Dissecting the loci underlying maturation timing in Atlantic salmon using haplotype and multi-SNP based association methods

Marion Sinclair-Waters[1,2], Torfinn Nome[3], Jing Wang[3,4], Sigbjørn Lien[3], Matthew P. Kent[3], Harald Sægrov[5], Bjørn Florø-Larsen[6], Geir H. Bolstad[7], Craig R. Primmer[1,2*] & Nicola J. Barson[3*]

[1]Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences University of Helsinki, Helsinki, Finland

[2]Institute of Biotechnology, Helsinki Institute of Life Sciences, University of Helsinki, Helsinki, Finland

[3]Centre for Integrative Genetics, Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway

[4]Key laboratory for Bio-Resources and Eco-Environment, College of Life Science, Sichuan University, Chengdu, China

[5]Rådgivende Biologer, Bergen, Norway

[6]Norwegian Veterinary Institute, Trondheim, Norway

[7]Norwegian Institute for Nature Research (NINA), Trondheim, Norway

*Shared last authors

ABSTRACT

Resolving the genetic architecture of fitness-related traits is key to understanding the evolution and maintenance of fitness variation. However, well-characterized genetic architectures of such traits in wild populations remain uncommon. In this study, we used haplotype-based and multi-SNP Bayesian association methods with sequencing data for 313 individuals from wild populations to further characterize known candidate regions for sea age at maturation in Atlantic salmon (*Salmo salar*). We detected an association at five loci (on chromosomes *ssa06*, *ssa09*, *ssa21*, and *ssa25*) out of 116 candidates previously identified in an aquaculture strain with maturation timing in wild Atlantic salmon. We found that at each of these five loci, variation explained by the locus was predominantly driven by a single SNP suggesting the genetic architecture of Atlantic salmon maturation includes multiple loci with simple, non-clustered alleles. This highlights the diversity of genetic architectures that can exist for fitness-related traits. Furthermore, this study provides a useful multi-SNP framework for future work using sequencing data to characterize genetic variation underlying phenotypes in wild populations.

INTRODUCTION

Understanding the genetic processes underlying fitness variation is a fundamental goal in evolutionary biology. Identifying genetic variants that underlie fitness-related traits is therefore crucial, yet remains challenging. Substantial effort has been made to characterize the genetic architecture of traits – i.e. Are there few or many loci involved? Are loci effects small or large? How are loci distributed across the genome? And what are the allele frequencies at these loci [1–5]? It is generally assumed that in most cases single genetic variants translate into only small changes in complex traits, and therefore follow a polygenic [6,7] or an omnigenic [3,8] model of inheritance.

Among genome-wide association studies published to date, many complex traits appear to be polygenic [9]. Although polygenicity is widespread, an increasing number of examples of major effect loci exist, whereby one locus explains a large proportion of the phenotypic variation [10,11]. In some cases, major effect loci can contain multiple tightly linked genes, coined “supergenes”, where localized reduction in recombination is often caused by larger chromosomal rearrangements. For example, this phenomenon is known to underlie phenotypic variation observed among ruff (*Philomachus pugnax*) mating morphs [12,13], Atlantic cod (*Gadus morhua*) [14,15] and rainbow trout (*Oncorhynchus mykiss*) migratory ecotypes [16], and *Heliconius* butterfly wing-pattern morphs [17]. More recent work has found that major effect loci can exist alongside a polygenic background where loci with a variety of effect sizes underlie trait variation [18,19]. Such mixed genetic architectures may be pervasive, but currently remain undetected due to the large sample sizes required for detecting loci with smaller effects [19] and it is possible that additional examples are to be found with future higher-powered studies. Although studies aimed at resolving genotype-phenotype links are mounting, well-characterized genetic architectures of fitness-related traits, particularly in natural populations, are still uncommon.

While some trait-associated loci have been identified, such findings lead to other crucial questions: How have trait-locus associations arisen? Has the locus arisen through a single or multiple new mutations? Or alternatively, did the locus emerge via recombination that gave rise to new combinations of existing variants? Numerous studies from the past decade have shown that major effect loci involve the cumulative effects of multiple mutations, rather than a single mutation, thus highlighting the relevance of considering the latter scenarios. For example, Bickle et al. [20] found that ~60% of variation in female abdominal pigmentation in *Drosophila melanogaster* can be explained by sequence variation at the *bab* locus, but a GWAS (genome-wide association study) analyzing the same trait did not identify a single SNP in *bab* that passed the genome-wide significance threshold. Alleles consisting of multiple SNPs were associated with high proportions of the variation, whereas, single SNPs had only small effects and were therefore missed in the single-SNP GWAS. Additionally, Linnen at al. [11] and Kerdaffrec et al. [21] also identify multiple mutations within a confined region that have cumulative effects on colour traits in deer mice and seed dormancy in *Arabidopsis thaliana*, respectively. In natural populations with gene flow such as in Linnen et al. [11] and Kerdaffrec et al. [21], this is perhaps not unexpected as theory predicts that clustered and major effect loci will evolve under such scenarios [22,23]. Given these findings, examining extended sequence haplotypes containing multiple SNPs, rather than each SNP independently, is important [24]. This can be achieved by using alternative strategies that look at combined effects of variants, rather than single-SNP methods typically used in GWAS.

Here we investigate the genetic basis of Atlantic salmon (*Salmo salar*) sea age at maturity – the number of years spent in the marine environment before reaching maturity and returning to the natal river (freshwater) to reproduce. Age at maturity is an important life history trait affecting fitness traits such as survival, size at maturity and reproductive success [25,26]. Substantial variation in Atlantic salmon sea age at maturity is maintained due to a trade-off between mating success at spawning grounds and survival, whereby individuals that mature later are larger and have higher reproductive success on the spawning grounds, but lower survival and thus lower chance of reaching reproductive age. In contrast individuals that mature early are smaller and have lower reproductive success, but higher survival and thus higher chance of reaching reproductive age [27,28].

Variation in maturation timing in Atlantic salmon is highly heritable [19,29,30] and consequently there is substantial interest in understanding the underlying genetic architecture. A large-effect locus on chromosome 25 explaining up to 39% of the variation in sea age at maturity was found in wild European populations [10] and domesticated salmon [31]. The primary candidate gene underlying the association of this locus is *vgll3* due to its close proximity to the associated SNP variation [10,31,32] and its known function in other species. The *vgll3* gene encodes a transcription cofactor that, amongst other things, regulates adipogenesis [33] and is associated with variation in puberty timing in humans [34,35]. In addition to *vgll3*, Sinclair-Waters et al. [19] identified 119 other candidate genes for male maturation in a GWAS including >11,000 males from the same Atlantic salmon aquaculture strain. Two particularly strong associations between maturation timing were found on chromosome 9 in close proximity to *six6* and chromosome 25, *vgll3*. The association of *six6* was also found by Barson et al. [10] in wild Atlantic salmon, however, the signal disappeared after correction for population structure. Interestingly, the *six6* gene is also associated with age at maturity in two Pacific salmon species [36], humans [35] and cattle [37]. However, Barson et al. [10] focused solely on single-SNP associations via GWAS without considering the possible influence of combined variant effects.

Studies using sequencing data to examine variation associated with important fitness-related traits in wild populations are limited. However due to developments in sequencing technologies and bioinformatics, studies using this approach are likely to rise in number. We therefore aim to provide a useful and timely framework for characterizing genetic variation underlying phenotypes in wild populations in the future. Here, we focus on further characterizing the association between the loci identified in Sinclair-Waters et al. [15] and sea age at maturity in wild Atlantic salmon. We integrate re-sequencing data and phenotype information for 313 individuals from 53 wild population of Atlantic salmon with alternative GWAS strategies that consider the combined effects of variants, rather than single-SNP effects. This approach can provide better resolution of the variants that are potentially involved in controlling fitness-related traits such as maturation timing in Atlantic salmon.

METHODS

Study material

Whole genome sequencing data was obtained for 313 wild individuals collected from 53 Norwegian and Finnish populations spanning the Norwegian coast and to the Barents sea in the north (59°N - 71°N) (Supplementary Table S1) previously reported in Bertolotti et al. [38]. The 313-individual dataset includes populations belonging to both the Atlantic and Barents/White sea phylogeographic groups. These regions were studied in Barson et al. [10] using SNP-array data and a single SNP approach, therefore missing variants and potentially combined variant effects. Individuals were categorized into three maturation categories based on the number of years spent at sea prior to their first return migration to rivers for spawning: 1 (one year spent at sea), 2 (two years spent at sea), or 3 (three or more years spent at sea). Only five individuals had spent four years and were therefore combined with three-year fish for all analyses.

SNP calling & filtering

Variant calling and the first round of filtering was done in a larger set of individuals described in Bertolotti et al. [38]. Raw Illumina reads were mapped to the Atlantic salmon genome (ICSASG_v2) [39] using *bcbio-nextgen v.1.1* [40]with the *bwa-mem aligner v.0.7.17* [41]. Genomic variation was identified using the Genome Analysis Toolkit (*GATK*) *v4.0.3.0.*, following *GATK*'s best practice recommendations. *Picard v2.18.7* [42] was used to mark duplicates and *GATK* was used for joint calling [43]. Variants were annotated using *SNPeff v. 4.3* [44]. Variant call were further filtered with GATK's variant filtration according to the following --*filterExpression*: "MQRankSum < -12.5 || ReadPosRankSum < -8.0 || QD < 2.0 || FS > 60.0 || (QD < 10.0 && AD[0:1] / (AD[0:1] + AD[0:0]) < 0.25 && ReadPosRankSum < 0.0) || MQ < 30.0". SNPs were then filtered using *SNPable* procedure [45], where 100 bp kmers are mapped to reference genome (ICSASG_v2) using Burrows-Wheeler Aligner (*bwa aln*) [46], and only SNPs within regions with reads that uniquely map are retained. We then removed additional SNPs with *vcftools* using the following criteria: --*min-alleles 2, --max-alleles 2, --maf 0.0000000001, --max-missing 0.7, --remove-indels, --minGQ 10,* and *–minDP 4.* A subset 313 individuals from wild populations was then extracted from this larger dataset using *vcftools* [47]. This reduced dataset was used for all subsequent analyses.

Principal component analysis

We produced a reduced SNP dataset by pruning one SNP from each SNP pair with a correlation coefficient (r^2) greater than 0.2 within a 50 kb block using the --*indep-pairwise 50 10 0.2* function implemented in *PLINK v1.9* [48]. This yielded 403,540 SNPs to examine population structure using a principal component analysis, *smartpca*, implemented in the EIGENSOFT *v5* software [49].

Data preparation

In this study, we focus on genomic regions containing the 116 candidate loci for age at maturity identified in Sinclair-Waters et al. [19]. We extracted SNP genotype data from 500 kb regions surrounding the 116 trait-associated SNPs identified in Sinclair-Waters et al. [19] using *vcftools'* [47] position filtering functions --*from-bp* and --*to-bp,* as well as allele filtering function --*mac 1* to keep only polymorphic sites. SNPs that were within 250 kb of an adjacent SNP were analyzed together by examining a region that extends 250 kb upstream of the first SNP to 250 kb downstream of the last SNP.

The current Atlantic salmon genome (ICSASG_v2) contains a known assembly error within the 500 kb region surrounding the known candidate loci *vgll3* [31]. A misplaced and misoriented scaffold currently placed downstream of *vgll3* belongs within a gap in the assembly just upstream of *vgll3* on ssa25. For this reason, we constructed a revised assembly for this chromosome. SNP calling was performed as described above. We then retained SNPs that had met the filtering criteria. A total of 8 candidate SNPs are located within regions of the genome that were moved. To find the position of these SNPs in the revised chromosome 25 sequence, we extracted 200 bp surrounding each of these SNPs from the current genome assembly (ICSASG_v2) using the *getfasta* function in *BEDTools* [50]. The 200 bp sequence was then blasted to the fixed assembly to determine the new position of each SNP using Blast's *blastn* function [51]. Using the new SNP positions, SNP genotypes within a 500 kb region surrounding the moved candidate SNPs were extracted from the fixed dataset using *vcftools*.

Association testing at candidate regions

We applied three association mapping methods to describe the genetic architecture underlying sea age at maturity at each of the candidate regions identified in Sinclair-Waters et al. [19]. First, a multi-SNP approach examining associations between phenotype and haplotypes was conducted using Bayesian linear regression implemented in *hapQTLv1.00* [52]. In this approach, a hidden Markov model is used to characterize haplotype structure and ancestry [53]. Haplotype sharing at each marker is then used to quantify genetic similarity among individuals. Haplotype associations are identified by testing for an association between genetic similarity at each marker and the phenotype [52]. Each of the extracted *vcf* files was converted to *bimbam* format using *PLINK 1.9* [54]. The resulting *bimbam* files were used as input for *hapQTL*. Second, single SNP associations were also identified using a Bayesian linear regression method implemented in *hapQTL* [55]. For all *hapQTL* association tests, sex and the six most significant principal components (see above) were included as covariates in the models. Each *hapQTL* run consisted of 2 EM runs (-e 2) with 40 steps (-w 40), 2 upper clusters (-C 2), 10 lower clusters (-c 10). Three replicate *hapQTL* runs were performed for each of the 116 selected regions. Based on recommendations from Jeffreys [56], Bayes factors greater than three were considered evidence for an association of either SNPs or haplotype with sea age at maturity phenotype.

Third, a multi-SNP approach aimed to estimate the number and identity of SNPs underlying trait variation at each candidate region using Bayesian Variable Selection regression implemented in *PiMASS* [55]. Due to computational restrictions, the *PiMASS* analysis was performed for only candidate regions that had a SNP or haplotype association with Bayes factor greater than 3. Prior to the *PiMASS* analysis, all missing genotypes were imputed in BIMBAM [55] as mean genotypes (-wmg) using default settings. Additionally, our phenotype values for sea age at maturity were adjusted to correct for confounding effects of sex and population structure by regressing the phenotype on sex and the six most significant principal components (see above) using the *lm* function in *R*. *PiMASS* was run with the residual phenotype values. We placed priors on the proportion of variance explained by SNP(s) (hmin = 0.001 and hmax = 0.999) and the number of SNPs in the model (pmin = $\log\frac{1}{N}$ and pmax = $\log\frac{300}{N}$, where N is the total number of SNPs). Each run consisted of a burn-in of 1000000 steps, followed by 2500000 steps where parameter values were recorded every 1000 steps. For each analysis, we examined the posterior inclusion probability for each SNP, the distribution of the number of included SNPs and the distribution of the proportions of variance explained per model. We also examined the path of estimated Bayes factors and parameter values (h, p, s) across all recorded iterations to check for convergence of runs.

To further assess whether more than one SNP in a candidate region was significantly associated with sea age at maturity, we regressed out the top-associated SNP from the residual phenotype values described above and reran *PiMASS* using the previously-used priors and settings. We then examined the posterior inclusion probability for each SNP, the distribution of the number of included SNPs, and the distribution of proportion of variance explained to determine whether there was evidence for multiple SNP associations within a given candidate region.

RESULTS

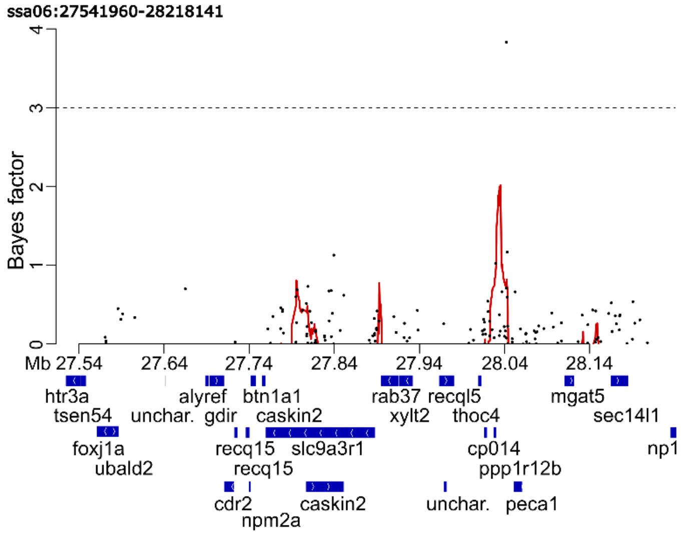
Principal component analysis

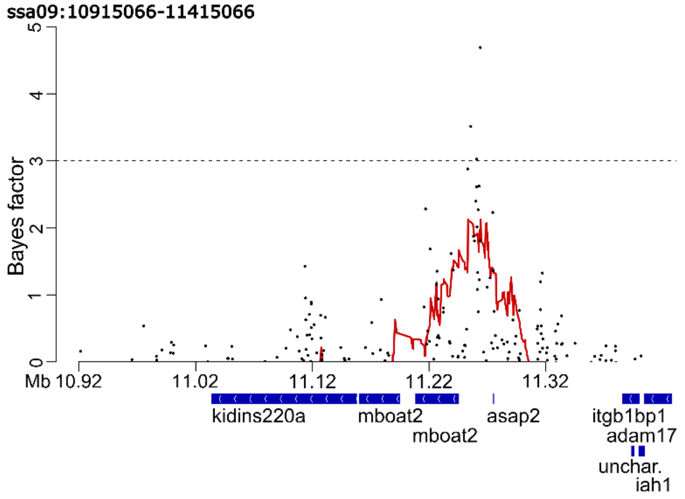
The first six principal components (PCs) calculated with the pruned SNP dataset explained 1.96%, 0.68%, 0.63%, 0.59%, 0.56% and 0.51% of the genetic variance, respectively (Supplementary Figure S1). These six PCs were included in subsequent association analyses to reflect population structure among samples.

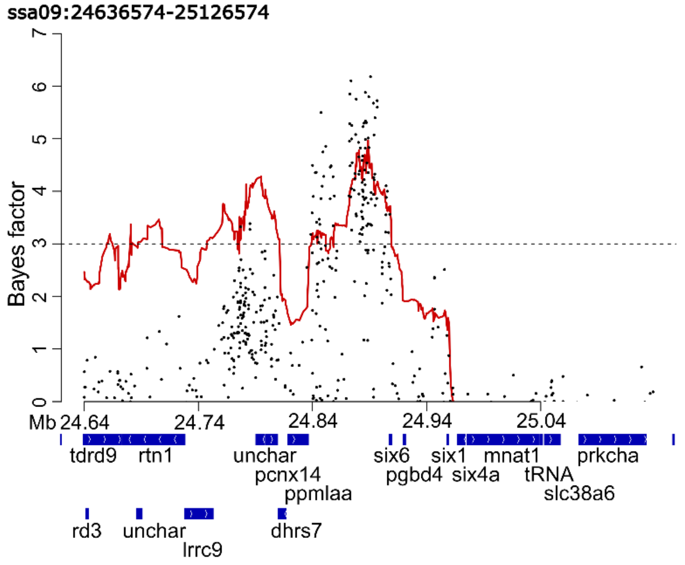
Associations identified with hapQTL

Single-SNP and haplotype association analyses with *hapQTL* revealed strong (Bayes factor > 3) association signals at 5 of the 116 candidate regions (Figure 1, Supplementary Figure S2). The strongest association observed within each region was with a single SNP, rather than an extended haplotype, suggesting a single mutation underlies the effect of each of these regions on maturation timing. However, exceptions occurred in the ssa09:24636574-25136574 and ssa25:28389273-28889273 regions, where second association signals were found upstream of the primary association signal and were most strongly linked to an extended haplotype. For instance, strong haplotype association scores (Bayes factor > 3) spanned a 26971 bp region (ssa09:24781742-24808713) containing an uncharacterized gene (LOC106610978) and *pcnx4*. In the ssa25:28389273-28889273 region, a strong haplotype signal was found within *edar* (Figure 1).

We find differences in the location of the top-associated SNPs found here and those identified in Sinclair-Waters et al. [19]. For regions ssa06:27541960-28218141, ssa09:10915066-11415066 and ssa25:28389273-28889273, the top-associated SNP was located further upstream than in Sinclair-Waters et al. [19]. Contrastingly, the strongest associated SNPs within the regions ssa09:24636574-25136574 and ssa21:49390687-49890687 differed only slightly (<5000 bp) between studies (Table 1).


ssa06:27541960-28218141
Bayes factor
0
1
2
3
4
Mb 27.54
27.64
27.74
27.84
27.94
28.04
28.14
htr3a
alyref
btn1a1
rab37
recql5
mgat5
tsen54
unchar.
gdir
caskin2
xylt2
thoc4
sec14l1
foxj1a
recq15
slc9a3r1
cp014
np1
ubald2
recq15
ppp1r12b
cdr2
caskin2
unchar.
peca1
npm2a


ssa09:10915066-11415066
Bayes factor
0
1
2
3
4
5
Mb 10.92
11.02
11.12
11.22
11.32
kidins220a
mboat2
asap2
itgb1bp1
mboat2
adam17
unchar.
iah1


ssa09:24636574-25126574
Bayes factor
0
1
2
3
4
5
6
7
Mb 24.64
24.74
24.84
24.94
25.04
tdrd9
rtn1
unchar
six6
six1
mnat1
prkcha
pcnx14
pgbd4
six4a
tRNA
ppmlaa
slc38a6
rd3
unchar
dhrs7
lrrc9

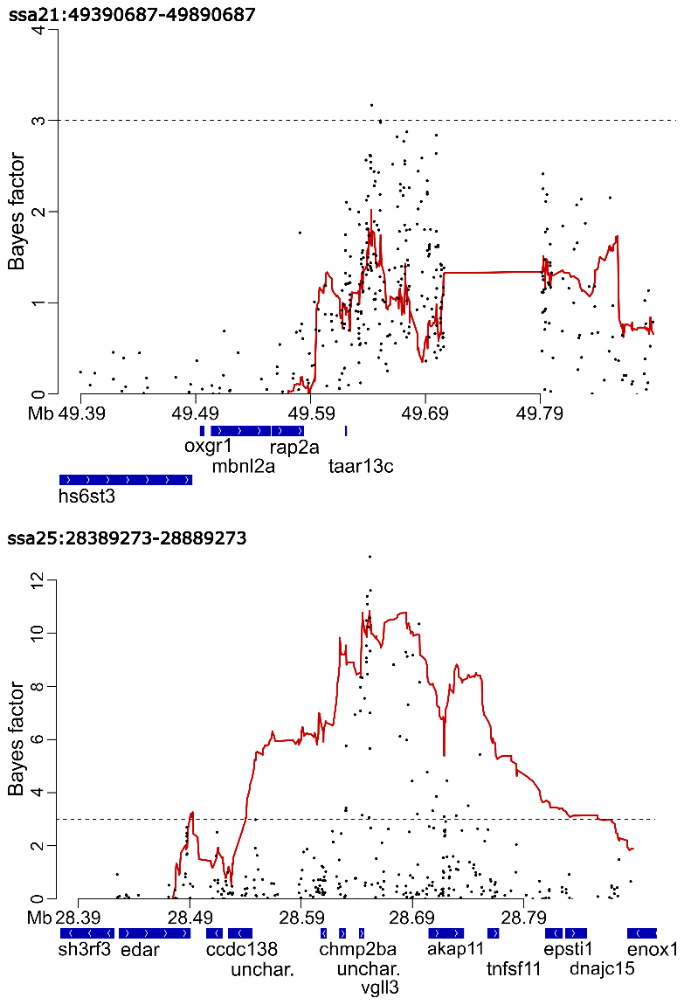


Figure 1. Plots displaying single SNP associations (black points) and haplotype associations (red line) scores from *hapQTL* for the five candidate regions with Bayes factors greater than 3. Y-axis shows the Bayes factor indicating the association strength. X-axis shows the position on the respective chromosomes.

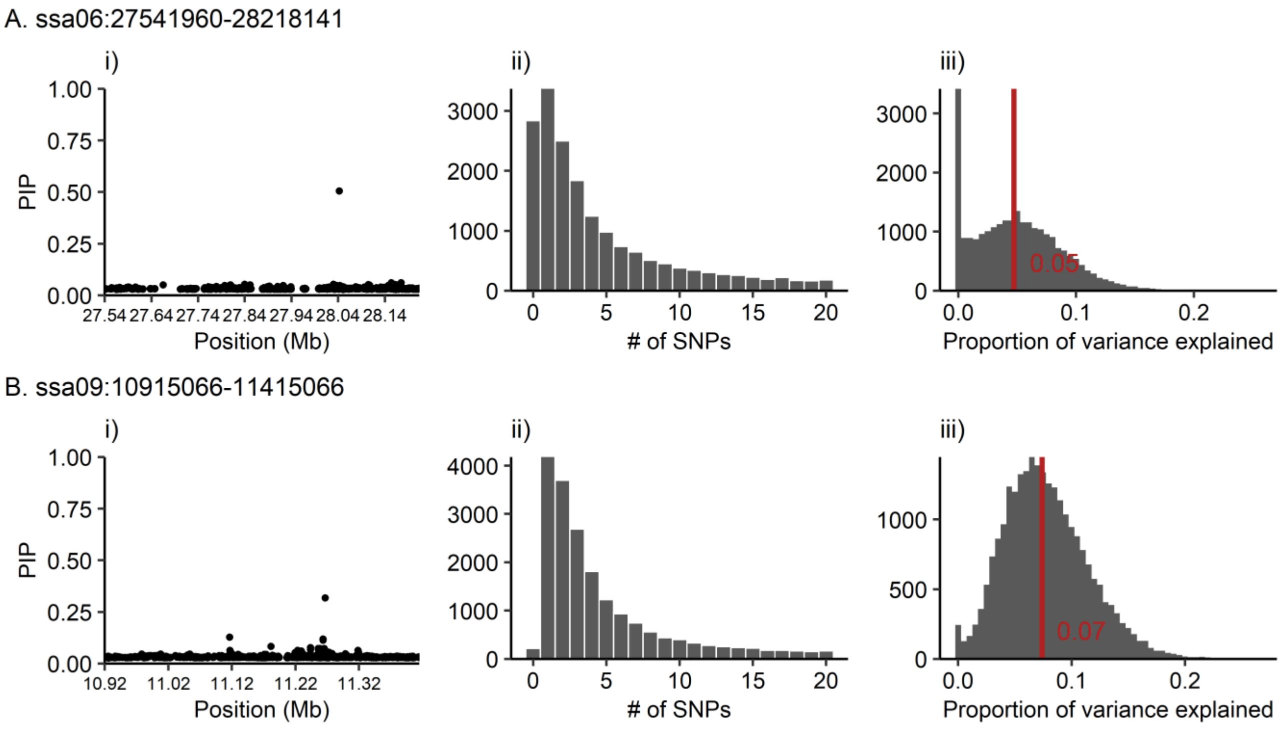
Table 1. Strongest association signals for each candidate region showing evidence of an association with sea age at maturity, the genes in closest proximity and association values from *hapQTL*. Top SNPs for each region from previous SNP-array study [19].

Candidate region	**Top signal**	**Closest gene**	**Bayes Factor**	**-log$_{10}$(*P*-value)**	**Allele frequency**	**Top SNP(s)[a]**	**Candidate gene(s)[a]**
ssa06:27541960-28218141	6:28045390 (SNP)	*pecam1* (intron)	3.835	5.107	0.320	6:27791960 6:27968141	*slc9a3r1* *recql5* LOC106606978
ssa09:10915066-11415066	9:11266848 (SNP)	*asap2a* (upstream)	4.696	5.434	0.074	9:11165066	*mboat2*
ssa09:24636574-25136574	9:24888841 (SNP)	*six6* (upstream)	6.184	4.242	0.425	9:24886574	*six6*
ssa21:49390687-49890687	21:49645222 (SNP)	*taar13c* (upstream)	3.172	4.649	0.464	21:49640687	*taar13c*
ssa25:28389273-28889273	25: 28651640 (SNP) [ICSASG_v2: 25:28669350]	*vgll3* (downstream)	12.893	6.406	0.358	25:28910202	*vgll3*

[a]From Sinclair-Waters et al. [19].

Multi-SNP associations identified using PiMASS

Multi-SNP association analysis with *PiMASS* showed that at four of five candidate regions, a single-SNP model was most commonly used to explain variation in sea age at maturity. At one candidate region, ssa09:24636574-25136574, a multi-SNP model including two SNPs was most commonly used to explain variation in sea age at maturity. Median proportion of variance explained by each candidate region ranged between 4% and 19% (Figure 2, Table 2). However, when the top-associated SNP was regressed out from the phenotype values, no SNPs were selected to explain sea age at maturity for all five candidate regions. Additionally, post-regression median proportion of variance was substantially lower – ranging between 0% and 1% (Supplementary Figure S3, Table 2). This would suggest that sea age variation explained by each of these regions is largely driven by a single mutation. We observe no obvious trends in parameter values or Bayes factors, suggesting models converged and burn-in period was adequate (Supplementary Figure S4 &S5).



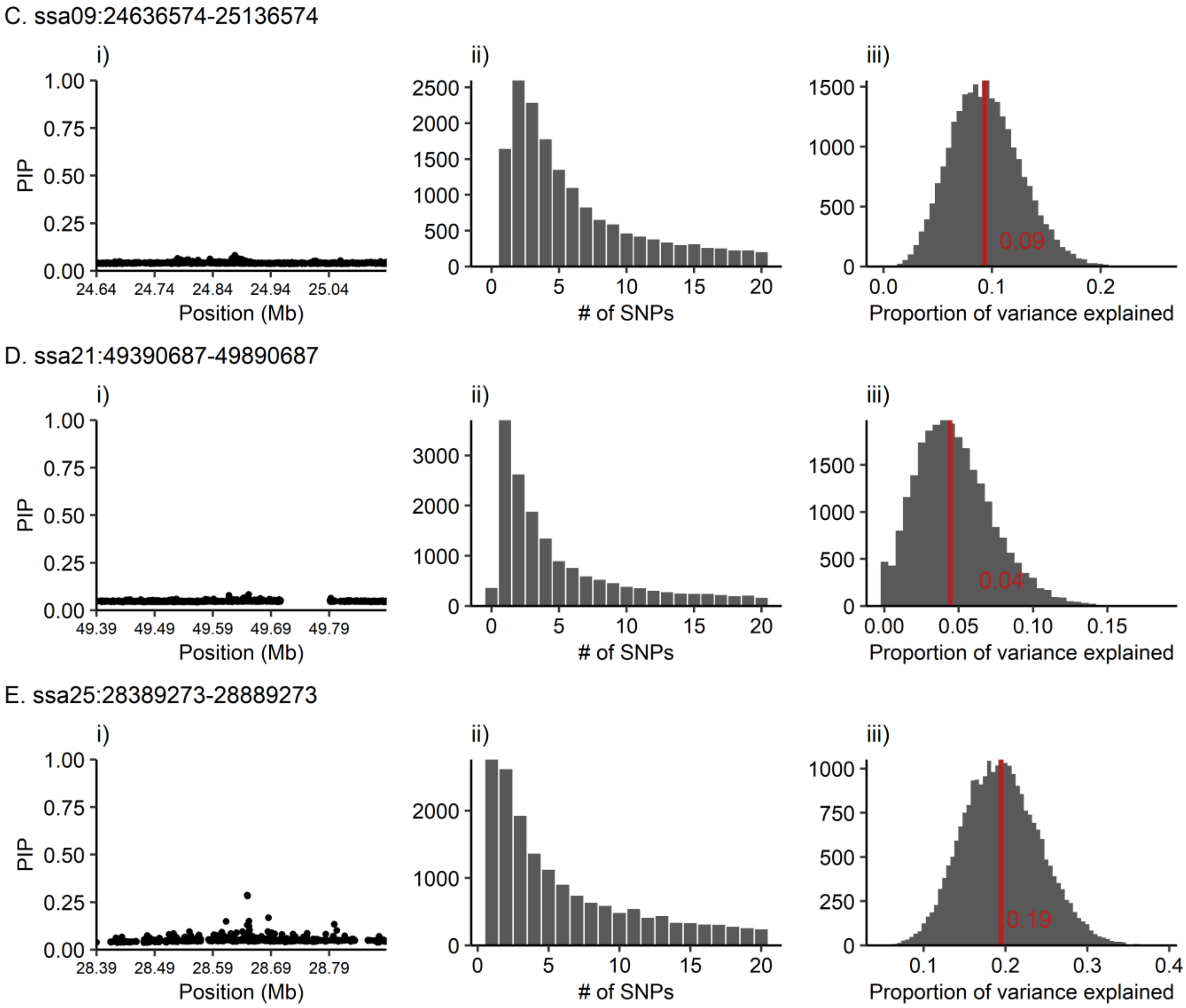


Figure 2. *PiMASS* results for each of the tested candidate regions: A. ssa06:27541960-28218141, B. ssa09:10915066-11415066 C. ssa09:24636574-25136574, D. ssa21:49390687-49890687, and E. ssa25:28389273-28889273. Plots display the following results for each candidate region: i) posterior inclusion probability (PIP) indicating the probability of a SNP being included in a model explaining sea age at maturity variation, ii) truncated distribution of the number of SNPs included in a model explaining sea age at maturity variation, and iii) distribution of proportion of variance explained per recorded iteration (2500). Red line indicates the median proportion of variance explained.

Table 2. *PiMASS* results prior to and after regression of top-associated SNP identified in the initial *PiMASS* analysis. These include the mode of the distribution of the number of SNPs and the median of the distribution of proportion of variance explained (PVE) for a model explaining sea age at maturity.

Candidate region	Mode # of SNPs	Median PVE	Mode # of SNPs (post-regression)	Median PVE (post-regression)
ssa06:27541960-28218141	1	0.05	0	0
ssa09:10915066-11415066	1	0.07	0	0.01
ssa09:24636574-25136574	2	0.09	0	0.01
ssa21:49390687-49890687	1	0.04	0	0
ssa25:28389273-28889273	1	0.19	0	0.01

DISCUSSION

Despite that combined effects of multiple variants at trait-associated loci are playing an important role in controlling fitness traits across a variety of species [11,20,21], our results indicate that sea age at maturation in Atlantic salmon is predominantly associated with single SNP variation at candidate regions. Using resequencing data to analyse 116 candidate loci and an analytical framework aimed at detecting multi-SNP associations, we find that single SNPs explain the variation in sea age at maturity in almost all cases. This work targeting candidate genes identified in aquaculture salmon strains suggests a mixed genetic architecture where a combination large-effect loci and smaller-effect loci also underlies age at maturity in wild Atlantic salmon populations. Two core loci, *vgll3* and *six6*, likely play a key role in determining age at maturity and additional smaller effect loci may be important for fine-tuning the trait across heterogeneous environments.

Theoretical modelling predicts that clustering of tightly linked adaptive mutations will occur under gene flow and selection in populations inhabiting spatially and/or temporally heterogeneous environments [22,23]. Although this seems to be a plausible scenario under which the genetic architecture of age at maturity has evolved in Atlantic salmon, our work suggests that the association in each of the candidate regions is driven by a single mutation. We cannot rule out, however, the possibility that the examined regions have pleiotropic effects and contain SNPs controlling other adaptive traits that have weak or no correlation with maturation timing. It is also possible that we did not have sufficient power to detect additional SNPs in these regions with small effects or with rare alleles. However, previous empirical studies have found few, but complex, loci with clusters of adaptive mutations [11,20,21], thus motivating our investigation of multi-SNP and haplotypic effects. Remington [24] also highlights the importance of distinguishing between allelic effects and single mutational effects when examining the genetic architecture of adaptive variation and its evolution. Our findings, however, suggest that alternative genetic architectures are feasible. One possible explanation could relate to the multiple whole genome duplication events that have occurred in Atlantic salmon and other salmonids [57]. The presence of multiple gene copies may impact the evolution of genetic architecture for traits such as age at maturity in Atlantic salmon. It is also possible that gene flow among Atlantic salmon populations is too restricted to neighbouring populations and/or strength of selection is insufficient for the establishment of linked mutations, as there is a rather specific balance of gene flow and selection required for clustered loci to arise [58]. Both an extension of models predicting genetic architecture and additional empirical studies – on a wider variety organisms and traits – are needed to evaluate the generality of particular architectures and to further understand the conditions under which they evolve.

We find additional evidence that a large-effect locus on ssa25, *vgll3*, largely underlies age at maturity in Atlantic salmon corroborating findings from a number of association studies on Atlantic salmon maturation [10,19,31,32,59]. The second strongest associated locus in this study is located in close proximity to *six6* on ssa09. This locus was previously found to be associated with early maturation in male farmed Atlantic salmon [19], with sea age at maturity in wild Atlantic salmon prior to population structure correction [10] and two species of Pacific salmon (Sockeye salmon and Steelhead trout) [36]. Additionally, we found another three loci associated with sea age at maturity: *pecam1*, *asap2aa* and *taar13c*. The handful of loci found here suggests that wild Atlantic salmon have a mixed genetic architecture where multiple loci, with a variety of effect sizes, control maturation timing – similar to what has been found in male farmed Atlantic salmon [19]. Knowledge of this mixed genetic architecture is highly relevant for how we predict the evolution of maturation timing in wild Atlantic salmon populations. A large body of work has shown the relevance of genetic architecture in determining evolutionary responses [60–68]. Recent works highlight the relevance of the genetic architecture underlying fitness traits when predicting a population’s response to environmental changes [69] and selective pressures such a fishing [70]. Future work elucidating how such mixed genetic architectures affect predicted evolution of traits, compared to that of omnigenic or polygenic architectures, will be valuable.

We find differences in locations of top-associated SNPs identified here and in Sinclair-Waters et al. [19]. This is not surprising given that we are examining sequence data that captures more SNP variation compared to SNP-array data used in Sinclair-Waters et al. [19]. Furthermore, we failed to find associations between sea age at maturity and many of the candidate regions identified in Sinclair-Waters et al. [19]. For example, several candidate regions on ssa03 and ssa04 displayed particularly strong association signals in aquaculture salmon, however, no signals at these regions were found here. Additionally, only one association peak at ssa06:27541960-28218141 was found here, whereas two independent associations within this region were found in aquaculture salmon [19]. Such differences may reflect changes in the genetic architecture of the trait evolving since the domestication of Atlantic salmon. Although, we would not expect large changes to occur given the domestication is relatively recent, just 10 to 15 generations ago [71]. Furthermore, this study is likely under-powered to detect all previously identified loci, particularly those with smaller effect sizes or rare alleles, due to smaller sample size. Additionally, there could be differences in genetic architecture among environments [72] and/or genotype by environment interactions giving rise to distinct genetic architectures in wild populations versus aquaculture strains.

We do not find strong evidence of multi-SNP associations at candidate loci examined in this study, however, we cannot yet disregard the utility of multi-SNP association methods for further resolving the genetic architecture of Atlantic salmon maturation. First, we do not examine the entire genome due to computational restrictions, rather, we focussed on 116 previously identified candidate regions. Second, the Atlantic salmon genome is highly complex [39] and therefore errors in the assembly that may be disruptive for haplotype-based analysis could exist. As new and improved versions of the Atlantic salmon genome are published, our ability to test for haplotypic associations will improve. Furthermore, in a few cases (ssa09:10915066-11415066, ssa09:24636574-25136574, ssa25:28389273-28889273) the *PiMASS* analyses post-regression of the top SNP selected no SNPs for a model explaining sea age at maturity variation, however, the median proportion of variance explained across all iterations was greater than zero. This may suggest that a weak signal was present, but was being missed due to insufficient power. Although this is largely speculative, it suggests that ruling out the possibility of multi-SNP associations at these particular candidate regions may be premature. Higher-powered studies (i.e. more individuals per population) may help to resolve this in the future.

In conclusion, our analytical framework, combining both single and multi-SNP association methods, reveals that single SNP variation is sufficient for explaining the association of previously identified candidate loci for Atlantic salmon maturation timing. Previous empirical and theoretical work have described trait-associated loci that have complex alleles with multiple variants, our findings therefore demonstrate the diversity of genetic architectures for fitness-related traits. Additional data, and a greater diversity of species and traits, will serve to better understand why this diversity of genetic architectures exists and how these particular genetic architectures evolve. The analytical framework used here will be a valuable resource for accomplishing this as individual-level resequencing data for wild species with phenotyped individuals becomes increasingly available.

Acknowledgements

Funding was provided by Academy of Finland (grant numbers 307593, 302873 and 327255), the Research Council of Norway (NFR-275310 and NFR-275862) and a Natural Sciences and Engineering Research Council of Canada postgraduate scholarship. Wild Atlantic salmon genome sequencing was funded by the Research Council of Norway (The Aqua Genome project; ref: 221734). We would like to acknowledge Terese Andersstuen, Dr Mariann Árnyasi and Hanna Hellerud Hansen from CIGENE for their work in organising the sequencing of samples. We thank Gunnel Østborg (NINA), Kurt Urdal (Rådgivende Biologer) and Natural Resources Institute Finland (LUKE) for their work collecting phenotype data. We also acknowledge the Aqua Genome project for providing access to data prior to public release. The Orion Computing Cluster at CIGENE-NMBU and CSC – IT Center for Science, Finland are acknowledged for computational resources. Storage resources were provided by the Norwegian National Infrastructure for

Research Data (NIRD, project NS9055K). Phenotype data was provided by the Norwegian Institute for Nature Research (NINA).

Data availability

Genome re-sequencing data for individuals used in this study are available in the European Nucleotide Archive (ENA) or NCBI with the project accession code PRJEB38061 [38].

Contributions

CRP, NJB, MSW conceived the study. TN developed the variant calling workflow and constructed the fixed assembly of *ssa25*. JW developed the variant filtering criteria. MSW performed all downstream analyses with input from NJB. MPK played key role in generating whole genome sequencing data. SL led the whole genome sequencing work as part of the AquaGenome project. HS, GHB, BFL, CRP coordinated Atlantic salmon sampling and provided phenotypic information. MSW, CRP, NJB drafted the manuscript. All authors commented on and approved the final manuscript.

Competing interests

There are no competing interests.

References

1. Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. Rare and low-frequency coding variants alter human adult height. Nature. 2017;542(7640):186–90.

2. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat Rev Genet. 2017;

3. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169(7):1177–86.

4. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. PLoS Genet. 2015;11(4):1–22.

5. Loh PR, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet. 2015;47(12):1385–92.

6. Fisher R. The correlations between relatives on the supposition of mendelian inheritance. Philos Trans R Soc Edinburgh. 1918;52:399–433.

7. Pritchard JK, Di Rienzo A. Adaptation - not by sweeps alone. Nat Rev Genet. 2010;11(10):665–7.

8. Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell. 2019;177(4):1022-1034.e6.

9. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet [Internet]. 2017;101(1):5–22. Available from: http://dx.doi.org/10.1016/j.ajhg.2017.06.005

10. Barson NJ, Aykanat T, Hindar K, Baranski M, Bolstad GH, Fiske P, et al. Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. Nature. 2015 Dec 17;528(7582):405–8.

11. Linnen CR, Poh Y-P, Peterson BK, Barrett RDH, Larson JG, Jensen JD, et al. Adaptive evolution of multiple traits through multiple mutations at a single gene. Science (80-). 2013;339(6125):1312–6.

12. Lamichhaney S, Fan G, Widemo F, Gunnarsson U, Thalmann DS, Hoeppner MP, et al. Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). Nat Genet. 2015;48(1):84–8.

13. Küpper C, Stocks M, Risse JE, Dos Remedios N, Farrell LL, McRae SB, et al. A supergene determines highly divergent male reproductive morphs in the ruff. Nat Genet. 2015;48(1):79–83.

14. Kirubakaran TG, Grove H, Kent MP, Sandve SR, Baranski M, Nome T, et al. Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. Mol Ecol. 2016;25:2130–43.

15. Sinclair-Waters M, Bradbury IR, Morris CJ, Lien S, Kent MP, Bentzen P. Ancient chromosomal rearrangement associated with local adaptation of a post-glacially colonized population of Atlantic Cod in the northwest Atlantic. Mol Ecol [Internet]. 2017;(October):1–13. Available from: http://doi.wiley.com/10.1111/mec.14442

16. Pearse DE, Barson NJ, Nome T, Gao G, Campbell MA, Abadía-Cardoso A, et al. Sex-dependent dominance maintains migration supergene in rainbow trout. bioRxiv [Internet]. 2018;504621. Available from: https://www.biorxiv.org/content/early/2018/12/22/504621.article-metrics

17. Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, et al. polymorphic supergene controlling butterfly mimicry. Nature. 2011;

18. Sinnott-Armstrong N, Naqvi S, Rivas MA, Pritchard JK. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. bioRxiv. 2020;2020.04.20.051631.

19. Sinclair-Waters M, Ødegård J, Korsvoll SA, Moen T, Lien S, Primmer CR, et al. Beyond large-effect loci: large-scale GWAS reveals a mixed large-effect and polygenic architecture for age at maturity of Atlantic salmon. Genet Sel Evol [Internet]. 2020;52(1):9. Available from: https://doi.org/10.1186/s12711-020-0529-8

20. Bickel RD, Kopp A, Nuzhdin S V. Composite effects of polymorphisms near multiple regulatory elements create a major-effect QTL. PLoS Genet. 2011;7(1):1–8.

21. Kerdaffrec E, Filiault DL, Korte A, Sasaki E, Nizhynska V, Seren Ü, et al. Multiple alleles at a single locus control seed dormancy in Swedish Arabidopsis. Elife [Internet]. 2016 Dec 14;5(3):1–24. Available from: http://elifesciences.org/lookup/doi/10.7554/eLife.22502

22. Yeaman S, Whitlock MC. The genetic architecture of adaptation under migration-selection balance. Evolution (N Y). 2011;65(7):1897–911.

23. Yeaman S. Genomic rearrangements and the evolution of clusters of locally adaptive loci. Proc Natl Acad Sci U S A [Internet]. 2013;110:E1743-51. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3651494&tool=pmcentrez&rendertype=abstract

24. Remington DL. Alleles versus mutations: Understanding the evolution of genetic architecture requires a molecular perspective on allelic origins. Evolution (N Y). 2015;69(12):3025–38.

25. Stearns SC. Life history evolution: Successes, limitations, and prospects. Naturwissenschaften. 2000;87(11):476–86.

26. Mobley KB, Aykanat T, Czorlich Y, House A, Kurko J, Miettinen A, et al. Maturation in Atlantic salmon (Salmo salar, Salmonidae): a review of ecological, genetic, and molecular processes. FEBS Lett. 2020;(November):1–58.

27. Fleming IA, Einum S. Reproductive ecology: a tale of two sexes. In: Atlantic Salmon Ecology. 2011. p. 35–65.

28. Mobley KB, Granroth-Wilding H, Ellmén M, Orell P, Erkinaro J, Primmer CR. Time spent in distinct life history stages has sex-specific effects on reproductive fitness in wild Atlantic salmon. Mol Ecol. 2020;29(6):1173–84.

29. Gjerde B. Response to individual selection for age at sexual maturity in Atlantic salmon. Aquaculture. 1984;38(3):229–40.

30. Reed TE, Prodöhl PA, Bradley C, Gilbey J, McGinnity P, Primmer CR, et al. Heritability estimation via molecular pedigree reconstruction in a wild fish population reveals substantial evolutionary potential for sea-age at maturity, but not size within age-classes. Can J Fish Aquat Sci [Internet]. 2018;cjfas-2018-0123. Available from: http://www.nrcresearchpress.com/doi/10.1139/cjfas-2018-0123

31. Ayllon F, Kjærner-Semb E, Furmanek T, Wennevik V, Solberg MF, Dahle G, et al. The vgll3 Locus Controls Age at Maturity in Wild and Domesticated Atlantic Salmon (Salmo salar L.) Males. PLoS Genet. 2015;11(11):1–15.

32. Sinclair-Waters M, Piavchenko N, Ruokolainen A, Aykanat T, Erkinaro J, Primmer CR. Refining the genomic location of SNP variation affecting Atlantic salmon maturation timing at a key large-effect locus. bioRxiv [Internet]. 2021; Available from: https://www.biorxiv.org/content/early/2021/04/26/2021.04.26.441431

33. Halperin DS, Pan C, Lusis AJ, Tontonoz P. Vestigial-like 3 is an inhibitor of adipocyte

differentiation. J Lipid Res. 2013;54(2):473–81.

34. Day FR, Thompson DJ, Helgason H, Chasman DI, Finucane H, Sulem P, et al. Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. Nat Genet. 2017;49(6):834–41.

35. Perry JRB, Day F, Elks CE, Sulem P, Thompson DJ, Ferreira T, et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature. 2014 Jul 23;514:92.

36. Waters CD, Clemento A, Aykanat T, Garza JC, Naish KA, Narum S, et al. Heterogeneous genetic basis of age at maturity in salmonid fishes. Molcular Ecol. 2021;

37. Cánovas A, Reverter A, DeAtley KL, Ashley RL, Colgrave ML, Fortes MRS, et al. Multi-tissue omics analyses reveal molecular regulatory networks for puberty in composite beef cattle. PLoS One. 2014;9(7):1–17.

38. Bertolotti AC, Layer RM, Gundappa MK, Gallagher MD, Pehlivanoglu E, Nome T, et al. The structural variation landscape in 492 Atlantic salmon genomes. Nat Commun [Internet]. 2020;11(5176). Available from: https://doi.org/10.1101/2020.05.16.099614

39. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016 May 12;533(7602):200–5.

40. Chapman B, Kirchner R, Pantano L, Smet M De, Beltrame L, Khotiainsteva T, et al. bcbio/bcbio-nextgen: v1.2.3. 2020 Apr 7 [cited 2020 Sep 17]; Available from: https://zenodo.org/record/3743344

41. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [Internet]. 2013 Mar 16 [cited 2020 Sep 17]; Available from: https://arxiv.org/abs/1303.3997

42. Picard toolkit. Broad Institute, GitHub repository. Broad Institute; 2019.

43. Depristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–501.

44. 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(2):80–92.

45. Li H. SNPable Regions [Internet]. 2009. Available from: http://lh3lh3.users.sourceforge.net/snpable.shtml

46. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul;25(14):1754–60.

47. Danecek P, Auton A, Abecasis G, Albers C, Banks E, DePristo M. The variant call format and vcftools. Bioinformatics. 2011;27(15):2156–2158.

48. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D. Plink: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet [Internet]. 2007;81. Available from: https://doi.org/10.1086/519795

49. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006;2(12).

50. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.

51. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC Bioinformatics. 2009;10:1–9.

52. Xu H, Guan Y. Detecting local haplotype sharing and haplotype association. Genetics. 2014;197(3):823–38.

53. Guan Y. Detecting structure of haplotypes and local ancestry. Genetics. 2014;196(3):625–42.

54. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015 Feb;4(1):7.

55. Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann Appl Stat. 2011;5(3):1780–815.

56. Harold Jeffreys. The Theory of Probability. 2020. 470 p.

57. Allendorf FW, Thorgaard GH. Tetraploidy and the Evolution of the Salmonid Fishes. Springer. Monographs in Evolutionary Biology. Boston; 1984. 55–93 p.

58. Yeaman S, Aeschbacher S, Bürger R. The evolution of genomic islands by increased establishment probability of linked alleles. Mol Ecol [Internet]. 2016 Jun 1;25(11):2542–58. Available from: https://doi.org/10.1111/mec.13611

59. Ayllon F, Solberg MF, Glover KA, Mohammadi F, Kjærner-semb E, Fjelldal PG, et al. The influence of vgll3 genotypes on sea age at maturity is altered in farmed mowi strain Atlantic salmon. BMC Genet. 2019;20(44):1–8.

60. Barton NH, Turelli M. Natural and sexual selection on many loci. Genetics. 1991 Jan;127(1):229–55.

61. Turelli M. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. Theor Popul Biol. 1984 Apr;25(2):138–93.

62. Turelli M, Barton NH. Polygenic Variation Maintained by Balancing Selection: Pleiotropy, Sex-Dependent Allelic Effects and G × E Interactions. Genetics. 2004;166(2):1053–79.

63. Turelli M, Barton NH. Dynamics of polygenic characters under selection. Theor Popul Biol. 1990;38(1):1–57.

64. Lande R. The maintenance of genetic variability by mutation in a polygenic character with linked loci. Genet Res. 2009/04/14. 1975;26(3):221–35.

65. Bulmer MG. The genetic variability of polygenic characters under optimizing selection, mutation and drift. Genet Res (Camb). 1972;19(1):17–25.

66. Débarre F, Yeaman S, Guillaume F. Evolution of Quantitative Traits under a Migration-Selection Balance : When Does Skew Matter ?*. Am Nat. 2015;186.

67. Fisher R. The genetical theory of natural selection. Clarendon, Oxford; 1930.

68. Yeaman S. Local Adaptation by Alleles of Small Effect *. Am Nat. 2015;186.

69. Kardos M, Luikart G. The genetic architecture of fitness drives population viability during rapid environmental change. Am Nat. 2021;

70. Oomen RA, Kuparinen A, Hutchings JA. Consequences of Single-Locus and Tightly Linked Genomic Architectures for Evolutionary Responses to Environmental Change. J Hered [Internet]. 2020;319–32. Available from: https://academic.oup.com/jhered/article/111/4/319/5867197

71. Gjerde B, Gjedrem T. Estimates of phenotypic and genetic parameters for carcass traits in Atlantic salmon and rainbow trout. Aquaculture. 1984;36(1–2):97–110.

72. Yan W, Wang B, Chan E, Mitchell-Olds T. Genetic architecture and adaptation of flowering time among environments. New Phytol [Internet]. 2021 Jan 23;n/a(n/a). Available from: https://doi.org/10.1111/nph.17229