Title: Diversification of large-effect loci in a duplicated genomic region leads to complex phenotypes

Authors: Patrick Barry[1], Natasha Howe[2], Gregory L. Owens[3], Diana Baetscher[1], Katie D'Amelio[1], Scott Vulstek[1], Josh Russell[1], Elizabeth Lee[4], Andrew Barclay[4], Anthony Gharrett[5], David A. Tallmon[6], Wes Larson[1]*

Affiliations:

[1] National Oceanic and Atmospheric Administration; Juneau, AK, 99801, USA

[2] Pacific States Marine Fisheries Commission; Seattle, WA, 98115, USA

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

[4] Alaska Department of Fish and Game; Anchorage, AK, 99518, USA

[5] University of Alaska Fairbanks, Department of Fisheries; Juneau, AK, 99801, USA

[6] University of Alaska Southeast, Biology and Marine Biology Program; Juneau, AK, 99801, USA

*Corresponding author. Email: wes.larson@noaa.gov

Abstract:

Whole genome duplication provides evolutionary opportunities to increase biological complexity and phenotypic diversity by selective retention of duplicated genes. We identify a small genomic region associated with complex phenotypic variation on duplicate copies of the same ancestral chromosome in salmon (Salmonidae). Across six species, three large-effect loci involved in migration timing, reproduction, and maturation are located in this genomic region. Life history and whole genome resequencing data support lineage-specific evolution of polymorphism related to early- and late-migrating phenotypes. The high frequency of transcription factors and lack of non-synonymous mutations suggests this region influences gene expression rather than modifying gene function. Repeated evolution of large-effect loci in multiple salmon species provides evidence of adaptive evolution facilitated by the salmonid specific whole genome duplication.

Main Text:

Whole genome duplication (WGD) events are common throughout eukaryote evolution and have been associated with adaptive radiations and the reduced risk of extinction (*1*). WGDs provide a source of newly duplicated genes (ohnologs) that can evolve novel functions without disrupting the function of previously existing genes (*2–4*). Commonly, rapid deterioration of function and gene loss occurs in one copy of ohnologs through the accumulation of mutations under relaxed purifying selection (*3*). However, the fates of duplicate gene copies are variable; and genes involved in transcriptional regulation, signal transduction, and development (*e.g.*, *Hox* gene cluster) are often retained at higher frequency (*5–8*). Duplicated genes, such as those involved in regulatory processes, could play a major role in shaping phenotypic variation, but few studies of wild populations have identified the evolutionary significance of duplicated genes.

The reshaping of genomes following WGD varies substantially, which influences the evolutionary processes that can create phenotypic diversity. Duplicated genomes often undergo rediploidization, the return to diploid meiotic pairing, through genome rearrangements or sequence divergence to mitigate detrimental effects of polyploidy (*9*, *10*). Rediploidization can occur over tens of millions of years and in parallel with speciation events (*11*). The history of duplication and rediploidization can influence gene function and evolutionary trajectories (*11*). Salmonids provide a useful model to study the evolutionary significance of a WGD and subsequent rediploidization process because the most recent common ancestor underwent a WGD ~106 million years ago (MYA) and has since diversified into >100 species across 10 genera with extensive life history variation (*12*, *13*). Pacific salmon and their close relatives have not completed redipolidization; residual tetrasomic inheritance on eight chromosome arm pairs is conserved across much of Salmoninae (*14–16*). Salmon genomes represent a mosaic of genes, characterized by diverse post-duplication histories; this diversity has likely played a key role in the evolution of complex phenotypes in response to varying demographic, environmental, and ecological regimes (*11*, *13*).

Most life history diversity is encoded by many alleles of low effect spread across the genome (polygenic traits) and few of relatively large effect size (*17*). Previous studies on the genetic architecture underlying phenotypic variation in salmon have focused mostly on polygenic traits with some trait-associated loci described in duplicated regions, potentially supporting the adaptive importance of WGD (*18*, *19*). However, it is difficult to form strong conclusions about evolutionary significance of alleles associated with polygenic trait variation because each site has a small phenotypic impact (but see (*20*)). Recent research has indicated that some complex phenotypes in salmon (*e.g.*, age-at-maturity, anadromy vs. residency, migration timing) are largely controlled by a single or small number of genomic regions (large-effect loci (*21*)), but this research did not identify links between large-effect loci and WGD. Our study identifies two large-effect loci that are associated with migration timing in multiple salmon species and located

on ohnologs. These large-effect loci provide an unprecedented opportunity to study how the genetic architecture underlying biological complexity can evolve from WGD.

Migration timing is associated with a duplicated genomic region

Adult migration timing ('run timing' hereafter), the transition of maturing salmon from marine to freshwater habitats prior to spawning, has extensive impacts on salmon populations and the ecosystems they inhabit. Run timing in salmon can be remarkably consistent across years within a single population; yet incredibly variable across populations and species (*22*). Run timing is synchronized with long-term average environmental conditions to maximize individual lifetime fitness. Within population variation in run timing buffers interannual variability in optimal return timing, while among population variation contributes to the portfolio of life history diversity in salmon that is vital for their resilience despite constantly changing environments (*23*). In addition to individual and population level effects, variation in salmon run timing has large cascading ecological impacts; it influences the spatial and temporal distribution of a large variety of species reliant on the delivery of marine-derived nutrients to freshwater and terrestrial ecosystems (*23*, *24*).

We investigated the genetic basis of run timing in four species of Pacific salmon (*Oncorhynchus sp.*) with whole genome sequencing. Collections within our study, 2-4 per species, encapsulate a range of run timing phenotypes (Table S1): the midpoint of early and late runs are separated by approximately three weeks (pink, *O. gorbuscha*); six weeks (sockeye, *O. nerka*); three months (chum, *O. keta*); and four months (coho, *O. kisutch*). In addition to differences in run timing, each species in the collection represents different scales of spatial overlap, including samples collected from the same stream pool (pink), tributaries at the inlet and outlet of the same lake (coho), two habitat types (beach/stream) within a series of interconnected lakes (sockeye), and populations separated by >1500 km in the same river (chum; Fig. S1).

Analysis of whole genome sequencing data for the four species revealed two primary peaks that diverged between early/late runs shared by two or more species. All species were aligned to the chum salmon genome to simplify genomic comparisons (chromosome numbers refer to this genome (*25*)). Genetic differentiation between runs was estimated with F_{ST}. Mean genome-wide F_{ST} between run timing groups ranged from 0.002 in pink salmon to 0.024 in chum salmon (Table S2). The greatest concentration of highly diverged loci spanned ~250 kb on Chromosome (Chr) 35 and diverged between runs of pink, sockeye, and chum salmon (mean F_{ST} 0.360 - 0.544). The second highest concentration spanned ~750 kb on Chr 29 and diverged between runs of chum and coho salmon (mean F_{ST} 0.315 - 0.709; Fig. 1).

The two chromosome arms that have the peaks of divergence associated with run timing are duplicated copies of the same ancestral chromosome, which arose from the salmonid WGD ~106 MYA (Fig. 2A). The peaks of divergence occurred within the same small ~850 kb region of the

ancestral genome (Fig. 2), which now represents chromosome arms 29a and 35a in chum salmon (*16*). These arms have undergone fusions with other chromosomes and rediploidized since the WGD and no longer recombine with each other (*25*); however chromosome arm 29b displays residual tetrasomic inheritance with Chr 16 (*14*). The area of highest divergence between early and late run salmon on Chr 35 was centered on the leucine-rich repeat containing 9 gene (*lrrc9*) and the peak on Chr 29 was distributed near the estrogen receptor beta gene (*esrb*; Fig. 1). Prior research on sockeye salmon linked variation at *lrrc9* to divergence between stream- and beach-spawning ecotypes (*26*). The duplicated copy of *esrb* (*esrb-L* on Chr 35, rather than the *esrb* gene on Chr 29) has been associated with distinct reproductive strategies (single vs. multiple spawning) in steelhead trout (*O. mykiss* (*27*)). Other genes in this region include *six6a*, a gene linked to age-at-maturity and run timing in Atlantic salmon (*Salmo salar*; (*28*, *29*)) and age-at-maturity in steelhead (*30*), and *syne2*, a gene that diverges between migration runs of herring and cod (*31*, *32*). Additional peaks were present in all species except pink salmon, suggesting this single genomic region may control run timing in pink salmon (Fig. 1). Additional peaks of divergence were generally unique to each species; however, Chr 8 had a shared peak in coho and sockeye and Chr 28 had a shared peak in coho and chum. The peak on Chr 8 was previously identified in sockeye, but there was no clear correlation between genotype and migration phenotype (*33*).

Sequence alignments of the ohnologs of pink, sockeye, coho, and chum salmon indicated that re-establishment of diploid inheritance of the region occurred before speciation of Pacific salmon but also provided evidence of species-specific patterns of diversification. The average pairwise sequence similarity between ohnologs within a given species were significantly greater than the distances among species for an individual ohnolog copy (t-test *lrrc9* $t = 14.5$, $P < 0.001$; *esrb* $t = 19.8$ $P < 0.001$; Fig. 2B-C). These patterns, consistent between genes, are indicative of rediploidization prior to species divergence (*11*). Rediploidization post-species divergence would have resulted in ohnologs grouping by species in gene trees. Within each set of ohnolog pairs, sequence divergence was asymmetric with greater sequence divergence in copies not associated with run timing (*lrrc9-L* and *esrb-L*).

Alleles evolved independently following species divergence

Shared trait-genotype associations between closely related species indicate either shared use of ancestral haplotypes or the independent evolution of functional alleles. We compared sequences of each species to distinguish these scenarios. First, to identify individual genotypes, we used PCAs constrained to the gene boundaries of *lrrc9* (NC_068455.1:28128954-28169980; Fig. 3C). Early- and late-run genotypes were delineated along the first principal component in three groups representing homozygous early individuals (possessing two copies of the early allele, EE), heterozygous individuals (one copy of each allele, EL), and homozygous late individuals (two copies of the late allele, LL; Fig. 3A). Pink, sockeye, and chum salmon had consistent differences in *lrrc9* allele frequencies between early and late run timing phenotypes (Fig. 3A).

However, each species had unique sequence polymorphisms that distinguish its early and late alleles (Fig. 4). All run-timing groups had some fraction of individuals that were heterozygous, with the exception of the chum salmon late-run collection that was entirely homozygous for the late allele (Fig. 3A). On average, EE individuals were 15 times more likely than LL individuals, and 2 times more likely than EL individuals to return early (Fig. S4). Comparisons between EE and LL individuals displayed disparate patterns in the breadth of divergence across species. Across the 310 kb region around *lrrc9* for which any species showed elevated divergence (F_{ST} >0.5*)*, pink salmon had the narrowest peak while sockeye had the widest, which included an additional five genes (Fig. 3C). Of the 1389 variable sites across the three species in this region, most sites were unique to a single species. Only 11 polymorphic sites were shared among multiple species, two of which were highly divergent for sockeye and chum salmon (Table S4).

The early- and late-run coho and chum salmon collections, which have the largest timing differences between runs in our dataset, diverge at *esrb*. Local PCAs, constrained to the highly divergent region around *esrb* (NC_068449.1:25414060-25501622; Fig. 3D), were used to assign genotypes within each species (Fig. 3B). No late allele homozygous (LL) and very few heterozygous (EL) chum salmon were observed in fish that returned early (Fig. 3B). Chum with EE genotypes were 1.3 times more likely to return early compared to EL fish. The lack of homozygous early genotypes (EE) in late returning fish was consistent between *esrb* and *lrrc9*. Coho salmon were nearly fixed for alternate *esrb* alleles, with only one heterozygote observed in the late run (Fig. 3B). Fixed differences at *esrb* for chum and coho between EE and LL imply a very large effect size, but the relative risk was undefined because the homozygous genotypes were entirely segregated by run timing (Fig. S4B). Patterns of divergence between EE and LL genotypes in the 1.4 Mb region around *esrb* differed by species (Fig. 3D). Chum salmon displayed a narrow peak centered around *esrb* and *syne2* compared to the broad peak observed for coho salmon. Consistent with the independent diversification and evolution of run timing alleles, only nine of the 5314 polymorphic sites around *esrb* were shared between chum and coho salmon and none were highly divergent in both species (Table S4).

Within and among species comparisons of both *lrrc9* and *esrb* suggest that despite associations with the same trait in the same gene, the evolution of early and late alleles occurred independently in each lineage. Phylogenetic trees indicate that alleles at both *lrrc9* and *esrb* are more similar within than among species, and patterns of genomic divergence at each locus also vary among species, with few shared polymorphisms (Fig. 4, Table S4). This pattern of repeated genetic divergence after speciation is consistent with other large-effect loci in salmon (*e.g., greb1l, six6a; 21*). The higher divergence among rather than within species, similar to *greb1l* (*34*), indicates that the split between alleles likely occurred after the speciation of the five Pacific salmon ~6–20 MYA (*35*). The selective landscape and demographic characteristics likely shaped lineage-specific genetic architecture underlying run timing variation; higher rates of gene flow between populations with different selective optima favor few loci of large effect (*36*). A single

locus was identified in pink salmon that spawn in the same stretch of creek with peak run timing separated by weeks, whereas three shared peaks on independently assorting chromosomes were identified in chum salmon where spawning is separated by ~1500 km and months. Our example illustrates how additional phenotypic complexity emerges from the ancestrally duplicated chromosomal region among multiple species with substantially different scales of run timing differences.

The duplicated genomic region is a putative master regulatory region

We showed that variation at a small duplicated region can influence complex migration phenotypes across four species of salmon. This region makes up only ~0.05% of the salmon genome but includes approximately half of the large-effect loci discovered in salmon to date and three of the four large-effect loci shared across multiple salmon species (all but *greb1l*). In addition to *lrrc9* and *esrb*, this region includes multiple *six* genes, transcription factors that control development and organogenesis in many animals (*37*), and genes that encode estrogen receptors, which function as hormone-dependent transcription factors and mediate the effects of female sex hormones (*38*). Although the fate of genes after WGD can include diversification or loss of function (*3*, *39*), the function of the duplicated region containing *lrrc9* and *esrb* appears not to have diverged because both chromosomal regions are involved in run timing polymorphism. It may instead reflect the co-opting of a secondary function of the gene region when it existed in singleton form (*40*). The presence of two large-effect loci associated with run timing in this duplicated region provides evidence of the adaptive importance of the WGD in salmon.

Gene function predisposes ohnologs to be retained or restored to singleton status following WGD (*41*). Copies of genes involved in transcriptional regulation or protein complexes are thought to be retained immediately following WGDs because variation in the number of these genes would cause a dosage imbalance that would disrupt the stoichiometry between gene products (*3*, *41*). The retention of these genes likely does not initially alter phenotypes, but, as selection to maintain dosage balance decreases over time, the duplicated gene copies can evolve new functions that can increase biological complexity (*41*). Our data are consistent with diversification at *lrrc9* and *esrb*. The evolutionary path from WGD to the diverse migration timing phenotypes that manifest across species might have progressed as follows: (1) *esrb* and *lrrc9* were descended from the same ancestral chromosome during WGD, (2) the chromosomes were functionally tetraploid and recombined, (3) divergence between chromosomes increased and the two chromosomes became functionally diploid, (4) salmon species diverged, and (5) new alleles evolved at *esrb* and *lrrc9* that contributed to the formation and maintenance of complex migration and reproduction phenotypes in multiple species. While it is difficult to speculate about the functional differences between *lrrc9* and *esrb*, or among alleles at these loci, it is clear that new variation at these loci, or in the small genomic region surrounding them, has arisen at least seven times in association with salmon life history variation (at *esrb* in chum and coho

salmon; at *lrrc9* in sockeye, pink, and chum salmon, and at *six6a* in Atlantic salmon and steelhead trout). The repeated association of this region with complex phenotypes emphasizes its evolutionary importance.

The ~800 kb duplicated region identified here contains two large-effect loci associated with run timing which likely regulate gene expression. The absence of shared polymorphisms across species suggests that this region evolved independently and selection repeatedly promoted the formation of new phenotypes. The few non-synonymous mutations and highest associations with phenotypes in intergenic regions observed among species, is consistent with other large-effect loci in Pacific and Atlantic salmon (*greb1l* and *vgll3*, respectively). This suggests that much of the variation at large-effect loci in salmon that has been identified to date is associated with mutations in non-coding regions near transcription factors that influence gene expression rather than modify protein structure. Given their association with complex phenotypes that involve multiple physiological and behavioral processes, these large-effect loci likely regulate multiple genes. While research on the functional significance of large-effect loci in salmon is limited, recent gene expression data from *vgll3* in Atlantic salmon support the hypothesis that variation in small intergenic regions can influence a large number of regulatory pathways and physiological processes (*42*). Alternatively, this region could encode an unrecognized gene not included in genome annotations such as a long noncoding RNA (lncRNA), which are transcribed from the genome but not translated into proteins and can play a role in regulating gene expression (*43*). If ohnologs of a single regulatory gene, like a lncRNA, spanned the region we identified, then the nearby but separate association signals we observe across species may reflect causative variation on duplicated copies of a single gene.

It has been hypothesized that WGD plays a large role in shaping biological complexity within and among species (*41*). However, most phenotypic variation is controlled by many loci of small effect which makes it difficult to investigate how WGD influences specific phenotypes and requires surveys of hundreds-to-thousands of genes (*e.g.*, *20*). Here, we provide concrete evidence that variation arising from a WGD was integral in the evolution of phenotypic complexity by linking evolutionary history and life history variation to two duplicated large-effect loci originating from the same ancestral genomic region. Comparing divergence between populations of four species of salmon with discrete migration timings, we demonstrate that WGD guided the genomic architecture and subsequently the evolution of large-effect loci linked to important adaptive phenotypic variation. The large-effect loci described here have a profound effect on salmon run timing and influence their resilience to environmental variation. Future research into the specific patterns of divergence and function of these loci will provide important information on the evolutionary role of WGDs as well as informing the management and conservation of salmon.

References

1. K. D. Crow, G. P. Wagner, What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? *Mol. Biol. Evol.* **23**, 887–892 (2006).

2. Y. Van de Peer, S. Maere, A. Meyer, The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725–732 (2009).

3. J. A. Birchler, H. Yang, The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *Plant Cell* **34**, 2466–2474 (2022).

4. M. Lynch, Genomics. Gene duplication and evolution. *Science* **297**, 945–947 (2002).

5. P. W. H. Holland, J. Garcia-Fernàndez, N. A. Williams, A. Sidow, Gene duplications and the origins of vertebrate development. *Development* **1994**, 125–133 (1994).

6. T. Blomme, K. Vandepoele, S. De Bodt, C. Simillion, S. Maere, Y. Van de Peer, The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* **7**, R43 (2006).

7. S. Maere, S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu, M. Kuiper, Y. Van de Peer, Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci.* **102**, 5454–5459 (2005).

8. F. G. Brunet, H. R. Crollius, M. Paris, J.-M. Aury, P. Gibert, O. Jaillon, V. Laudet, M. Robinson-Rechavi, Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.* **23**, 1808–1816 (2006).

9. L. Comai, The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846 (2005).

10. A. L. Hufton, G. Panopoulou, Polyploidy and genome restructuring: a variety of outcomes. *Genomes Evol.* **19**, 600–606 (2009).

11. F. M. Robertson, M. K. Gundappa, F. Grammes, T. R. Hvidsten, A. K. Redmond, S. Lien, S. A. M. Martin, P. W. H. Holland, S. R. Sandve, D. J. Macqueen, Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biol.* **18**, 111 (2017).

12. S. Lien, B. F. Koop, S. R. Sandve, J. R. Miller, M. P. Kent, T. Nome, T. R. Hvidsten, J. S. Leong, D. R. Minkley, A. Zimin, F. Grammes, H. Grove, A. Gjuvsland, B. Walenz, R. A. Hermansen, K. von Schalburg, E. B. Rondeau, A. Di Genova, J. K. A. Samy, J. Olav Vik, M. D. Vigeland, L. Caler, U. Grimholt, S. Jentoft, D. Inge Våge, P. de Jong, T. Moen, M. Baranski, Y. Palti, D. R. Smith, J. A. Yorke, A. J. Nederbragt, A. Tooming-Klunderud, K. S. Jakobsen, X. Jiang, D. Fan, Y. Hu, D. A. Liberles, R. Vidal, P. Iturra, S. J. M. Jones, I. Jonassen, A. Maass, S. W. Omholt, W. S. Davidson, The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205 (2016).

13. M. K. Gundappa, T.-H. To, L. Grønvold, S. A. M. Martin, S. Lien, J. Geist, D. Hazlerigg, S. R. Sandve, D. J. Macqueen, Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of Salmonid evolution. *Mol. Biol. Evol.* **39**, msab310 (2022).

14. R. K. Waples, L. W. Seeb, J. E. Seeb, Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (Oncorhynchus keta). *Mol. Ecol. Resour.* **16**, 17–28 (2016).

15. F. W. Allendorf, S. Bassham, W. A. Cresko, M. T. Limborg, L. W. Seeb, J. E. Seeb, Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived Salmonid fishes. *J. Hered.* **106**, 217–227 (2015).

16. B. J. G. Sutherland, T. Gosselin, E. Normandeau, M. Lamothe, N. Isabel, C. Audet, L. Bernatchez, Salmonid chromosome evolution as revealed by a novel method for comparing RADseq linkage maps. *Genome Biol. Evol.* **8**, 3600–3617 (2016).

17. H. A. Orr, The population genetics of adaptation: the distribution of factors fixed during evolution. *Evolution* **52**, 935–949 (1998).

18. M. A. Campbell, E. C. Anderson, J. C. Garza, D. E. Pearse, Polygenic basis and the role of genome duplication in adaptation to similar selective environments. *J Hered* **112**, 614–625 (2021).

19. S. J. Salisbury, G. R. McCracken, R. Perry, D. Keefe, K. K. S. Layton, T. Kess, C. M. Nugent, J. S. Leong, I. R. Bradbury, B. F. Koop, M. M. Ferguson, D. E. Ruzzante, Limited genetic parallelism underlies recent, repeated incipient speciation in geographically proximate populations of an Arctic fish (Salvelinus alpinus). *Mol. Ecol.* **29**, 4280–4294 (2020).

20. K. E. Dougan, A. J. Bellantuono, T. Kahlke, R. M. Abbriano, Y. Chen, S. Shah, C. Granados-Cifuentes, M. J. H. van Oppen, D. Bhattacharya, D. J. Suggett, M. Rodriguez-Lanetty, C. X. Chan, Whole-genome duplication in an algal symbiont bolsters coral heat tolerance. *Sci. Adv.* **10**, eadn2218.

21. R. S. Waples, M. J. Ford, K. Nichols, M. Kardos, J. Myers, T. Q. Thompson, E. C. Anderson, I. J. Koch, G. McKinney, M. R. Miller, K. Naish, S. R. Narum, K. G. O'Malley, D. E. Pearse, G. R. Pess, T. P. Quinn, T. R. Seamons, A. Spidle, K. I. Warheit, S. C. Willis, Implications of large-effect loci for conservation: A review and case study with Pacific salmon. *J. Hered.* **113**, 121–144 (2022).

22. T. P. Quinn, *The Behavior and Ecology of Pacific Salmon and Trout* (University of Washington Press, ed. 2, 2018).

23. D. E. Schindler, R. Hilborn, B. Chasco, C. P. Boatright, T. P. Quinn, L. A. Rogers, M. S. Webster, Population diversity and the portfolio effect in an exploited species. *Nature* **465**, 609–612 (2010).

24. S. M. Gende, R. T. Edwards, M. F. Willson, M. S. Wipfli, Pacific salmon in aquatic and terrestrial ecosystems: Pacific salmon subsidize freshwater and terrestrial ecosystems through several pathways, which generates unique management and conservation issues but also provides valuable research opportunities. *BioScience* **52**, 917–928 (2002).

25. E. B. Rondeau, K. A. Christensen, H. A. Johnson, D. Sakhrani, C. A. Biagi, M. Wetklo, C. A. Despins, R. A. Leggatt, D. R. Minkley, R. E. Withler, T. D. Beacham, B. F. Koop, R. H. Devlin, Insights from a chum salmon (Oncorhynchus keta) genome assembly regarding whole-genome duplication and nucleotide variation influencing gene function. *G3 GenesGenomesGenetics* **13**, jkad127 (2023).

26. A. J. Veale, M. A. Russello, An ancient selective sweep linked to reproductive life history evolution in sockeye salmon. *Sci Rep* **7**, 1747 (2017).

27. S. Willis, J. Stephenson, A. Pierce, L. Medeiros, L. Jenkins, D. R. Hatch, S. Narum, A genomic region associated with iteroparous spawning phenology is linked with age-at-maturity in female steelhead trout. *Evol. Appl.* **17**, e13622 (2024).

28. E. Cauwelier, J. Gilbey, J. Sampayo, L. Stradmeyer, S. J. Middlemas, Identification of a single genomic region associated with seasonal river return timing in adult Scottish Atlantic salmon (Salmo salar), using a genome-wide association study. *Can. J. Fish. Aquat. Sci.* **75**, 1427–1435 (2018).

29. N. J. Barson, T. Aykanat, K. Hindar, M. Baranski, G. H. Bolstad, P. Fiske, C. Jacq, A. J. Jensen, S. E. Johnston, S. Karlsson, M. Kent, T. Moen, E. Niemelä, T. Nome, T. F. Næsje, P. Orell, A. Romakkaniemi, H. Sægrov, K. Urdal, J. Erkinaro, S. Lien, C. R. Primmer, Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature* **528**, 405–408 (2015).

30. C. D. Waters, A. Clemento, T. Aykanat, J. C. Garza, K. A. Naish, S. Narum, C. R. Primmer, Heterogeneous genetic basis of age at maturity in salmonid fishes. *Mol. Ecol.* **30**, 1435–1456 (2021).

31. G. V. Clucas, R. N. Lou, N. O. Therkildsen, A. I. Kovach, Novel signals of adaptive genetic variation in northwestern Atlantic cod revealed by whole-genome sequencing. *Evol. Appl.* **12**, 1971–1987 (2019).

32. S. Lamichhaney, A. P. Fuentes-Pardo, N. Rafati, N. Ryman, G. R. McCracken, C. Bourne, R. Singh, D. E. Ruzzante, L. Andersson, Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. *Proc. Natl. Acad. Sci.* **114**, E3452–E3461 (2017).

33. W. A. Larson, T. H. Dann, M. T. Limborg, G. J. McKinney, J. E. Seeb, L. W. Seeb, Parallel signatures of selection at genomic islands of divergence and the major histocompatibility complex in ecotypes of sockeye salmon across Alaska. *Mol. Ecol.* **28**, 2254–2271 (2019).

34. N. F. Thompson, E. C. Anderson, A. J. Clemento, M. A. Campbell, D. E. Pearse, J. W. Hearsey, A. P. Kinziger, J. C. Garza, A complex phenotype in salmon controlled by a simple change in migratory timing. *Science* **370**, 609–613 (2020).

35. D. R. Montgomery, Coevolution of the Pacific salmon and Pacific Rim topography. *Geology* **28**, 1107–1110 (2000).

36. S. Yeaman, M. C. Whitlock, The genetic architecture of adaptation under migration-selection balance. *Evol. Int. J. Org. Evol.* **65**, 1897–1911 (2011).

37. K. Kawakami, S. Sato, H. Ozaki, K. Ikeda, Six family genes–structure and function as transcription factors and their roles in development. *Bioessays* **22**, 616–626 (2000).

38. S. Nilsson, J.-Å. Gustafsson, Biological role of estrogen and estrogen receptors. *Crit. Rev. Biochem. Mol. Biol.* **37**, 1–28 (2002).

39. M. A. Campbell, M. C. Hale, G. J. McKinney, K. M. Nichols, D. E. Pearse, Long-term conservation of ohnologs through partial tetrasomy following whole-genome duplication in Salmonidae. *G3 GenesGenomesGenetics* **9**, 2017–2028 (2019).

40. G. C. Conant, K. H. Wolfe, Turning a hobby into a job: How duplicated genes find new functions. *Nat. Rev. Genet.* **9**, 938–950 (2008).

41. Y. Van de Peer, E. Mizrachi, K. Marchal, The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).

42. J.-P. Verta, J. E. Moustakas-Verho, I. Donner, M. Frapin, A. Ruokolainen, P. V. Debes, J. Erkinaro, C. R. Primmer, A complex mechanism translating variation of a simple genetic architecture into alternative life histories. *Proc. Natl. Acad. Sci.* **121**, e2402386121 (2024).

43. J. S. Mattick, P. P. Amaral, P. Carninci, S. Carpenter, H. Y. Chang, L.-L. Chen, R. Chen, C. Dean, M. E. Dinger, K. A. Fitzgerald, T. R. Gingeras, M. Guttman, T. Hirose, M. Huarte, R. Johnson, C. Kanduri, P. Kapranov, J. B. Lawrence, J. T. Lee, J. T. Mendell, T. R. Mercer, K. J. Moore, S. Nakagawa, J. L. Rinn, D. L. Spector, I. Ulitsky, Y. Wan, J. E. Wilusz, M. Wu, Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat. Rev. Mol. Cell Biol.* **24**, 430–447 (2023).

44. W. W. Smoker, A. J. Gharrett, M. S. Stekoll, Genetic variation of return date in a population of pink salmon: A consequence of fluctuating environment and dispersive selection. *Alsk. Fish. Res. Bull.* **5**, 46–54 (1998).

45. A. J. Gharrett, J. Joyce, W. W. Smoker, Fine-scale temporal adaptation within a salmonid population: mechanism and consequences. *Mol. Ecol.* **22**, 4457–4469 (2013).

46. J. D. Echave, C. V. Manhard, W. W. Smoker, M. D. Adkison, A. J. Gharrett, Out crosses between seasonally different segments of a Pacific salmon population reveal local adaptation. *Environ. Biol. Fishes* **100**, 1469–1481 (2017).

47. C. V. Manhard, M. D. Adkison, J. J. Hard, W. W. Smoker, A. J. Gharrett, Local adaptation of phenology revealed in outcrosses between spawning segments of a salmonid population. *Mol. Ecol.* **27**, 4698–4710 (2018).

48. P. J. Lisi, D. E. Schindler, K. T. Bentley, G. R. Pess, Association between geomorphic attributes of watersheds, water temperature, and salmon spawn timing in Alaskan streams. *Geomorphology* **185**, 78–86 (2013).

49. P. T. Euclide, W. A. Larson, Y. Shi, K. Gruenthal, K. A. Christensen, J. Seeb, L. Seeb, Conserved islands of divergence associated with adaptive variation in sockeye salmon are maintained by multiple mechanisms. *Mol. Ecol.* **00**, 1–21 (2023).

50. O. W. Johnson, W. S. Grant, R. G. Kope, K. Neely, F. W. Waknitz, R. K. Waples, "Status review of chum salmon from Washington, Oregon, and California" (NOAA Technical Memorandum NMFS-NWFSC-32, National Marine Fisheries Service, 2725 Montlake Blvd. E., Seattle, WA 98112-2097., 1997).

51. N. O. Therkildsen, S. R. Palumbi, Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Mol. Ecol. Resour.* **17**, 194–208 (2017).

52. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

53. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

54. M. A. Alexandrou, B. A. Swartz, N. J. Matzke, T. H. Oakley, Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. *Mol. Phylogenet. Evol.* **69**, 514–523 (2013).

55. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

56. G. Jun, M. K. Wing, G. R. Abecasis, H. M. Kang, An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. *Genome Res.*, doi: 10.1101/gr.176552.114 (2015).

57. M. G. DeSaix, M. D. Rodriguez, K. C. Ruegg, E. C. Anderson, Population assignment from genotype likelihoods for low-coverage whole-genome sequencing data. *Methods Ecol. Evol.* **15**, 493–510 (2024).

58. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).

59. H. Wickham, *Ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016; https://ggplot2.tidyverse.org).

60. G. Bhatia, N. Patterson, S. Sankararaman, A. L. Price, Estimating and interpreting FST: the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).

61. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

62. N. van der Velden, *Geneviewer: Gene Cluster Visualizations* (2025; https://github.com/nvelden/geneviewer).

63. E. Paradis, K. Schliep, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

64. G. Yu, Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinforma.* **69**, e96 (2020).

65. J. Meisner, A. Albrechtsen, Inferring population structure and admixture proportions in low-depth NGS data. *Genetics* **210**, 719–731 (2018).

66. S. Durinck, P. T. Spellman, E. Birney, W. Huber, Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

67. A. Untergasser, I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, S. G. Rozen, Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115–e115 (2012).

68. N. R. Campbell, S. A. Harmon, S. R. Narum, Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Mol. Ecol. Resour.* **15**, 855–867 (2015).

69. T. Magoč, S. L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).

Acknowledgments: We thank the numerous graduate students, contractors, and NMFS staff who have provided foundational research and collected samples at the Auke Creek weir, especially John Joyce, Sidney Taylor, Chris Manhard, and Bill Smoker. We also thank Alaska Department of Fish and Game U.S. Fish and Wildlife employees for collecting samples and facilitating the transfer of DNA from their archives. We thank Curry Cunningham, Nicolas Lou, Daniel Schindler, Stephanie Carlson, and Peter Sudmant for useful discussions about the paper, and Devon Pearse and Fred Allendorf for providing valuable comments on manuscript drafts. This work was also supported in part by the high-performance computing and data storage resources SEDNA operated by National Marine Fisheries Service.

Funding: None declared.

Author contributions: Conceptualization: WL, DT, AJG, PB. Methodology: WL, DT, PB, NH, KDA, GLO, DSB. Investigation: PB, NH, WL, DT, GLO, SV, JR, AB, EL, AJG. Visualization:

NH, PB, WL, DT. Funding acquisition: WL. Project administration: WL. Supervision: WL, DT. Writing – original draft: WL, PB, DT, NH, Writing – review & editing: WL, DT, PB, NH, GLO, DSB, KDA, SV, JR, EL, AB, AJG.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: Fastq files for pink, chum, coho, and sockeye (whitefish) will be submitted to NCBIs Sequence Read Archive (SRA). Sequences from sockeye populations Teal Creek and Anvil Beach are available at NCBI SRA (PRJNA1006708). Amplicon sequencing data for pink salmon will be uploaded to DRYAD. All code and non-sequencing data will be made available at https://github.com/AFSC-Genetics/Salmon_Run_Timing.

Supplementary Materials:

Materials and Methods
Figs. S1 to S4
Tables S1 to S4
References (*44-69*)

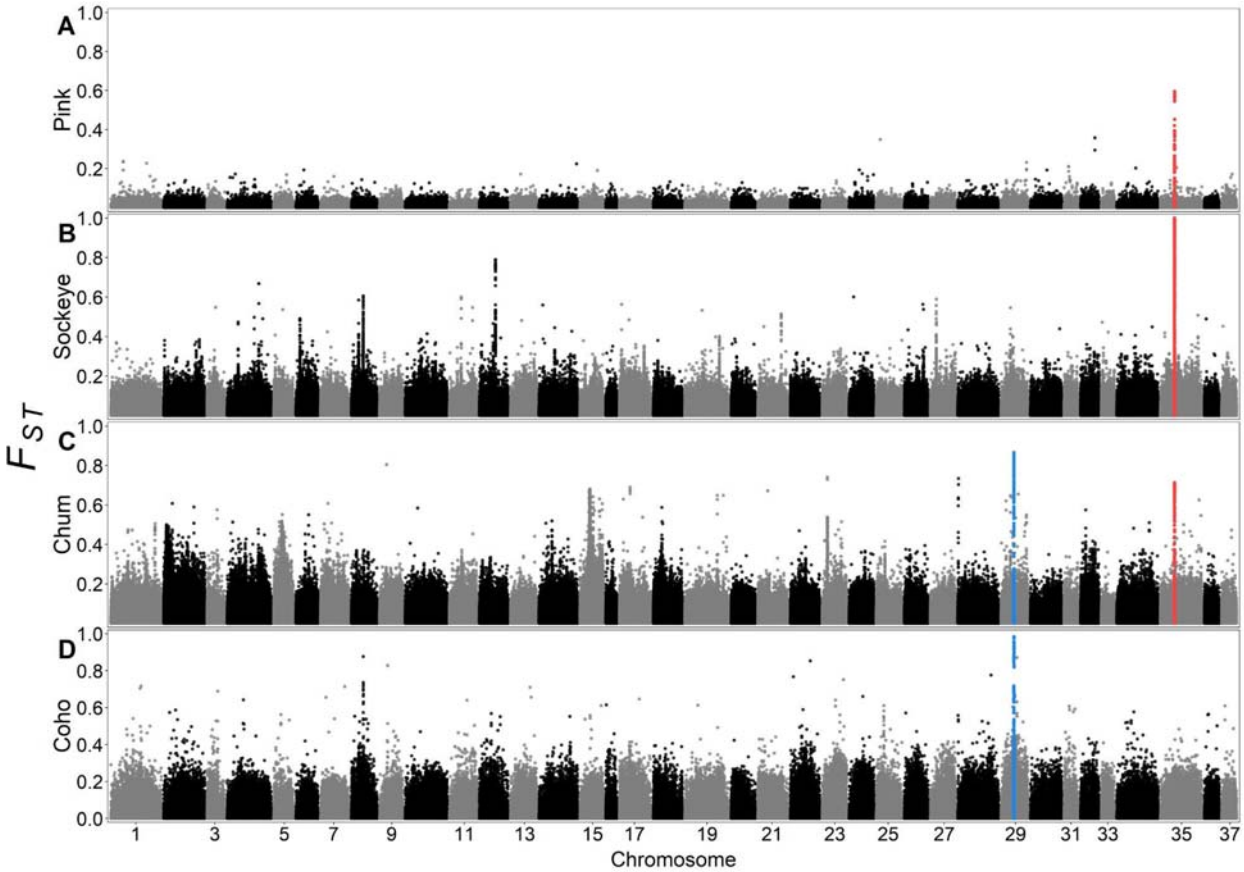

Figure 1: **Peaks of divergence associated with migration timing in four species of salmon.** Early and late run salmon showed high genetic divergence (F_{ST}) in a duplicated genomic region present on Chromosome 29 (blue peak) and/or Chromosome 35 (red peak). Pink salmon runs were sampled from the same stream pool and the runs were separated by ~3 weeks while all other species were sampled in discrete habitats and were separated by 6 weeks to 4 months. All species were aligned to the chum salmon genome. Colors alternate by chromosome.

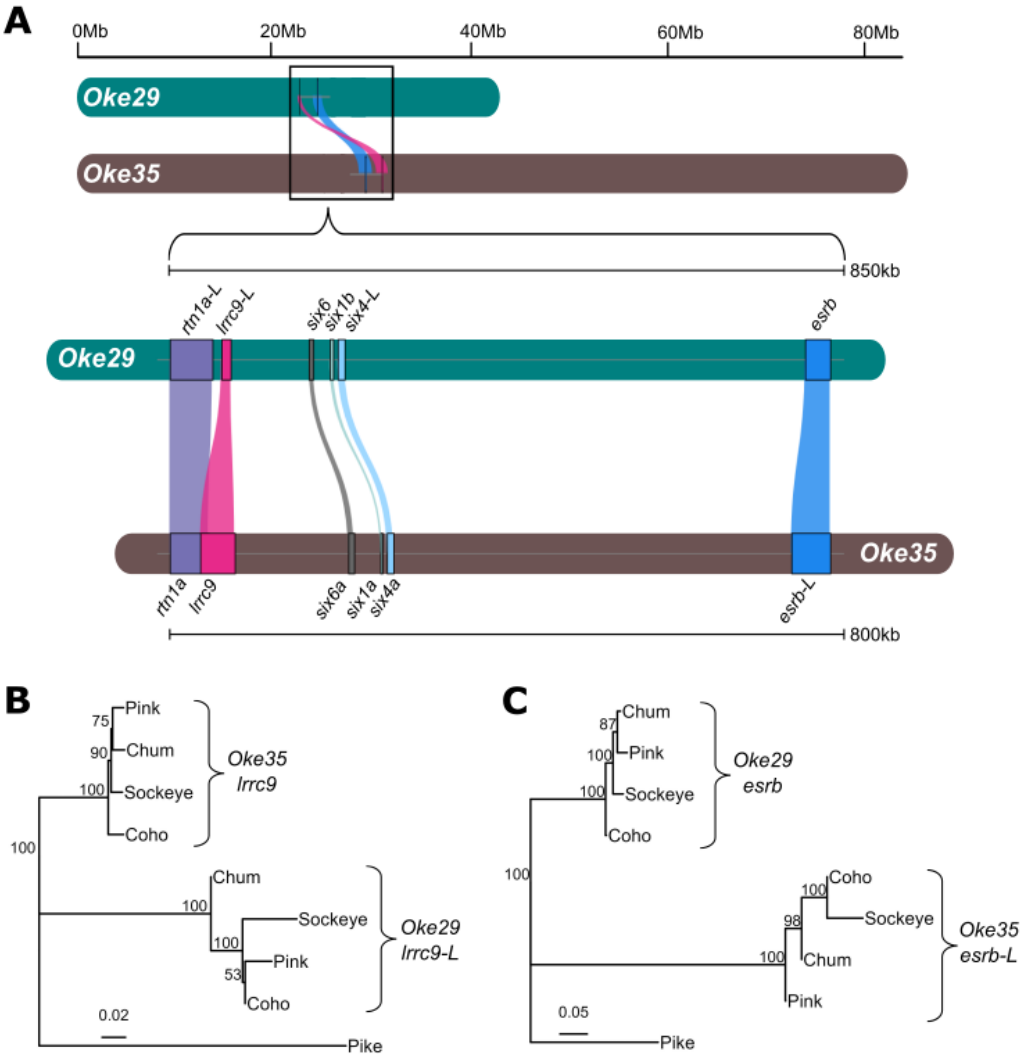

Figure 2: **Phylogenetic analysis indicates ohnologs diverged before speciation.** (A) Comparison of an ~850 kb duplicated region and specific genes containing loci most associated with run timing on Chromosomes 29 (*esrb*) and 35 (*lrrc9*). A previously identified gene associated with age-at-maturity and run timing in Atlantic salmon, and age-at-maturity in steelhead trout (*six6a*) occurs within this region. The orientation of the genes on Oke35 have been reversed because of a large (~40Mb) and likely ancient inversion that may have played a role in rediploidization. Neighbor joining trees with bootstrap support based on Nei's distance between ohnologs, of (B) *lrrc9* and (C) *esrb* with Northern Pike (a close relative to the recent common ancestor of salmonids) used as an outgroup. Trees show genes grouped by ohnolog rather than by species, consistent with divergence before speciation. Sequence data for trees available in supplemental Table S3.

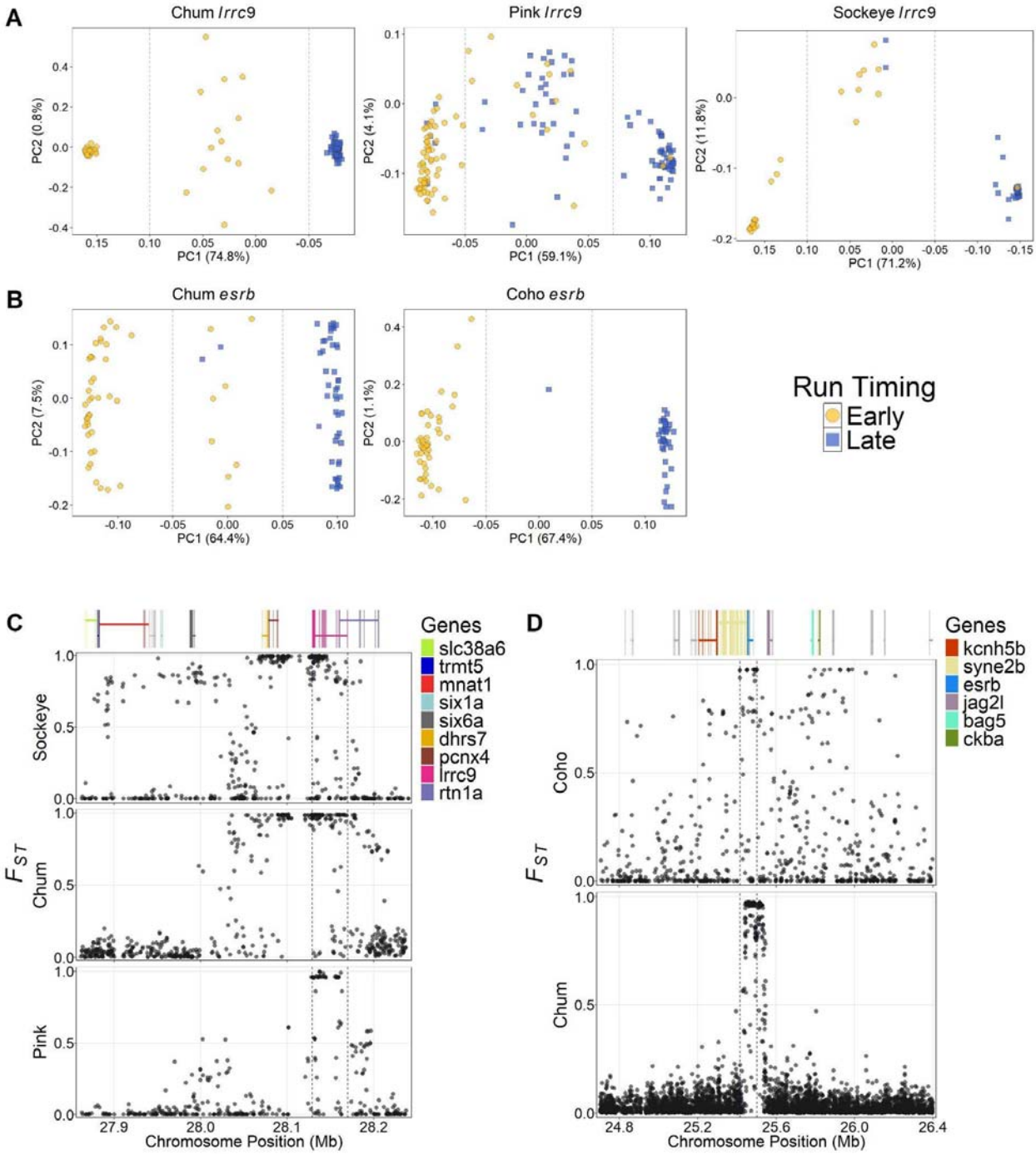

Figure 3: **Variation at two large-effect loci is associated with migration timing.** Association between salmon run timing phenotypes (early and late) with genotypes at *lrrc9* on Chromosome 35 in chum, pink, and sockeye salmon and *esrb* on Chromosome 29 in chum and coho salmon. (Top panels) Local PCAs for (A) *lrrc9* and (B) *esrb* where each point represents an individual salmon and genotypes are separated by dashed lines: homozygous early (EE, left), heterozygous (EL, middle), and homozygous late (LL, right). (Bottom panels) Genetic differentiation (F_{ST}), calculated between EE and LL individuals for (C) *lrrc9* and (D) *esrb*. Dashed lines in (C) and (D) represent the boundaries of the regions used to construct the PCAs above. Genes and their exons from the chum reference genome are denoted.

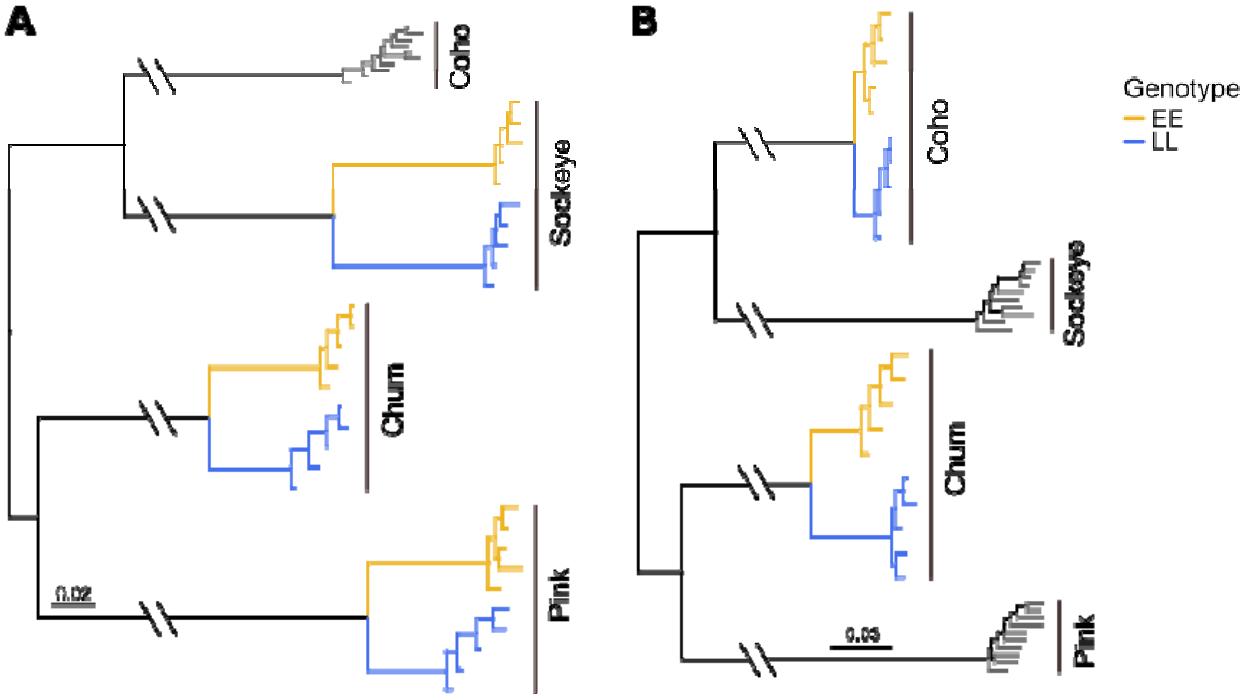

Figure 4: **Recurrent evolution of alleles associated with migration timing at two major effect loci.** Early and late run timing alleles at (A) *lrrc9* and (B) *esrb* evolved independently following speciation in Pacific salmon. Phylogenetic trees included ten fish for each species, five individuals from each homozygous genotype defined by PCA (Fig. 3) if associated with run timing and 10 individuals with the highest coverage from species that did not show an association. Branch lengths were shortened by 0.075 for *lrrc9* and 0.091 for *esrb*.