List of Figures

List of Tables

Figures

Map of sampling locations including run timing

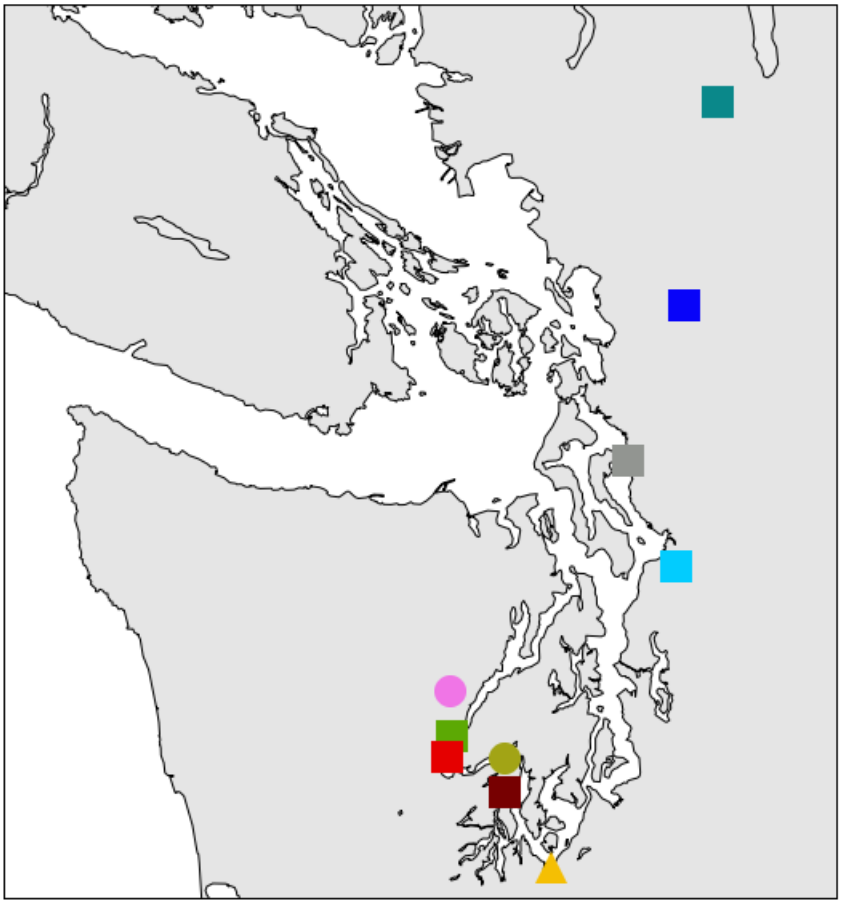

Figure 1: Collection locations and runtiming of chum salmon sampled near Puget Sound.

Linkage map

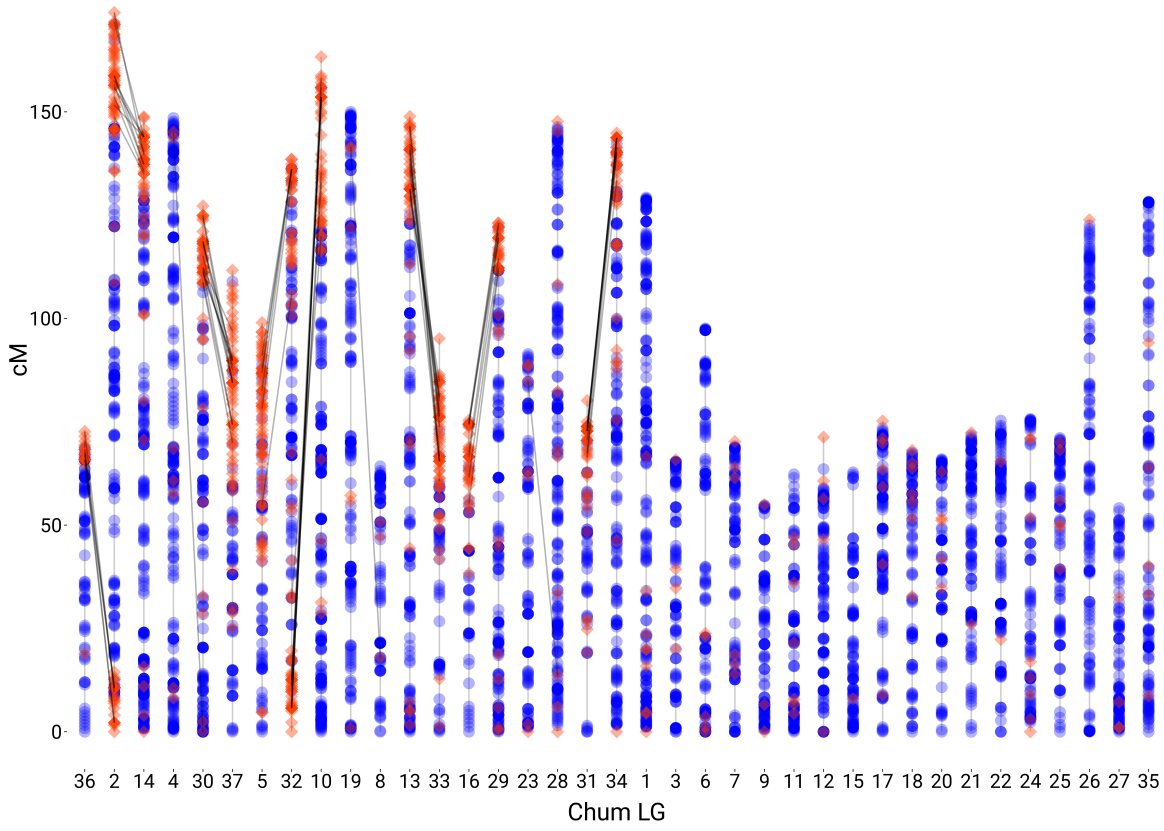

Figure 2: 37 linkage groups, likely corresponding to the 37 chromosomes in the chum salmon karyotype. Paralogous loci are shown as red diamonds, non-paralogs are blue circles. Black lines connect confounded catalog entries that have been resolved into two paralogous loci. The 16 distal concentrations of paralogs form 8 pairs of homeologous chromosomes. Notice LGs 2 and 32 have distinct ancestral relationships on each end.

Ascertainment bias

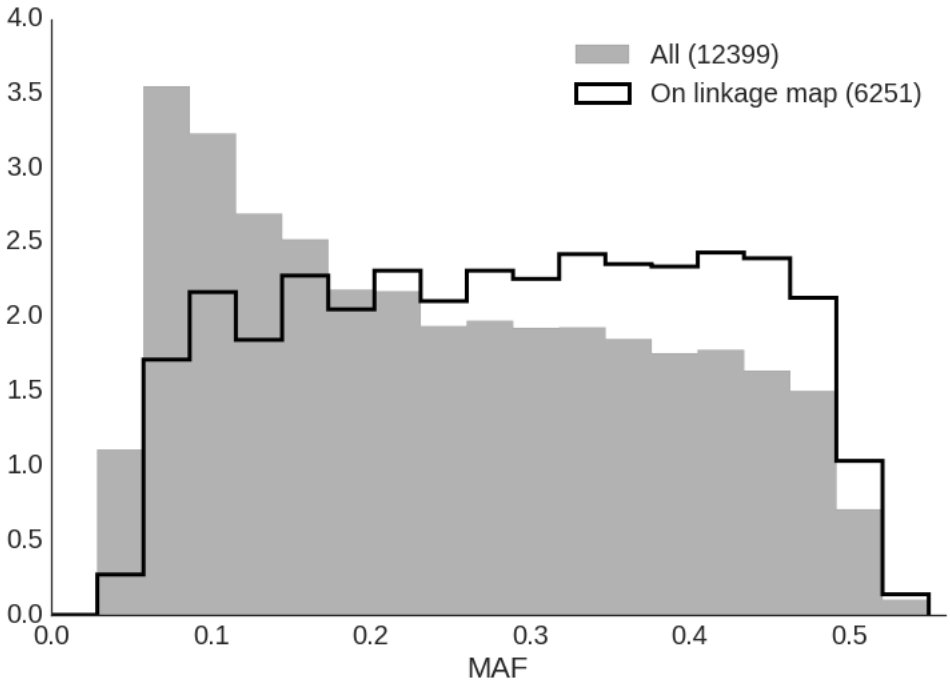

Figure 3: Folded minor allele frequency (MAF) for all loci (grey) and the subset of loci placed on the linkage map (black outline). The rightward shift in the MAF distribution shows the effect of ascertainment bias. Notice the y-axis is density-scaled to accommodate differing number of loci in each set.

Population structure

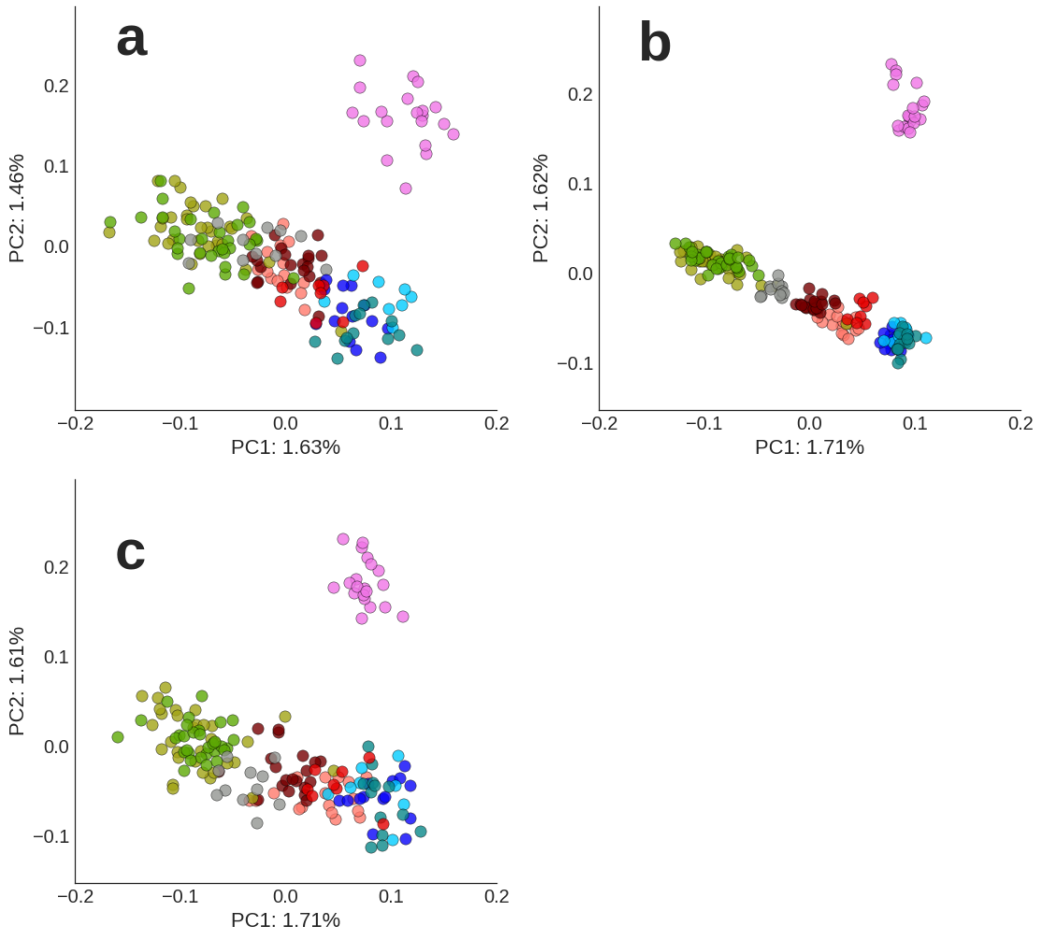

Figure 4: Population structure - Individual-based PCA from ten populations (colors) of chum salmon from Puget Sound. Population structure obtained from paralogs (a) is similar to that obtained by non-paralogs (b), especially after down-sampling to match the number of loci (c).

Manhattan plot

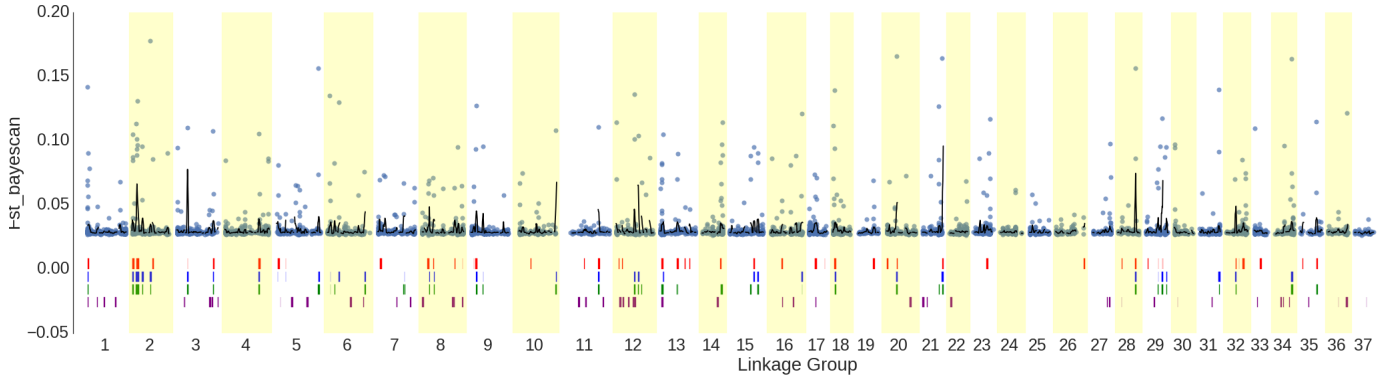

Figure 5: Manhattan plot of differentiation across the 37 linkage groups of chum salmon. Points are Bayescan Fst values for single loci. Population genetic statistics were calculated at each cM position by calculating an inverse distance-weighted average value from loci within a 5cM wide window centered on each position. Black outlines show loci selected as life history outliers in Bayescan. Black Lower shaded regions show genomic regions in the upper 99%, as determined by bootstrap permutation. Color codes for shaded regions: Red: LFMM qvalue, Blue: Bayescan qvalue, Green: Bayescan Fst, Purple: Weir Fst.

Tables

Sequencing and genotyping

Table 1: Sample sizes. usable sequences, and genotyping rates

		Aligned sequences		Genotyping rate	
Collection	n	mean	std	mean	std
Hamma Hamma	20	1,419,541	1,427,760	0.87	0.08
Lilliwaup Creek	20	2,760,125	999,141	0.98	0.01
Nisqually Kalama Creek	17	2,270,022	1,432,866	0.96	0.03
Sherwood River Fall	32	3,235,188	966,091	0.96	0.04
Sherwood River Summer	31	2,504,974	1,183,089	0.91	0.07
Skookum Creek	11	1,644,932	637,844	0.95	0.09
Snohomish River	14	1,135,085	495,888	0.94	0.07
Squakum Creek	8	999,084	650,927	0.86	0.08
Stillaguamish River	13	710,538	269,873	0.91	0.06
Hoodsport Hatchery*	8	509,422	148,391	0.85	0.08

*paired-end sequencing.

Genetic diversity

Table 2: Genetic diversity

	Heterozygosity	Ne
Hamma Hamma	0.30	339
Lilliwaup Creek	0.34	5,959
Nisqually Kalama Creek	0.32	161
Sherwood River Fall	0.33	319
Sherwood River Summer	0.31	145
Skookum Creek	0.33	1,788
Snohomish River	0.32	2,122
Squakum Creek	0.31	∞*
Stillaguamish River	0.30	1,001
Hoodsport Hatchery	0.30	∞*

*small sample size (<10)

Outline

- Genotyping duplicates
  - Legacy of the salmonid WGD
  - uncharacterized regions of the genome
  - first approach in salmon using next-gen seq data.
- Genome scan
  - Puget Sound chum salmon populations
    - Population structure
    - chum salmon anadromous life history -
    - ESA listing - summer chum ESU
    - Effective population size
  - map-assisted, paired population design
  - draw on synteny / orthology to interpret results
- Linkage Map
  - consensus map
  - synteny
  - annotation

Abstract

The common ancestor of salmonids underwent a whole genome duplication (WGD) approximately 100 million years ago. Understanding the genetic legacy of this event is critical to the conservation and management of these economically and socially important fish species. In contrast to most animals with strictly disomic inheritance, regions of the salmon genome undergo tetrasomic inheritance. Loci in these regions are often excluded from genetic analyses owing to increased complexity during both genotype assignment and subsequent analyses. Here I develop methods to better characterize the tetrasomically-inherited regions of salmonid genomes.

Introduction

Genotyping duplicates

Legacy of the salmonid WGD - Uncharacterized regions of the genome - retained through unknown mechanisms.

This is the first time that high-throughput sequencing data has been applied to score duplicated loci within salmonids.

Genome scan

map-assisted genome scan with a paired population design. The linkage map will be used to interpret the population genomic data. They provide information on the genomic adjacency of loci and facilitate the assessment of statistical independence between alleles. This can address the persistent problem of pseudoreplication, such as during the estimation of effective population size (e.g., Larson et al. (2014) or for marker development for mixed stock analysis. Kernel smoothing and bootstrapping (e.g., Hohenlohe et al. (2010) will be used to identify genomic regions with elevated level of divergence.

Despite the limited geogrpahic range of this studym

LFMM and other methods of environmental correlation have increased power to detect selection from standing genetic variation (Pritchard2010, Frichot2013)

Puget Sound chum salmon

Chum salmon (Oncorhynchus keta) have the widest distribution of any Pacific salmonid, from Korea, around the Pacific Rim, to Oregon (Salo, 1991). Chum salmon are abundant and are utilized by tribal and non-tribal fishers and comprise the dominant commercial fishery in Washington State. Recently some chum salmon populations have undergone drastic declines. National Oceanic and Atmospheric Administration (NOAA) Fisheries recognizes four evolutionarily significant units (ESUs) of chum salmon in the Pacific Northwest. Of the four, two are listed as threatened under the Endangered Species Act: the Hood Canal summer-run ESU and the Columbia River ESU. The Hood Canal summer-run ESU is composed of 16 historic populations, 7 of which are extinct (Good et al., 2005) and is the earliest-returning chum salmon stock in the Americas.

Two evolutionary significant units (ESU) - Hood canal summer-run vs the rest. 'Genetically and ecologically distinct' threatened under the ESA. Notice which populations where supplemented by hatchery programs? Chum salmon stray at similar rates to other pacific salmon Small et al. (2014).

Salmonids in the Pacific Northwest have a rich variety of life histories, with variation in run-timing, straying rates, age at maturity, freshwater residence, and many other dimensions (Quinn, 2005). This species and population-level diversity adds resilience to ecosystems and to species of conservation and economic concern (Schindler et al., 2010), especially in the face of climate change (reviewed in Schindler and Hilborn (2015)). Life history differs across populations within Puget Sound. Generally, eggs are deposited in November - December. Embryos develop and hatch after 4 months and migrate to sea, with survival and growth very dependent on favorable estuarine and marine conditions (Quinn, 2005). Chum salmon return to freshwater to spawn at 3-5 years of age, and generally spawn within 100km of the ocean.

"genetically similar populations with dissimilar life histories and morphology may provide insights at the onset of ecological speciation and reproductive isolation (Hendry 2009)." - from Aykanat et al.

Methods

Salish Sea collections

Colections of chum salmon were made across the Salish Sea, of the Pacific Ocean on the border of the USA and Canada from. Collection locations and timings were designed to capture the existing diversity in genetic structure and life history, including the latest-spawning chum salmon population (Nisqually) and the ESA-listed Hood Canal summer spawning population (Hamma Hamma). Fin clips were taken from adult fish were collected from each location during a single sampling event and preserved in ethanol.

DNA was extracted from tissues using DNEasy-96 kits from Qiagen (Venlo, Netherlands). A total of 200 adult samples were prepared for RAD sequencing with the SbfI restriction enzyme as per Baird et al. (2008) and Etter et al. (2011a). DNA from each sample was uniquely barcoded (6bp), pooled into libraries of 10-30 samples, and sequenced for 101 base cycles at the Genomics Core Facility at the University of Oregon, on a Illumina HiSeq2000. Hoodsport samples were sequenced using a paired-end protocol, all other collections were sequenced using a single-end protocol. Data from 240 gynogentic haploid offspring, used in construction of the consensus linkage map and originally reported in Waples et al. (2015), were sequenced in a similar manner.

Sequence analysis

Genetic variation was quantified with a reference-based approach using the `Stacks` software pipeline (Catchen et al., 2013). A chum salmon reference was constructed from by conducting an all-by-all self-alignment of the catalog from (Waples et al., 2015) using bowtie2 (Langmead and Salzberg, 2012) allowing up to three mismatches. Catalog entries placed on the linkage map of Waples et al. (2015) were retained, otherwise only a single sequence was retained from each group of aligned sequences. Reads from each individual were demultiplexed and quality filtered with 'process-radtags' and then aligned to the constructed reference with BWA-mem (version 0.7.5a-r405) (Li, 2013). Alignments containing indels or with a mapping quality ¡ 20 were removed. Stacks components `pstacks`, `cstacks`, `sstacks`, and `populations` were used to identify and genotype genetic variants for each individual. 20 individuals, 2 from each collection location, were used to ascertain genetic

variation, other samples were genotyped at these loci.

The initial set of genotypes and individuals was assessed prior to further analysis. Loci and individuals with more than 25% missing data were removed. Loci with a minor allele frequency (MAF) below 5% were removed as they are more difficult to distinguish from sequencing errors (Nielsen et al., 2011) . Hardy-Weinberg equilibrium (HWE) was tested within each collection using the mid p-value statistic (Graffelman and Moreno, 2013). Loci with HWE rejected in more than 5 collections were removed. Finally, within each locus, only the single SNP with the largest minor allele frequency was retained, to reduced pseudo-replication caused by physical linkage. All filters were applied in `PLINK` (v1.90beta) (Chang et al., 2014). Note that for the PCA analyses of population structure (see below), allelic haplotypes of all SNPs within each locus were utilized.

Linkage map

We constructed a consensus linkage map from three families of gynogenetic haploid offspring (family sizes 175, 34, 31) using the software LEPmap (Rastas et al., 2013). This linkage map here builds on the map presented in Waples et al. (2015) with the addition of two additional families and the placement of centromeric regions. As the linkage map is constructed from gynogenetic haploid offspring, it reflects only the recombination events that occur within the female lineage. As with many other species, there are sex-specific differences in recombination rates not reflected in this map. Regions of each chromosome likely to contain the centromere are estimated by measuring recombination fractions along chromosomes **cite Limborg**.

Paralogous loci confounded by alignment were identified and resolved using their segregation pattern within the gynogenetic offspring as in Waples et al. (2015) and were included when constructing the linkage map. For all loci that were variable in at least one offspring, the observed allelic segregation pattern was fit to the predicted segregation patterns under different possible parental genotypes. The parental genotype was selected as the genotype that was most likely to produced the observed segregation pattern, accounting for genotyping error. This parental genotype was used identified segregating loci suitable for inclusion on the linkage map.

Population structure and diversity

Allele frequencies, Heterozygosity, and Fst (Weir and Cockerham, 1984) were calculated for each locus in `PLINK`.

Principal component analyses (PCAs) were conducted on genotype matrices with `EIGENSOFT` (v6.0.1) (Patterson et al., 2006), including tests for population structure by comparing largest eigenvalues to the Tracey-Widom distribution (Tracy and Widom, 1994). Genotypes at paralogous loci were scored for the presence/absence of each allelic haplotype using the dominance coding suggested by Patterson et al. (2006). PCAs were compared with a Procrustes analysis. When supplied with two PCA projections, this method attempts to find an optimal superimposition, achieved by translation, rotation, reflection and scaling. After this transformation is complete, the remaining difference in shape is a measure of the Procrustes distance between the PCA projections. (Peres-Neto and Jackson, 2001).

Effective population size (Ne) was estimated for each population using the LD method implemented in the `LDNe` software package (Waples and Do, 2010). The LD method estimates the average correlation of alleles at pairs of loci (r2). The mean pairwise r2 value across independently-assorting loci provides an estimate of contemporary effective population size. Physically linked loci can downwardly bias Ne estimates; to avoid this potential bias, only loci placed on the linkage map were included in this analysis, and r2 measurements between loci co-located on a chromosome were also excluded.

Genome scans

Rolling-means of populations and test statistics were calculated at each cM position using a sliding-window analysis. At each focal cM a weighted-mean value was calculated across all loci within 2.5cM on either side, with weights for each locus within the window inversely proportional to squared distance to the focal point. At each focal cM, bootstrapped upper 99%? intervals were calculated by permuting random loci 1000 times into the windowed positions.

LFMM

Genome scans. Bayescan all populations - LFMM models with run-timing data

Synteny - relation to genetic resources

By design, RADseq generates sequence data exclusively near restriction enzyme cut sites; alignments of RAD contigs to genomic resources relate RAD data to much larger genomic sections. These resources often have functional annotations, whole gene sequences, and reading frame information that is unavailable to RADseq projects, expanding my ability to interpret genetic differentiation in a biological context.

Results

Sequencing and genotyping

table of sequencing results - by collection location, including family and wild samples. Genotyping rate

Linkage map

Here we present a consensus map placing 7795 loci onto 37 linkage groups. These 37 linkage groups correspond 1:1 with those reported in (Waples et al., 2015) and likely have a 1:1 correspondence with the 37 chromosomes in the most common chum salmon karyotype(Phillips and Rab, 2001). Of the 13,407 loci scored in the wild, 6,251 were placed on the linkage map.

A total of **xxx** paralogs were identified and placed onto the linkage map. The location of these paralogs were were concentrated on the distal ends of three chromosomes, consistent with results found in other Salmonid species (Brieuc et al., 2014; Kodama et al., 2014; Waples et al., 2015). These eight pairs of homeologous chromosome arms have elevated levels of sequence identity, likely due to ongoing residual tetrasomic inheritance.

placement of centromeres

paralogs

syntenic/orthologous relationships, see supplemental figure xx.

Ascertainment Bias

The linkage map was Initial population variant discovery proceeded on two individuals from each population. When these disocered snps are filtered by inclusion on the linkage map constructed from three females from the

Hoodsport collection there is a notable right-shift in the allele frequency spectrum (Figure 3). This occurs as rare variants from other populations are excluded as they were not variable in any of the three parents used for linkage mapping.

Genetic diversity

Hamma hamma, the Endangered species act (ESA)-listed collection, and member of a distinct population segment, did not show reduced diversity as measured by hetereozygosity or contemporary Ne.

Two populations has infinite point estimatates of Ne, these two populations were also the two populaitons with the smallest sample sizes, both below ten. Small sample sizes with

Population structure

Population structure represented by the PCA projections was consistent with Small et al. (2014)?. Population structure, as reflected by the PCA projections, was similar whether measured with the paralogs or non duplicated loci: Procrustes similarity xxx.

The distinctness of the Hamma Hamma summer run chum salmon was well supported in our analyses of population structure. In all PCAs, Hamma Hamma individuals were separated from the rest of the indidivuals along one of the primary axes (Figure 4).

Population structure ave similar neutral patterns of population structure

"PCA uses the mean allele frequencies as the first factor, and then two factors that represent deviations from this mean in two orthogonal directions (e.g., the diagonals of the square). As a result the PCA loadings on the second and third factors effectively recapitulate the geography of the space, as previously observed. ... In summary, the fact that the first factor in PCA represents the mean allele frequencies is responsible both for the fact that it produces less interpretable factors in the discrete case and more interpretable results in the continuous case. " Engelhardt

Discussion

The potential for bias in genetic inference resulting from this ascertainment bias.

Recent studies have shown the benefits of polyploidy in other species (Selmecki et al., 2015). Possible benefits are unknown but could include reduced inbreeding depression in small isolated populations typical of salmonids.

Compared to Small et al. (2014) the effective sizes (N_e) are larger, this could be due to the downward bias removed by utilizing the linkage map.

discuss population vs individual based results

Supplemental Figures

- Figures
    1. Linkage map
    2. Manhattan plots - genome scans statistics
    3. Cross-validation error of inferred ancestry - how to select K
    4. Chinook synteny oxford grid
    5. PCA of mapped/unmapped loci to address ascertainment bias?
- Tables
    1. Chum reference sequence (FASTA)
    2. Locus-specific info:
        - Map position
        - Allele frequencies
        - F-statistics
        - Bayescan q value and alphas
        - LFMM q value
    3. PCA projections, SNP loadings, Procrustes analysis, Tracy-Widom statistics
    4. LFMM inferred latent factors and environmental variables
    5. Smoothed genomic statistics from genome scan -

Figure 6: Centromere placemet - y plots

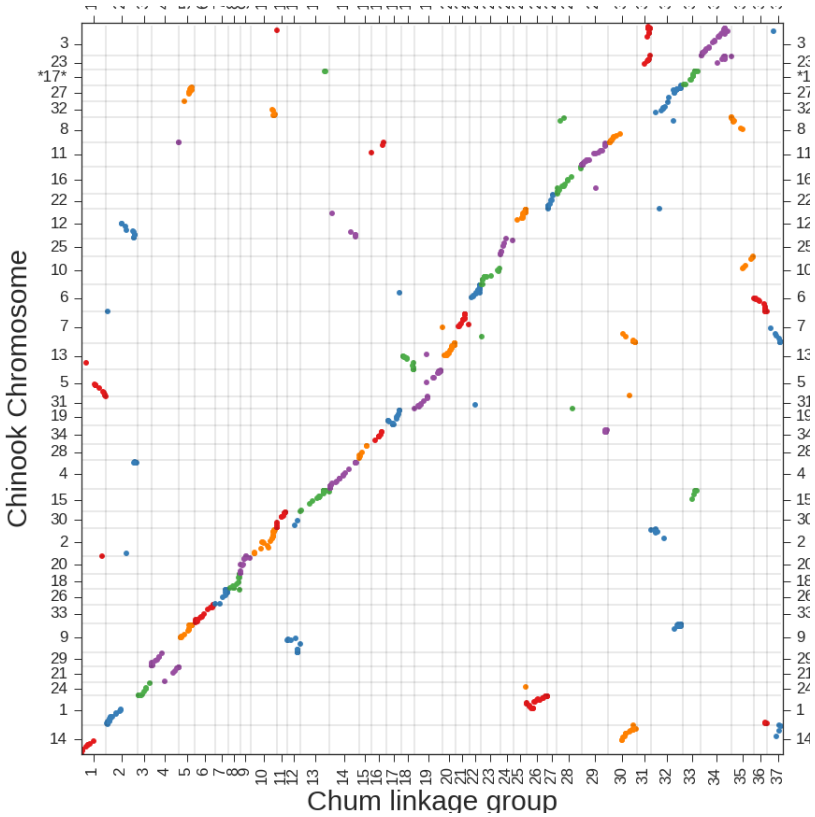

Figure 7: Oxford grid - Chum and Chinook linkage groups. Loci are colored by their LG assignment in chum salmon and positioned according to the order within each genome.

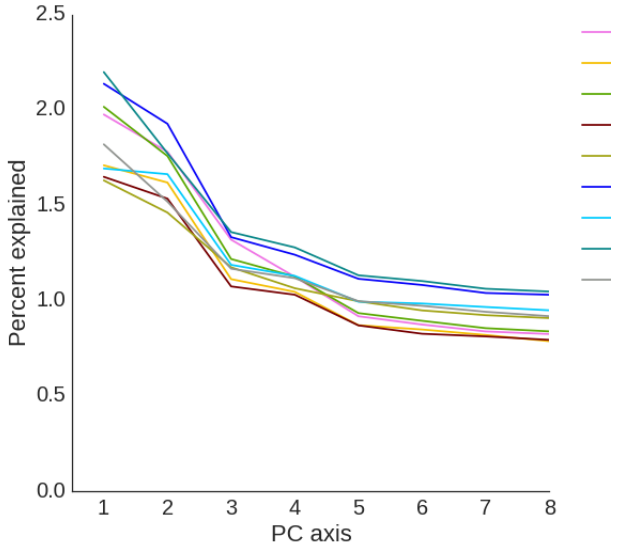

Figure 8: Percent variance explained (eigenvalue) for the first eight PC axes of each locus set. Notice the similarity between the two bi-allelic sets and the two haplotypic sets.

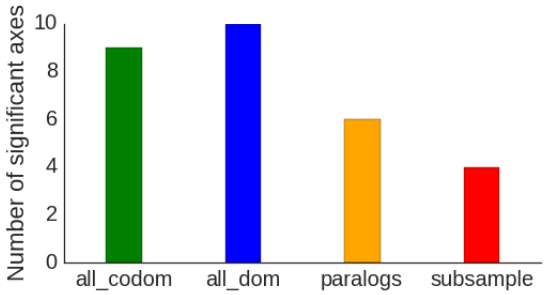

Figure 9: Number of significant PC axes as determined by the Tracey-Widom test.

Figure 10: Centromere placemet - y plots

Figure 11: Ancestry coefficients

Table 3: Centromere placement

Table 4: Procrustes analysis

References

Brieuc, M. S., Waters, C. D., Seeb, J. E., and Naish, K. A. (2014). A dense linkage map for chinook salmon (oncorhynchus tshawytscha) reveals variable chromosomal divergence after an ancestral whole genome duplication event. *G3 (Bethesda)*, 4(3):447–60.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Mol Ecol*, 22(11):3124–40.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2014). Second-generation plink: rising to the challenge of larger and richer datasets. *arXiv preprint arXiv:1410.4803*.

Good, T. P., Waples, R. S., and Adams, P. (2005). Updated status of federally listed esus of west coast salmon and steelhead. Report.

Graffelman, J. and Moreno, V. (2013). The mid p-value in exact tests for hardy-weinberg equilibrium. *Stat Appl Genet Mol Biol*, 12(4):433–48.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., and Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. *PLoS Genet*, 6(2):e1000862.

Kodama, M., Brieuc, M. S. O., Devlin, R. H., Hard, J. J., and Naish, K. A. (2014). 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-Genes Genomes Genetics*, 4(9):1717–1730.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359.

316 Larson, W. A., Seeb, L. W., Everett, M. V., Waples, R. K., Templin, W. D.,
317 and Seeb, J. E. (2014). Genotyping by sequencing resolves shallow popu-
318 lation structure to inform conservation of chinook salmon (oncorhynchus
319 tshawytscha). *Evol Appl*, 7(3):355–69.

320 Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs
321 with BWA-MEM. *ArXiv e-prints*.

322 Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype
323 and snp calling from next-generation sequencing data. *Nat Rev Genet*,
324 12(6):443–51.

325 Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and
326 eigenanalysis. *PLoS Genet*, 2(12):e190.

327 Peres-Neto, P. R. and Jackson, D. A. (2001). How well do multivariate data
328 sets match? the advantages of a procrustean superimposition approach
329 over the mantel test. *Oecologia*, 129(2):169–178.

330 Phillips, R. and Rab, P. (2001). Chromosome evolution in the salmonidae
331 (pisces): an update. *Biol Rev Camb Philos Soc*, 76(1):1–25.

332 Quinn, T. P. (2005). *The behavior and ecology of Pacific salmon and trout*.
333 UBC Press.

334 Rastas, P., Paulin, L., Hanski, I., Lehtonen, R., and Auvinen, P. (2013). Lep-
335 map: fast and accurate linkage map construction for large snp datasets.
336 *Bioinformatics*, 29(24):3128–34.

337 Salo, E. O. (1991). Life history of chum salmon (oncorhynchus keta).

338 Schindler, D. E. and Hilborn, R. (2015). Sustainability. prediction, precau-
339 tion, and policy under global change. *Science*, 347(6225):953–4.

340 Schindler, D. E., Hilborn, R., Chasco, B., Boatright, C. P., Quinn, T. P.,
341 Rogers, L. A., and Webster, M. S. (2010). Population diversity and the
342 portfolio effect in an exploited species. *Nature*, 465(7298):609–12.

343 Selmecki, A. M., Maruvka, Y. E., Richmond, P. A., Guillet, M., Shoresh, N.,
344 Sorenson, A. L., De, S., Kishony, R., Michor, F., Dowell, R., et al. (2015).
345 Polyploidy can drive rapid adaptation in yeast. *Nature*, 519(7543):349–352.

Small, M. P., Johnson, T. H., Bowman, C., and Martinez, E. (2014). Genetic assessment of a summer chum salmon metapopulation in recovery. *Evol Appl*, 7(2):266–85.

Tracy, C. A. and Widom, H. (1994). Level spacing distributions and the bessel kernel. *Communications in Mathematical Physics*, 161(2):289–309.

Waples, R. K., Seeb, L. W., and Seeb, J. E. (2015). Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (oncorhynchus keta). *Mol Ecol Resour*.

Waples, R. S. and Do, C. (2010). Linkage disequilibrium estimates of contemporary n e using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evol Appl*, 3(3):244–62.

Weir, B. S. and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, pages 1358–1370.

Acknowledgements

WDFW - chum salmon expertise. Carita Pascal - lab work and library prep. Samples sources? Rachel Hovel - multivariate stat advice.