Norwegian University
of Life Sciences

Master's Thesis 2021 60 ECTS
Faculty of Biosciences

Assessing the effects of the new Atlantic salmon (*Salmo salar*) genome assembly on imputation accuracy

Domniki Manousi
Animal Breeding and Genetics

Table of contents

Special thanks

I would like to first thank my supervisors, Simen Rød Sandve and Tim Martin Knutsen for their continuous help, support and guidance throughout this experience, I am very fortunate and grateful to have worked by your side. Also, thanks to the companies CIGENE and AQUAGEN for hosting my master's project and providing me with the computational power as well as the genotyping information to conduct the present study. I thank Kristine S. R. Stenløkk for kindly providing me with data on Atlantic salmon structural variation signals but also for taking the time to discuss them with me and answer all of my questions. Finally, special thanks go to my friends and family for their unconditional patience, love and support, during my times of stress, self-doubt and frustration.

June 1st , 2021

Summary

The Atlantic salmon is one of the most economically important species in modern-day aquaculture. For this reason, a lot of effort has been put into implementation and improvement of breeding programs for this species, achieving vast genetic process in a considerably short period of time. Improvements in sequencing technologies have facilitated the use of genomic selection, integrating molecular genetic information and increasing selection response for key production traits of polygenic architecture. However, implementation of genomic selection requires large, densely genotyped populations, which can prove challenging, especially considering aquatic populations. Genotype imputation therefore, constitutes a cost-efficient method that amplifies the genotyping density of large populations, allowing them to be analyzed in low-density and cost genotyping platforms. Although at the time of the first Atlantic salmon genome assembly leading sequencing and bioinformatic methods were applied to assemble the genome reference, the high genomic complexity of the species severely impacted the quality of the produced assembly. Assembly errors are expected to primarily affect genotyping quality and consequently all downstream analyses. The recent release of a new genome assembly for Atlantic salmon (NCBI GeneBank reference: GCA_905237065.2), constructed using long-read sequencing technologies, is expected to improve our understanding of salmon genetics and genomics as well as contribute to the application of higher-quality genomic data in salmon breeding. In this study we explored the improvements achieved in the new genome assembly, as these were realized through a genotype imputation analysis using a small sample of immediate relatives. We report large structural changes occurring in the new genome assembly and discuss their impact on imputation accuracy as well as on currently available genotyping platforms. We also provide potential considerations regarding local heterogeneity of imputation accuracy in relationship to salmon's high genomic complexity as well as occurrence of structural variation elements. Finally, we discuss possible strengths and weaknesses of different imputation approaches relative to our experimental sample limitations.

Introduction

Responding to the ever-increasing world demand for food, aquaculture has dramatically improved in productivity and intensity over the past decades (Ross D. Houston, 2017; R. D. Houston & Macqueen, 2019). The increase in production came as a result of selectively breeding individuals for traits of economic interest. Traditionally, improvement of such traits was achieved by estimating the genetic contribution of the breeding candidates through pedigree information and phenotypic expression records (Ross D. Houston, 2017; T. Meuwissen et al., 2016)

Dawning of the genomics era facilitated Genome Wide Association studies (GWAS) to identify and integrate molecular genetic information into animal breeding, increasing selection response. Genetic loci highly associated with traits of interest (QTL) were first identified through GWAS. Analysis of these QTL provided insight about the genetic makeup of key production traits, narrowing down genetic and environmental variance and ultimately increasing precision of selection (T. Meuwissen et al., 2016). The method under which the effects of such variants were considered in the estimation of breeding values (EBV) was termed Marker Assisted Selection (MAS). Applying MAS led to improvements in aquaculture, especially regarding disease control (Gonen et al., 2015; Moen et al., 2009; Moen et al., 2015). Although powerful as a method, application of MAS came with certain limitations since expression of most key-production traits is not the result of a single yet powerful QTL but rather the cumulative effect of several loci of small contribution each. Due this polygenic architecture, a large portion of genetic variation for important traits could not be explained by methods such as MAS (T. Meuwissen et al., 2016).

Supported by the advances in sequencing technologies, the discovery and utilization of a large number of SNP markers facilitated the establishment of the Genomic Selection (GS) theory in 2001 (T. H. Meuwissen et al., 2001). GS had a drastic impact in aquaculture, leading to the re-design of thus far applied breeding schemes (Fernandez et al., 2014; Meuwissen et al., 2001; Tsai, Hamilton, Guy, et al., 2015; Tsai, Hamilton, Tinch, et al., 2015). Due to its several benefits, the GS method was widely adopted and has been routinely implemented since. To accurately estimate the genetic merit of selection candidates, GS depends on a densely genotyped reference population with phenotypic records on a trait of interest. Using the information obtained from the reference population, the phenotypic effects of each genotype are estimated. These marker effects are then applied on a population of genotyped only selection candidates and genomic predictions are established, evaluating the candidates' breeding potential (GEBV). Intuitively, precision of GS relies heavily upon reference population size and genotyping density of the reference individuals, but also on the genotyping density of the selection-candidates (Calus et al., 2014).

Improvement in sequencing technologies resulted in increasing accessibility to genotype information. However, genotyping large populations using high-density platforms is still very costly and practically challenging, especially considering the fecundity and family size of aquatic species such as salmon. For this reason, the necessity for establishment of cost-efficient ways to extract a maximum of genomic information from big cohorts, becomes evident. One solution is to use genotype imputation - a cheap and efficient strategy to estimate unknown genotype information and increase genotyping density in large populations (Kijas et al., 2017; Tsai et al., 2017; Tsairidou et al., 2020).

Genotype imputation

Genotype imputation is a key aspect of modern breeding programs. It is an *in-silico* genotype inference method that provides access to genomic information of large cohorts using a smaller, densely genotyped reference population (Whalen et al., 2018). The basic framework behind imputation is the exploitation of linkage disequilibrium (LD) between markers and/or family information. Implementation of imputation differs depending on the assumption of relatedness between the reference and target populations. Family-based

imputation is mainly applied in animal population studies due to the strong family structure observed in commercially bred populations and uses linkage and Mendelian segregation rules as well as pedigree information to infer missing genotypes (Sargolzaei et al., 2014). On the other hand, population-based imputation methods consider all participating individuals to be unrelated and thus use probabilistic methods to model haplotype frequency between genotyped markers (Browning et al., 2018; Loh et al., 2016). In addition, there are hybrid methods that perform imputation by combining population and family based approaches (Hickey et al., 2011; Sargolzaei et al., 2014)

Result of genotype imputation is the inference of unknown genotypes in the genomes of the target individuals. In this manner, genotype imputation is used as a “stepping-stone” to provide sparsely genotyped populations with the high -even sequence level- genotype resolution required to carry out a series of downstream analyses, such as association studies and genomic selection.

Imputation approaches

1. Family based imputation using FImpute

Family based imputation methods assume the presence of relationships between individuals in the reference and target populations. Under this assumption, identical by descent (IBD) haplotype segments (haplotype segments inherited from a common ancestor) can be identified within populations (Y. Li et al., 2009). Depending on the level of relationship, longer haplotype segments are shared between closely related individuals, while considerably shorter IBD haplotype segments are shared between more distant relatives (Sargolzaei et al., 2014). In the context of genotype imputation, relatedness can be defined through pedigree information input or by phasing all individuals’ genotypes (i.e. estimating haplotypes) followed by identification of haplotype similarities. In 2008, Kong et al. introduced the heuristic method of Long-Range Phasing (LRP), establishing the concept of surrogate parents (Kong et al., 2008). Using related or “seemingly unrelated” individuals, LRP allowed the construction of relationships through pooling of IBD segments identified within populations. By constructing relationships, the LRP method allowed family-based imputation to be applied even when pedigree input was either ambiguous or incomplete. One commonly used imputation software that infers missing genotypes by utilizing the LRP method in addition to pedigree information is FImpute (Sargolzaei et al., 2014).

FImpute first uses the pedigree information to identify haplotype segments (haplotype discovery) shared between related individuals in the target and reference populations. It then infers missing alleles between these haplotypes, taking into consideration potential crossover events. Iterating and rotating haplotype discovery between parents and offspring, haplotype information is gathered and used to construct a “ haplotype library”. After all reported pedigree relationships have been analyzed, FImpute proceeds to search for potential relationships between individuals without pedigree information (parentage discovery). To do that, the entire genome is assessed; A progressively decreasing overlapping sliding window is used in chromosomal order to detect linkage across haplotypes, based on the theory that closer relatives share longer and less frequently observed haplotype segments than more distantly related individuals. The haplotype library is constantly updated throughout the process and is finally used to estimate haplotype frequencies. Using the estimated frequencies, the remaining missing alleles are finally inferred. As a result, FImpute can accommodate for both the presence and absence of pedigree information, which is of utmost importance in the animal breeding and aquaculture industry. In instances where pedigree information is not provided, identification of extant relationships between target and reference individuals can also be reconstructed using parentage assignment methods (Grashei et al., 2018; Vandeputte & Haffray, 2014).

2. Population based imputation using Beagle

Mainly targeted at human studies, population-based imputation approaches assume that the reference and target populations consist of either very distantly or even unrelated individuals. Population-based methods exploit the non-random association (LD) between alleles of individual markers that reside in close proximity. Based on exploitation of LD between markers, fine-scale recombination maps are highly recommended in order to increase inference accuracy. To model haplotype frequencies, population-based methods operate under the main framework of the Hidden Markov Model (HMM), as this model was proposed by Li and Stephens (Li & Stephens, 2003; Sargolzaei et al., 2014; Whalen et al., 2018). The way the HMM model is implemented however, greatly varies across different imputation algorithms (S. R. Browning & Browning, 2011). One of the most commonly used population-based imputation software is Beagle (B. L. Browning et al., 2018). Beagle performs haplotype phasing and genotype imputation in two separate steps. A brief description of the two steps in Beagle is provided here.

At its first step, haplotype phasing, Beagle induces a localized haplotype cluster model to define the haplotypes constituting the target and reference population individuals (S. R. Browning & Browning, 2007). The haplotype cluster model can be represented as a leveled acyclic graph consisting of several nodes and edges. The graph begins as a single -root- node and ends at a single -terminal- node including several nodes in-between the root and terminal nodes; all in-between nodes are sectioned in levels. Considering a sample of haplotypes, each level of the graph denotes a marker for which the haplotypes are genotyped. The nodes included within each level represent the alleles observed in a given genotyped marker. Moving through levels, nodes split and merge depending on the alleles observed in each marker, while edges (clusters) connect consecutive nodes, estimating the predictive likelihood of observing specific combinations of genotyped alleles, i.e. haplotypes, in the sample. After assessing each genotyped marker of the haplotype sample, the model reaches the terminal node, completing the acyclic model. Under this model, each haplotype in a given sample can be visualized as an individual path from the root till the terminal node of the graph, passing through one node per level. The haplotype cluster model can thus be considered as a special class of HMM (N. Li & Stephens, 2003). In the case of Beagle, the local haplotype cluster graph is extended to a diploid form in order to assess the unphased genotypes of individuals. The phasing step iterates, sampling and evaluating genotypes given the haplotype cluster model and the sampled population. The last iteration outputs the most-likely pair of haplotypes for each genotyped individual, conditional to the model and the individuals' genotype (S. R. Browning & Browning, 2007).

The second step of Beagle uses the already phased populations to perform genotype imputation through the implementation of an additional HMM model (B. L. Browning et al., 2018). To impute the missing genotypes, Beagle uses the pre-phased information to first concatenate tightly linked markers (0.005 cM) of target haplotypes and through the HMM estimates haplotype probabilities. Then, the algorithm creates haplotype "mosaics" (*composite haplotypes*) of the reference population, based on segmental identity between the reference and target population haplotypes' alleles (IBS). Long IBS segments sharing identity between target and reference individuals are considered to contain at least one segment of identical alleles due to segregation (IBD). Composing reference haplotypes that conclude the target population constitutes an effective way to reduce the HMM state space and increase computational speed (B. L. Browning et al., 2018). Finally, Beagle uses linear extrapolation in the HMM model to impute markers in the target population according to the reference (B. L. Browning & Browning, 2016).

Although initially designed for humans, the population-based approach was eventually introduced into animal breeding and aquaculture as the best-fitting option for genotype imputation of unrelated individuals or for instances of unavailable, incomplete or unreliable pedigree information (Bolormaa et al., 2019).

Quantification of imputation accuracy

The success and precision of all analyses following imputation relies heavily upon the accuracy to which the ungenotyped information is inferred in populations. For that reason, evaluation of imputation accuracy

is of utmost importance. Genotype imputation can be assessed through different perspectives (marker-wise or individual-wise accuracy) and different metrics, depending on the availability of genotype information as well as the objective of the imputation analysis (Calus et al., 2014). Most commonly used metrics for imputation accuracy assessment are concordance, Imputation Quality Score (IQS) and squared correlation (r^2) between the imputed and true genotypes.

Concordance is among the simplest measures of imputation accuracy assessment. This metric estimates the rate of correctly imputed genotypes relative to the total genotypes inferred in a given marker. Although the output of concordance can prove valuable, estimation of this metric heavily relies on allele frequency (Calus et al., 2014). Due to this dependency, concordance results tend to be inflated for markers with low Minor Allele Frequency (MAF) as the metric does not account for the possibility of correct inference of the minor allele due to chance. Therefore concordance is considered as a useful yet quite unreliable measure of imputation accuracy (Rowan et al., 2019). Although concordance can prove relatively inaccurate, its implementation into more sophisticated methods can increase reliability of accuracy assessment (P. Lin et al., 2010).

Imputation Quality Score (IQS) is another often used metric of imputation accuracy (P. Lin et al., 2010). IQS estimation utilizes concordance but additionally adjusts imputation accuracy for the probability of correctly imputing genotypes by chance as well as for the non-random imputation errors that can occur when combining datasets from different genotyping platforms. In this manner, IQS proves to be a more robust metric for imputation accuracy against low MAF and therefore a more reliable accuracy assessment approach.

The squared correlation (r^2) between imputed and true genotypes is the most commonly used metric for assessing imputation accuracy. Unlike concordance, r^2 is independent of allele frequencies and can therefore account for low MAF estimates (Calus et al., 2014). Ultimately, r^2 is an estimate of how well the imputation analysis is to infer an alternative genotype when the minor allele for that genotype is less frequently -or even rarely- observed. Squared correlation is a powerful and highly reliable metric and for that reason it is often used as a quality threshold for imputed genotype information prior to downstream analyses (Yoshida & Yáñez, 2021).

Genotype imputation implications and weaknesses

Mentioned previously, imputation delivers to study individuals the genotype resolution required for a series of analyses (Marchini & Howie, 2010). Among the biggest advantages of the method is that the increase in genotype density is achieved in a significantly cost-efficient manner in comparison to investment in densely genotyping individuals, thus allowing larger cohorts to participate into studies. One of the most valuable uses of genotype imputation has been the combination of data from multiple studies that have used different genotyping platforms. Data are collected and combined in order to increase the testing population sample size and density, improving resolution of putative QTL and thus providing substantially more power in analyses such as association testing (Bernardes et al., 2019; Lin et al., 2010). GWAS analyses are used in order to identify the variant(s) affecting – or in association with - complex traits of economic and functional importance. To reliably identify causative variants, GWAS requires large number of meticulously phenotyped and genotyped individuals. In this manner, imputation has been extensively used to provide the high -or even whole genome sequence- density required for such analyses and increase the probability of defining rare markers associated with traits of interest (Bernardes et al., 2019; Tsai et al., 2015).

Genomic selection (GS) is another field benefitted from genotype imputation. Explained previously, GS heavily relies on genotyping density and population size to accurately estimate the effects for marker genotypes in the reference population and assess the genetic merit of selection candidates. Due to the family structure and size of breeding populations however, genotyping large cohorts using high density platforms can prove challenging (Tsairidou et al., 2020). Genotype imputation can therefore be used to

amplify the desirable genotyping density of animals typed using considerably low density -and consequently cost- genotyping panels.

There is a number of factors that can affect the accuracy and consequent impact of imputation. Most of these factors are related to characteristics of the reference population: size, genotyping density and composition. Discussed among accuracy quantification metrics, low MAF can have a negative impact on imputation accuracy. This is often attributed to sampling errors and software bias as markers with low MAF tend to be imputed according to the reference, regardless of their true genotype (Shi et al., 2018). To limit the negative impact of very low MAF on inference accuracy, appropriate quality control (QC) filtering is usually applied prior to the imputation process (S. Lin & Zhao, 2010).

Regarding the reference population, size and composition play a very important role primarily on genotype imputation and consequently in all analyses utilizing the inferred information. When performing family-based imputation in large cohorts, the existence of immediate or close relatives defines the rate of accuracy, while addition of a pedigree file can significantly enhance imputation performance (Sargolzaei et al., 2014). When closely related animals are not available, imputation accuracy will rely on the size of the reference population as well as the level of relationships between animals to estimate the LD between markers and correctly infer genotypes. When the reference population is closely related to the target individuals however, size of the reference population has a smaller effect (Calus et al., 2014). Conversely, the degree of relatedness loses impact with increasing genotyping density, as increasing the genotyping information intuitively improves haplotype resolution (Sargolzaei et al., 2014).

Composition of the reference genotypes is equally important; composition includes density and distribution of the markers on the genotyping panel. The density of the reference panel – and consequent SNP coverage of the population's genome - defines the resolution to which target individuals are analyzed. Lower array densities are not as able to provide reliable information regarding rare variants or variants in weaker LD with the genotyped markers (Bolormaa et al., 2019). Adding to that, the markers' distribution on the genotyping array is directly associated with genotyping coverage and ultimately accuracy of imputation, as LD decreases with increased distance between genotyped markers (Kijas et al., 2017). Finally, genotyping and mapping errors stemming from wrongly called genotypes and false mapping of genotyping markers in the genome reference, additionally leads to poor imputation results (van den Berg et al., 2019). False annotation of markers due to assembly errors in the genome reference can lead to loss of LD and decrease of imputation power. Results of this are less informative association analyses and compromised results during downstream analyses. As genotyping array markers are annotated according to reference genome assemblies, occurrence of such errors highlights the importance of improving the sequencing and assembly methods to yield higher quality and accuracy.

Reference genome and sequencing technologies

Over the past two decades, sequencing technologies have made vast progress in terms of throughput and efficiency (Ghosh et al., 2018; Goodwin et al., 2016; Shendure et al., 2017). Massively parallel -next generation- sequencing (NGS) enabled rapid and cheap DNA sequencing (Pareek et al., 2011). Along with advances in bioinformatic methods, NGS revolutionized our ability of *de novo* sequencing and assembly as well as re-sequencing of genomes, with large implications in animal breeding and aquaculture (Ghosh et al., 2018; Lien et al., 2016; Robledo et al., 2018; You et al., 2020). Cheap and fast sequencing allowed for the efficient discovery of a large number of single nucleotide polymorphisms (SNP) and construction of high resolution linkage maps (Gonen et al., 2014) which made it possible to design high density SNP genotyping arrays, used for powerful GWAS studies. (Houston et al., 2014; Yanez et al., 2016)

A major limitation of next-generation sequencing methods (from now on referred to as 2nd generation sequencing methods) however, is the relatively short length of the sequences this technology produces.

Short sequence lengths impact genome assembly contiguity, especially for genomic regions containing highly repetitive content as well as for genomes characterized by high complexity. The latter instance can be particularly encountered in species that have experienced a relatively recent Whole Genome Duplication (WGD) (Robledo et al., 2018; You et al., 2020). Consequently, genomic regions with high repeat content, structural variation occurring between haplotypes or with duplicated copies in several places in the genome, were often erroneously assembled (Bickhart et al., 2017; Lien et al., 2016).

To overcome the limitations of 2nd generation sequencing methods, third generation sequencing technology has been developed (Bickhart et al., 2017; Sedlazeck et al., 2018). Sequence reads produced by these instruments can vary from 1 Kbp up to even 1 Mbp, drastically improving the contiguity of genome assemblies (Goodwin et al., 2016; Sedlazeck et al., 2018). However, these longer reads have also been characterized by prohibitive cost as well as increased sequencing error rates, due to low sequencing coverage and also homopolymer tracts (Kraft & Kurth, 2020; Shafin et al., 2020). Since the encountered sequencing errors are randomly distributed, different solutions have been developed to minimize the negative impact; both in terms of computational and sequencing technology (Sedlazeck et al., 2018). Presently, the two major 3rd generation sequencing technologies on the market have read base accuracy >95% and assembly base quality (provided that the sequencing coverage is satisfactory) that produces genome assemblies with error rates comparable to the 2nd generation genome assemblies (Jain et al., 2016; Roberts et al., 2013).

The Atlantic salmon

Atlantic salmon (*Salmo salar L.*), hereby referred to as salmon, is one of the most economically and biologically important species of modern-day aquaculture (R. D. Houston & Macqueen, 2019). In addition, salmon is in the scientific forefront when it comes to implementing breeding programs and biotechnological tools such as marker assisted selection and more recently genomic selection (Gjedrem & Rye, 2018; Gonen et al., 2014; Ross D. Houston, 2017; R. D. Houston & Macqueen, 2019).

In 2016, the first ever chromosome level genome assembly for Atlantic salmon was published (NCBI GeneBank reference: GCA_000233375.4) (Lien et al., 2016). At the time of assembly leading sequencing and computational methods were employed, combining longer (600bp) sanger sequencing reads with short (150bp) 2nd generation sequencing reads. As a result, the assembled reference genome of salmon consisted of more than 300,000 pieces. Consequent to the high fragmentation of the sequenced genome, certain portion of the sequence was not able to be assigned to a chromosome (Lien et al., 2016), while regions harboring duplication and repeat elements are believed to be at high risk of being absent from the genome assembly or falsely positioned. One of the largest sources for genome assembly errors are related to the whole genome duplication (WGD) event the Salmonidae family underwent 80-100 Mya (Lien et al., 2016; Robertson et al., 2017). Although this event is old and most duplicated chromosomal regions from this event are distinctly different today (85-90% sequence similarity), almost 25% of the salmonid genome has experienced delayed and incomplete rediploidization. Such genomic segments are referred to as regions of lineage specific ohnolog resolution (LORe) (Robertson et al., 2017). As LORe regions share much higher sequence similarity across duplicated segments compared to the genome average, they have been very challenging to assemble using short read sequencing technologies.

To improve this salmon genome assembly, particularly in the LORe genomic regions, a new genome assembly for Atlantic salmon was constructed using Oxford Nanopore long-read sequencing technologies (NCBI GeneBank accession GCA_905237065.2). This genome assembly is expected to provide new and exciting opportunities, further improving our understanding of salmon genetics and genomics. In addition, the new genome assembly is anticipated to help make a significant leap in the application of genomic data in salmon breeding.

Aims and objectives

The main objective of this study was to explore the features and improvements of the newly released Atlantic salmon genome assembly, through the scope of genotype imputation, an analysis of routine application and high significance in modern aquaculture. This study is organized as follows. First, we carry out an imputation software comparison to define the best performing imputation strategy when a considerably small sample of immediate relatives is only available (parents-offspring). Then we perform genotype imputation using the current and recently released genome assembly versions (referred to as ‘old’ and ‘new’, respectively) to investigate the features of the new assembly, as these are reflected on genotype imputation accuracy. Finally we examine the impact of imputation analysis parameters (reference population size, genotyping density and filtering quality) on imputation accuracy, given the new genome assembly enhancements.

Methods

Animals and SNP genotyping data

No animal experiments were conducted in this study, therefore we did not deal with ethical consideration requirements during the conduction of the following analyses. For all conducted analyses, genotype information from 1,310 Atlantic salmon individuals was provided by the aquaculture company AquaGen AS. The cohort consisted of fish from the breeding nucleus reared between 1998 and 2012. No pedigree information was provided prior to the analysis.

All individuals have been genotyped using a proprietary xHD Affymetrix Axiom array (~930.000 SNP markers). This array is hereafter referred to as high-density SNP (HighD-SNP) panel. In addition, a subset of the study cohort, 226 fish reared in the year class of 2008, were genotyped with the proprietary 70kv1 Affymetrix Axiom array (~70,000 SNP markers). This array is hereafter referred to as low-density SNP (LowD-SNP) panel.

Genome assemblies

We performed genotype imputation using two different genome assemblies referred to as the 'old' and 'new' assembly. The old assembly (Lien et al., 2016), was generated by a combination of sanger- and Illumina short read sequencing and scaffolded using linkage information (NCBI GeneBank accession GCF_000233375.1). This assembly has a size of 2.61 Gbp and consists of 368,060 contigs. The new assembly (NCBI GeneBank accession GCA_905237065.2) is based on Oxford Nanopore long read sequencing technology (Jain et al., 2016) and is scaffolded using a combination of linkage information and HiC-data. This assembly has a size of 2.76 Gbp and consists of 4,000 contigs. The new assembly therefore has a 90-fold improvement in assembly contiguity and is expected to contain less assembly errors, therefore representing a significantly improved model for the Atlantic salmon genome. A detailed comparison of the assembly quality per se was out of scope for this thesis.

Reposition of SNPs on the new genome assembly

Because of an increase in chromosome anchored genome sequence and intragenomic reordering of sequence in the new genome assembly, we had to re-assign genomic positions for the SNP markers present on the genotyping arrays that we examined. To achieve this, we extracted the genomic sequence flanking the SNPs (35 bp on each side) and mapped these SNP-containing sequences to the new genome using the Burrows – Wheeler Aligner (BWA, version 0.7.17)(Li, 2013). For each sequence, BWA reports the mapping coordinates in the genome along with a summary statistics score (MAPQ) that reflects the overall mapping quality of the sequence. MAPQ has a range between 0 and 37, indicating the probability that a given query sequence is falsely mapped. This probability can be calculated as:

$$P = 10^{\left(-\frac{i}{10}\right)},$$

where *i* represents the MAPQ quality score. A mapping quality of 37 translates to P=0.0001 or a 0.01 percent chance that the examined sequence has been incorrectly aligned on the genome reference. All analyses performed in this study used only markers mapped with MAPQ score equal to 37. Chromosome number, physical position coordinates and mapping quality information for these markers was used to assign genomic SNP positions in the new assembly for the tested genotype arrays. A software pipeline for SNP-array marker re-mapping is available from AquaGen AS upon request. BWA output filtering and array update

were performed using the software PLINK v1.9 (function --*update-chr*) and v2.0 (function ---*update-map*), in a Linux cluster environment (Chang et al., 2015; Purcell et al., 2007).

Data quality control (QC) filtering

Quality filtering

Quality control (QC) filtering was performed using PLINK v1.9 and v2.0 (Chang et al., 2015; Purcell et al., 2007). We used PLINK v1.9 to remove markers having the same physical position, reference and alternative alleles, while the rest of quality filtering was carried out using PLINK v2.0. The reason behind using two different software versions of the same software is because between earlier (v1.9) and later (v2.0) developed versions, utility of certain analysis tools has been modified. Using PLINK v2.0 (Chang et al., 2015) we identified and discarded variants with more than 2% missing genotypes and minor allele frequency less than 2.5%. Individuals with more than 10% missing genotypes were removed, as well as variants deviating from the Hardy Weinberg equilibrium (HWE) with a p-value < 10^{-10}. Variants with duplicated physical positions were identified using the function *duplicated()* in R (R core team) and were discarded through the function --*exclude* in PLINK v2.0.

Matching markers across SNP panels and chromosome strands

For analyses utilizing multiple SNP panels, it is important that all genotype information is assessed according to the same reference. For this reason, SNP markers in the target panel that are not present in the reference must be discarded. In addition, alleles of common markers between SNP panels have to be called from the same chromosome strand. To fulfill these requirements, we employed the strand-alignment software conform-gt (version 24May16.cee.jar; Browning, 2016) in order to :

a) Identify the common markers between SNP panels and discard markers found only on the target density SNP panel

b) Identify the chromosome strand for markers commonly present between genotyping panels and discard markers whose strand cannot be defined

c) Match reference and alternative alleles for the same markers between reference and target density SNP panels

We assigned the alleles of markers in the LowD-SNP panel using the HighD-SNP panel as reference for the two genome assemblies and identified matching markers based on their mapping position in the two panels. As conform-gt examines each chromosome individually, we used the *BCFtools* genomic analysis toolkit from the SAMtools software suite (Li, 2011; Li et al., 2009) to merge all chromosome output into a single file.

Finally, in order to compare imputation accuracy between different genome assemblies, we examined only common SNP markers between the LowD- and HighD-SNP panels that were assigned to a unique genome location in both versions of the Atlantic salmon genome assembly (i.e. only a single match in the genome in the BWA sequence mapping). These common variants were first identified between assembly SNP panels using the R function *inner_join()* and then retained using the function --*extract* in PLINK v2.0.

Quality control of genomic relationships

The FImpute algorithm tested in this thesis (Sargolzaei et al., 2014), can utilize pedigree information to accurately detect shared haplotype segments between more or less related individuals. Since pedigree information for the study individuals was not available *a priori*, pedigree relationships were re-constructed

based on parentage assignment methods using genomic relationship likelihoods (GRL) (Grashei et al., 2018). GRL can be obtained from AquaGen AS upon request. In our study, the candidate parents consisted of 204 fish from the 2005 year class while the putative offspring group included 216 fish from the 2008 year-class.

Genotype imputation

Evaluation of SNP genotype imputation tools for genome assembly comparisons

We performed an imputation experiment to assess the performance of two routinely used imputation software, namely FImpute v3 (Sargolzaei et al., 2014) and Beagle v5.2 (B. L. Browning et al., 2018; S. R. Browning & Browning, 2007). The two imputation algorithms are based on different theoretical approaches to perform haplotype identification (phasing) and infer missing genotypes. Details about how these imputation algorithms work are outlined in the introduction section "*Imputation approaches*". The software comparison was carried out by imputing 450,368 genotypes in 195 individuals (from LowD-SNP panel up to a HighD-SNP panel), using their respective parents (90) genotyped on the HighD-SNP panel as the reference population.

The two imputation algorithms required different curation of input data. For Beagle we used the Variant Call Format (.vcf) output file, produced by PLINK v2.0 during SNP QC filtering. On the other hand, for FImpute the software PLINK v2.0 was used to first convert the SNP QC output into R-acceptable allelic dosage format (.traw). A custom R script was then applied to adjust the required information into format compatible with FImpute. FImpute analysis was executed using the software's default parameters (Sargolzaei et al., 2014). For Beagle, some default parameters required modification due to density and mapping coordinates of the LowD-panel variants. Window size was set to 39.0 cM and overlap to 5.0 cM. In addition, effective population size for the analysis was adjusted to ne=200. Although Beagle does not utilize pedigree information, inclusion of a recombination map is highly recommended. In this study however, a fine-scale recombination map of Atlantic salmon was not provided so, analyses were performed using the default recombination parameters (1cM/ 1Mb). Both imputation software analyses were executed using parallel processing and 8 CPU units in a Linux cluster environment. Imputation performance of the tested software was assessed in terms of imputation accuracy and execution run-time (minutes to complete the analysis). The imputation results from the best performing imputation tool were used for comparative analyses between the old and new genome assemblies.

Genotype imputation using a large reference population and cross validation

An alternative imputation strategy to using individuals genotyped on separate SNP panels for the target and reference populations, is to use a large population genotyped on a single SNP panel and subset portions to construct the target population, irrespective of relationship between individuals. To evaluate this imputation approach we performed a 10-fold cross-validation (CV) analysis by iteratively looping through the following steps:

(i) random subsampling without replacement of 10% of a large reference pool of individuals genotyped with the HighD-SNP chip

(ii) down sampling of the SNP genotyping density for the subsampled individuals to construct the target population

(iii) use of the reference and target populations to perform genotype imputation

(iv) assessment of imputation accuracy by comparing the imputed with real genotypes for the subsampled individuals.

We performed two different CV analyses using the HighD-SNP panel with SNPs positioned in the new genome assembly. In the first analysis we down sampled SNP density according to the markers present in the LowD-SNP panel (62,106 markers), described in section “Animals and SNP genotyping data ”. For the second experiment we down sampled markers to match those included in a medium-density -relative to the HighD density- SNP panel (proprietary 220k Affymetrix Axiom array, 203,335 markers), thereby referred to as MediumD.

A total of 1,293 individuals with genotype information on the HighD-SNP panel were included in the reference pool for the two CV analyses. Remapping of SNPs for the HighD-SNP panel followed the mapping-coordinates update steps described in section “*Reposition of SNPs on the new genome assembly”*. However, to test SNP panel quality and integrity, we applied more stringent QC restrains to the SNP data of the HighD-SNP panel. Ensuring that SNPs positioned in the new assembly were of highest mapping quality, only markers identified by the Affymetrix platform as reliable on their genotyping integrity were retained for the CV analyses (Affymetrix power tools, 2018). Additional quality control (QC) of SNP genotype data through PLINK 2.0 (Chang et al., 2015), removed markers with more than 1% missing genotypes, minor allele frequency below 0.5% as well as individuals with more than 5% of their genotypes missing. In addition, variants deviating from the Hardy Weinberg equilibrium with a p-value < 10^{-6} were also discarded.

We generated ten independent subsamples for each CV analysis (i.e. 10-fold CV) of either 129 or 130 individuals by random sampling without replacement using the sample() function in R. For each of the ten CV iterations, the remainder of the individuals (1,163 or 1,164) were used as the reference population. For each individual in each subsample we then down sampled genotypes retaining those present in the MediumD- and LowD-SNP panels (depending on CV analysis), using the function *--extract* in PLINK v2.0. Variants of the LowD- and MediumD- panels that were not present in the HighD panel, were automatically excluded by PLINK v2.0. Genotype imputation was performed using Beagle v5.2 (B. L. Browning et al., 2018; S. R. Browning & Browning, 2007), with modified parameters and parallelized processing described in section “*Evaluation of SNP genotype imputation tools*”.

Assessment of imputation accuracy

Average imputation accuracy

To assess imputation accuracy of single SNPs in the parent-offspring design, we compared the imputed and true genotypes for each marker by estimating rate of concordance (CR) and the square of Pearson’s correlation coefficient (r^2). Concordance is a very useful measure of imputation accuracy, indicating the proportion of correctly imputed genotypes. However, described in introduction section “*Quantification of imputation accuracy*”, CR tends to overestimate imputation accuracy considering minor allele frequency (MAF). For that reason we additionally employed the squared correlation (r^2) between true and imputed genotypes, a highly robust and consequently reliable imputation accuracy metric.

In the case of the CV imputation experiments, we only used r^2 and calculated accuracy by averaging the r^2 estimate of each marker across all 10 CV replicates.

Local genomic accuracy assessment

To investigate differences in imputation performance between genomic regions we used a “rolling window” approach to compute mean imputation accuracy for SNPs in the old and new genome assemblies. Each window contained 100 SNP markers while consecutive windows overlapped by 50 SNP. In order to visualize the local genomic differences in imputation accuracy throughout chromosomes, the average r^2 of each rolling window was plotted against the physical position of the first SNP included in each constructed window.

Results

Data filtering

Immediate relatives imputation design

SNP quality (QC) filtering was performed on each SNP panel in the two genome assemblies, summarized in Table i. This resulted in 43,013 high quality SNPs on the LowD-SNP panel that were commonly found in both genome assemblies and were also included in the HighD-SNP panel. Filtering of the HighD-SNP panel resulted in 493,381 SNPs commonly found in the two genome assemblies.

From a total of 1,310 animals passing the QC filtering criteria, 216 were candidate offspring genotyped on both LowD- with HighD-SNP panels. Of these candidates, 195 offspring (F_2 generation) were able to be assigned to 90 parents (F_1 generation). A total of 159 trios (offspring with two known parents) and 36 duos (offspring with one known parent) were identified, originating from crosses (families) between 51 sires and 39 dams. On average, each dam was mother to 5 offspring and sires fathered 4 offspring, while family size varied between 1 and 10 individuals. The 21 individuals present in the target population dataset (LowD-SNP panel) with no identified parents were discarded from the immediate relatives imputation experiments.

Table i. Quality control (QC) filtering of SNP markers for the LowD- (70kv1) and HighD-SNP (xHD) panels in the old and new genome assembly, respectively.

Assembly version	*Genotyping Array*	*Initial SNP on array*	*Quality filtered SNP*	*Common high-quality SNPs between assemblies*
Old	LowD	62,583	54,355	43,013
New	LowD	63,624	62,571	
Old	HighD	723,656	532,270	493,381
New	HighD	661,227	499,132	

Cross-validation (CV) imputation design

QC filtering of the HighD-SNP panel resulted in a total of 1,293 animals with 409,019 high-quality SNPs assigned to the new genome assembly (Figure i). In total, the target populations contained genotype information from 183,582 and 44,201 SNP markers for the MediumD- and LowD-SNP panel density designs, respectively.

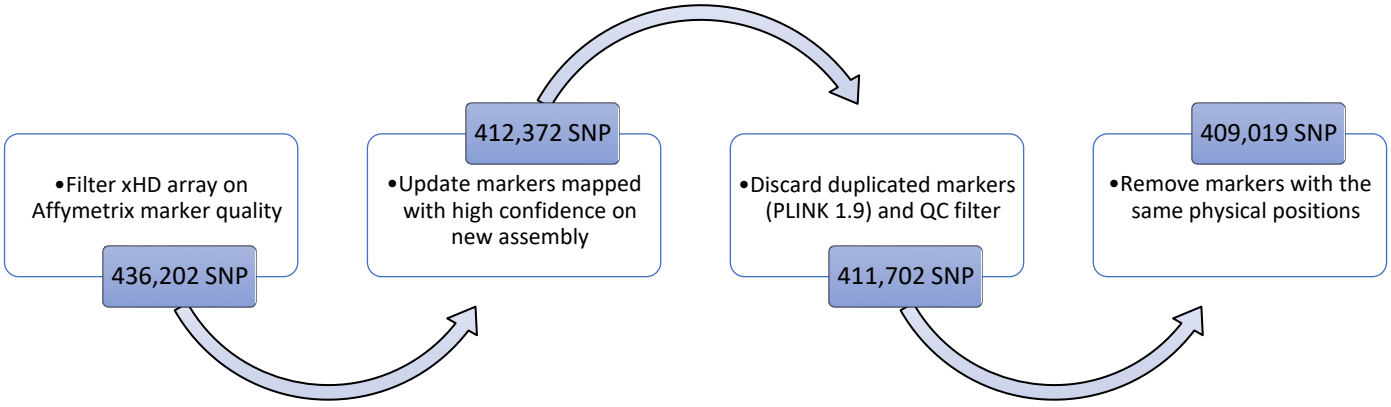

Figure i. HighD-SNP panel update and quality filtering for the cross-validation analyses

Comparing imputation software

Before comparing imputation accuracy between genome assemblies, we first tested two different imputation algorithms, namely FImpute v3 and Beagle v5.2 (B. L. Browning et al., 2018; Sargolzaei et al., 2014). For FImpute we provided pedigree information (see above section "*Immediate relatives imputation design*") and employed a population and family imputation approach. Beagle being a purely population-based algorithm, did not require pedigree information. Imputation performance results are presented below (Table 1). Although the rate of concordance (CR) did not show big discrepancies across imputation algorithms (approx. 99% of all missing markers were imputed correctly), average imputation accuracy (r^2) was significantly better (Wilcoxon signed rank test, p-value < 2.2e-16) for both assembly versions with Beagle (Table 1).

Table 1 Software comparison between imputation algorithms FImpute v3 and Beagle v2.5. Software performance is assessed in terms of imputation accuracy (CR and r^2) and computational cost (minutes to perform the analysis)

	FImpute v3 (pedigree)		*Beagle v5.2*	
	Old assembly	*New assembly*	*Old assembly*	*New assembly*
Concordance (CR)	0.986	0.988	0.993	0.991
Average accuracy (r^2)	0.830	0.834	0.842	0.851
Run time (mins)	3.58	4.02	8.07	7.83

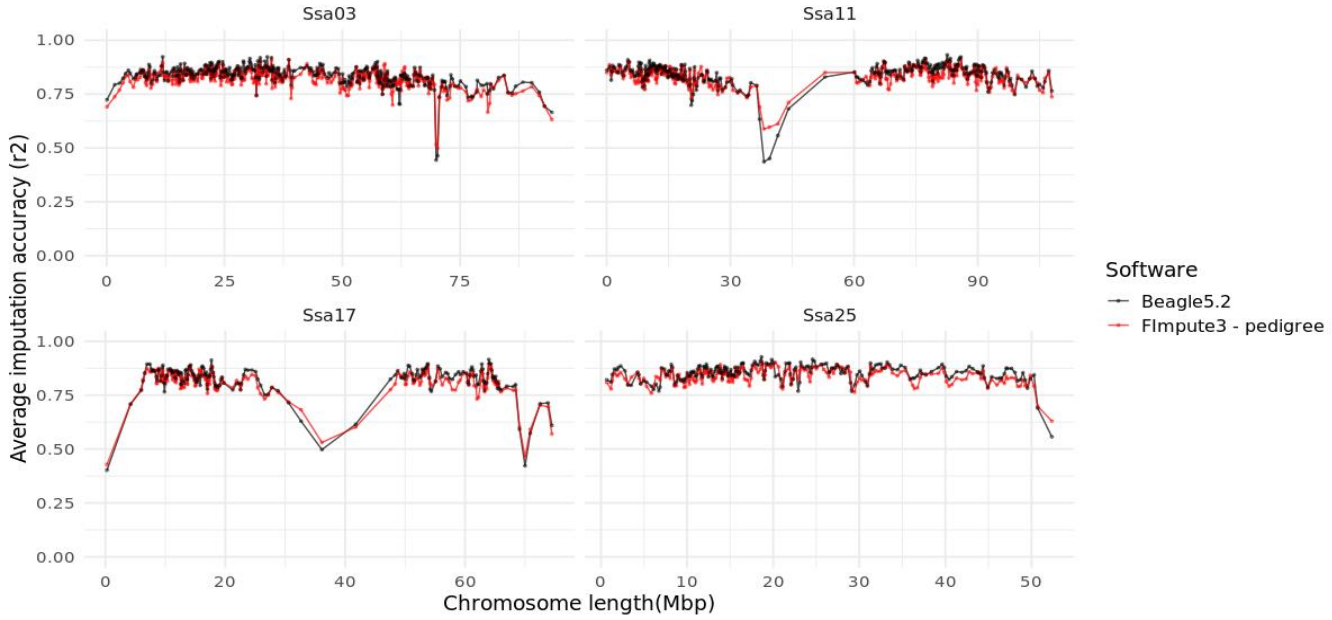

Figure 1 Local accuracy comparison between Beagle v5.2 (Black lines and points) and FImpute v3 (red lines and points) imputation software using the rolling window method in selected chromosomes. Genome-wide comparison is presented in figure S1 (supplementary material).

Investigating the local imputation accuracy across chromosomes (Figure 1), the two software mostly performed poorly in the same genomic regions. Local imputation accuracy comparisons for all chromosomes can be found in Figure s1 of supplementary material, however they show similar trends as the example chromosomes highlighted in Figure 1. In genomic regions of poor average imputation accuracy (r^2 < 0.50), Beagle performed worse than FImpute (e.g. Ssa11 and Ssa17 on Figure 1, Ssa25 on Figure

s1). Assessing computational efficiency of each software, Beagle was much slower than FImpute, requiring approximately twice the computational time (Table 1). Despite the slower processing time, Beagle only took ~8 minutes to perform the analysis and yielded overall better imputation accuracy. We therefore chose Beagle v5.2 to perform all imputation analyses shown in this study.

Comparing imputation accuracy between old and new genome assemblies

Genome wide assembly comparison

To compare the impact of an improved long-read reference genome assembly on imputation accuracy, we compared results of ten-fold genotype imputation (from ~43k to ~493k SNPs) using SNPs positioned on the new against the old genome assembly. Assessing concordance, 99% of total genotypes were correctly imputed in both genome assemblies (Table 2). Overall average imputation accuracy -measured as r^2- was relatively low for both assemblies; In comparison to the old however, average imputation accuracy using the new genome assembly was significantly higher (Table 2, Wilcoxon signed rank test, p-value < 2.2e-16). It is important to note that although the average r^2 estimates were below 0.90 for both assemblies, majority of markers were imputed with an r^2 value close to 1.0 (red and black dashed lines in Figure 2). Our results showed that using SNP positions placed on the new long-read genome assembly, resulted in small yet significant (p-value < 2.2e-16) improvement in imputation accuracy at the whole genome level.

Table 2 Average imputation accuracy results (estimated as concordance (CR) and squared correlation (r^2)) for a ten-fold genotype imputation analysis using the old and new genome assemblies

	Beagle v5.2	
Accuracy metric	*Old assembly*	*New assembly*
Concordance (CR)	0.993	0.991
Average accuracy (r^2) (± st.dev)	0.842 (0.24)	0.851 (0.22)

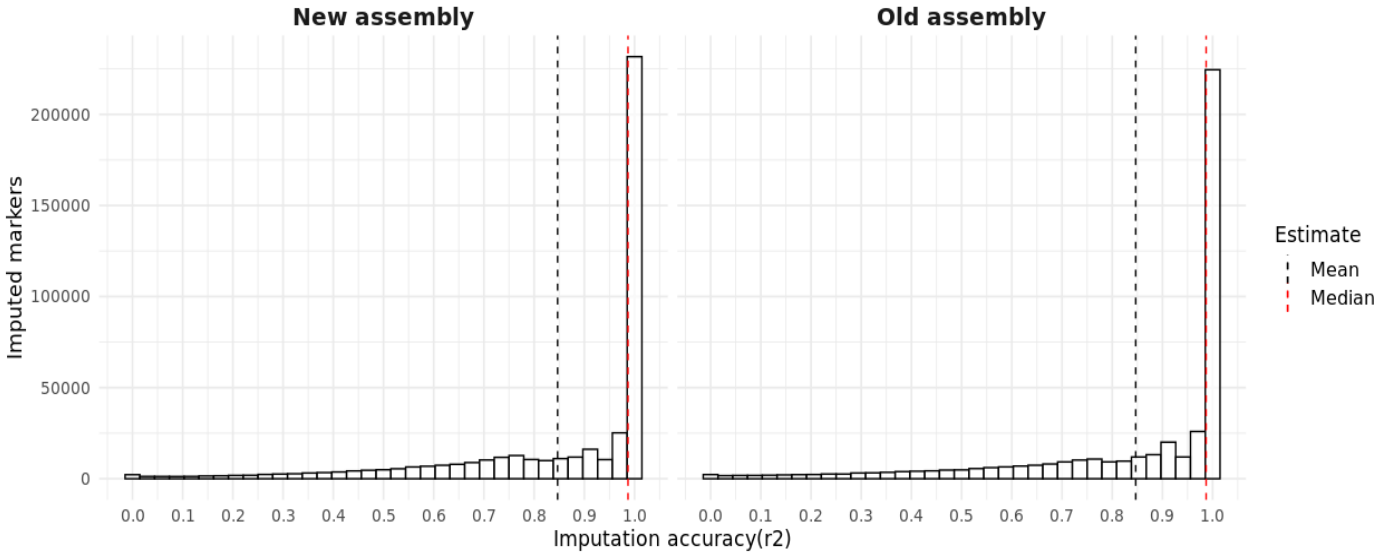

Figure 2 Histogram of r^2 values for imputed markers in the new (left) and old (right) genome assembly. The black and red dashed lines on each plot represent the mean and median accuracy (r^2) value, respectively

Local genomic differences in imputation accuracy between assemblies

Although the genome-wide average imputation accuracy for the two assemblies did not show dramatic differences, it is still possible for the two assemblies to drastically differ across certain genomic regions. To investigate this possibility, we applied the rolling window approach and estimated the average r^2 in windows of 100 SNP width.

Figure 3 shows the rolling window analysis for each of the two genome assemblies in selected chromosomes (see Figure s2 in *supplementary material* for all chromosomes). For majority of chromosomes, a general accuracy decrease was observed towards the ends of chromosomes (Figure s2), however there are also instances of drastic accuracy drop at the center of chromosomes, for example in SSa01, Ssa11 and Ssa18 (Figure s2).

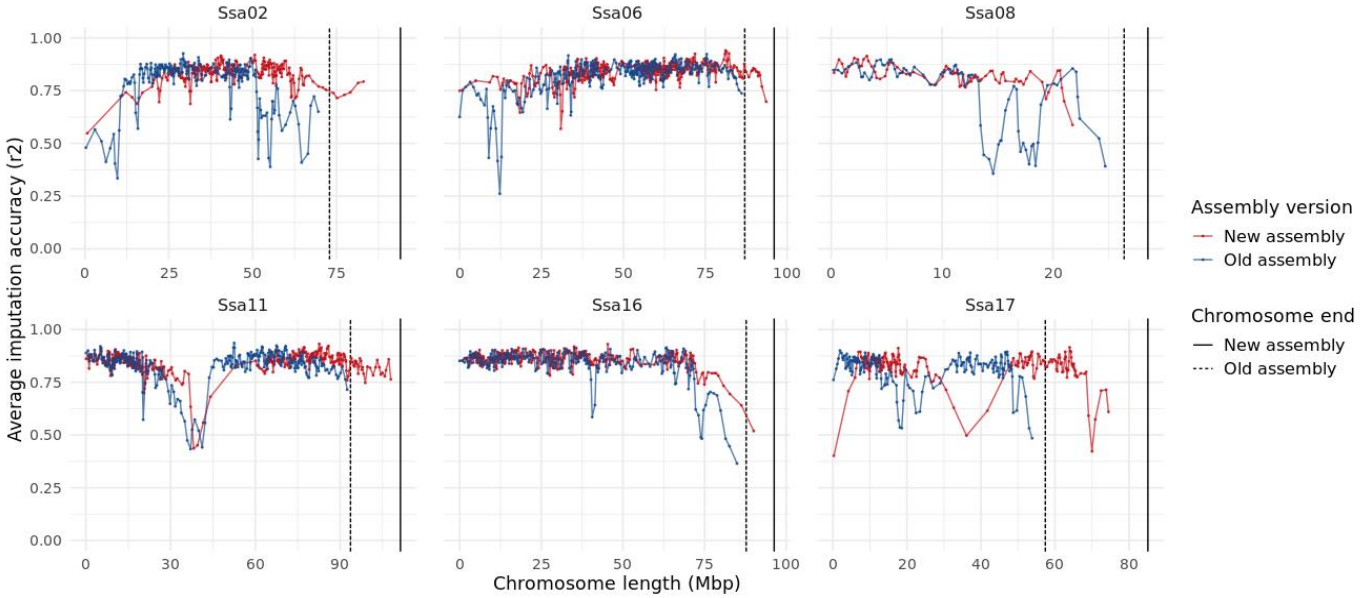

Figure 3 Local imputation accuracy assessment of the old (blue) and new (red) genome assemblies using the rolling windows method in selected chromosomes. Each rolling window represents the average imputation performance of 100 SNP markers in chromosomal order. The two vertical lines indicate the physical length of each chromosome for the old (dashed line) and new (solid line) genome assemblies. A genome wide comparison is provided in Figure s2 of supplementary material.

While most genomic regions performed similarly between the two genome assemblies, there are some striking accuracy differences observed between genomic regions. For example, in chromosomes Ssa06 and Ssa08 (Figures 3 above and s2 in *supplementary material*) several hotspots of poor imputation accuracy seem to be "resolved" by the new assembly version. In addition, smaller but substantial improvements can also be observed across several chromosomal segments (see Ssa02, Ssa11 and Ssa16 in Figure 3), aligning with the small improvement in average genome-wide imputation accuracy in the new assembly (Table 2). Only a few genomic regions show a large accuracy drop in the new assembly compared to the old version; for example on chromosome Ssa17 (Figure 3) as well as towards the ends of chromosomes Ssa15, Ssa18 and Ssa25 (Figure s2, *supplementary material*).

Impact of duplicate similarity on imputation accuracy

Based on the observed heterogeneity in local imputation accuracy between the two genome assemblies, we asked whether this could, at least partially, be attributed to the duplicated nature of the salmonid genome. Although most of the duplicated salmon chromosome segments started diverging 80-100 Mya, approximately 25% of the genome experienced delayed rediploidization (Lien et al., 2016; Robertson et al., 2017). The segments that experienced delayed rediploidization, referred to as Lineage specific Ohnolog Resolution regions (LORe-regions), share very high sequence similarity between duplicated loci and thus tend to be very difficult to assemble correctly. These LORe regions are represented by red bands in Figure 4. On the other hand, timely-diverging duplicated regions referred to as regions of Ancestral specific Ohnolog Resolution (AORe-regions), contain more distinct genomic sequence. AORe regions in the Atlantic salmon genome can be seen in Figure 4 as grey bands, respectively.

We therefore hypothesized that regions with high duplicate sequence similarity (LORe regions) would undergo a much larger assembly improvement in the new genome assembly compared to AORe regions, reflecting this improvement as increase in imputation accuracy. The results, presented in the heatmap tracks a and b in Figure 5, show a strong association between accuracy decrease and genomic regions of high sequence similarity. In line with our hypothesis, for markers residing within the LORe regions (highlighted red bands in the center plots of Figures 4 and 5), imputation accuracy is considerably lower. Interestingly, this difference was more pronounced for the old genome assembly (track b), compared to the new assembly version (track a). In addition, low imputation accuracy can also be observed in small segments outside LORe regions and towards the ends of chromosome Ssa15 as well as in two central regions in chromosomes Ssa01 and Ssa18 (Figure 5). For chromosome Ssa01, although accuracy in the old genome assembly shows a big local decrease in comparison to the rest the chromosome's performance, no evidence of high sequence similarity has previously been identified for this region (Lien et al., 2016). This region is represented by a small gap at approximately $100*10^6$ bp in chromosome Ssa01 of Figure 4.

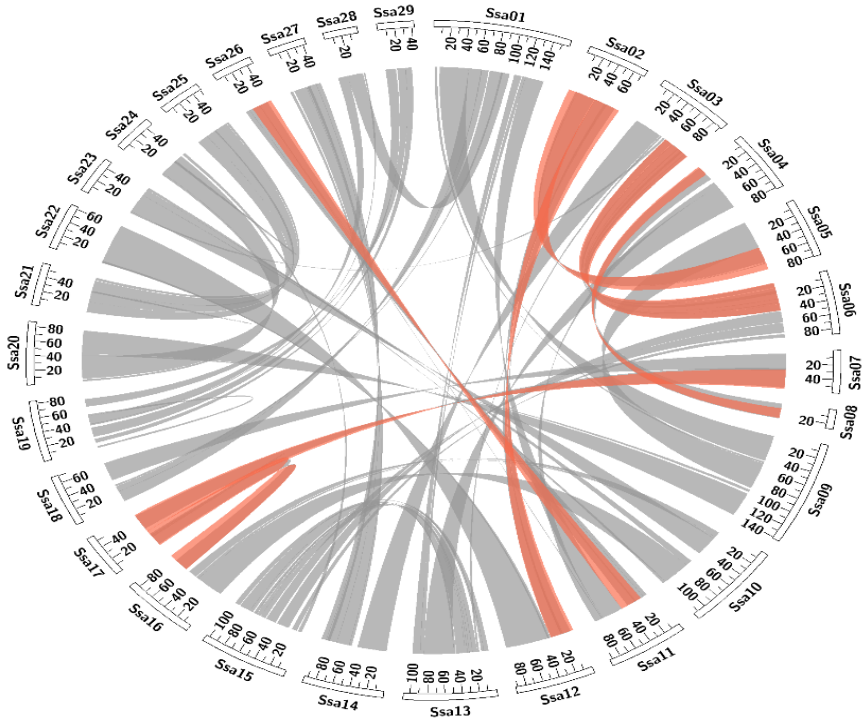

Figure 4 Circos plot (Krzywinski et al., 2009) of the salmonid genome showing the high duplicate sequence similarity regions (LORe) as red ribbons connecting duplicate regions. In grey, the early diverged regions (AORe) are shown. (Lien et al., 2016; Robertson et al., 2017)

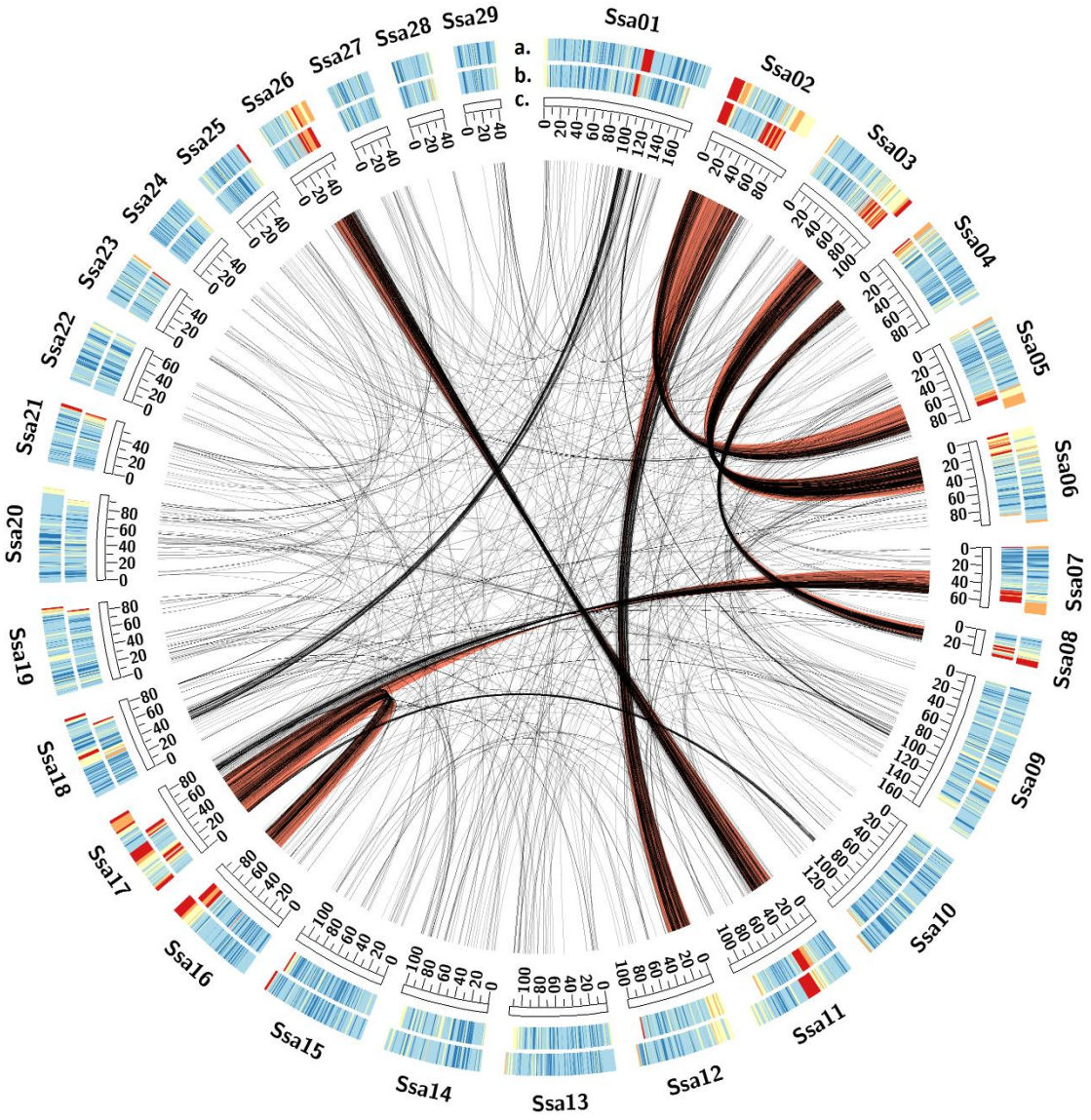

Figure 5 Circos plot (Krzywinski et al., 2009) showing the genome-wide imputation accuracy performance (rolling windows method) between the new (track a) and old (track b) genome assemblies. Heatmap tracks a and b range in colors from deep red to deep blue indicating the lowest and highest accuracy values (r^2), respectively. Track c shows the physical span of each chromosome. In the center plot, imputed markers positioned in a different chromosome according to the new assembly (inter-chromosome positioned) are presented as black lines connecting the exchanging chromosomes, while the previously identified LORe regions (Lien et al., 2016) are indicated as bright red bands.

Impact of marker position rearrangement on imputation accuracy

A main reason for the improved imputation accuracy in the new genome assembly is likely related to marker position rearrangements. Comparison of the imputed markers' physical position between genome assemblies, shows differences regarding the physical span of chromosomes in the new assembly version. Chromosomes show physical extension ranging from 35Kb up to 27Mb (average length= 8.4Mb). Such differences can be observed between the tracks a and b in Figure 5 as well as in Figures 3 above and s2 in supplementary material. Table s1 in supplementary material provides further details regarding the differences in chromosomal length between the two genome assemblies.

Although we expected virtually all SNPs to have a different physical position in the new genome assembly due to changes in genome length and intragenomic sequence reordering, 3,833 SNPs were assigned to a different chromosome in the new versus the old assembly. Seen in Figure 6, these inter-chromosomal exchanges had a drastic impact on imputation accuracy. The average imputation accuracy for inter-chromosome repositioned markers' performance increased from 0.269 in the old genome assembly to 0.674 in the new assembly version. On the contrary, markers that were repositioned within the same chromosome showed only slight accuracy improvement between assemblies (0.006 increase in average accuracy in the new assembly, shown in Figure 6). Aligning to our anticipations, majority of inter-chromosomal rearranged markers were exchanged between LORe regions (2.864 markers, ~75% of total inter-chromosome repositions). These markers are shown as black lines in the central plot of Figure 5, connecting the chromosomes between which they were exchanged. On the other hand, inter-chromosome repositioned markers moving to AORe regions were substantially fewer (979 markers, 25% of total inter-chromosome repositions). Interesting inter-chromosomal exchanges between non-LORe regions constitute the genomic segment exchanged between chromosome pairs Ssa01 and Ssa18, as well as one smaller but very impactful exchange between chromosomes Ssa10 and Ssa17, shown in the center plot of Figure 5. Table s1 in *supplementary material* presents the number of inter-chromosome repositioned markers that were either removed from or introduced to each chromosome due to the new genome assembly rearrangements.

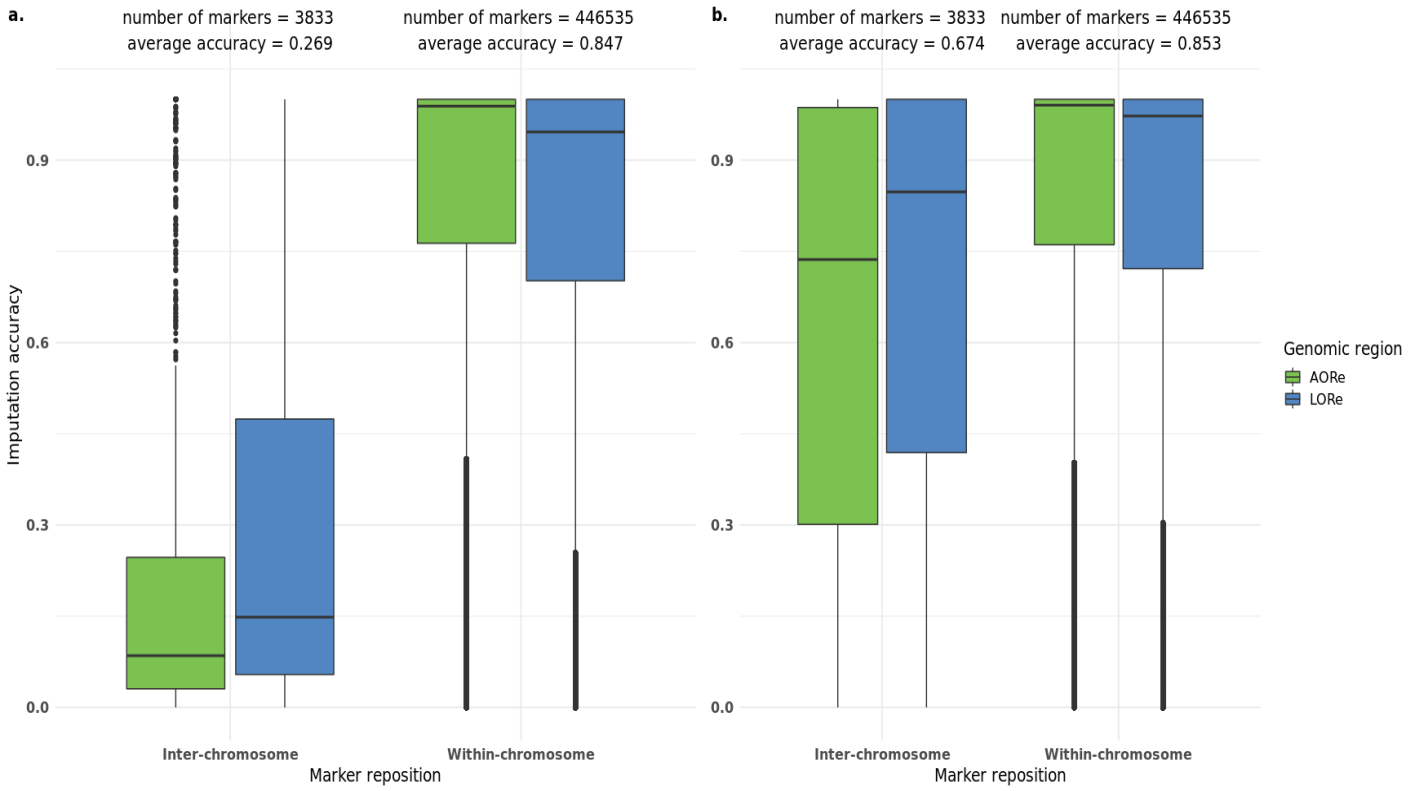

Figure 6 Accuracy performance of imputed markers depending on their reposition in the new genome assembly. X axis defines the direction of repositioned markers while the number and mean imputation accuracy of markers included in each reposition class is defined at the top. Subplots a and b show the imputation accuracy of the repositioned markers according to their old and new mapping coordinates, respectively. The Imputation accuracy performance of markers repositioned in AORe and LORe regions is shown in green and blue colors, respectively.

Further exploration of the LORe regions in the new genome assembly shows that less markers could be positioned in the high duplicate similarity regions compared to the old assembly version. In the old genome assembly 68,958 high quality markers resided within LORe regions, while in the new version these markers decreased to 63,114. Of total LORe region positioned markers, only 61,746 markers were commonly identified in both genome assemblies.

In addition, comparison of imputation performance between the old and new genome assemblies (heatmap tracks a and b in Figure 5) indicates that despite the overall improvement, the new genome assembly did not manage to completely resolve the low imputation accuracy across LORe regions. Regions of similarly low imputation accuracy between genome assemblies as well as small LORe segments of worse imputation performance in the new compared to the old genome assembly are seen in chromosomes Ssa02, Ssa17 and Ssa26 in Figure 5.

Structural variation and local imputation accuracy performance

Structural variation (SV) describes a broad range of DNA sequence rearrangements (Freeman et al., 2006) that vary in form, complexity and impact (Bertolotti et al., 2020; Charlier et al., 2012; Ho et al., 2020; Schutz et al., 2016). Genotype imputation of such elements can be particularly challenging as structural variants do not occur uniformly within a given population, potentially interfering with accurate estimation of haplotype frequencies and consequently imputation accuracy. We therefore hypothesized that the local drops in imputation accuracy that we observed could be associated with occurrence of segregating structural variation in the Atlantic salmon breeding nucleus. To test this hypothesis, we obtained a dataset of 274,196 structural variants identified by comparing long-read sequencing signals from 4 wild Norwegian salmon individuals against the new genome assembly (unpublished data shared by Kristina S. R. Stenløkk). A short description of this dataset content is provided in Table s2 and Figure s3 in supplementary material.

We associated the genomic position of 397 large SV elements (length > 30Kbp) with the local imputation accuracy of the new assembly, as this was estimated by the rolling windows method. In total, 217 SV elements residing within the physical intervals of at least one rolling window were examined (Table 3). Figure 7 shows the occurrence of structural variation (red points) in association to local imputation accuracy, in selected chromosomes (genome-wide results are provided in Figure s4 of *supplementary material*).

Overall, occurrence of structural variation was detected in regions of both low as well as high local imputation accuracy, e.g. regions in chromosomes Ssa01 and Ssa11 (Figure 7).However, majority of regions that show substantial decrease in local imputation accuracy overlap with structural variation signals, as seen on chromosomes Ssa01, Ssa02 and Ssa17 in Figure 7. These regions are mostly the same LORe regions that experienced inter-chromosomal repositioning of SNPs. What is more, regions with poor imputation accuracy overlapping with structural variation, appear to have much lower SNP density (longer genomic segments are required in order to capture the accuracy of 100 SNPs, shown in Figure 7).

Since low SNP density can impact phasing and consequently imputation performance, this indicates that structural variation cannot serve as a reliable diagnostic for assessment of imputation performance per se. Rather, structural variants seem to be associated with lower accuracy indirectly -potentially through further underlying genomic features linked to the similarity between duplicated regions.

Table 3 Length details for 217 identified structural variants with length above 30 Kbp that overlapped the physical intervals of at least one rolling window. Due to length restrictions, no insertions were retained for analysis. Details of all structural variants provided to this study are provided in Table s2 of supplementary material.

SV element	Number	Min. length (bp)	Max. length (bp)	Average Length (bp)
Duplications	49	30,706	1,849,972	30,706
Deletions	131	30,132	2,620,731	30,132
insertions	-	-	-	-
inversions	37	30,137	411,146	30,137

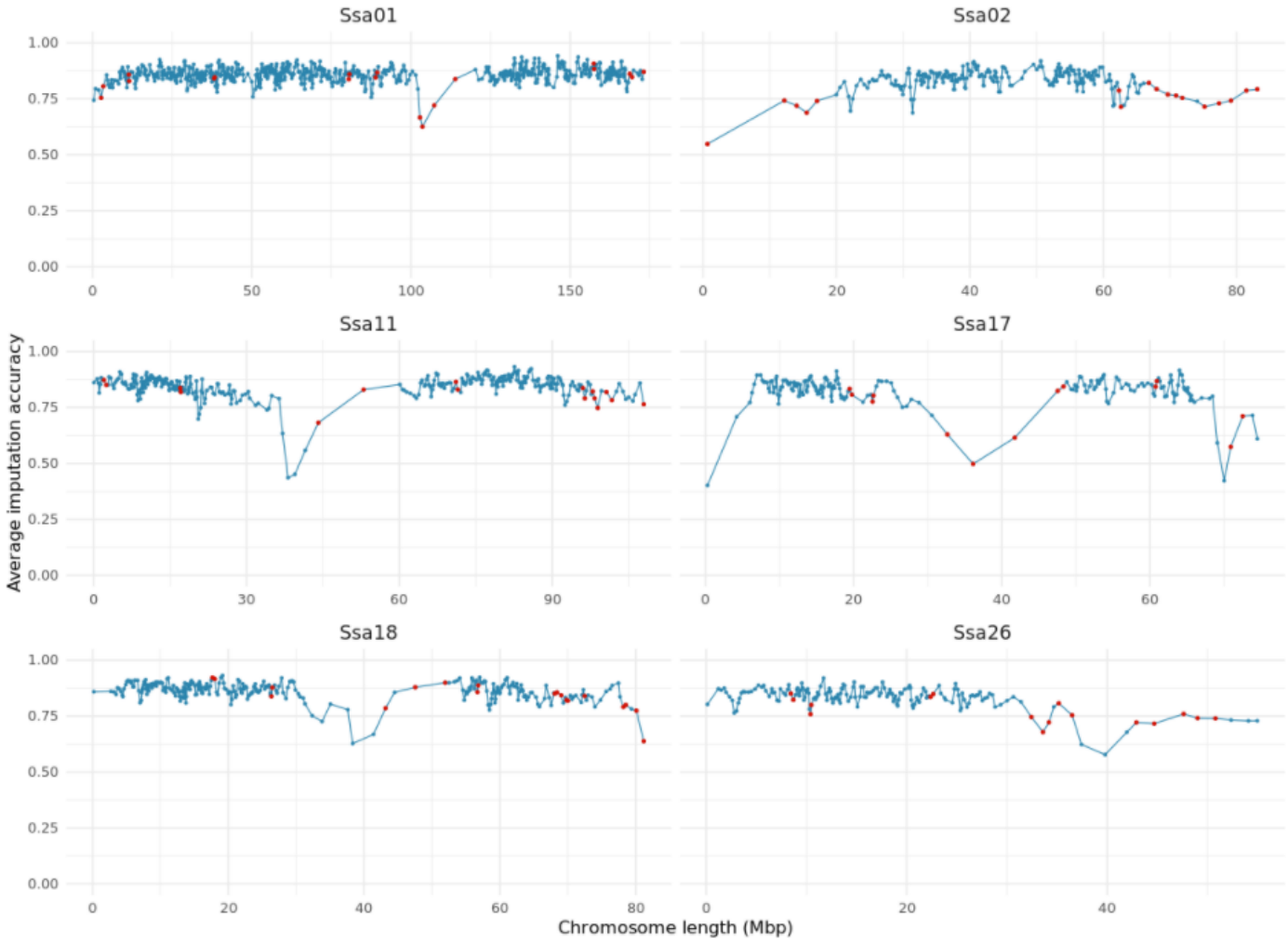

Figure 7 Association of structural variation presence with local imputation accuracy using the rolling window approach in selected chromosomes. Rolling windows are shown in blue, as points across the line. Red points indicate the presence of at least one SV element residing within a rolling window. A genome wide comparison is provided in Figure s4 in supplementary material

Genotype imputation accuracy using alternative experimental designs

So far, we have shown that an improved genome assembly can increase imputation accuracy. However, the mean accuracy value in our study (r^2=~0.85) is among the lower bound of imputation accuracy performance compared to other salmon related studies (Kijas et al., 2017; Tsairidou et al., 2020; Yoshida et al., 2018). We therefore asked whether our low imputation accuracy was the result of a sub-optimal imputation experiment design; using a small reference population (90 parents) while attempting to increase SNP density as much as 10-fold with relatively relaxed SNP quality filtering restrictions (see *methods* section "*Quality filtering*"). To test this hypothesis, we designed a new imputation accuracy experiment where we implemented more stringent SNP QC and tested both 10-fold (LowD to HighD) but also 2-fold (MediumD to HighD) genotype imputation. In addition, we used a larger reference population of randomly sampled individuals instead of a small sample of immediate relatives.

Shown in Table 4, using a large reference population clearly improved imputation accuracy compared to our immediate relatives design (~0.93 against ~0.85, respectively). Imputing from the LowD or MediumD

SNP panels did not have a high impact on genome-wide imputation accuracy (0.002 difference in average r^2 accuracy shown in Table 4) yet, there was large difference in local imputation performance; Imputation using a higher target SNP density (MediumD-to-HighD), performed markedly better in regions of poor local imputation accuracy (Figure 8). However, even when target population SNP density increased from ~44,000 (LowD) to ~184,000 markers (MediumD), regions of poor imputation accuracy still yielded considerably low accuracy results (e.g. chromosomes Ssa17 and Ssa26 in Figure 8 and chromosomes Ssa02 and Ssa03 in Figure s5 of *supplementary material*). This observation highlights that currently available genotyping panels face a certain weakness regarding SNP density and distribution in the new genome assembly.

Table 4 Imputation results for two cross-validation analyses, using a large population of randomly sampled individuals, stringent QC thresholds and two different target population SNP densities

Cross validation analysis design	*Target SNP density*	*Reference SNP density*	*Mean accuracy (r^2) (± st.dev)*
MediumD- to High-D SNP panel	183,582 SNP	409,019 SNP	0.928 (0.08)
LowD- to High-D SNP panel	44,201 SNP	409,019 SNP	0.926 (0.08)

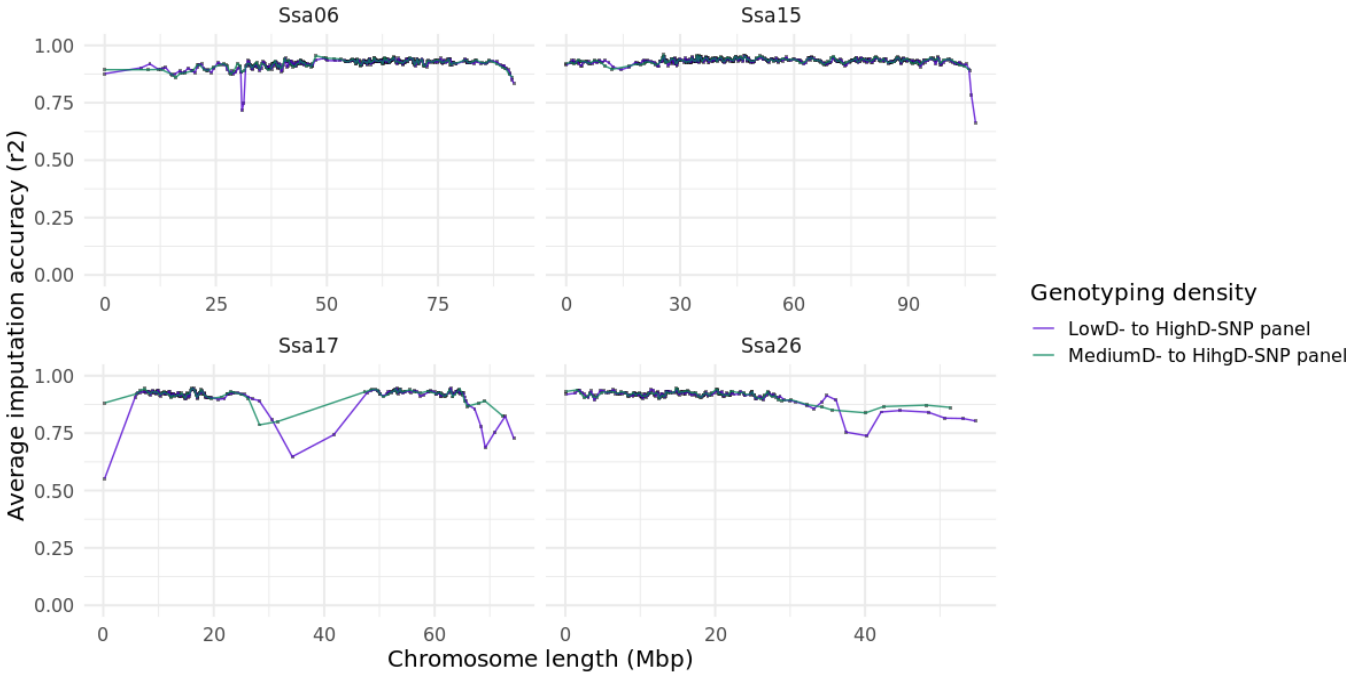

Figure 8 Comparison of local imputation accuracy for two cross-validation analyses using stringent QC thresholds and two target population SNP densities in selected chromosomes. Local imputation accuracy for the LowD to HighD (in purple) and MediumD to HighD (in green) analyses was assessed using the rolling windows method. A genome wide comparison is provided in Figure s5 of supplementary material

The CV analyses performing 10-fold and 2-fold genotype imputation yielded similar accuracy results (~0.93), indicating that target population SNP density is not the main reason behind the substantial accuracy discrepancy between the CV (r^2=~0.93) and immediate relatives (r^2=~0.85) experimental design. In the case of the CV analysis however, more stringent QC filtering was applied, described in methods section "*Genotype imputation using a large reference population and cross validation*". As genotyping errors and poor quality SNP markers can severely affect imputation accuracy, we wanted to quantify the effect of

similarly stringent QC filtering in the immediate relatives design. For that reason, we further restricted QC filtering -as performed in the CV analyses- in the immediate relatives design (parents-offspring) and performed 10-fold genotype imputation (LowD to HighD). We then compared average accuracy results to those obtained from the CV analysis when imputing from the LowD to HighD panel densities. The average overall and local imputation accuracy results between the two experimental designs are presented in Table 5 as well as Figures 9 (below) and s6 in supplementary material.

Increasing QC stringency improved average imputation accuracy for the immediate relatives design but also decreased the standard deviation of imputed values as shown in Table 5. Similarly, local genomic regions with lower imputation accuracy also showed notable improvement in imputation performance (Figure 9). These observations indicate the impact of genotyping errors and inclusion of low quality genotyping data on imputation accuracy. Still though, average imputation accuracy of the immediate relatives analyses was substantially lower than the performance of CV analysis designs (r^2=~0.89 versus r^2=~0.93, respectively).

Table 5. Comparison of imputation accuracy between two quality filtering restriction instances using the immediate relatives experimental design (parents-offspring).

Quality control filtering	*Target population density*	*Reference population Density*	*Average accuracy (r^2) (± st. dev.)*
Strict QC	44,201	409,019	0.886 (0.18)
Relaxed QC	43,013	493,381	0.851 (0.22)

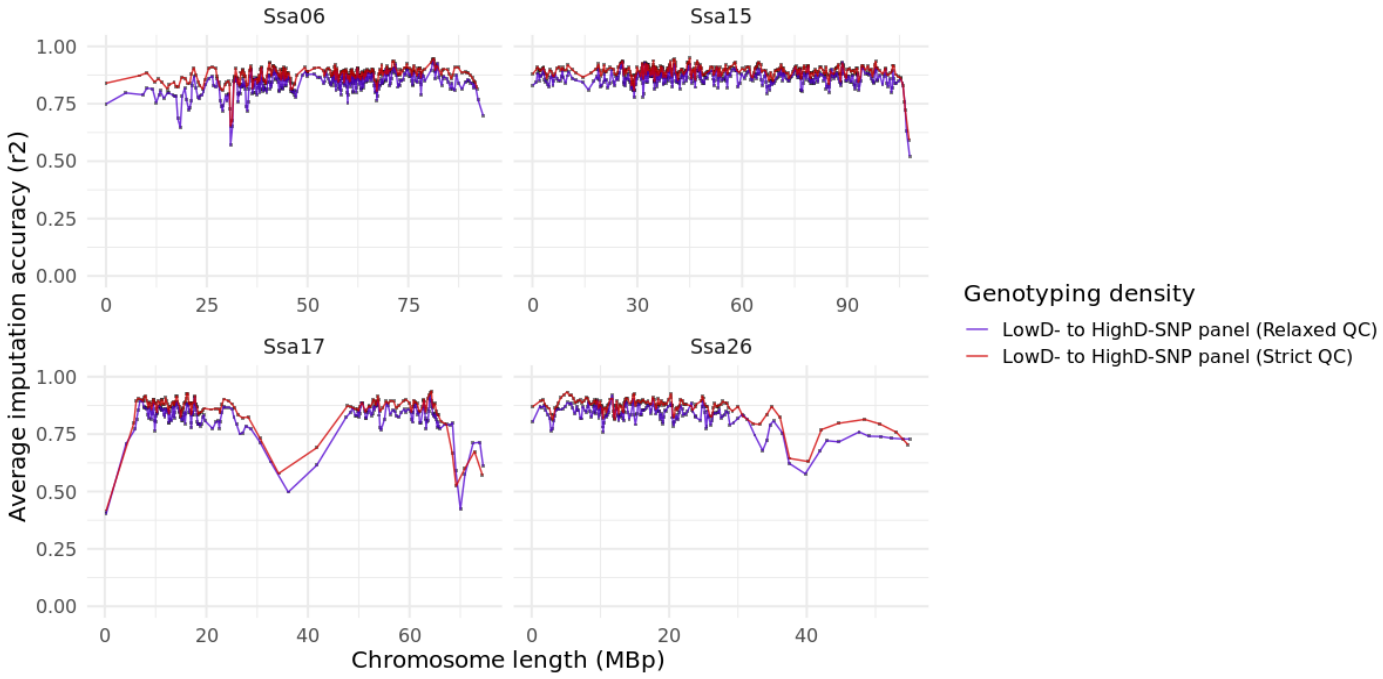

Figure 9 Comparison of imputation performance for the parents-offspring experimental design using stringent (red lines and points) and more relaxed (purple lines and points) QC thresholds, in selected chromosomes. Assessment of local imputation accuracy was performed with the rolling windows method. A genome wide comparison is provided in Figure s6 in supplementary material.

Discussion

Assembly comparison

New assembly improvements

In this thesis we clearly demonstrate that the new genome assembly can significantly improve imputation accuracy, particularly in genomic regions where the duplicate similarity is high (Figures 5 and 6 in *results* section). Much of this improvement is likely a consequence of the assembly contiguity that can be achieved through long read sequencing technologies. The new salmon reference genome is made up by 90 times fewer contigs (4,000) than the old assembly (368,060); this discrepancy is expected to largely affect the error rate of the scaffolding process when combining contigs to whole chromosome sequences (Sedlazeck et al., 2018). In line with this, we found that repositioning of SNPs between duplicated chromosome regions with high similarity, resulted in very large accuracy increase (Figure 6 in *results* section). These findings make apparent that the new genome assembly managed to accurately locate the misassembled complex genomic regions and correctly assign them to their genuine mapping position, thus enhancing resolution and precision.

Interestingly, improvement in the new genome assembly was not as drastic for markers rearranged within the same chromosome. Average imputation accuracy of such markers only slightly improved in the new genome assembly (r^2=0.853 in new versus r^2=0.847 in the old genome assembly). Majority of these markers (~95%) typically belonged to low similarity duplicate regions (AORe) (Lien et al., 2016; Robertson et al., 2017). The sequence content of such regions is more distinct, making AORe regions easier to assemble and order into chromosome sequences. We therefore conclude that the imputation improvements seen in the new assembly is mostly due to genomic sequences that have been misplaced within duplicate chromosome regions in the old assembly.

Genomic regions of poor imputation accuracy in the new assembly

While mean imputation performance increased in the new assembly, we also identified genomic regions that had lower imputation accuracy compared to the old genome assembly (see chromosomes Ssa11, Ssa17 and Ssa18 in Figure 5 of results section). Such regions coincide with high duplicate sequence similarity, also shown in Figure 5 of results section. There are three likely reasons as to why the new genome assembly performed worse than the old version in these genomic segments.

Firstly, the genomic rearrangement that occurred in the new assembly heavily influenced the regions of high duplicate similarity by moving erroneously positioned markers between as well as within chromosomes (Figure 6 in *results* section). The rearrangement impaired SNP-coverage for these regions, observed in Figure 7 of results section. Genotyping density plays a crucial role in imputation (Antolin et al., 2017; Kijas et al., 2017; Shi et al., 2018) and for that reason fluctuations in local SNP coverage could be the explanation behind the accuracy drops that we observe in regions of higher sequence similarity. The decrease in number of markers residing within genomic regions of higher complexity, indicates a certain weakness of currently available SNP genotyping panels to cater to certain features of the new genome assembly with high precision. Such issues, related to marker density and distribution, need to be addressed and accounted for in future design and production of genotyping panels.

Secondly, LORe regions contain many large structural variants known to segregate in Norwegian Atlantic salmon. One possible explanation therefore is that the reference population used in this study was segregating for alternative structural variant haplotypes with high allele frequencies, potentially interfering with imputation accuracy. Accuracy of genotype inference depends on the correct estimation of haplotype frequencies (S. R. Browning & Browning, 2011). Most commercial genotyping panels however, assess structural variation positionally rather than in the form of long genomic segments (Rafter et al., 2020). Therefore, occurrence of large variation elements could disturb systematically observed haplotypes within populations, ultimately impacting

imputation performance. Although structural variants overlapped regions of decreased imputation accuracy and high genomic similarity, their occurrence was not restricted to low-accuracy regions and thus they could not prove indicative of imputation performance in our analysis.

Finally, it is also possible that these few genomic regions that performed worse in the new assembly represent assembly errors that were not present in the old assembly. Such problems could relate to low sequence coverage, an issue faced by long-read sequencing, or possible setbacks of the medians used during the scaffolding process (Sedlazeck et al., 2018).

Under such considerations, assessment of true impact of the SV elements overlapping our study findings is considered inconclusive. Although earlier studies have explored the impact of structural variation in aquaculture species (Bertolotti et al., 2020; Liu et al., 2021), the divergence of structural variation between farmed and wild salmon and the consequent effect of augmented or exclusive presence of SV elements in farmed cohorts has not been thoroughly examined or fully understood yet. As a result there is a lack of databases listing SV elements specifically segregating within aquaculture strains such as the one used in this study. At the same time, the need for recognition of variation elements using genotyping arrays and construction of SV databases, as well as the interest in the potentials of variation-aware (graph) reference genomes (Crysnanto & Pausch, 2020) becomes more apparent. As linear genome references provide the foundations upon which more complex structures are built, the refinement and improvements brought by the new salmon genome assembly constitute a big step forward.

Imputation software comparison

The use of pedigree and imputation accuracy

Our first goal was to compare the performance of two routinely used imputation software in an experimental design where only immediate relatives were available (90 parents, 195 offspring). Our results show that Beagle yielded higher average imputation accuracy compared to FImpute (r^2 > 0.84 against r^2= ~0.83, respectively), even though FImpute was provided with a pedigree and Beagle does not exploit relationship information (see Table 1 in *results* section).

Earlier studies have shown that the use of pedigree information improves imputation accuracy in analyses assessing close relatives (Gualdron Duarte et al., 2013; Huang et al., 2012; Sargolzaei et al., 2014). Our findings, although using similarly high relationship levels, did not yield as high imputation accuracy as the mentioned studies. This could be attributed to our sample size and composition. Contrary to the mentioned studies, where further ancestry was included in the reference population, our study had available only a single generation of individuals (parents). Intuitively, the difference in impact of including pedigree information between our study and previously published findings could be explained by the limited information depth of the pedigree that we used. In addition, our sample size of individuals was considerably smaller (a total 285 fish), which could have also played an important role in our fairly low accuracy outcome, as imputation accuracy increases with population size. Similar observations to ours using small populations of immediate relatives only, have been made elsewhere (Tsai et al., 2017). Comparison of our findings with studies that use a wider range of relationships between individuals, highlights the importance of exploiting larger, multigenerational reference populations and accounting for their respective relationships in the pedigree.

Although inclusion of pedigree information can significantly increase imputation accuracy, pedigree information availability and most importantly reliability often constitutes a problem, especially in aquaculture (Hayes et al., 2012; Vandeputte & Haffray, 2014). Our imputation software comparison results indicate that given the sample size and relatedness restrains, population-based methods can prove equally if not more capable of imputing close relatives, allowing for pedigree information to be omitted. In our study, Beagle's higher overall imputation accuracy could be attributed to our target sample's composition as we included offspring with either one or both parents genotyped in the reference population. While family based methods and pedigree information can accurately define the haplotypes for closely related individuals, assessing

individuals between populations as unrelated, determined more accurately the haplotype phase for the offspring with ungenotyped parents. Arguably, FImpute provides the ability of inferring ungenotyped parents provided that a certain number of their offspring are included in sample. However, this option cannot be exploited unless the parent is known -though ungenotyped. Our findings thus highlight the advantage and potential in incorporating probabilistic approaches into family-based imputation methods when populations with high family structure are assessed (Antolin et al., 2017).

Conversely, in genomic regions where both software show substantial decrease in imputation accuracy (mean accuracy $r^2 < 0.50$), FImpute performed better than Beagle. Family-based imputation approaches assume the existence of relationships between populations and thus try to define long shared haplotype segments between related individuals. On the other hand, probabilistic approaches focus on identification of short haplotype segments that segregate between populations under the assumption of unrelatedness (Antolin et al., 2017). In our analysis, the low accuracy regions were also characterized by very low SNP marker density. It is therefore possible that due to the haplotype-assessment differences of the two approaches (described in *introduction* section "*Imputation approaches*"), combined with use of a partially complete pedigree, FImpute was able to assess long haplotype segments and better infer the genotypes in regions of sparse marker density and consequently decreased LD between SNPs.

Computational cost and efficiency

High accuracy is the ultimate goal of genotype imputation, however choosing the most appropriate imputation strategy for an analysis requires further parameters to be considered. Due to the population size and volume of data handled in aquaculture, computational efficiency is highly sought after. Earlier studies in livestock have associated higher computational burden with population-based imputation methods (Antolin et al., 2017; Piccoli et al., 2014; Sargolzaei et al., 2014). In line with this, in our study FImpute was significantly faster than Beagle, reducing the computational time required for the same analysis by 50% (4 against 8 minutes, respectively, see Table 1 in *results* section). However, described in methods section "*Evaluation of SNP genotype imputation tools for genome assembly comparisons*", we used the latest versions of both imputation algorithm, employing parallelized processing and utilizing several (8) CPU cores. In addition, the sample that we used consisted of very few individuals (285 fish in total). In that respect, our analyses required little processing time overall, allowing us to use Beagle in all genotype imputation analyses for this thesis.

Although quicker in computation, FImpute required software specific formatting of the input information (described in *methods* section "*Evaluation of SNP genotype imputation tools*"). This requirement made the imputation analysis time consuming, user-dependent and thus error-prone. On the contrary, Beagle accepts more conventional file formats (Variant Call Format, .vcf), allowing the use of output from routinely used genomic analysis tools such as PLINK (used in this study) to be directly implemented into the imputation analysis pipeline. In this manner Beagle, although computationally costly, proves less user-dependent and thus more efficient. Imputation software is constantly improving in power and efficiency, however one needs to take into account the strengths and weaknesses of different imputation methods relative to the analysis design; population size, composition and pedigree availability are some points of consideration.

Impact of reference population size, SNP density and quality filtering on imputation accuracy

As the last objective of this thesis, we assessed the impact of (i) population size, (ii) target genotyping density and (iii) quality filtering on imputation accuracy (Figures 1, 8 ,9 and ,Tables 1, 4 and 5 in *results* section).

We show that using a reference population of 1,293 fish yielded significantly higher imputation accuracy than the experimental design where 285 individuals with an almost complete pedigree were only included (r^2=~0.93 against r^2=~0.85 seen in tables 1 and 4 and Figures 1 and 8 in *results* section, respectively). Imputation accuracy in our analysis was evaluated using squared correlation (r^2), a metric that rewards correct inference of a minor allele compared to the common allele (Calus et al., 2014). In that respect, by increasing the population size, rare alleles become more frequently observed and thus more accurately defined (Tsai et al., 2017). In addition, by including more individuals, segregation of non-IBD shared haplotype segments can be better determined, further contributing to better imputation performance (Antolin et al., 2017). This is especially the case for population-based software, such as Beagle in our study, that do not account for relationships between populations and thus naturally benefit from larger population sizes (B. L. Browning et al., 2018). Lastly, although the reference and target groups of the CV analysis consisted of randomly sampled individuals, the cohort we used in our study originated from a single breeding nucleus. In breeding schemes random mating is restricted, meaning that our sample included, to some extent, multigenerational family structures. Discussed elsewhere (Gualdron Duarte et al., 2013; Tsai et al., 2017), the existence of multiple generations within the reference population can increase imputation accuracy by improving the reference haplotype estimation through shared allele segregation between related individuals.

Previous studies have shown that increasing SNP density improves imputation accuracy (Huang et al., 2012; Tsairidou et al., 2020) as higher genotyping densities can provide better SNP-coverage, reduce the number of missing genotypes and thus the possibility of erroneous haplotype inference (Bernardes et al., 2019; Browning & Browning, 2011). Interestingly, our cross-validation experiment showed negligible genome wide accuracy differences yet large local genomic improvement (e.g. chromosomes Ssa03, Ssa17 and Ssa26 in Figure 8) in imputation accuracy when the target population genotyping density increased from 44,201 to 183,582 SNP markers. In addition, human imputation studies have observed that after reaching a certain genotyping coverage, the impact of increasing genotyping density begins to saturate (Shi et al., 2018). Our results, presented in Figure 8 of results section as well as in Figure s5 of supplementary material, support the previous notion by indicating that cost-efficient genotyping strategies can be implemented through currently available SNP panels and yet yield considerably high imputation accuracy (r^2= ~0.93, Table 4 in *results* section). Most importantly, differences in local imputation accuracy between the target genotyping panels we used, were observed in regions of low genotyping coverage and high genomic complexity. These findings underline certain weaknesses of currently available arrays and also highlight the importance and necessity of uniform SNP coverage across the genome, with particular emphasis given towards genomic regions of higher genomic complexity.

Lastly, genotyping errors are an issue for imputation as erroneously called genotypes as well as mispositioned markers interfere with haplotype phasing and consequently impact imputation accuracy (S. R. Browning & Browning, 2011). This issue becomes more evident considering the genomic complexity of the Atlantic salmon genome since SNP marker annotation is additionally challenged by genomic regions of ancestral duplication (Houston et al., 2014). Further restriction of SNP marker and genotype quality in our analysis increased average imputation accuracy from 0.851 to 0.885 in the parents-offspring design (Table 5 and figure 9 in *results* section). These results show that quality restriction constitutes and efficient way to significantly improve (p-value < 2.2e-16) imputation accuracy. On the other hand, despite applying more stringent quality restrictions and increasing accuracy, using only immediate relatives for analysis still yielded lower imputation accuracy compared to the CV analysis. This comparison further supports the notion that better imputation performance in our analyses is tightly associated to differences in populations' size and composition. On the other hand, increasing quality restrains comes at the cost of discarding more, potentially valuable information. While genotypes with very low minor alleles frequencies and considerable deviation from the Hardy Weinberg equilibrium indicate putative genotyping errors, in certain cases they can be associated with causative genes of variation signals (Rafter et al., 2020). In this manner, although strict quality filtering can increase imputation accuracy, it is important to adjust the filtering restrains according to the analysis objectives and allow for certain amount of rare allele variants to be represented.

References

1. Antolin, R., Nettelblad, C., Gorjanc, G., Money, D., & Hickey, J. M. (2017). A hybrid method for the imputation of genomic data in livestock populations. *Genet Sel Evol, 49*(1), 30. doi:10.1186/s12711-017-0300-y
2. Bernardes, P. A., Nascimento, G. B. D., Savegnago, R. P., Buzanskas, M. E., Watanabe, R. N., de Almeida Regitano, L. C., Coutinho, L. L., Gondro, C., & Munari, D. P. (2019). Evaluation of imputation accuracy using the combination of two high-density panels in Nelore beef cattle. *Sci Rep, 9*(1), 17920. doi:10.1038/s41598-019-54382-w
3. Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., Robledo, D., Kent, M. P., Rosaeg, L. L., Holen, M. M., Mulugeta, T. D., Ashton, T. J., Hindar, K., Saegrov, H., Floro-Larsen, B., Erkinaro, J., Primmer, C. R., Bernatchez, L., Martin, S. A. M., Johnston, I. A., Sandve, S. R., Lien, S., & Macqueen, D. J. (2020). The structural variation landscape in 492 Atlantic salmon genomes. *Nat Commun, 11*(1), 5176. doi:10.1038/s41467-020-18972-x
4. Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., Lee, J., Lam, E. T., Liachko, I., Sullivan, S. T., Burton, J. N., Huson, H. J., Nystrom, J. C., Kelley, C. M., Hutchison, J. L., Zhou, Y., Sun, J., Crisa, A., Ponce de Leon, F. A., Schwartz, J. C., Hammond, J. A., Waldbieser, G. C., Schroeder, S. G., Liu, G. E., Dunham, M. J., Shendure, J., Sonstegard, T. S., Phillippy, A. M., Van Tassell, C. P., & Smith, T. P. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet, 49*(4), 643-650. doi:10.1038/ng.3802
5. Bolormaa, S., Chamberlain, A. J., Khansefid, M., Stothard, P., Swan, A. A., Mason, B., Prowse-Wilkins, C. P., Duijvesteijn, N., Moghaddar, N., van der Werf, J. H., Daetwyler, H. D., & MacLeod, I. M. (2019). Accuracy of imputation to whole-genome sequence in sheep. *Genet Sel Evol, 51*(1), 1. doi:10.1186/s12711-018-0443-5
6. Browning B. conform-gt.24May16.cee.jar. 2016. https://faculty.washington.edu/browning/conform-gt.html
7. Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet, 84*(2), 210-223. doi:10.1016/j.ajhg.2009.01.005
8. Browning, B. L., & Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet, 98*(1), 116-126. doi:10.1016/j.ajhg.2015.11.020
9. Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet, 103*(3), 338-348. doi:10.1016/j.ajhg.2018.07.015
10. Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet, 81*(5), 1084-1097. doi:10.1086/521987
11. Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nat Rev Genet, 12*(10), 703-714. doi:10.1038/nrg3054
12. Calus, M. P., Bouwman, A. C., Hickey, J. M., Veerkamp, R. F., & Mulder, H. A. (2014). Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. *Animal, 8*(11), 1743-1753. doi:10.1017/S1751731114001803
13. Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience, 4*, 7. doi:10.1186/s13742-015-0047-8
14. Charlier, C., Agerholm, J. S., Coppieters, W., Karlskov-Mortensen, P., Li, W., de Jong, G., Fasquelle, C., Karim, L., Cirera, S., Cambisano, N., Ahariz, N., Mullaart, E., Georges, M., & Fredholm, M. (2012). A deletion in the bovine FANCI gene compromises fertility by causing fetal death and brachyspina. *PLoS One, 7*(8), e43085. doi:10.1371/journal.pone.0043085
15. Charlier, C., Coppieters, W., Rollin, F., Desmecht, D., Agerholm, J. S., Cambisano, N., Carta, E., Dardano, S., Dive, M., Fasquelle, C., Frennet, J. C., Hanset, R., Hubin, X., Jorgensen, C., Karim, L., Kent, M., Harvey, K., Pearce, B. R., Simon, P., Tama, N., Nie, H., Vandeputte, S., Lien, S., Longeri, M., Fredholm, M., Harvey, R. J., & Georges, M. (2008). Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nat Genet, 40*(4), 449-454. doi:10.1038/ng.96
16. Crysnanto, D., & Pausch, H. (2020). Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. *Genome Biol, 21*(1), 184. doi:10.1186/s13059-020-02105-0
17. Fernandez, J., Toro, M. A., Sonesson, A. K., & Villanueva, B. (2014). Optimizing the creation of base populations for aquaculture breeding programs using phenotypic and genomic data and its consequences on genetic progress. *Front Genet, 5*, 414. doi:10.3389/fgene.2014.00414

18. Fisher T. Affymetrix power tools. 2018. https://www.thermofisher.com/no/en/home/life-science/microarray-analysis/microarray-analysis-partners-programs/affymetrix-developers-network/affymetrix-power-tools.html. Accessed 13 Mar 2020
19. Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., Carter, N. P., Scherer, S. W., & Lee, C. (2006). Copy number variation: new insights in genome diversity. *Genome Res, 16*(8), 949-961. doi:10.1101/gr.3677206
20. Garcia-Ruiz, A., Cole, J. B., VanRaden, P. M., Wiggans, G. R., Ruiz-Lopez, F. J., & Van Tassell, C. P. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection (vol 113, pg E3995, 2016). *Proceedings of the National Academy of Sciences of the United States of America, 113*(33), E4928-E4928. doi:10.1073/pnas.1611570113
21. Ghosh, M., Sharma, N., Singh, A. K., Gera, M., Pulicherla, K. K., & Jeong, D. K. (2018). Transformation of animal genomics by next-generation sequencing technologies: a decade of challenges and their impact on genetic architecture. *Crit Rev Biotechnol, 38*(8), 1157-1175. doi:10.1080/07388551.2018.1451819
22. Gjedrem, T., & Rye, M. (2018). Selection response in fish and shellfish: a review. *Reviews in Aquaculture, 10*(1), 168-179. doi:10.1111/raq.12154
23. Gonen, S., Baranski, M., Thorland, I., Norris, A., Grove, H., Arnesen, P., Bakke, H., Lien, S., Bishop, S. C., & Houston, R. D. (2015). Mapping and validation of a major QTL affecting resistance to pancreas disease (salmonid alphavirus) in Atlantic salmon (Salmo salar). *Heredity (Edinb), 115*(5), 405-414. doi:10.1038/hdy.2015.37
24. Gonen, S., Lowe, N. R., Cezard, T., Gharbi, K., Bishop, S. C., & Houston, R. D. (2014). Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing. *BMC Genomics, 15*, 166. doi:10.1186/1471-2164-15-166
25. Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet, 17*(6), 333-351. doi:10.1038/nrg.2016.49
26. Grashei, K. E., Odegard, J., & Meuwissen, T. H. E. (2018). Using genomic relationship likelihood for parentage assignment. *Genet Sel Evol, 50*(1), 26. doi:10.1186/s12711-018-0397-7
27. Gualdron Duarte, J. L., Bates, R. O., Ernst, C. W., Raney, N. E., Cantet, R. J., & Steibel, J. P. (2013). Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. *BMC Genet, 14*, 38. doi:10.1186/1471-2156-14-38
28. Hayes, B. J., Bowman, P. J., Daetwyler, H. D., Kijas, J. W., & van der Werf, J. H. (2012). Accuracy of genotype imputation in sheep breeds. *Anim Genet, 43*(1), 72-80. doi:10.1111/j.1365-2052.2011.02208.x
29. Hayes, B., & Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. *Genome, 53*(11), 876-883. doi:10.1139/G10-076
30. Heng Li. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2
31. Hickey, J. M., Kinghorn, B. P., Tier, B., Wilson, J. F., Dunstan, N., & van der Werf, J. H. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet Sel Evol, 43*, 12. doi:10.1186/1297-9686-43-12
32. Ho, S. S., Urban, A. E., & Mills, R. E. (2020). Structural variation in the sequencing era. *Nat Rev Genet, 21*(3), 171-189. doi:10.1038/s41576-019-0180-9
33. Houston, R. D. (2017). Future directions in breeding for disease resistance in aquaculture species. *Revista Brasileira de Zootecnia, 46*(6), 545-551. doi:10.1590/s1806-92902017000600010
34. Houston, R. D., & Macqueen, D. J. (2019). Atlantic salmon (Salmo salar L.) genetics in the 21st century: taking leaps forward in aquaculture and biological understanding. *Anim Genet, 50*(1), 3-14. doi:10.1111/age.12748
35. Houston, R. D., Taggart, J. B., Cezard, T., Bekaert, M., Lowe, N. R., Downing, A., Talbot, R., Bishop, S. C., Archibald, A. L., Bron, J. E., Penman, D. J., Davassi, A., Brew, F., Tinch, A. E., Gharbi, K., & Hamilton, A. (2014). Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). *BMC Genomics, 15*, 90. doi:10.1186/1471-2164-15-90
36. Huang, Y., Hickey, J. M., Cleveland, M. A., & Maltecca, C. (2012). Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. *Genet Sel Evol, 44*, 25. doi:10.1186/1297-9686-44-25
37. Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol, 17*(1), 239. doi:10.1186/s13059-016-1103-0
38. Kijas, J., Elliot, N., Kube, P., Evans, B., Botwright, N., King, H., Primmer, C. R., & Verbyla, K. (2017). Diversity and linkage disequilibrium in farmed Tasmanian Atlantic salmon. *Anim Genet, 48*(2), 237-241. doi:10.1111/age.12513
39. Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D. F., Stefansson, H., & Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet, 40*(9), 1068-1075. doi:10.1038/ng.216

40. Kraft, F., & Kurth, I. (2020). Long-read sequencing to understand genome biology and cell function. *Int J Biochem Cell Biol, 126*, 105799. doi:10.1016/j.biocel.2020.105799
41. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res, 19*(9), 1639-1645. doi:10.1101/gr.092759.109
42. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), 2987-2993. doi:10.1093/bioinformatics/btr509
43. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352
44. Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics, 165*(4), 2213-2233.
45. Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annu Rev Genomics Hum Genet, 10*, 387-406. doi:10.1146/annurev.genom.9.081307.164242
46. Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R. A., von Schalburg, K., Rondeau, E. B., Di Genova, A., Samy, J. K., Olav Vik, J., Vigeland, M. D., Caler, L., Grimholt, U., Jentoft, S., Vage, D. I., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D. R., Yorke, J. A., Nederbragt, A. J., Tooming-Klunderud, A., Jakobsen, K. S., Jiang, X., Fan, D., Hu, Y., Liberles, D. A., Vidal, R., Iturra, P., Jones, S. J., Jonassen, I., Maass, A., Omholt, S. W., & Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature, 533*(7602), 200-205. doi:10.1038/nature17164
47. Lin, P., Hartz, S. M., Zhang, Z., Saccone, S. F., Wang, J., Tischfield, J. A., Edenberg, H. J., Kramer, J. R., A, M. G., Bierut, L. J., Rice, J. P., & Coga Collaborators Cogend Collaborators, G. (2010). A new statistic to evaluate imputation reliability. *PLoS One, 5*(3), e9697. doi:10.1371/journal.pone.0009697
48. Lin, S., & Zhao, H. (2010). *Handbook on Analyzing Human Genetic Data*.
49. Liu, S., Gao, G., Layer, R. M., Thorgaard, G. H., Wiens, G. D., Leeds, T. D., Martin, K. E., & Palti, Y. (2021). Identification of High-Confidence Structural Variants in Domesticated Rainbow Trout Using Whole-Genome Sequencing. *Front Genet, 12*, 639355. doi:10.3389/fgene.2021.639355
50. Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Y, A. R., H, K. F., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & A, L. P. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet, 48*(11), 1443-1448. doi:10.1038/ng.3679
51. Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet, 11*(7), 499-511. doi:10.1038/nrg2796
52. Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics, 157*(4), 1819-1829.
53. Meuwissen, T., Hayes, B., & Goddard, M. (2016). Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers, 6*(1), 6-14. doi:10.2527/af.2016-0002
54. Moen, T., Baranski, M., Sonesson, A. K., & Kjoglum, S. (2009). Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar): population-level associations between markers and trait. *BMC Genomics, 10*, 368. doi:10.1186/1471-2164-10-368
55. Moen, T., Torgersen, J., Santi, N., Davidson, W. S., Baranski, M., Odegard, J., Kjoglum, S., Velle, B., Kent, M., Lubieniecki, K. P., Isdal, E., & Lien, S. (2015). Epithelial Cadherin Determines Resistance to Infectious Pancreatic Necrosis Virus in Atlantic Salmon. *Genetics, 200*(4), 1313-+. doi:10.1534/genetics.115.175406
56. Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *J Appl Genet, 52*(4), 413-435. doi:10.1007/s13353-011-0057-x
57. Piccoli, M. L., Braccini, J., Cardoso, F. F., Sargolzaei, M., Larmer, S. G., & Schenkel, F. S. (2014). Accuracy of genome-wide imputation in Braford and Hereford beef cattle. *BMC Genet, 15*, 157. doi:10.1186/s12863-014-0157-9
58. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi:10.1086/519795
59. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Comp uting Vienna, Austria.(https://www.R-project.org/)
60. Rafter, P., Gormley, I. C., Parnell, A. C., Kearney, J. F., & Berry, D. P. (2020). Concordance rate between copy number variants detected using either high- or medium-density single nucleotide polymorphism genotype panels and the potential of imputing copy number variants from flanking high density single nucleotide polymorphism haplotypes in cattle. *BMC Genomics, 21*(1), 205. doi:10.1186/s12864-020-6627-8
61. Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biol, 14*(7), 405. doi:10.1186/gb-2013-14-6-405

62. Robertson, F. M., Gundappa, M. K., Grammes, F., Hvidsten, T. R., Redmond, A. K., Lien, S., Martin, S. A. M., Holland, P. W. H., Sandve, S. R., & Macqueen, D. J. (2017). Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biol, 18*(1), 111. doi:10.1186/s13059-017-1241-z
63. Robledo, D., Palaiokostas, C., Bargelloni, L., Martinez, P., & Houston, R. (2018). Applications of genotyping by sequencing in aquaculture breeding and genetics. *Rev Aquac, 10*(3), 670-682. doi:10.1111/raq.12193
64. Rowan, T. N., Hoff, J. L., Crum, T. E., Taylor, J. F., Schnabel, R. D., & Decker, J. E. (2019). A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. *Genet Sel Evol, 51*(1), 77. doi:10.1186/s12711-019-0519-x
65. Sargolzaei, M., Chesnais, J. P., & Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics, 15*, 478. doi:10.1186/1471-2164-15-478
66. Schutz, E., Wehrhahn, C., Wanjek, M., Bortfeld, R., Wemheuer, W. E., Beck, J., & Brenig, B. (2016). The Holstein Friesian Lethal Haplotype 5 (HH5) Results from a Complete Deletion of TBF1M and Cholesterol Deficiency (CDH) from an ERV-(LTR) Insertion into the Coding Region of APOB. *PLoS One, 11*(4), e0154602. doi:10.1371/journal.pone.0154602
67. Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet, 19*(6), 329-346. doi:10.1038/s41576-018-0003-4
68. Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., Vollger, M. R., Monlong, J., Garrison, E., Eichler, E. E., Salama, S., Haussler, D., Green, R. E., Akeson, M., Phillippy, A., Miga, K. H., Carnevali, P., Jain, M., & Paten, B. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol, 38*(9), 1044-1053. doi:10.1038/s41587-020-0503-6
69. Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature, 550*(7676), 345-353. doi:10.1038/nature24286
70. Shi, S., Yuan, N., Yang, M., Du, Z., Wang, J., Sheng, X., Wu, J., & Xiao, J. (2018). Comprehensive Assessment of Genotype Imputation Performance. *Hum Hered, 83*(3), 107-116. doi:10.1159/000489758
71. Tsai, H. Y., Hamilton, A., Guy, D. R., Tinch, A. E., Bishop, S. C., & Houston, R. D. (2015). The genetic architecture of growth and fillet traits in farmed Atlantic salmon (Salmo salar). *BMC Genet, 16*, 51. doi:10.1186/s12863-015-0215-y
72. Tsai, H. Y., Hamilton, A., Tinch, A. E., Guy, D. R., Gharbi, K., Stear, M. J., Matika, O., Bishop, S. C., & Houston, R. D. (2015). Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. *BMC Genomics, 16*, 969. doi:10.1186/s12864-015-2117-9
73. Tsai, H. Y., Matika, O., Edwards, S. M., Antolin-Sanchez, R., Hamilton, A., Guy, D. R., Tinch, A. E., Gharbi, K., Stear, M. J., Taggart, J. B., Bron, J. E., Hickey, J. M., & Houston, R. D. (2017). Genotype Imputation To Improve the Cost-Efficiency of Genomic Selection in Farmed Atlantic Salmon. *G3 (Bethesda), 7*(4), 1377-1383. doi:10.1534/g3.117.040717
74. Tsairidou, S., Hamilton, A., Robledo, D., Bron, J. E., & Houston, R. D. (2020). Optimizing Low-Cost Genotyping and Imputation Strategies for Genomic Selection in Atlantic Salmon. *G3 (Bethesda), 10*(2), 581-590. doi:10.1534/g3.119.400800
75. van den Berg, S., Vandenplas, J., van Eeuwijk, F. A., Bouwman, A. C., Lopes, M. S., & Veerkamp, R. F. (2019). Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide association studies. *Genet Sel Evol, 51*(1), 2. doi:10.1186/s12711-019-0445-y
76. Vandeputte, M., & Haffray, P. (2014). Parentage assignment with genomic markers: a major advance for understanding and exploiting genetic variation of quantitative traits in farmed aquatic animals. *Front Genet, 5*, 432. doi:10.3389/fgene.2014.00432
77. Whalen, A., Gorjanc, G., Ros-Freixedes, R., & Hickey, J. M. (2018). Assessment of the performance of hidden Markov models for imputation in animal breeding. *Genet Sel Evol, 50*(1), 44. doi:10.1186/s12711-018-0416-8
78. Yanez, J. M., Naswa, S., Lopez, M. E., Bassini, L., Correa, K., Gilbey, J., Bernatchez, L., Norris, A., Neira, R., Lhorente, J. P., Schnable, P. S., Newman, S., Mileham, A., Deeb, N., Di Genova, A., & Maass, A. (2016). Genomewide single nucleotide polymorphism discovery in Atlantic salmon (Salmo salar): validation in wild and farmed American and European populations. *Mol Ecol Resour, 16*(4), 1002-1011. doi:10.1111/1755-0998.12503
79. Yoshida, G. M., & Yáñez, J. M. (2021). Increased accuracy of genomic predictions for growth under chronic thermal stress in rainbow trout by prioritizing variants from GWAS using imputed sequence data. *Evolutionary Applications*. doi:10.1111/eva.13240
80. Yoshida, G. M., Carvalheiro, R., Lhorente, J. P., Correa, K., Figueroa, R., Houston, R. D., & Yáñez, J. M. (2018). Accuracy of genotype imputation and genomic predictions in a two-generation farmed Atlantic salmon population using high-density and low-density SNP panels. *Aquaculture, 491*, 147-154. doi:10.1016/j.aquaculture.2018.03.004

81. You, X., Shan, X., & Shi, Q. (2020). Research advances in the genomics and applications for molecular breeding of aquaculture animals. *Aquaculture, 526*. doi:10.1016/j.aquaculture.2020.735357

Supplementary material

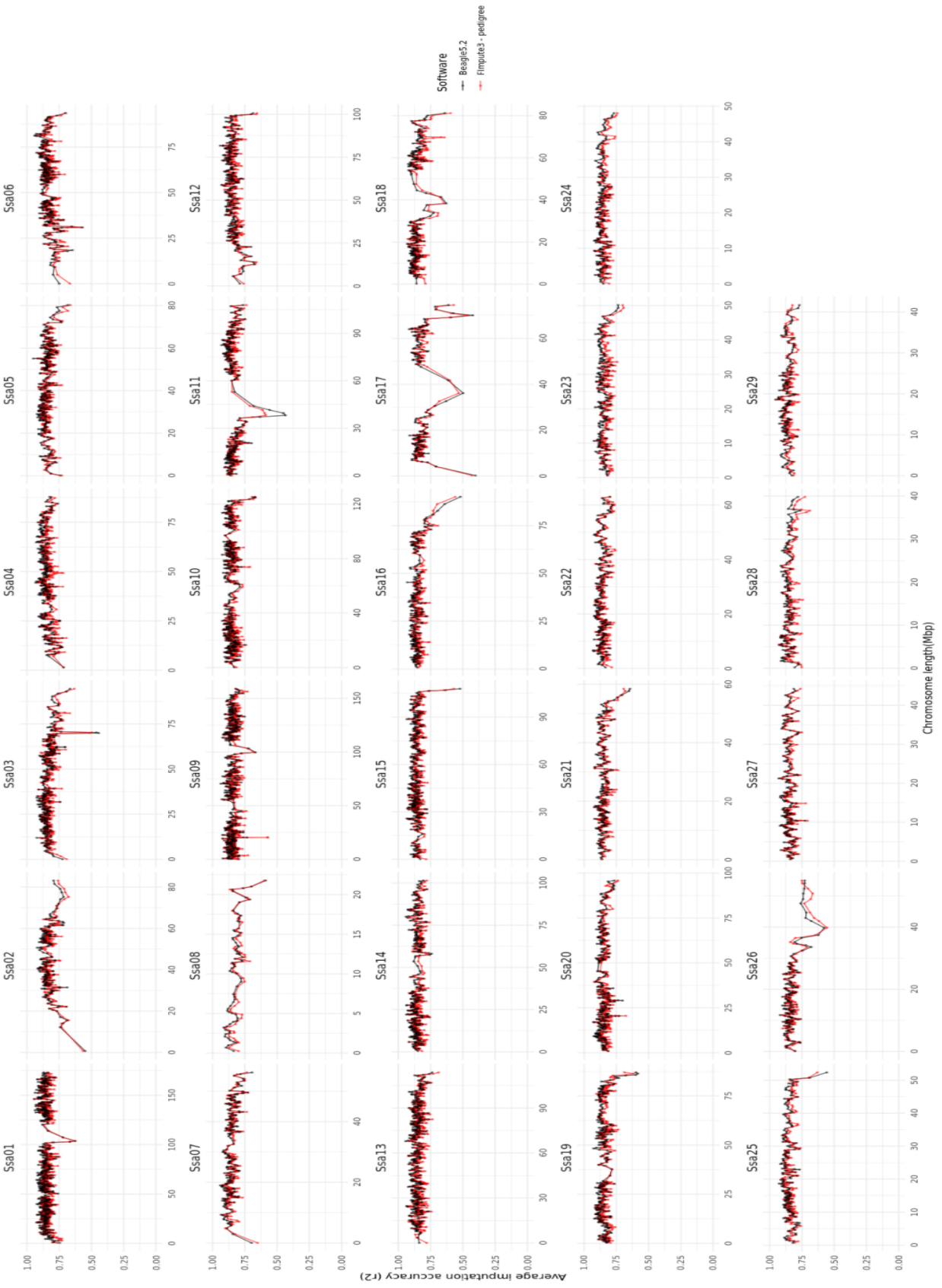

Figure s1 Local accuracy comparison between Beagle v5.2 (Black lines and points) and FImpute v3 (red lines and points) imputation software using the rolling window method.

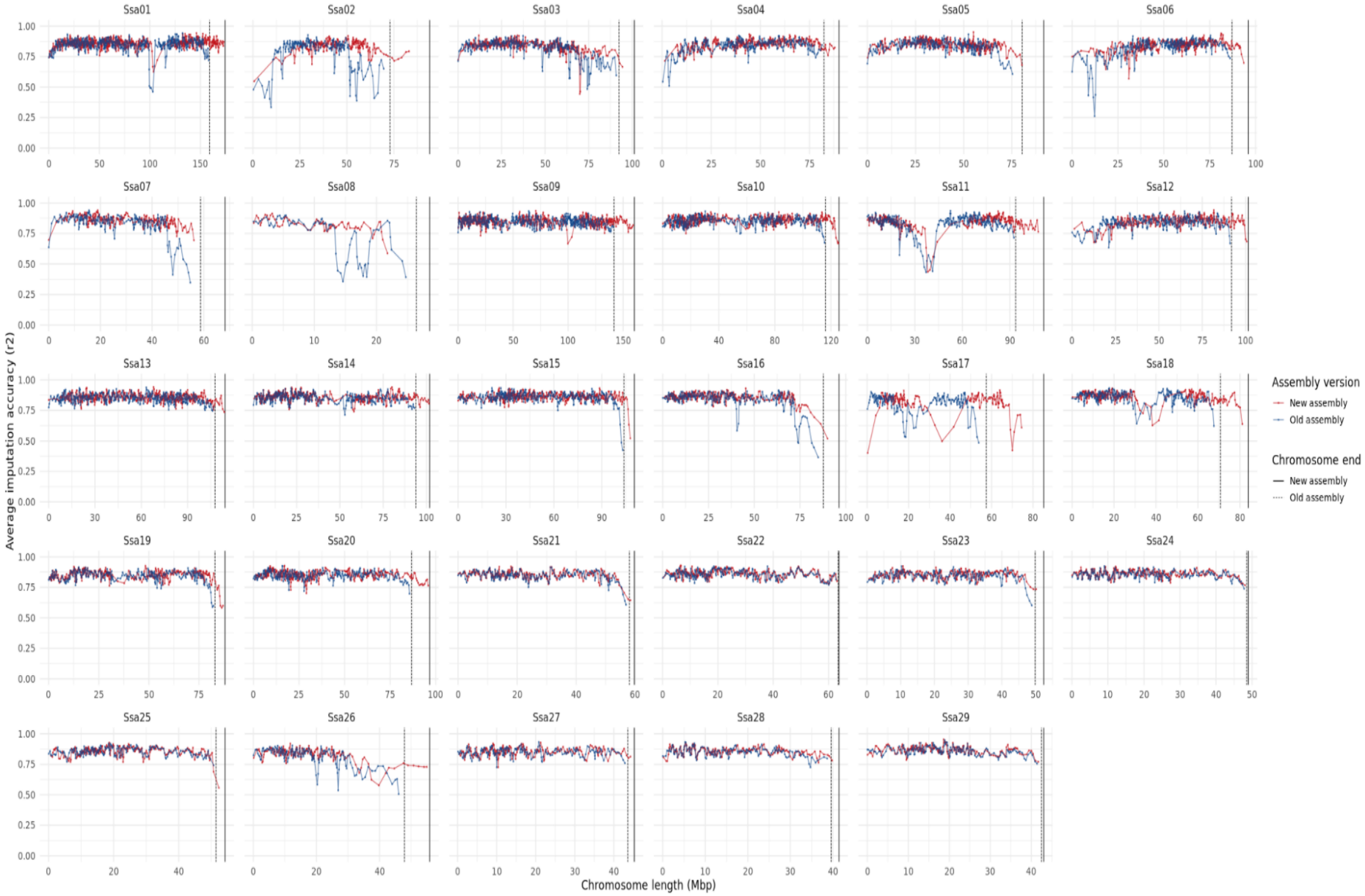

Figure s2 Local imputation accuracy assessment of the old (blue) and new (red) genome assemblies using the rolling windows method. Each rolling window represents the average imputation performance of 100 SNP markers in chromosomal order. The two vertical lines indicate the physical length of each chromosome for the old (dashed line) and new (solid line) genome assemblies.

Table s1 Inter-chromosome re-arrangement and elongation details per chromosome number between the two assembly versions

Chromosome number	Markers migrated	Markers introduced	Length in old assembly (Mbp)	Length in new assembly (Mbp)	Length difference (Mbp)
Ssa01	191	69	159.03	174.37	15.34
Ssa02	526	145	72.92	94.07	21.15
Ssa03	474	360	92.48	101.26	8.78
Ssa04	133	421	82.39	90.08	7.69
Ssa05	104	218	80.42	91.55	11.13
Ssa06	356	476	87.04	96.03	8.99
Ssa07	158	68	58.76	68.24	9.48
Ssa08	414	122	26.41	28.55	2.14
Ssa09	36	66	141.71	160.19	18.48
Ssa10	83	19	116.11	125.77	9.66
Ssa11	251	279	93.69	111.37	17.68
Ssa12	91	326	91.88	101.58	9.7
Ssa13	20	30	107.74	114.22	6.48
Ssa14	66	25	93.87	101.85	7.98
Ssa15	31	13	103.89	110.43	6.54
Ssa16	226	168	87.75	96.31	8.56
Ssa17	221	481	57.4	85.13	27.73
Ssa18	51	132	70.69	84.03	13.34
Ssa19	24	27	82.97	87.99	5.02
Ssa20	27	36	86.78	96.71	9.93
Ssa21	14	13	58.01	59.7	1.69
Ssa22	11	4	63.42	63.77	0.35
Ssa23	19	11	49.85	52.39	2.54
Ssa24	9	11	48.51	48.99	0.48
Ssa25	11	29	51.48	54.26	2.78
Ssa26	257	254	47.88	55.88	8
Ssa27	14	15	43.57	45.24	1.67
Ssa28	3	1	39.59	41.43	1.84
Ssa29	12	14	42.47	43.01	0.54

Table s2 Physical position details for 274,196 structural variation elements, identified in 4 wild Norwegian salmon individuals, categorized by variation type

Element	*Number*	*Min. length (bp)*	*Max. length (bp)*	*Average Length (bp)*
Duplications (DUP)	2,242	51	4,859,508	19,248
Deletions (DEL)	146,355	51	4,842,359	844
Insertions (INS)	124,285	50	16,804	360
Inversions (INV)	1,314	51	4,396,921	24,421

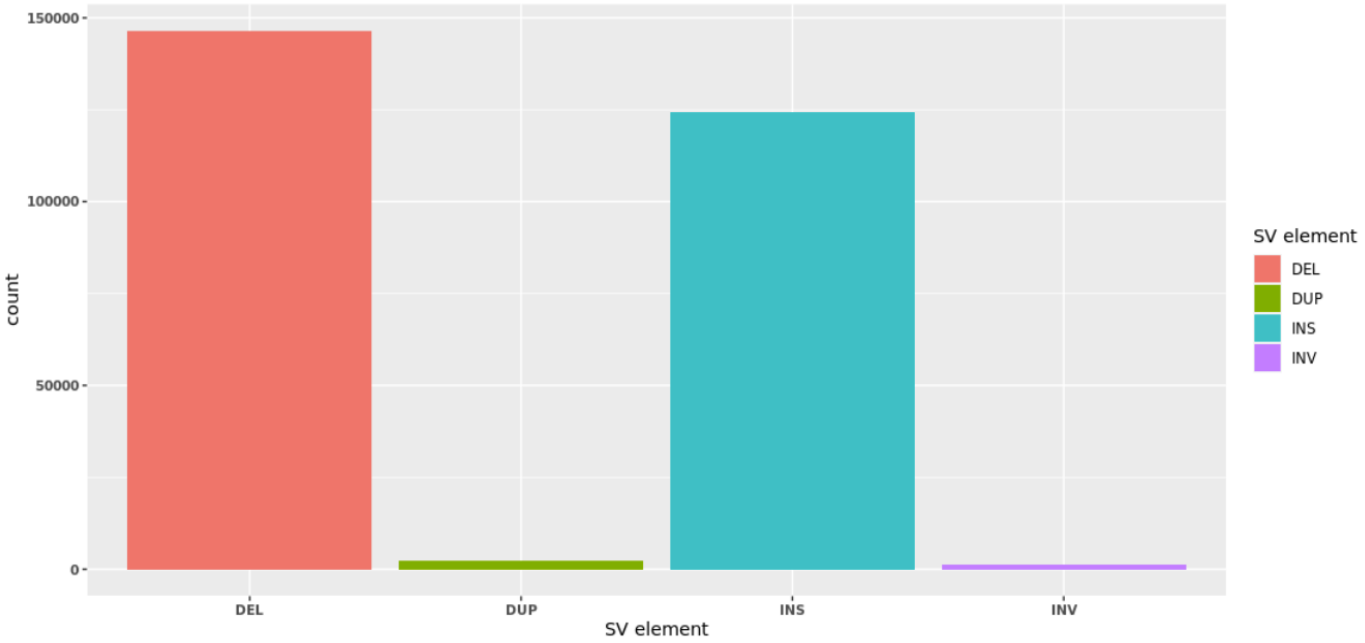

Figure s3 Composition of a dataset including for 274,196 structural variation elements, identified in 4 wild Norwegian salmon individuals, categorized by variation type.

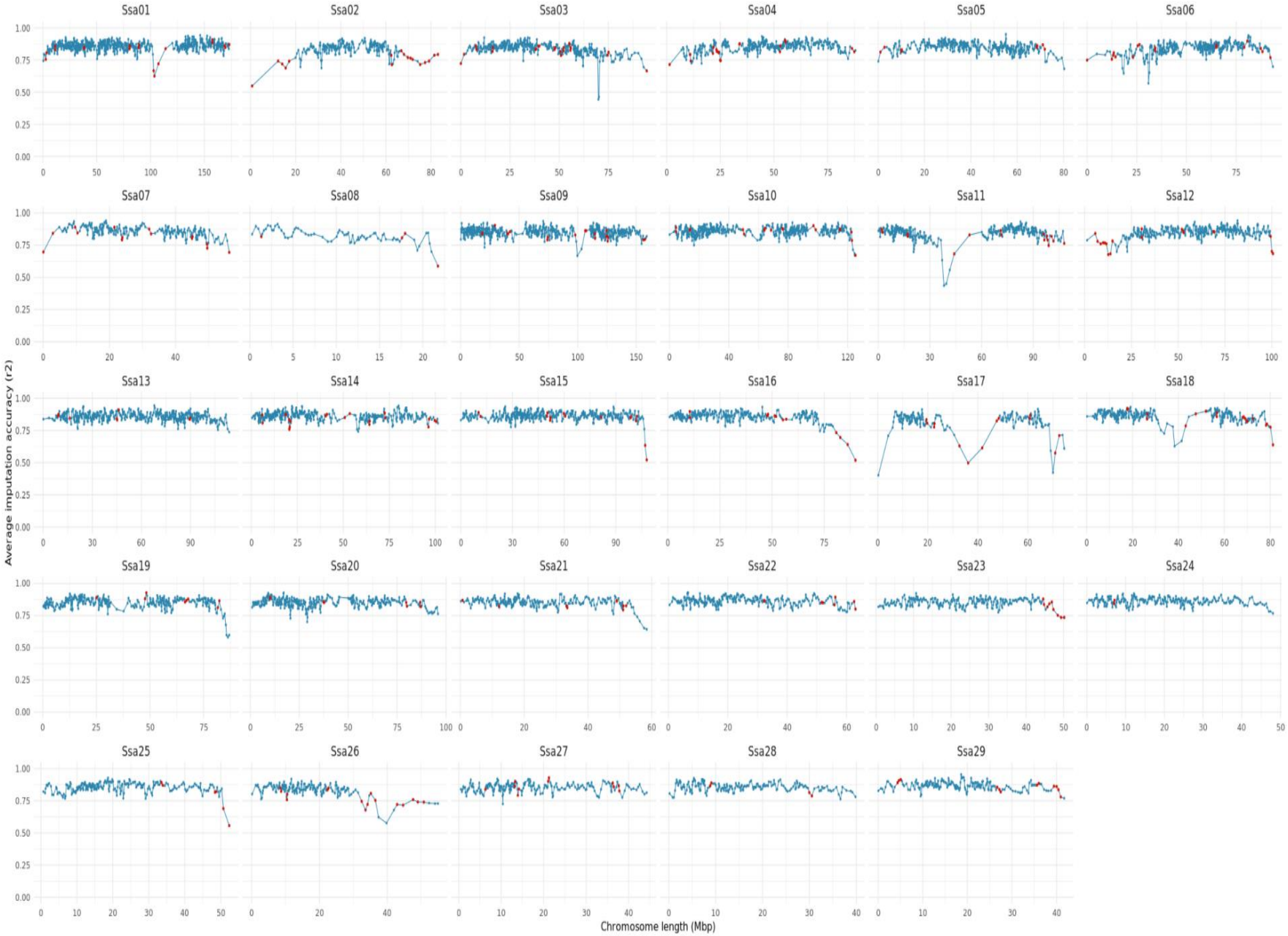

Figure s4 Association of structural variation presence with local imputation accuracy using the rolling window approach. Rolling windows are shown in blue, as points across the line. Red points indicate the presence of at least one SV element residing within a rolling window.

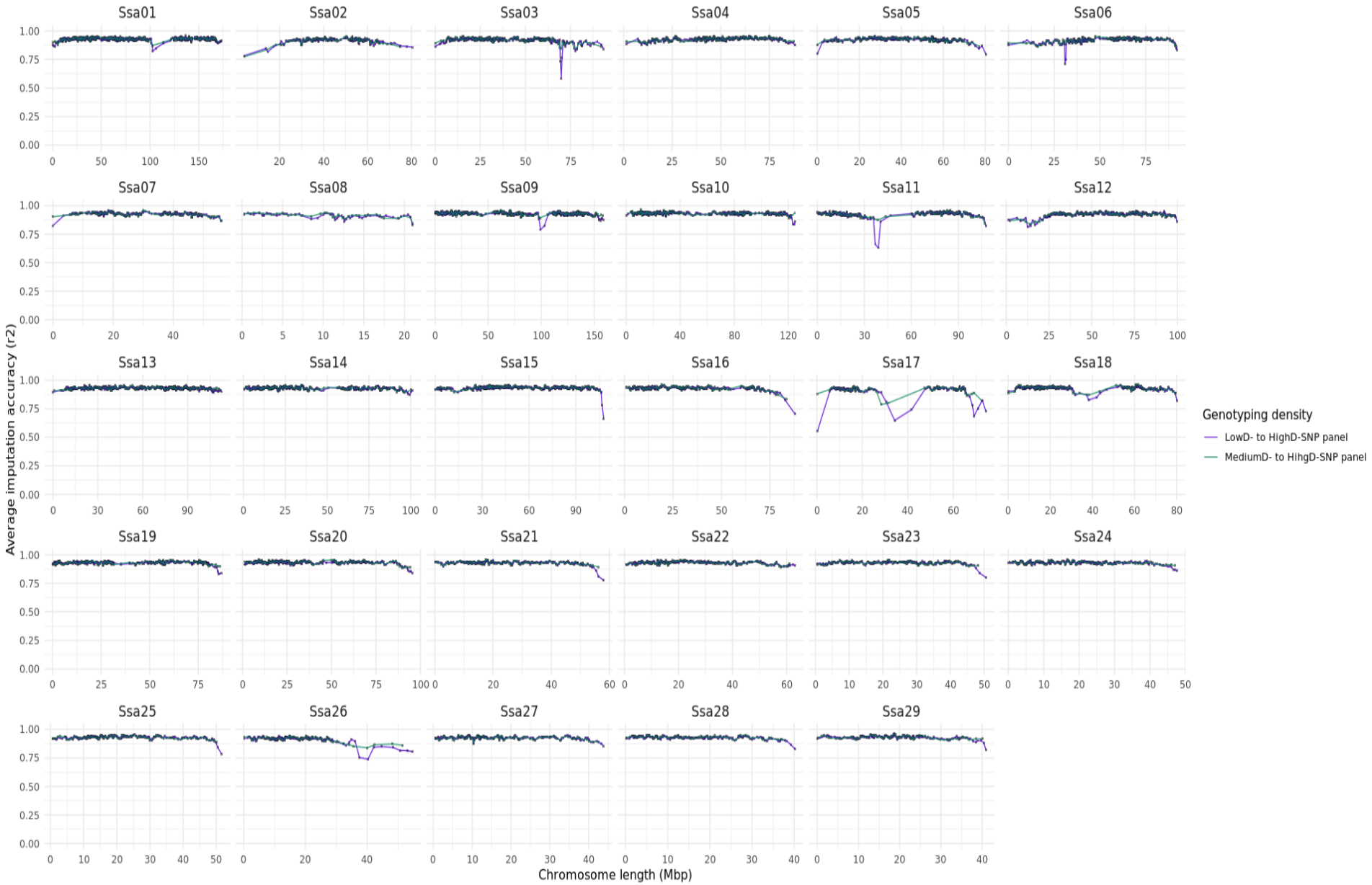

Figure s5 Comparison of local imputation accuracy for two cross-validation analyses using stringent QC and two target population SNP densities. Local imputation accuracy for the LowD to HighD (in purple) and MediumD to HighD (in green) analyses was assessed using the rolling windows method.

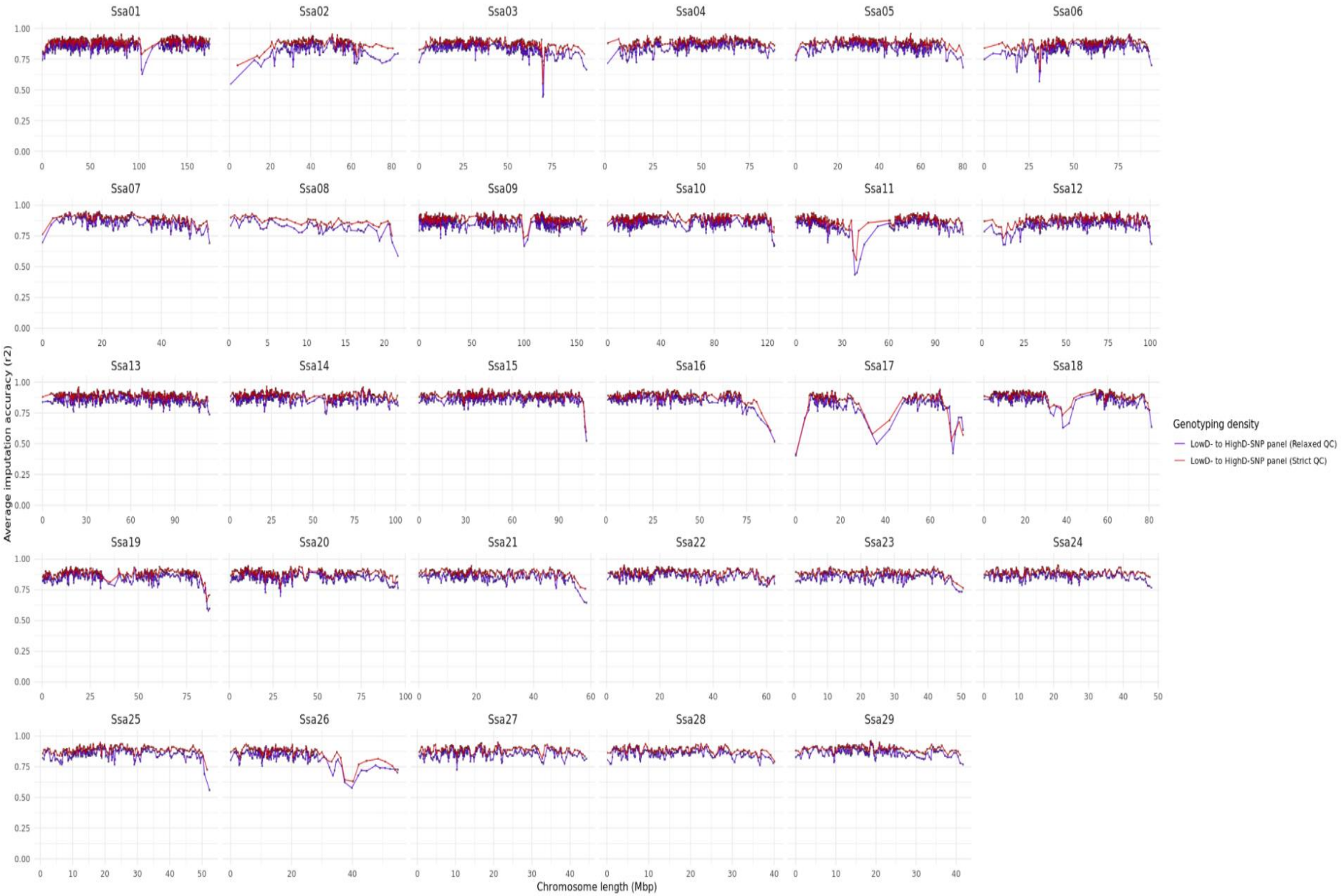

Figure s5 Comparison of imputation performance for the parents-offspring experimental design using stringent (red lines and points) and more relaxed (purple lines and points) QC thresholds. Assessment of local imputation accuracy was performed with the rolling windows method.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway