001 was used. Here, a total of 69 pairwise comparisons between populations from the major regional groups was performed in order to test for a prevailing pattern. Here, the regional groups considered were all previously unglaciated areas during the LGM, namely California, Cascadia, Alaska (although some samples were likely under ice at different time periods) and Haida Gwaii. For each model, 20 independent replicate runs were performed and only models with the lowest AIC and ΔAIC were kept. A model was classified as "ambiguous" and not used for parameter estimation if ΔAIC between the best model and second-best model was below 10. The whole pipeline is available at https://github.com/QuentinRougemont/DemographicInference.

Analyses based on whole genome resequencing data.

We used 55 individuals representing 11 populations from California to Alaska (S2 Table). Each individual was sequenced on an Illumina platform using paired-end 150 bp reads. Reads were processed using fastp for trimming [105], bwa mem v0.7.13 [106]for mapping, samtools v1.7 requiring a minimum quality of 10, and

picard to remove duplicates (http://broadinstitute.github.io/picard/). Then SNP calling was performed using GATK [107]. Genotypes were filtered for depth between 10 and 100 reads to remove low confidence genotypes including potential paralogs, displaying high coverage. Then following GATK Best Practices we excluded all sites that did not match the following criterion: MQ < 30, QD < 2, FS > 60, MQRankSum < -20, ReadPosRankSum < 10, ReadPosRankSum > 10. We also generated a vcf file using the samtools mpileup pipeline, merging individuals with bcftools and performing the same stringeant filtering as with the vcf constructed with GATK. Finally, we also generated a separate vcf file using the emit-all-site option to call variable and invariable sites across the whole genome. This vcf file was used in the sliding windows analysis below to test the effect of linked selection. The whole pipeline is available on github (https://github.com/QuentinRougemont/gatk_haplotype).

SMC++ [53] was used to infer changes in population size through time. SMC++ works similarly to the PSMC model but takes advantage of information contained in the Site Frequency Spectrum (SFS) in addition to linkage disequilibrium, and is particularly well suited to analyse large sample sizes. Estimates of population size changes were performed for all 11 populations. Splitting time was estimated between all pairs of samples based on the joint SFS. A generation time of 3.5 years and a mutation rate of $8e^{-9}$ mutation/bp/generation were applied. The pipeline to reproduce the analysis is available on github (https://github.com/QuentinRougemont/smcpp_input)

We computed pairwise linkage disequilibrium using vcftools with the r^2 statistics calculated in all 11 populations separately. A window of 1,000,000 base pairs was used and all SNPs were included. To reduce the number of SNPs, we allowed no missing data and used a MAF of 10% and a p-value of Hardy-Weinberg disequilibrium of 0.05 in each population, keeping between 2 and 4 million SNPs depending on the population. We then estimated LD decay by plotting LD against physical distance measured in base pairs and using smoothing functions implemented in ggplot2 [108] package in R. vcftools was also used to identify singletons, for which the distribution were counted by localities, as for the GBS dataset.

We used LDHat software [109] to estimate effective recombination rates ($\rho=4.Ne.r$ where r represents the recombination rate per generation and *Ne* is the effective population size) along the genome. Unphased genotypes were converted into LDHat format using vcftools with a minimum MAF of 10% since only common variants are useful for such inferences [109]. Following the authors' guidelines, the genome was split in chunks of 2,000 SNPs with overlapping windows of 500 SNPs to compute recombination rate and data were then merged together. We measured recombination rates for each river as well as globally including all populations except the population from the Thompson R. watershed that was too divergent from the remaining samples. The pipeline to reproduce the analysis is available on github (https://github.com/QuentinRougemont/LDhat_workflow).

Clarifying the role of linked selection

The genetic landscape of divergence was measured by estimating levels of nucleotide diversity (π), gene density, levels of genetic differentiation (F_{ST}), levels of divergence (Dxy), and scaled recombination rate along the genome. Estimates of π, F_{ST}, and Dxy were computed into 500kb windows using Python scripts

available from [110]. Gene density was computed directly from the gff file of the Coho salmon genome v1.0 (NCBI ftp file: ftp://ftp-trace.ncbi.nih.gov/genomes/Oncorhynchus_kisutch/GFF/; Rondeau et al. In prep) and measured in 500 kb windows. Recombination rates were averaged into 250 and 500-kb windows using a Python script after being estimated with LDhat as described above. We used a PCA to obtain a synthetic view across all 11 π estimates, and all pairwise 55 Dxy and 55 F_{ST} separately. This allowed capturing the common variation affecting these three estimates. First, simple Spearman correlations based on linear models testing each (z-transformed) variable projected on the PC1 axis separately were carried out to produce fig 5. Then Mixed Linear Models were used to test for correlations between either π, Dxy, or F_{ST} and either recombination landscape (rho) or gene density windows as explanatory variables, with and without interaction. Correlations were also calculated considering each π, Dxy, and F_{ST} by population separately without PCA transformation (S9 Table) which returned patterns that were congruent with those observed with PCA results.

Measuring GC content, Non-Synonynous and Synonymous diversity

The pipeline developed by [111] was used to compute Tajima's π estimator of nucleotide diversity and the GC content over non-overlapping 10-kb windows. Then we concatenated the gene into different classes according to their length and compute π_N and π_S over 4 mb windows, which allows circumventing the problem associated to low π_s values. We then verified the correlation between populations scaled recombination rate (ρ), measured over 250-kb windows and GC at the third codon position (approximately neutral). To do so, we assigned a ρ value for each gene according to its position into each 250-kb windows. Then we averaged ρ values over the same length class as GC3 classes and compared the median population recombination rate estimates and GC3s over 4-mb windows. Finally, we measured the correlation between GC3s and πN/πS using linear models.

Genetic load estimation from the GBS data set

Estimating ancestral and derived alleles

We used the derived allele count as an estimator of deleterious allele count. We used the genomes of three outgroup species, the chinook salmon, the rainbow trout, and the Atlantic salmon, to classify SNPs as ancestral or derived. Whole genome data for the chinook salmon (n = 3 individuals) were provided by one of us (B. Koop, unpublished), whereas rainbow trout (n = 5) and Atlantic salmon (n = 5) data were downloaded from NCBI Sequence Read Archive (rainbow trout, SRA, Bioproject : SRP117091; *Salmo salar* SRA Bioproject: SRP059652). Each individual was aligned against the Coho salmon V1 reference genome (GCF_002021745.1) using GATK UnifiedGenotyper and calling every SNP using the EMIT_ALL_SITES modes. We then constructed a Python script and determined the ancestral state of the GBS SNPs if 1) the SNP was homozygous in at least two of the three outgroups, and 2) match one of the two alleles identified in Coho salmon. Otherwise, the site was inferred as missing and not used in subsequent analyses.

Measuring damaging impact of non-synonymous alleles

We tested differences in mutation load among populations as follows. The software Provean [112] was used to predict whether a given non-synonymous mutation was deleterious with a threshold score of -2.5 or less using the pipeline available at https://github.com/QuentinRougemont/gbs_synonymy_with_genome. We analysed data in two ways: first we counted the total number of putative homozygous deleterious alleles per individual as well as the total number of deleterious alleles (both in homozygous and heterozygous states) using: Ntotal = 2 X N_{homo} + N_{hetero} [34]. These individual values were then averaged per population and major regional groups (i.e., California, Cascadia, British Columbia, Haida Gwaii, Thompson, and Alaska). We then computed derived allele frequencies (DAF) in all sampling locations and across mutation categories (synonymous, non-synonymous, and non-synonymous deleterious) and tested for significant allele frequency differences among populations in non-synonymous and non-synonymous deleterious mutations using Wilcoxon rank sum tests. DAF spectra were constructed for all population separately. For the ease of visualisation, we also constructed DAF spectra by region for a sample of size n = 100 individuals (Fig 7A). This size was chosen according to the smallest sample size of the three combined Haida Gwaii populations. Finally, we tested for a preferential enrichment of "hotspots or coldspots" of recombination in deleterious mutations. To define a coldspot, we first computed a lower bound that we defined as the $mean_{RHO}$ - 5*standard errors for each chromosome separately. Similarly, 'hotspots" of recombination were identified using a upper bound defined as $mean_{RHO}$ + 5 standard errors. We then tested if the average recombination rate of each of the 250 kb windows was falling below the threshold (for coldspot) or above it (for hotspots). We then tested if recombination hostpots and coldspots contained more or less putatively deleterious mutations than "normally" recombining regions using χ^2 tests. Linear mixed effects models were further used to test if there was a relationship between recombination rate and the distribution of putatively deleterious mutations. The response variable was the deleterious state considered binomial (0 = non deleterious, 1 = deleterious) and the explanatory variable was the recombination rate. The chromosome identity was included as a random effect. This model was compared against a null model excluding recombination rate. The analysis was replicated but using the synonymous and non-synonymous state as the response variable instead of the putatively deleterious state of the considered mutation. Models were carried out in R using the LME4 package [113]. Finally, we used SnpEff v3.4 [114] to obtain the functional annotations of the putatively deleterious variants. The annotations were comprised of mis-sense variants, non-coding transcripts, 3' and 5' untranslated regions, 5kb up- and down-stream variants, intergenic and intronic, splice acceptor and splice region, stop gained and start loss (S14 Table). We found 35% of deleterious mutations to be missense variants, 38% to be non-coding transcript and 22% to be either upstream or downstream gene variants, with the remaining being spread over the categories.

Ethic Statement

A permit number SIRUL 111722 was obtained to work on DNA sequences.

Raw data will be deposit on NCBI together with Short Read Archive (SRA) accession number.

Acknowledgements:

We thank B. Bougas, A. Perrault-Payette, C. Hernandez for laboratory support and K. Wellband, H. Cayuela, F. Hartmann for their constructive comments. TL and QR thank Benoit Nabholz for discussions regarding the deleterious mutation loads and Hill-Robertson effects. Computations were performed on Colosse (Calcul Quebec), Graham and Cedar (Compute Canada) servers. This research was carried out in conjunction with EPIC4 (Enhanced Production in Coho: Culture, Community, Catch), a project supported by the government of Canada through Genome Canada, Genome British Columbia, and Genome Quebec. The authors declare no conflicts of interest.

Figure Legend

Figure 1) Sampling design

A) Sampling locations of 58 Coho salmon populations distributed across the species' North American range of distribution. Each dot represents a sampling location. Inset: Map showing the extend of ice-sheet during the Last Glacial Maximum 13 KyA. Data modified from [115]

Figure 2) Genetic diversity and differentiation

A) Linear relationship between expected heterozygosity and distance from the southernmost population located in California. B) Linear increase in genetic differentiation as measured by β_{ST} as a function of the distance from the southernmost population located in California. Negative values indicate the most likely ancestral population. The relationship in A and B was tested using linear models. The grey vertical bar in panel B) indicates the approximate location of the southern limit of the ice-sheet at the end of the last glacial maxima.

Figure 3) Genetic structure and gene flow

A) Principal Component Analysis (PCA) summarising population genetic structure among 1,957 individuals based on the principal component axes 1 and axes 2. Each point represents an individual and the colours represent the major regional groups. B) Inference of population splits and mixture by Treemix with 4 statistically significant migration events ($p < 0.0001$). Bootstrap supports and migration weights are indicated according to the colour scale.

Figure 4) Inferences of demographic history using ∂a∂i and smc++

A) Estimates of divergence time (in years) between each major region as inferred by ∂a∂i under the best model (displayed in blue) based on the GBS SNP data set. B) Estimates of effective population size *Ne* for each major region as inferred by ∂a∂i under each best demographic model based on the GBS SNP data set. C) Estimates of migration rate among populations in neutral regions (*m)* of the genome and in areas of restricted recombination (*me*) based on the GBS SNP data set. D) Estimates of effective population size (*Ne)* change through time (in years) for the whole genome data using SMC++.

Figure 5) Genome-wide landscapes in Coho salmon with: A) landscape of differentiation, B) landscape of genetic diversity C) divergence, D) recombination and E) gene density. Plots are averaged over 500-kb windows. Only the first 16 chromosomes are displayed for simplicity. The same figure with all chromosomes is available in Fig. S15.

Figure 6) Linked selection revealed by correlation among different metrics

Correlations between recombination rate ($\rho = 4Ne.r$) genetic diversity (π), divergence (Dxy) and differentiation (F_{ST}) statistics (A-C), and correlation between gene density and π, Dxy, F_{ST} statistics (D-F). Correlations between π, Dxy and F_{ST} statistics (G-I). The Spearman correlation values are plotted but mixed linear models were fitted to further test interaction among explanatory variables (recombination rate and gene density).

Figure 7) Correlation between π_N/π_S and GC3 content for each population separately

π_N/π_S is used here as a proxy for the efficacy of purifying selection with smaller values reflecting more efficient purifying selection. All correlations are significant ($p < 0.001$).

Figure 8) Analysis of deleterious load in the Coho salmon genome

A) Derived allele frequency spectra across major groups. The non-synonymous variants and putatively deleterious variants frequency are shown. Data are normalized for a sample of size n = 102 corresponding to the smallest size for the combined samples in Haida Gwaii. B) Distribution of the count of homozygous derived deleterious alleles in each major group. C) Distribution of the count of total derived deleterious alleles in each major group. D) Correlation ($p = 0.0034$, $R^2=0.13$) between geographic distance to the putative ancestral source and the distribution of homozygous derived putatively deleterious mutations.

References:

1. Allendorf FW. Evolution in a Toxic World. BioScience. 2017;67: 576–577. doi:10.1093/biosci/bix029

2. Ceballos G, Ehrlich PR, Dirzo R. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. Proc Natl Acad Sci. 2017;114: E6089–E6096. doi:10.1073/pnas.1704949114

3. Provan J, Bennett KD. Phylogeographic insights into cryptic glacial refugia. Trends Ecol Evol. 2008;23: 564–571. doi:10.1016/j.tree.2008.06.010

4. Hewitt GM. Post-glacial re-colonization of European biota. Biol J Linn Soc. 1999;68: 87–112. doi:10.1006/bijl.1999.0332

5. Bernatchez L, Wilson CC. Comparative phylogeography of Nearctic and Palearctic fishes. Mol Ecol. 1998;7: 431–452. doi:10.1046/j.1365-294x.1998.00319.x

6. Frankham R, Ballou JD, Briscoe DA. Introduction to Conservation Genetics by Richard Frankham. In: Cambridge Core [Internet]. Jan 2010 [cited 2 Jul 2019]. doi:10.1017/CBO9780511809002

7. Funk WC, McKay JK, Hohenlohe PA, Allendorf FW. Harnessing genomics for delineating conservation units. Trends Ecol Evol. 2012;27: 489–496. doi:10.1016/j.tree.2012.05.012

8. Noor M a. F, Bennett SM. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity. 2009;103: 439–444. doi:10.1038/hdy.2009.151

9. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 2014;23: 3133–3157. doi:10.1111/mec.12796

10. Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. Nat Genet. 2014;46: 220–224. doi:10.1038/ng.2896

11. Bierne N, Gagnaire P-A, David P. The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. Curr Zool. 2013;59: 72–86. doi:10.1093/czoolo/59.1.72

12. Barton N, Bengtsson BO. The barrier to genetic exchange between hybridising populations. Heredity. 1986;57: 357. doi:10.1038/hdy.1986.135

13. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974;23: 23–35.

14. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134: 1289–1303.

15. Kaplan NL, Hudson RR, Langley CH. The “hitchhiking effect” revisited. Genetics. 1989;123: 887–899.

16. Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background selection. Genet Res. 1996;67: 159–174.

17. Payseur BA, Nachman MW. Natural selection at linked sites in humans. Gene. 2002;300: 31–42.

18. Hudson RR, Kaplan NL. Deleterious Background Selection with Recombination. Genetics. 1995;141: 1605–1617.

19. Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. PLOS Biol. 2016;14: e2000234. doi:10.1371/journal.pbio.2000234

20. Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. Veeramah K, Wittkopp PJ, Gronau I, editors. eLife. 2018;7: e36317. doi:10.7554/eLife.36317

21. Comeron JM. Background selection as null hypothesis in population genomics: insights and challenges from Drosophila studies. Philos Trans R Soc Lond B Biol Sci. 2017;372. doi:10.1098/rstb.2016.0471

22. Simons YB, Sella G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. Curr Opin Genet Dev. 2016;41: 150–158. doi:10.1016/j.gde.2016.09.006

23. Charlesworth D, Willis JH. The genetics of inbreeding depression. Nat Rev Genet. 2009;10: 783–796. doi:10.1038/nrg2664

24. Kirkpatrick M, Jarne P. The Effects of a Bottleneck on Inbreeding Depression and the Genetic Load. Am Nat. 2000;155: 154–167. doi:10.1086/303312

25. Charlesworth B. The Effects of Deleterious Mutations on Evolution at Linked Sites. Genetics. 2012;190: 5–22. doi:10.1534/genetics.111.134288

26. Robinson JA, Räikkönen J, Vucetich LM, Vucetich JA, Peterson RO, Lohmueller KE, et al. Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. Sci Adv. 2019;5: eaau0757. doi:10.1126/sciadv.aau0757

27. Robinson JA, Ortega-Del Vecchyo D, Fan Z, Kim BY, vonHoldt BM, Marsden CD, et al. Genomic Flatlining in the Endangered Island Fox. Curr Biol CB. 2016;26: 1183–1189. doi:10.1016/j.cub.2016.02.062

28. Abascal F, Corvelo A, Cruz F, Villanueva-Cañas JL, Vlasova A, Marcet-Houben M, et al. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. Genome Biol. 2016;17: 251. doi:10.1186/s13059-016-1090-1

29. Dobrynin P, Liu S, Tamazian G, Xiong Z, Yurchenko AA, Krasheninnikova K, et al. Genomic legacy of the African cheetah, Acinonyx jubatus. Genome Biol. 2015;16: 277. doi:10.1186/s13059-015-0837-4

30. Yang Y, Ma T, Wang Z, Lu Z, Li Y, Fu C, et al. Genomic effects of population collapse in a critically endangered ironwood tree Ostrya rehderiana. Nat Commun. 2018;9: 5449. doi:10.1038/s41467-018-07913-4

31. Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, et al. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science. 2015;348: 242–245. doi:10.1126/science.aaa3952

32. Grossen C, Guillaume F, Keller LF, Croll D. Accumulation and purging of deleterious mutations through severe bottlenecks in ibex. bioRxiv. 2019; 605147. doi:10.1101/605147

33. Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, et al. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad Sci. 2016;113: E440–E449. doi:10.1073/pnas.1510805112

34. Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. Estimating the mutation load in human genomes. Nat Rev Genet. 2015;16: 333–343. doi:10.1038/nrg3931

35. Krkosek M, Ford JS, Morton A, Lele S, Myers RA, Lewis MA. Declining wild salmon populations in relation to parasites from farm salmon. Science. 2007;318: 1772–1775. doi:10.1126/science.1148744

36. Irvine JR, Fukuwaka M. Pacific salmon abundance trends and climate change. ICES J Mar Sci. 2011;68: 1122–1130. doi:10.1093/icesjms/fsq199

37. Gustafson RG, Waples RS, Myers JM, Weitkamp LA, Bryant GJ, Johnson OW, et al. Pacific Salmon Extinctions: Quantifying Lost and Remaining Diversity. Conserv Biol. 2007;21: 1009–1020. doi:10.1111/j.1523-1739.2007.00693.x

38. Smith CT, Nelson RJ, Wood CC, Koop BF. Glacial biogeography of North American coho salmon (Oncorhynchus kisutch). Mol Ecol. 2001;10: 2775–2785.

39. Beacham TD, Wetklo M, Deng L, MacConnachie C. Coho Salmon Population Structure in North America Determined from Microsatellites. Trans Am Fish Soc. 2011;140: 253–270. doi:10.1080/00028487.2011.558782

40. McPhail JD, Lindsey CC. Freshwater fishes of northwestern Canada and Alaska. Fisheries Research Board of Canada : available by mail from the Queen's Printer; 1970.

41. UBC Press | Pacific Salmon Life Histories, By Cornelis Groot, Leo Margolis and Leo Margolis. In: UBC Press [Internet]. [cited 1 Jul 2019]. Available: https://www.ubcpress.ca/pacific-salmon-life-histories

42. Hocutt CH, Wiley EO, editors. The Zoogeography of North American Freshwater Fishes. 1 edition. New York: Wiley-Interscience; 1986.

43. Mee JA, Moore J-S. The ecological and evolutionary implications of microrefugia. J Biogeogr. 2014;41: 837–841. doi:10.1111/jbi.12254

44. Warner BG, Mathewes RW, Clague JJ. Ice-free conditions on the queen charlotte islands, british columbia, at the height of late wisconsin glaciation. Science. 1982;218: 675–677. doi:10.1126/science.218.4573.675

45. Cubry P, Vigouroux Y, François O. The Empirical Distribution of Singletons for Geographic Samples of DNA Sequences. Front Genet. 2017;8. doi:10.3389/fgene.2017.00139

46. Weir BS, Goudet J. A Unified Characterization of Population Structure and Relatedness. Genetics. 2017;206: 2085–2103. doi:10.1534/genetics.116.198424

47. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984;38: 1358–1370. doi:10.2307/2408641

48. Ward RD, Woodwark M, Skibinski DOF. A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. J Fish Biol. 1994;44: 213–232. doi:10.1111/j.1095-8649.1994.tb01200.x

49. Pickrell JK, Pritchard JK. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLOS Genet. 2012;8: e1002967. doi:10.1371/journal.pgen.1002967

50. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLOS Genet. 2009;5: e1000695. doi:10.1371/journal.pgen.1000695

51. Roux C, Tsagkogeorga G, Bierne N, Galtier N. Crossing the species barrier: genomic hotspots of introgression between two highly divergent Ciona intestinalis species. Mol Biol Evol. 2013;30: 1574–1587. doi:10.1093/molbev/mst066

52. COSEWIC assessment and status report on the coho salmon Oncorhynchus kisutch (Interior Fraser population) in Canada - Species at Risk Public Registry [Internet]. [cited 1 Jul 2019]. Available: https://wildlife-species.canada.ca/species-risk-registry/document/default_e.cfm?documentID=105

53. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 2017;49: 303–309. doi:10.1038/ng.3748

54. Haenel Q, Laurentino TG, Roesti M, Berner D. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. Mol Ecol. 2018;27: 2477–2497. doi:10.1111/mec.14699

55. Eyre-Walker Adam. Recombination and mammalian genome evolution. Proc R Soc Lond B Biol Sci. 1993;252: 237–243. doi:10.1098/rspb.1993.0071

56. Galtier N, Piganeau G, Mouchiroud D, Duret L. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics. 2001;159: 907–911.

57. Christensen KA, Leong JS, Sakhrani D, Biagi CA, Minkley DR, Withler RE, et al. Chinook salmon (Oncorhynchus tshawytscha) genome and transcriptome. PLOS ONE. 2018;13: e0195461. doi:10.1371/journal.pone.0195461

58. Gao L, Jia J, Kong X. A SNP-Based Molecular Barcode for Characterization of Common Wheat. PLOS ONE. 2016;11: e0150947. doi:10.1371/journal.pone.0150947

59. Yáñez JM, Naswa S, López ME, Bassini L, Correa K, Gilbey J, et al. Genomewide single nucleotide polymorphism discovery in Atlantic salmon (Salmo salar): validation in wild and farmed American and European populations. Mol Ecol Resour. 2016;16: 1002–1011. doi:10.1111/1755-0998.12503

60. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319: 1100–1104. doi:10.1126/science.1153717

61. Wakeley J. Coalescent Theory: An Introduction. 1st edition. Greenwood Village, Colo: W. H. Freeman; 2008.

62. Alcala Nicolas, Vuilleumier Séverine. Turnover and accumulation of genetic diversity across large time-scale cycles of isolation and connection of populations. Proc R Soc B Biol Sci. 2014;281: 20141369. doi:10.1098/rspb.2014.1369

63. Shafer ABA, Cullingham CI, Côté SD, Coltman DW. Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. Mol Ecol. 2010;19: 4589–4621. doi:10.1111/j.1365-294X.2010.04828.x

64. O'Reilly P, Reimchen TE, Beech R, Strobeck C. MITOCHONDRIAL DNA IN GASTEROSTEUS AND PLEISTOCENE GLACIAL REFUGIUM ON THE QUEEN CHARLOTTE ISLANDS, BRITISH COLUMBIA. Evol Int J Org Evol. 1993;47: 678–684. doi:10.1111/j.1558-5646.1993.tb02122.x

65. Pruett CL, Topp CM, Maley JM, McCracken K, Rohwer S, Birks S, et al. Evidence from the genetics of landbirds for a forested pleistocene glacial refugium in the haida gwaii area. Condor. 2013;115: 725–737. doi:10.1525/cond.2013.120123

66. Geraldes A, Askelson KK, Nikelski E, Doyle FI, Harrower WL, Winker K, et al. Population genomic analyses reveal a highly differentiated and endangered genetic cluster of northern goshawks (Accipiter gentilis laingi) in Haida Gwaii. Evol Appl. 2019;12: 757–772. doi:10.1111/eva.12754

67. Gómez A, Lunt DH. Refugia within Refugia: Patterns of Phylogeographic Concordance in the Iberian Peninsula. In: Weiss S, Ferrand N, editors. Phylogeography of Southern European Refugia: Evolutionary perspectives on the origins and conservation of European biodiversity. Dordrecht: Springer Netherlands; 2007. pp. 155–188. doi:10.1007/1-4020-4904-8_5

68. Brown LR, Moyle PB, Yoshiyama RM. Historical Decline and Current Status of Coho Salmon in California. North Am J Fish Manag. 1994;14: 237–261. doi:10.1577/1548-8675(1994)014<0237:HDACSO>2.3.CO;2

69. Petit RJ, Aguinagalde I, Beaulieu J-L de, Bittkau C, Brewer S, Cheddadi R, et al. Glacial Refugia: Hotspots But Not Melting Pots of Genetic Diversity. Science. 2003;300: 1563–1565. doi:10.1126/science.1083264

70. Waples RS, Teel DJ, Myers JM, Marshall AR. Life-History Divergence in Chinook Salmon: Historic Contingency and Parallel Evolution. Evolution. 2004;58: 386–403. doi:10.1111/j.0014-3820.2004.tb01654.x

71. Burri R. Linked selection, demography and the evolution of correlated genomic landscapes in birds and beyond. Mol Ecol. 2017;26: 3853–3856. doi:10.1111/mec.14167

72. Stankowski S, Chase MA, Fuiten AM, Rodrigues MF, Ralph PL, Streisfeld MA. Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. bioRxiv. 2019; 342352. doi:10.1101/342352

73. Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, et al. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Res. 2015;25: 1656–1665. doi:10.1101/gr.196485.115

74. Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, et al. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J Evol Biol. 2017;30: 1450–1477. doi:10.1111/jeb.13047

75. Burri R. Dissecting differentiation landscapes: a linked selection's perspective. J Evol Biol. 2017;30: 1501–1505. doi:10.1111/jeb.13108

76. Rougemont Q, Bernatchez L. The demographic history of Atlantic salmon (Salmo salar) across its distribution range reconstructed from approximate Bayesian computations*. Evolution. 2018;72: 1261–1277. doi:10.1111/evo.13486

77. Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, et al. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. Nat Commun. 2016;7: 13195. doi:10.1038/ncomms13195

78. Van Doren BM, Campagna L, Helm B, Illera JC, Lovette IJ, Liedvogel M. Correlated patterns of genetic diversity and differentiation across an avian family. Mol Ecol. 2017;26: 3982–3997. doi:10.1111/mec.14083

79. Macqueen DJ, Johnston IA. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc R Soc B Biol Sci. 2014;281. doi:10.1098/rspb.2013.2881

80. Kodama M, Brieuc MSO, Devlin RH, Hard JJ, Naish KA. Comparative mapping between Coho Salmon (Oncorhynchus kisutch) and three other salmonids suggests a role for

chromosomal rearrangements in the retention of duplicated regions following a whole genome duplication event. G3 Bethesda Md. 2014;4: 1717–1730. doi:10.1534/g3.114.012294

81. Brieuc MSO, Ono K, Drinan DP, Naish KA. Integration of Random Forest with population-based outlier analyses provides insight on the genomic basis and evolution of run timing in Chinook salmon (Oncorhynchus tshawytscha). Mol Ecol. 2015;24: 2729–2746. doi:10.1111/mec.13211

82. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016;533: 200–205. doi:10.1038/nature17164

83. Sutherland BJG, Rico C, Audet C, Bernatchez L. Sex Chromosome Evolution, Heterochiasmy, and Physiological QTL in the Salmonid Brook Charr Salvelinus fontinalis. G3 Bethesda Md. 2017;7: 2749–2762. doi:10.1534/g3.117.040915

84. Allendorf FW, Bassham S, Cresko WA, Limborg MT, Seeb LW, Seeb JE. Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. J Hered. 2015;106: 217–227. doi:10.1093/jhered/esv015

85. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet. 2007;8: 610–618. doi:10.1038/nrg2146

86. Marsden CD, Vecchyo DO-D, O'Brien DP, Taylor JF, Ramirez O, Vilà C, et al. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc Natl Acad Sci. 2016;113: 152–157. doi:10.1073/pnas.1512501113

87. Roze D, Barton NH. The Hill-Robertson effect and the evolution of recombination. Genetics. 2006;173: 1793–1811. doi:10.1534/genetics.106.058586

88. Zhou Y, Massonnet M, Sanjak JS, Cantu D, Gaut BS. Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. Proc Natl Acad Sci. 2017;114: 11715–11720. doi:10.1073/pnas.1709257114

89. Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L. On the accumulation of deleterious mutations during range expansions. Mol Ecol. 2013;22: 5972–5982. doi:10.1111/mec.12524

90. Peischl S, Dupanloup I, Foucal A, Jomphe M, Bruat V, Grenier J-C, et al. Relaxed Selection During a Recent Human Expansion. Genetics. 2018;208: 763–777. doi:10.1534/genetics.117.300551

91. Laenen B, Tedder A, Nowak MD, Toräng P, Wunder J, Wötzel S, et al. Demography and mating system shape the genome-wide impact of purifying selection in Arabis alpina. Proc Natl Acad Sci. 2018;115: 816–821. doi:10.1073/pnas.1707492115

92. Robinson JA, Brown C, Kim BY, Lohmueller KE, Wayne RK. Purging of Strongly Deleterious Mutations Explains Long-Term Persistence and Absence of Inbreeding Depression in Island Foxes. Curr Biol CB. 2018;28: 3487-3494.e4. doi:10.1016/j.cub.2018.08.066

93. Kim BY, Huber CD, Lohmueller KE. Deleterious variation shapes the genomic landscape of introgression. PLOS Genet. 2018;14: e1007741. doi:10.1371/journal.pgen.1007741

94. Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. Genome Biol. 2007;8: R18. doi:10.1186/gb-2007-8-2-r18

95. Rodgers-Melnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, Mitchell SE, et al. Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc Natl Acad Sci. 2015;112: 3823–3828. doi:10.1073/pnas.1413864112

96. Torres R, Stetter MG, Hernandez RD, Ross-Ibarra J. The temporal dynamics of background selection in non-equilibrium populations. bioRxiv. 2019; 618389. doi:10.1101/618389

97. Moore J-S, Harris LN, Le Luyer J, Sutherland BJG, Rougemont Q, Tallman RF, et al. Genomics and telemetry suggest a role for migration harshness in determining overwintering habitat choice, but not gene flow, in anadromous Arctic Char. Mol Ecol. 2017;26: 6784–6800. doi:10.1111/mec.14393

98. Catchen JM, Hohenlohe PA, Bernatchez L, Funk WC, Andrews KR, Allendorf FW. Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. Mol Ecol Resour. 2017;17: 362–365. doi:10.1111/1755-0998.12669

99. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinforma Oxf Engl. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330

100. Goudet J. hierfstat, a package for r to compute and test hierarchical F-statistics. Mol Ecol Notes. 2005;5: 184–186. doi:10.1111/j.1471-8286.2004.00828.x

101. Frichot E, François O. LEA: An R package for landscape and ecological association studies. Methods Ecol Evol. 2015;6: 925–929. doi:10.1111/2041-210X.12382

102. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664. doi:10.1101/gr.094052.109

103. Dray S, Dufour A-B. The ade4 Package: Implementing the Duality Diagram for Ecologists. J Stat Softw. 2007;22: 1–20.

104. Rougemont Q, Gagnaire P-A, Perrier C, Genthon C, Besnard A-L, Launey S, et al. Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. Mol Ecol. 2017;26: 142–162. doi:10.1111/mec.13664

105. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinforma Oxf Engl. 2018;34: i884–i890. doi:10.1093/bioinformatics/bty560

106. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio. 2013; Available: http://arxiv.org/abs/1303.3997

107. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43: 491. doi:10.1038/ng.806

108. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. New York: Springer-Verlag; 2009. Available: https://www.springer.com/gp/book/9780387981413

109. McVean G, Awadalla P, Fearnhead P. A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences. Genetics. 2002;160: 1231–1241.

110. Martin S. ttps://github.com/simonhmartin/genomics_general. ttps://github.com/simonhmartin/genomics_general.

111. Leroy T, Anselmetti Y, Tilak M-K, Bérard S, Csukonyi L, Gabrielli M, et al. A bird's white-eye view on neosex chromosome evolution. bioRxiv. 2019; 505610. doi:10.1101/505610

112. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PloS One. 2012;7: e46688. doi:10.1371/journal.pone.0046688

113. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015;67: 1–48. doi:10.18637/jss.v067.i01

114. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6: 80–92. doi:10.4161/fly.19695

115. Government of Canada NRC. GEOSCAN Search Results: Fastlink [Internet]. 7 Dec 2015 [cited 26 Jul 2019]. Available: https://geoscan.nrcan.gc.ca/starweb/geoscan/servlet.starweb?path=geoscan/fulle.web&search1=R=214399

Supporting Information Legend

S1 Fig. Linear decrease in genetic diversity when considering πSNP as a function of the distance to the southernmost sample site. Each points represents a sample site and is coloured by region.

S2 Fig. Network of shared and private polymorphisms. The branch (grey) represent shared polymorphism between sample site and are proportional to levels of sharing. Each point represent the number of private polymorphisms and is coloured by region. Computation were based on a sample of size 100 in each region to enable comparison.

S3 Fig. Measure of linkage disequilibrium decay using whole genome sequencing data. LD between pairs of loci was measured as r^2 in windows of 1 million base pairs. Means smoothed value are display for each whole genome samples.

S4 Fig. Patterns of Isolation By Distance. Increasing F_{ST} as a function of the distance to the southermost site. Each points represents a sample site and is coloured by region.

S5 Fig. Summaries of F_{ST} values.
A. F_{ST}-based Hierarchical tree depicting relationship among samples. Colors represent the major region.
B. Heatmap of F_{ST} values among samples ordered from North to South on the X and Y-axis.

S6 Fig. Structure and Admixture inferences.
A. Admixture Barplot obtained from LEA for various K-values.
B. Progressive decrease of LEA cross-entropy criterion. Lower cross-entropy values indicates the number of cluster compatible with the data (here from 30 to 60).

S7 Fig. Multidimensional Scaling (MDS) plot depicting relationship among individuals. Each points represents an individual site and is coloured by region.

S8 Fig. Principal Component Analysis recapitulating the relationship among individuals. The Axis 3 and axis 4 are displayed.

S9 Fig. Treemix results
A. Proportion of variance explained (y-axis) as a function of the number of migration edge (x-axis)
B. Treemix tree infered without gene flow and residuals
C. Residuals for Treemix tree with four migration edges

S10 Fig. Compared Demographic Models
Strict Isolation (SI), Isolation with constant Migration (IM), Ancient Migration (AM) and Secondary Contact (SC). The models shared the following parameters: Tsplit: number of generation of divergence (backwards in time). *Nanc*, N_1, N_2: effective population size of the ancestral population, of the first and second daughter population. M_1 and M_2 represent the effective migration rates per generation with *m* the proportion of population made of migrants from the other population. T*sc* is the number of generations since gene flow started (secondary contact) after a period of isolation. T*am* is the number of generations since the two populations have diverged without gene flow. Each models are declined in alternative version allowing homogeneous or heterogeneous gene flow and homogeneous and heterogeneous effective size to account for the effect of linked selection (affecting *Ne)* and barrier to gene flow (affecting *m*).

S11 Fig. Estimate of effective population size across models with linked selection (2N suffix) and barriers to gene flow (2m suffix). AM2m = Ancient Migration with heterogeneous migration.
AM2N = Ancient Migration with heterogeneous effective population size. IM2m = Isolation with Migration with heterogeneous migration. IM2N2m = Isolation with Migration with heterogeneous migration and with heterogeneous effective population size, SC2m = Secondary Contact with heterogeneous migration. SC2N2m = Secondary Contact with heterogeneous migration and with heterogeneous effective population size.

S12 Fig. SMC++ estimates of effective population sizes for the 11 whole genome samples included.

S13 Fig. SMC++ estimates of divergence time between all possible pairs of joint Site Frequency Spectra based on WGS data. Two estimates of mutation rate were used. One standard estimate of 1.25e-8 µ/bp/generation and a second based on unpublished data.

S14 Fig. Negative Relationship between effective recombination rate (rho) and length of chromosome in bp.

S15 Fig. Genome-wide landscapes in coho salmon with (A) landscape of differentiation, B) landscape of genetic diversity C) divergence, D) recombination E) gene density, plots are averaged over 500kb windows. The figure is the same as figure 5 but all the 30 chromosomes are displayed.

S16 Fig. Correlation between π_N/π_S and effective population size (*Ne*) during postglacial time. Each point represent a sample locality and is color coded by region.

S17 Fig. DAF spectrum of nonsynonymous and putatively deleterious mutation in each region for all samples.

S18 Fig. Significant mean derived allele frequencies of deleterious mutation. Displayed are the mean dervied allele frequencies of polymorphic deleterious sites in each regions +/- 2 standard deviation.
x-axis = chromosome position (in bp) and y-axis = effective recombination landscape. Each red dots represents a candidate deleterious mutation.

S19 Fig. Location of putatively deleterious mutation along the effective recombination landscape.

S20 Fig : Positive correlation between chromosome length and occurence of deleterious variants along chromosome.

Table Legend:

S1 Table: Abbreviation, Region and coordinates (Longitude and Latitude) of each river used in the GBS data with the number of individuals provided (nb. Inds).

S2 Table: River Name, region and coordinates (Longitude and Latitude) of each river used in the whole genome resequencing data.

S3 Table: BST values along with 95% confidence intervals for each river from the GBS data (82 K SNPs). 95% confidence intervals obtained after 1000 bootstraps.

S4 Table: Distribution of singleton in the WGS data. The Thompson sample displays less singelton than southern samples.

S5 Table: model choice results for dadi. AIC and deltaAIC are provided for each pairwise comparison and model. AM = Ancient Migration, IM = Isolation with Migration, SI = Strict Isolation, SC = Secondary Contact, the simplest models assume homogeneous migration and homogeneous effective population size. 2N suffix = heterogeneous effective population size, 2Msuffix = heterogeneous migration.
Model with both suffix assumes that both effective population size and migration are heterogeneous.
Modil with a single suffix assumes that either migration or effective population size are heterogeneous.

S6 Table : Parameter estimates from GBS data obtained under the best demographic model with dadi.
Ne1 and Ne2, effective population size of the compared pair. m1 ← 2 and m2 ← 1, migration from population 2 to population 1 and migration from population 1 into population 2. me12 and me21, effective migration rate estimated in the most differentiated regions of the genome Ts: Split Time of the ancestral population in two population; Tsc: duration of the secondary contact P: proportion of the genome freely exchanged (1-P provides the proportion of the genome non-neutrally exchanged); Q: proportion of the genome with a reduced effective population size due to selection at linked sites; hrf = Hill-Robertson factor representing the reduction of Ne in the region Q with reduced Ne.

S7 Table: PCA loadings for Dxy, Fst and Pi.

S8 Table: A) Spearman correlation association to the comparison in Figure 6.
B) Linear models testing the combined effect of recombination (Rho) and gene density (Gene count).
Interaction terms were not significant for Pi and Dxy and were removed.
S9 Table: spearman correlation obtained between recombination and Dxy, Fst and Pi when considering each possible pairs of river separately (n = 55) or each river independently (for Pi only).

S10 Table: Results of liear models testing the correlation between piN/piS and GC3.

S11 Table: Summary of deleterious variation by region.
1)Derived Allele Frequency (DAF) of deleterious mutation, after averaging by rivers and then by major regional group. 2) Count of deleterious mutations in each rivers and then averaged by major regional group. 3) Number of homozyguous derived deleterious mutations by individual, after averaging by rivers and then by major regional group. 4) Number of heterozygous mutations by individuals, after averaging by rivers and then by major regional group 5) Total load of derived deleterious mutations by individuals, after averaging by rivers and then by major regional group.

S12 Table: Results of Wilcoxon test (Man-Whitney tests) for differences in derived allele frequencies among major groups for a sample of size 100.

S13 Table: Results of Wilcoxon test (Man-Whitney tests) for differences in count of derived homozygous variants and total load among individuals in each region.

S14 Table: SNPeff results classified by categories for the deleterious mutations identified in the GBS data.

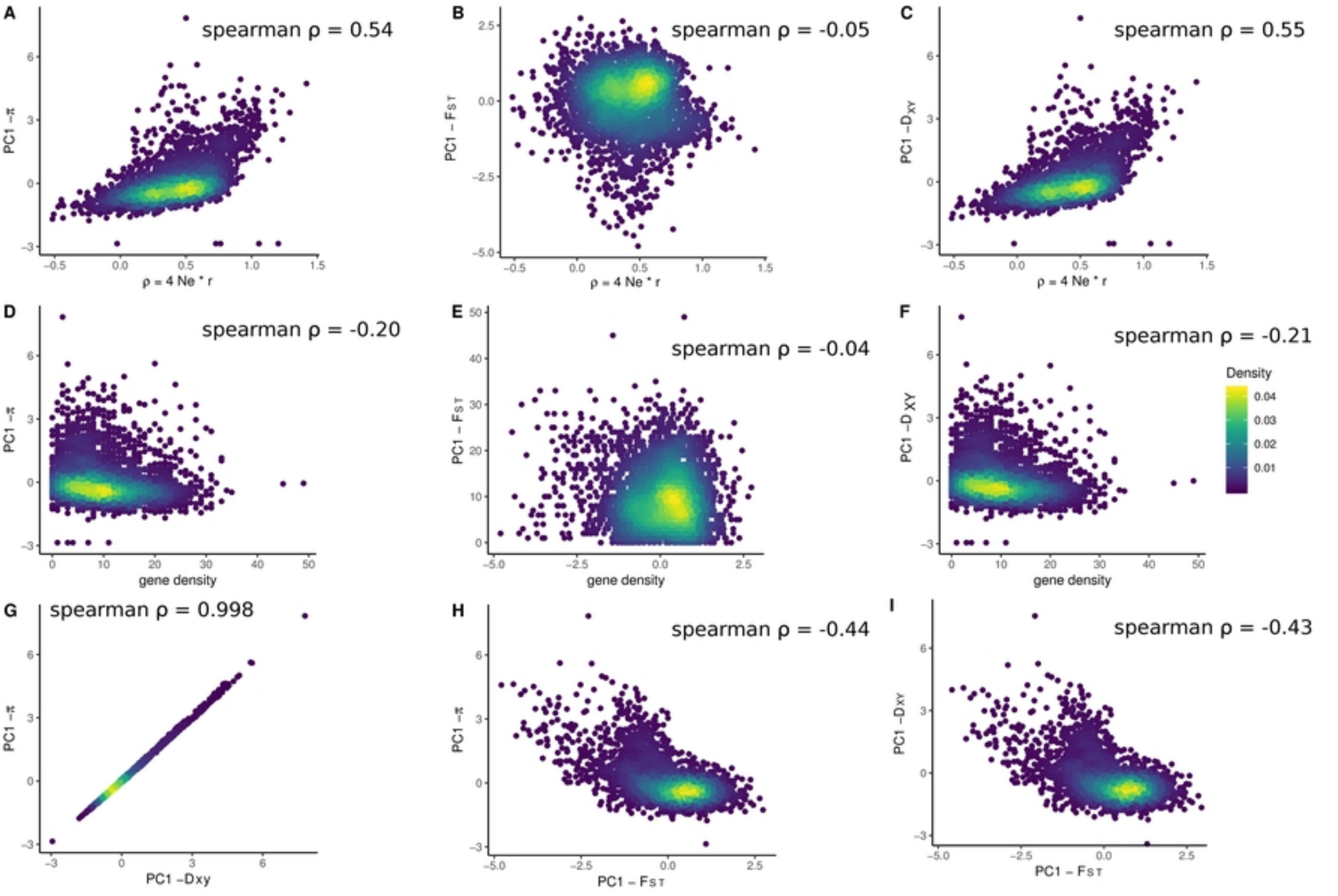

A
spearman ρ = 0.54
PC1 –π
ρ = 4 Ne * r
B
spearman ρ = -0.05
PC1 – FST
ρ = 4 Ne * r
C
spearman ρ = 0.55
PC1 –DXY
ρ = 4 Ne * r
D
spearman ρ = -0.20
PC1 –π
gene density
E
spearman ρ = -0.04
PC1 – FST
gene density
F
spearman ρ = -0.21
PC1 –D XY
gene density
Density
0.04
0.03
0.02
0.01
G
spearman ρ = 0.998
PC1 –π
PC1 –Dxy
H
spearman ρ = -0.44
PC1 –π
PC1 – FST
I
spearman ρ = -0.43
PC1 –DXY
PC1 – FST

Fig 6

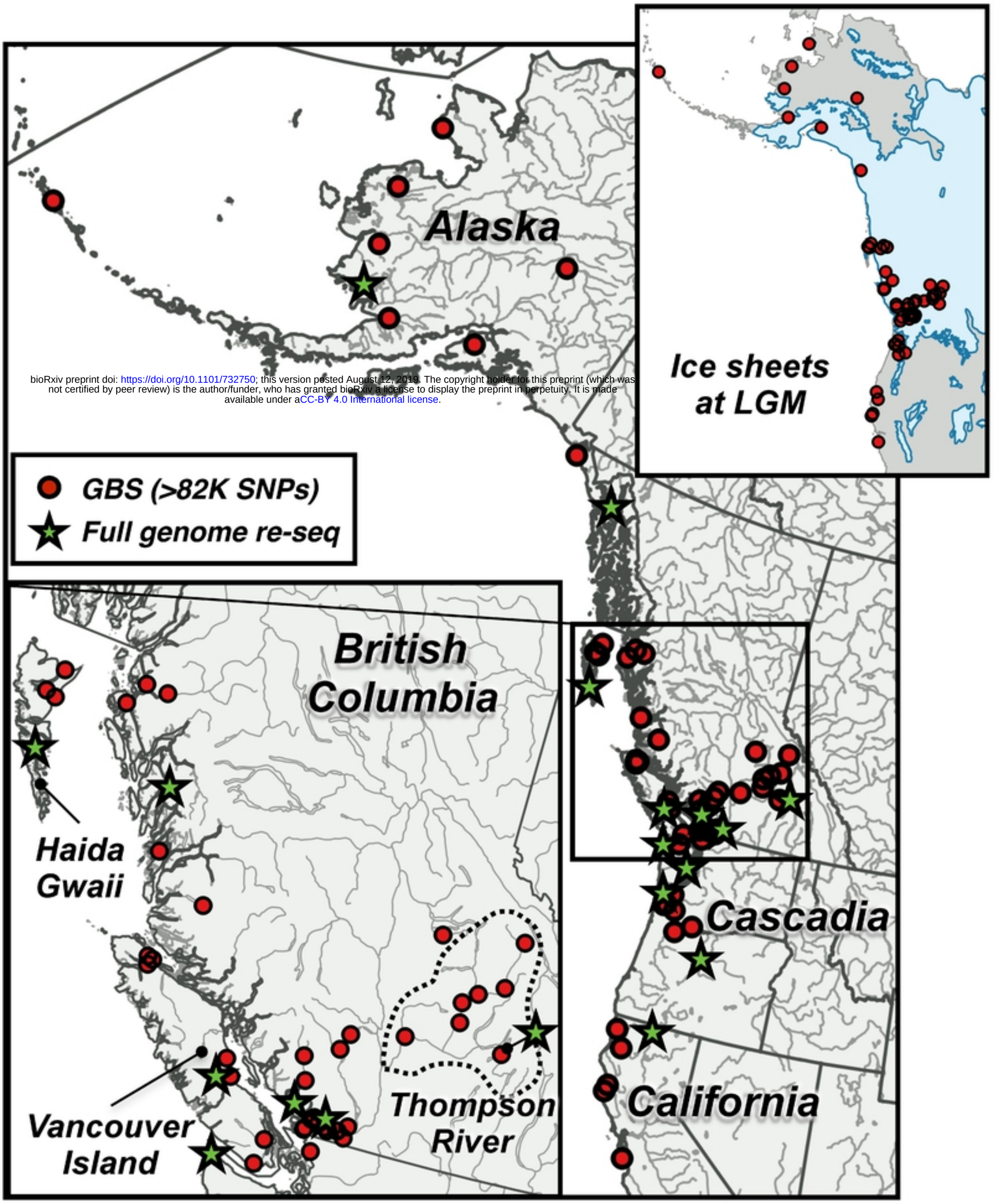
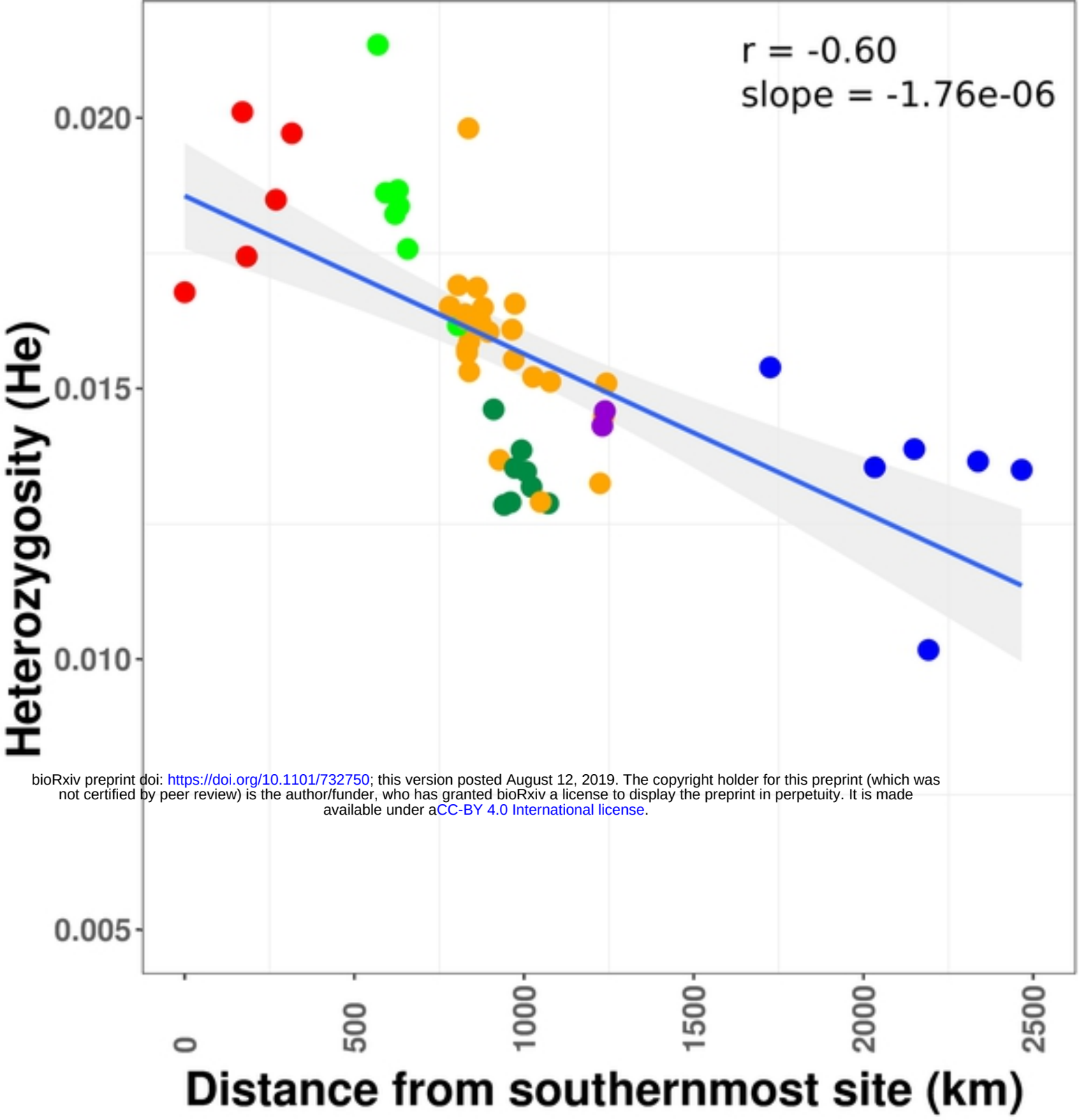


Fig 1

A



B

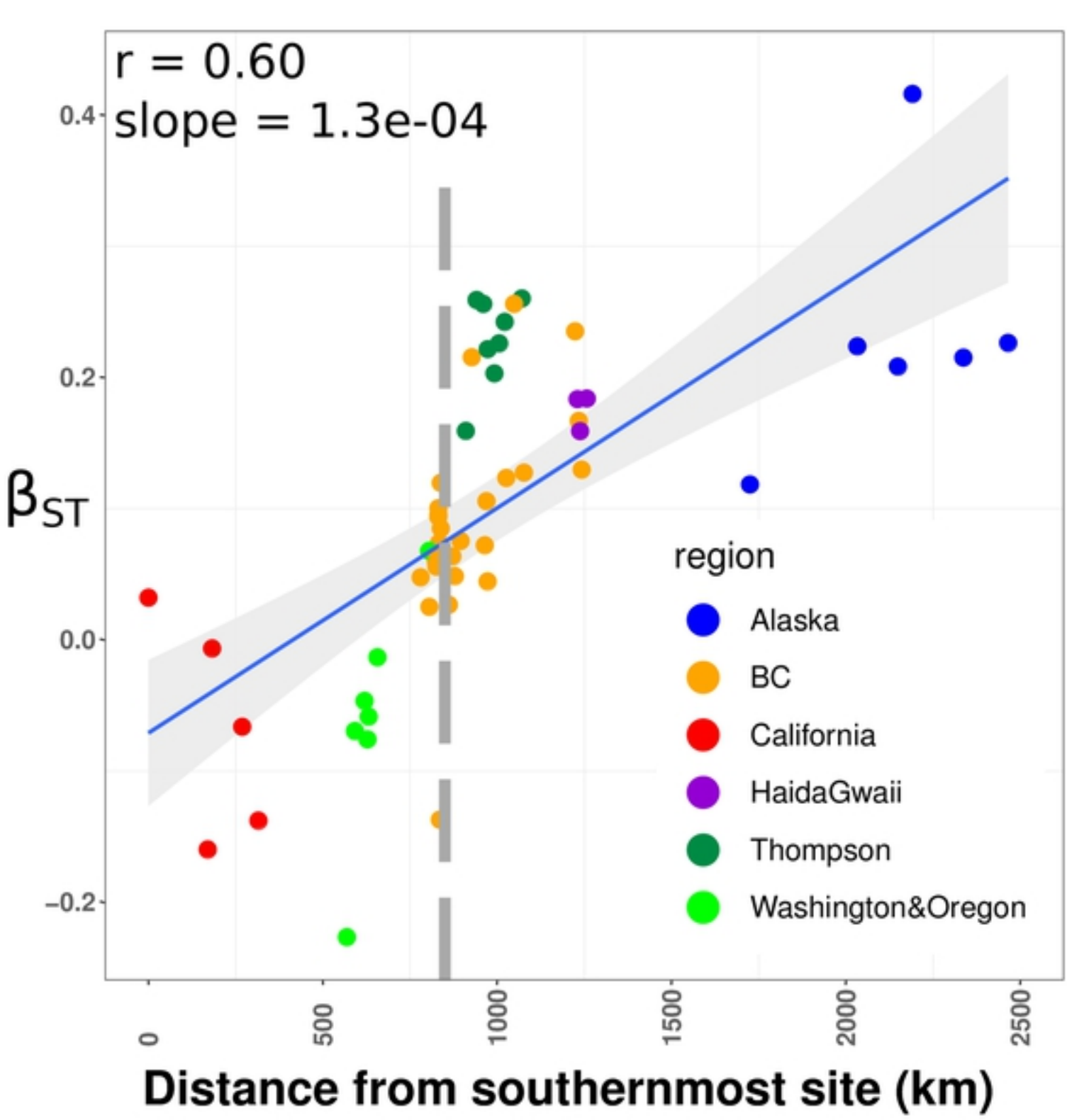
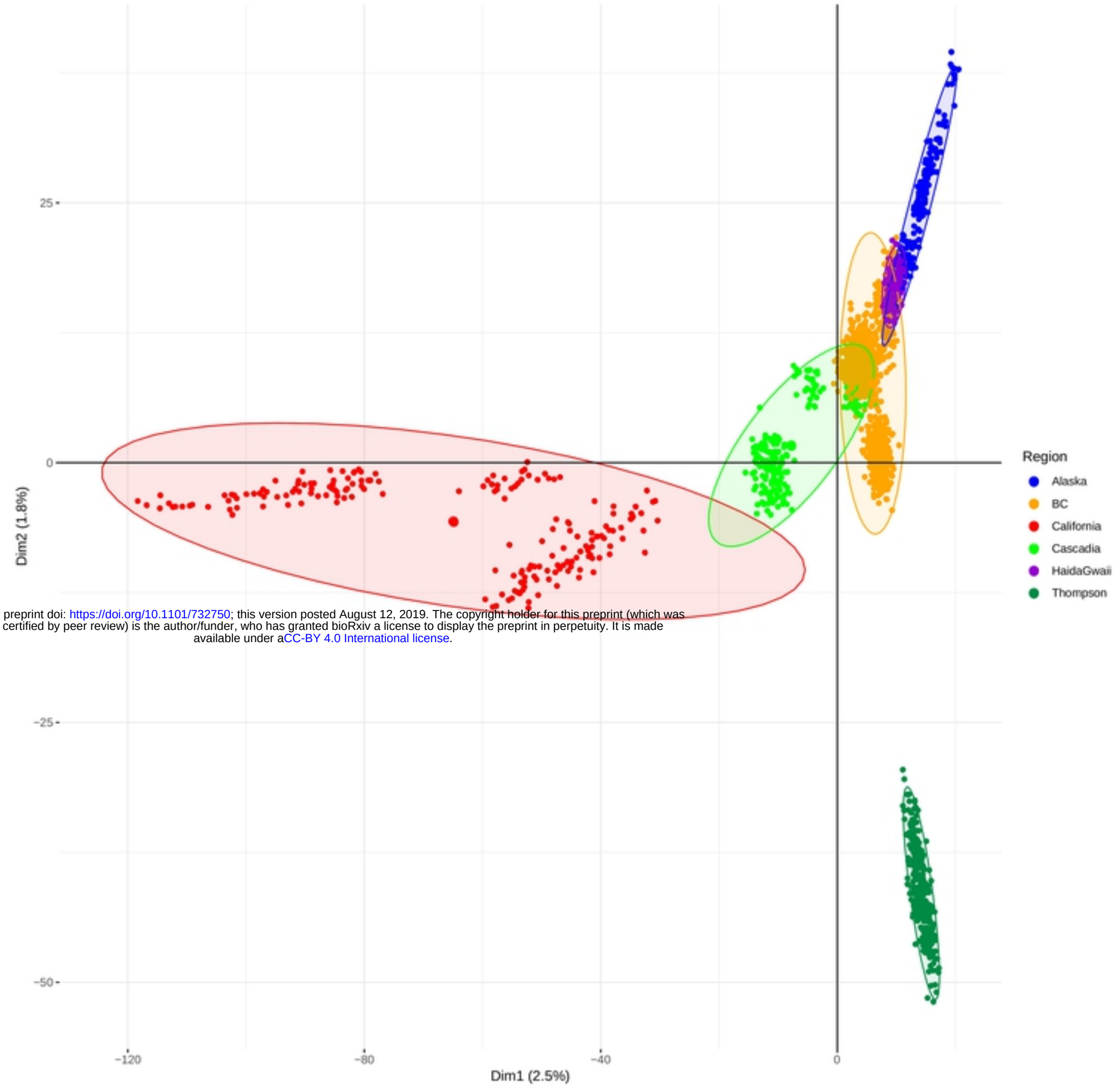


Fig 2

A



B

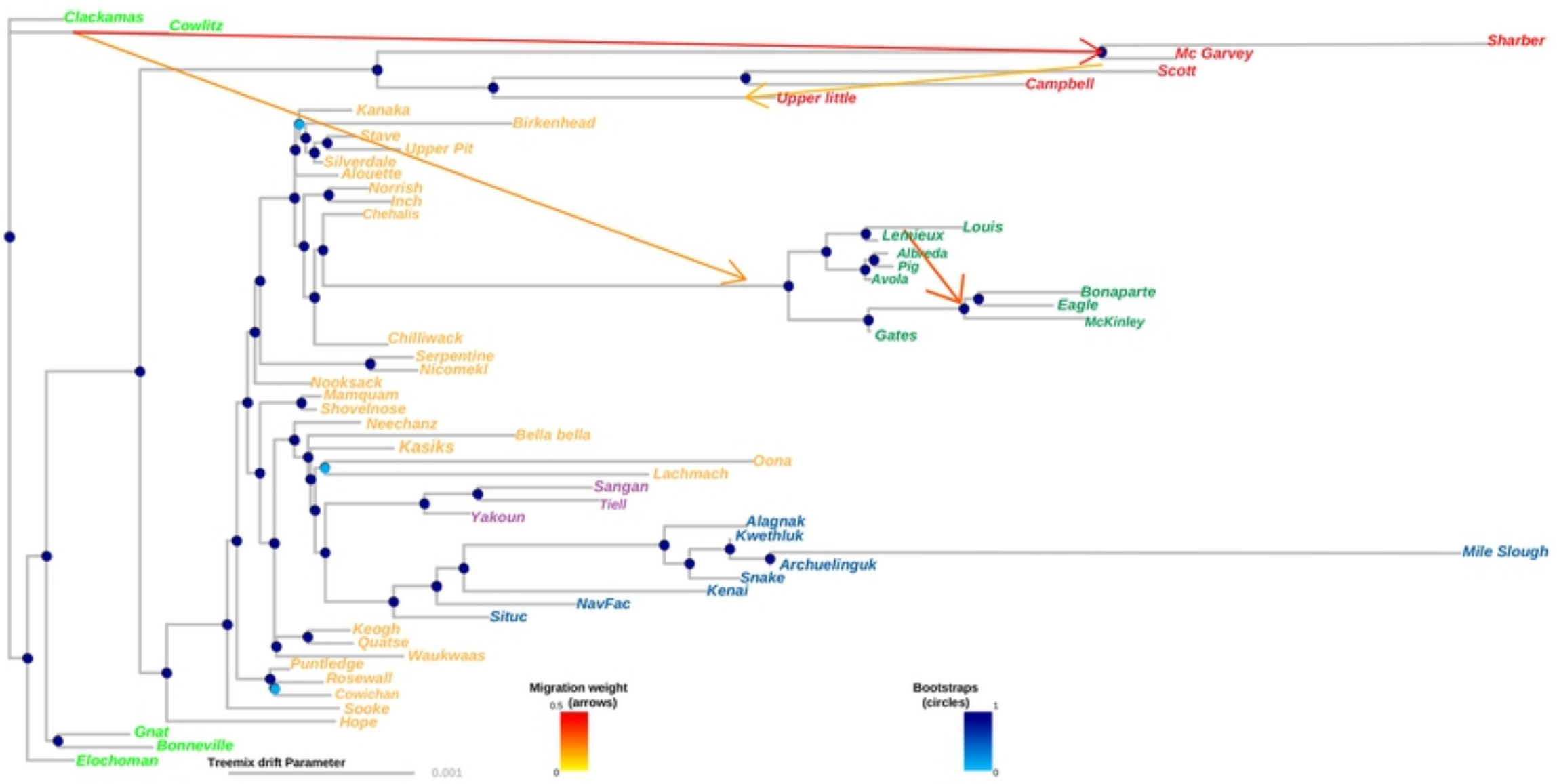


Fig 3

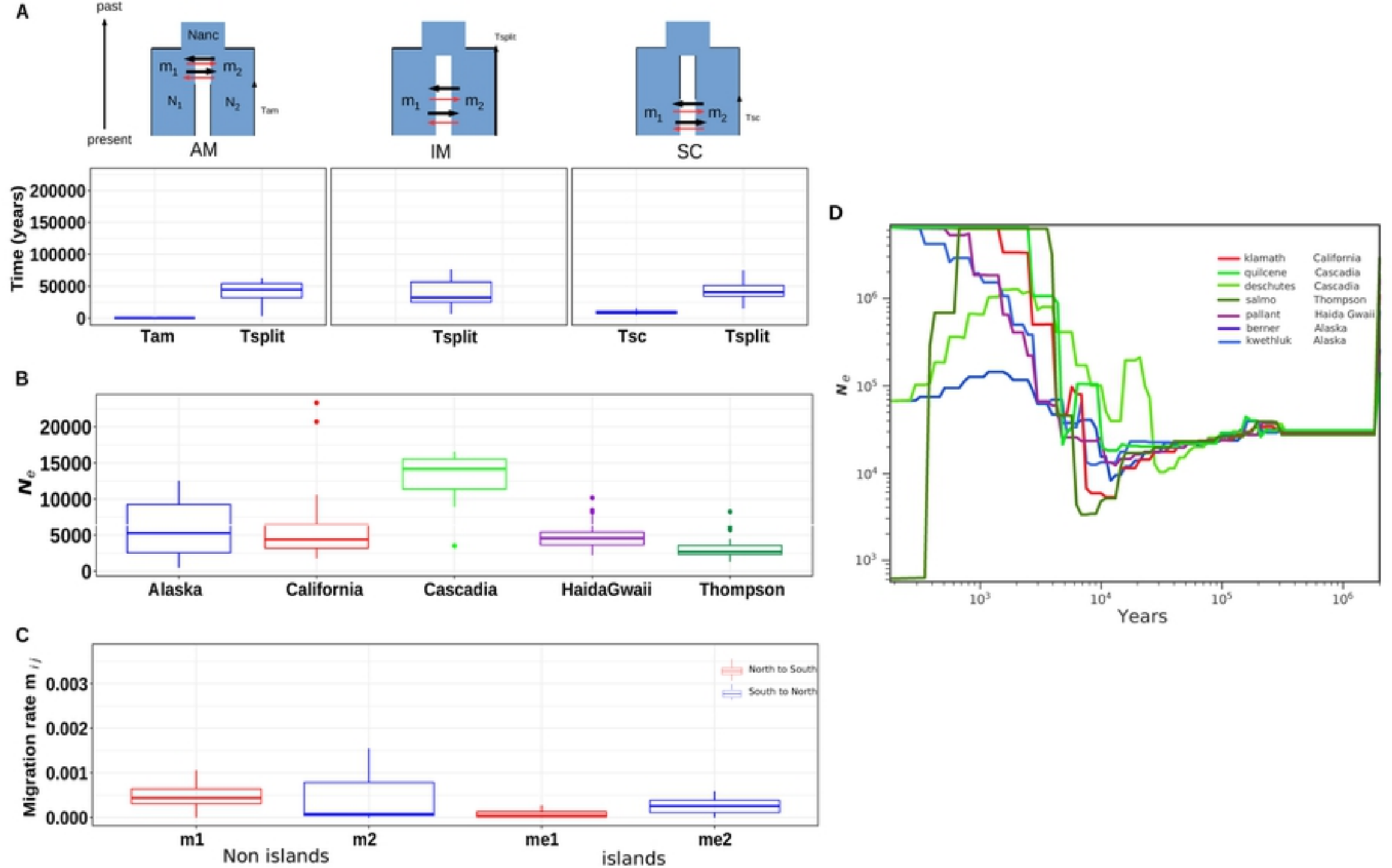

A
past
present
Nanc
m_1
m_2
N_1
N_2
Tam
AM
Tsplit
IM
Tsc
SC
Time (years)
200000
150000
100000
50000
0
Tam
Tsplit
Tsplit
Tsc
Tsplit
B
N_e
20000
15000
10000
5000
0
Alaska
California
Cascadia
HaidaGwaii
Thompson
C
Migration rate m_{ij}
0.003
0.002
0.001
0.000
North to South
South to North
m1
m2
me1
me2
Non islands
islands
D
N_e
10^6
10^5
10^4
10^3
klamath California
quilcene Cascadia
deschutes Cascadia
salmo Thompson
pallant Haida Gwaii
berner Alaska
kwethluk Alaska
10^3
10^4
10^5
10^6
Years

Fig 4

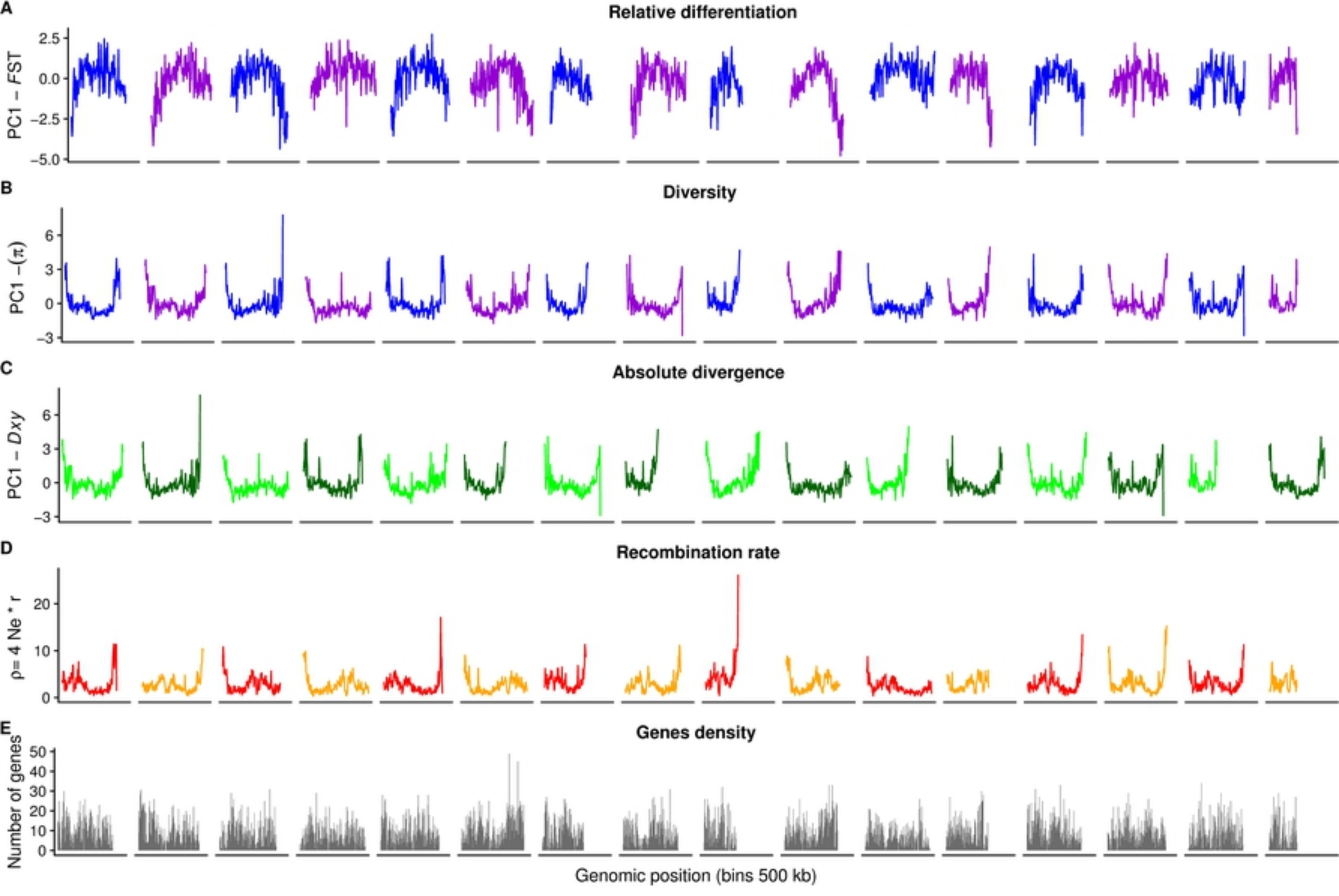

A
Relative differentiation
PC1 – FST
2.5
0.0
−2.5
−5.0
B
Diversity
PC1 –(π)
6
3
0
−3
C
Absolute divergence
PC1 – Dxy
6
3
0
−3
D
Recombination rate
ρ= 4 Ne * r
20
10
0
E
Genes density
Number of genes
50
40
30
20
10
0
Genomic position (bins 500 kb)

Fig 5

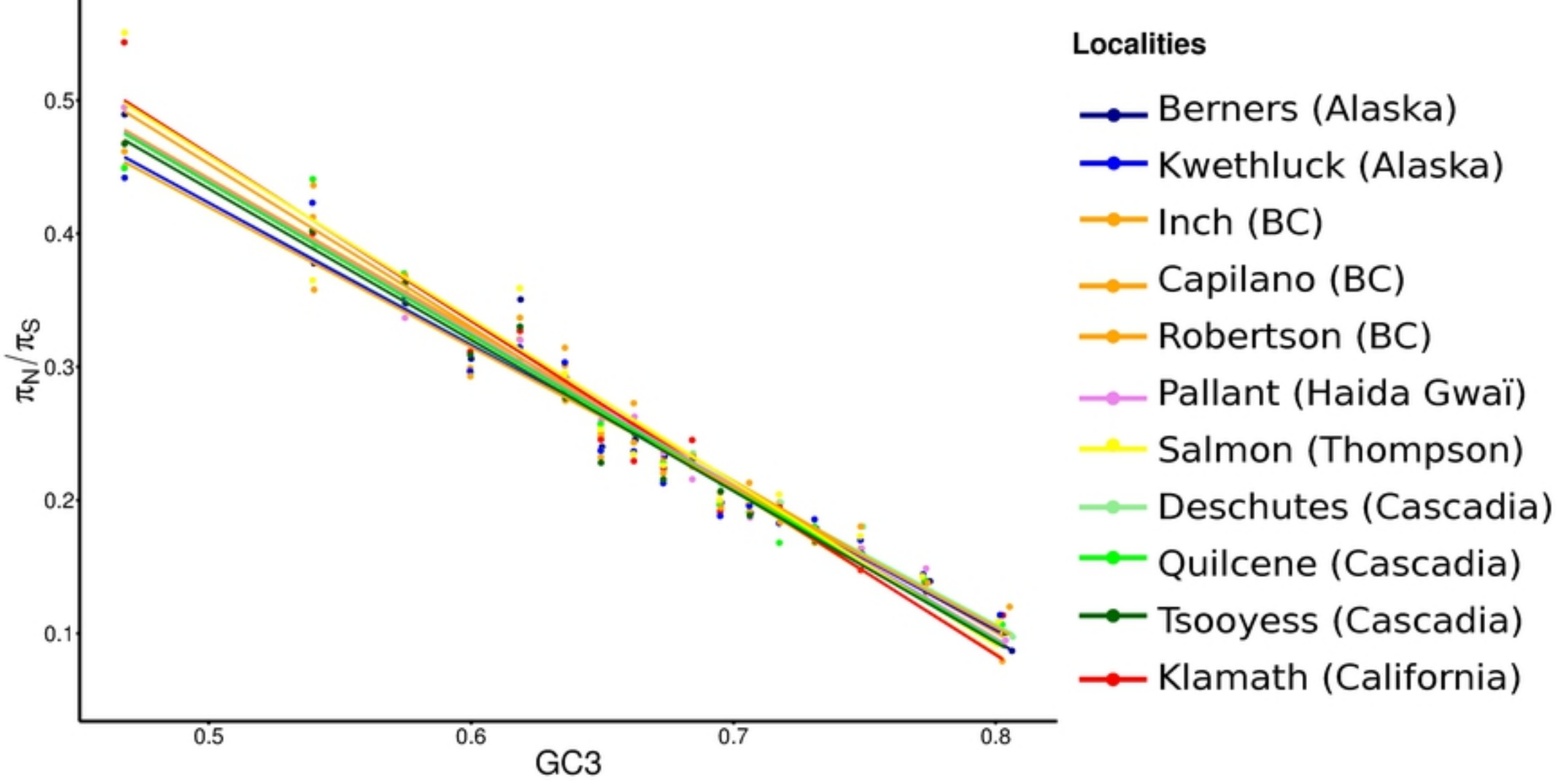

Localities
Berners (Alaska)
Kwethluck (Alaska)
Inch (BC)
Capilano (BC)
Robertson (BC)
Pallant (Haida Gwaï)
Salmon (Thompson)
Deschutes (Cascadia)
Quilcene (Cascadia)
Tsooyess (Cascadia)
Klamath (California)
0.5
0.4
0.3
0.2
0.1
π_N/π_S
0.5
0.6
0.7
0.8
GC3

Fig 7

A
Cascadia
California
BC
HaidaGwaii
Thompson
Alaska
Count
800
600
400
200
0
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency (DAF)
class
Non−synonymous
Deleterious
B
No. of homozygous deleterious alleles
10
5
0
Cascadia
California
BC
HaidaGwaii
Thompson
Alaska
C
Total No. of deleterious alleles
60
40
20
Cascadia
California
BC
HaidaGwaii
Thompson
Alaska
D
Homozygous derived deleterious allele count
6
4
2
0
0
500
1000
1500
2000
2500
Distance from the putative source in Kms

Fig 8