Edinburgh Research Explorer

Accuracy of genotype imputation and genomic predictions in a two-generation farmed Atlantic salmon population using highdensity and low-density SNP panels

OPEN ACCESS

Accepted Manuscript

Accuracy of genotype imputation and genomic predictions in a two-generation farmed Atlantic salmon population using high-density and low-density SNP panels

Grazyella M. Yoshida, Roberto Carvalheiro, Jean P. Lhorente, Katharina Correa, René Figueroa, Ross D. Houston, José M. Yáñez

Please cite this article as: Grazyella M. Yoshida, Roberto Carvalheiro, Jean P. Lhorente, Katharina Correa, René Figueroa, Ross D. Houston, José M. Yáñez , Accuracy of genotype imputation and genomic predictions in a two-generation farmed Atlantic salmon population using high-density and low-density SNP panels. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Aqua(2018), doi:10.1016/j.aquaculture.2018.03.004

Accuracy of genotype imputation and genomic predictions in a two-generation farmed Atlantic salmon population using high-density and low-density SNP panels

Grazyella M. Yoshida[a,b], Roberto Carvalheiro[b], Jean P. Lhorente[c], Katharina Correa[c], René Figueroa[c], Ross D. Houston[d], José M. Yáñez[a,b,e,*]

[a]Facultad de Ciencias Veterinarias y Pecuarias, Universidad de Chile, Av. Santa Rosa 11735, La Pintana, Santiago 8820808, Chile.

[b]School of Agricultural and Veterinarian Sciences, São Paulo State University (Unesp), Jaboticabal, Via de Acesso Prof. Paulo Donato Castellane, 14884-900, Jaboticabal, Brazil.

[c]Aquainnovo, Cardonal S/N, Puerto Montt, Chile.

[d]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian EH25 9RG, UK.

[e] Núcleo Milenio INVASAL, Concepción, Chile.

*Corresponding author: Telephone: +56 2 2978 55 33.

E-mail address: jmayanez@uchile.cl

Abstract

The objective of this study were: (i) to assess genotype imputation accuracy in different scenarios using genome-wide SNP data from a population comprising two generations farmed Atlantic salmon and (ii) to assess the accuracy of genomic prediction for a quantitative trait (body weight) using the imputed genotypes. The pedigree consisted of genotypes of 53 parents and 1,069 offspring genotyped using a high-density single nucleotide polymorphism (SNP) panel (50K). Two groups were created: i) Group A: 90% of the offspring were included into training and 10% into validation sets; ii) Group B: 10% of the offspring were included into training and 90% into validation sets. Different scenarios of available genotypic information from relatives were tested for the two groups previously described. Imputation was performed using three in silico low-density panels (0.5, 3 and 6 K) with all markers except the markers present on the low-density panel masked in the validation sets. The accuracy of genomic selection was tested using the imputation scenarios that resulted in the best and the worst imputation accuracy for the three low density panels and were compared to accuracy obtained from pedigree-based best linear unbiased prediction (PBLUP) and genomic predictions using the 50 K SNP panel. In general, imputation accuracy ranged from 0.74 to 0.98 depending on scenario. For the best scenario with the highest number of animals in reference population (Group A), the accuracy of imputation ranged from 0.95 to 0.98 depending on the low-density panel used. For the best scenario with the lowest number of animals in reference population (Group B), the accuracy of imputation ranged from 0.94 to 0.98 depending on the low-density panel used. In general, the number of SNPs in the low-density panels had a greater influence on the accuracy of imputation than the size of the reference set. The accuracies of genomic prediction using imputed genotypes, ranging from 0.71 to 0.73, outperformed PBLUP (0.66) and were identical

or very similar to the use of all true genotype data (0.73). The high imputation genomic prediction accuracy and suggest that the imputation of genotypes from low density (0.5 to 3K) to high density (50K) could be a cost-effective strategy for the feasibility of the practical implementation of genomic selection in Atlantic salmon.

1. Introduction

Recent advances in genotyping technology have facilitated the availability of high density genotyping panels, which can be used to accelerate the genetic progress of breeding programs by implementing genomic selection (Meuwissen et al., 2001). In fact, genomic predictions have shown to increase the accuracy of breeding value estimation for several traits in salmonids (Bangera et al., 2017; Correa et al., 2017; Tsai et al., 2016, 2015; Vallejo et al., 2016; 2017; Yoshida et al., 2018). These methodologies are expected to be increasingly used in aquaculture species (Yáñez et al., 2015), especially for the improvement of traits which are difficult to measure in the selection candidates, such as disease resistance and carcass quality traits (Sonesson and Meuwissen, 2009; Yáñez and Martínez, 2010; Yáñez et al., 2014a). However, the cost associated with genotyping may represent a limiting factor for the use of genomic selection (VanRaden, et al., 2011). An alternative for genomic applications would be to use a genotype imputation method for inferring missing genotypes that were not successfully called during genotyping, to infer the genotypes of ungenotyped parents and/or to infer genotypes for individuals genotyped with a low-density panel using a reference population genotyped for a high-density marker panel (Sargolzaei et al., 2009).

The accuracy of imputation is influenced by several factors, including proportion of genotypes to be imputed (Zhang and Druet, 2010; Hickey et al., 2012), number of individuals in reference set (Druet et al., 2010; Zhang and Druet, 2010), relatedness between validation and reference set (Carvalheiro et al., 2014; Cleveland and Hickey, 2014), chromosomal position (Duarte et al., 2013; Hozé et al., 2013) and minor allele frequency (Badke et al., 2013).

Imputation could be used in Atlantic salmon breeding programs to decrease the costs of genotyping (Tsai et al., 2017). For instance, dense genotypes (e.g. 50K) from parents might be used to impute the missing genotypes from lower to higher-density in the offsprings. However, low accuracies of imputation may be a limitation for the efficient use of lower-density panels. It is worth to mention that the cost for a low-density panel is considerably lower (e.g. USD $ 5 to 15 for a 500 SNP panel) than the cost for a higher-density one (e.g. USD $ 45 to 75 for a 50K SNP panel, depending on the number of samples). Thus, assessing the accuracy of imputation in different scenarios is crucial define an adequate genotyping strategy aiming at maximizing the genetic progress and minimizing the genotyping costs.

The objective of this study was to assess the accuracy of imputation and genomic predictions by testing different scenarios, using different densities of low-density SNP panels and number of animals in reference and validation set in a two-generation farmed Atlantic salmon population.

2. Material and Methods

2.1 Data

The Atlantic salmon population used in the current study belongs to the 2006 and 2010 year-classes of the breeding program of Salmones Chaicas (Puerto Montt, Chile). The origin, management of the fish and genotyping are described in detail by Correa et al. (2015; 2016; 2017), Bangera et al. (2017) and Yáñez et al. (2013; 2014b; 2016). Briefly, the eggs of each full-sib family were incubated and reared in separate tanks from fecundation until tagging. An average number of 32 fish/family (ranging from 26 to 40) were tagged and distributed in six different tanks, with an average of 160 fish/tank (ranging from 137 to 170). Fish were reared until they were an average of 25

months old and the trait body weight was recorded on each individual fish, with an average 331.2 g (SD = 121 g).

Genomic DNA was extracted from fin clip samples from 53 parents (19 sires and 34 dams) and 1,069 offspring, which were genotyped using a 50K Affymetrix SNP array (Correa et al., 2015; 2016; Yáñez et al., 2016), hereafter called the high-density (HD) panel. Before imputation, genotypes and samples were filtered according to the following exclusion criteria: Hardy-Weinberg Disequilibrium (p-value $< 1 \times 10^{-6}$), Minor Allele Frequency (MAF < 0.02) and genotyping rate for SNP and samples < 0.95. The SNPs and samples passed in the quality control were used for downstream analysis.

Three *in silico* low-density (LD) panels were constructed with SNP densities of 499 (LD0.5K), 2,928 (LD3K), and 5,878 (LD6K). The SNPs from the LD panels were initially selected based on a proportional number of SNPs to chromosome size. Then SNPs were selected, based on approximate even spacing within each chromosome, highest MAF (within those which passed quality control) and having unique position when performing BLAST of the 71 pb probes against the reference genome of Atlantic salmon (GenBank Accession no. GCA_000233375.4).

2.2 Imputation scenarios

Two different groups of individuals were created, varying in the proportion of offspring in reference and validation set. For the "Group A" analyses, 90% of the offspring was used as reference and 10% as validation set. The "Group B" analyses were run using 10% of the offspring as reference and 90% as validation set. The assignment of the offspring to the reference and validation sets was at random, and five replicates were used each time.

Five scenarios per group were investigated, each of which defined the validation set for imputation. Scenario 1 (A1 and B1) involved genotyping of all the parents and offspring using the HD panel. Scenario 2 (A2 and B2) and Scenario 3 (A3 and B3) was the same as Scenario 1 except that genotypes for the dams and sires, respectively, were removed from the validation set. Scenario 4 (A4 and B4) and Scenario 5 (A5 and B5) comprised genotyping only the parents and the sibs with the HD panel, respectively. In each scenario, a pedigree of 1,115 individuals was used for imputation, consisting of two generations of records for each genotyped animal. Imputation of genotypes was performed using the FImpute v2.2 software (Sargolzaei et al., 2014) and the accuracy of imputation was calculated as the correlation between true and imputed genotypes for the validation set.

2.3 Genomic predictions

Phenotypic data for body weight were available for animals from the 2010 year-class (Yoshida et al., 2017) and used to test the impact of imputation errors on the accuracy of genomic predictions. The accuracy of genomic predictions was evaluated only in Group B, because this is more proximate to realistic applications of genomic predictions using imputed genotype data. We used the imputed genotypes from the three low-density panels (0.5, 3 and 6K) for Scenario B1 and B5 (Group B) due the fact that these are the scenarios with the highest (B1) and lowest (B5) imputation accuracy. The breeding values (EBV) were estimated using both pedigree and genomic best linear unbiased prediction (PBLUP and GBLUP, respectively). The numerator (A) and genomic (G) relationship matrices were used to account for the kinship between animals in PBLUP and GBLUP, respectively (VanRaden, 2008). The statistical model fitted was as follows:

$y = X\beta + Zg + e$

where y is a vector of phenotypes (body weight), β is a vector of fixed effects (tank and age), g is a vector of additive genetic effects that follows a distribution $\sim N(0, A\sigma_g^2)$ or $\sim N(0, G\sigma_g^2)$, for PBLUP and GBLUP, respectively, where σ_g^2 is the additive genetic variance, and A and G are the pedigree and genomic relationship matrices, respectively. X and Z are incidence matrices for fixed and additive effects, respectively, and e is the vector of random residual with a distribution $\sim N(0, I\sigma_e^2)$, where σ_e^2 is the residual variance and I is an identity matrix.

We used the BLUPF90 software package (Misztal et al., 2016) to perform the genetic evaluations using pedigree and genomic information. Prediction accuracies were assessed using a five-fold cross validation scheme. Briefly, all phenotyped and genotyped animals (n = 963) were randomly divided into five validations sets (20% of the dataset; mean = 192 and SD = 4 animals), which were predicted one at a time by masking their phenotypes and using the remaining animals as a training set (80% of the dataset; n = 771 and SD = 4 animals) to estimate the marker effects. Prediction accuracies were calculated in the validation sets using the following formula:

$$r_{GEBV,BV} = \frac{r_{GEBV,y}}{h},$$

where $r_{GEBV,y}$ is the correlation between the EBV or GEBV of a given model (predicted for the validation set using information from the training set) and the phenotypic record, while h is the square root of the pedigree-based estimate of heritability.

3. Results

3.1 Accuracy of genotype imputation

A total of 37,259 SNPs and 1,122 samples passed the filtering criteria. We observed that genotype imputation accuracy increased with increasing marker density of

the LD genotyping panels and with increasing proportions of close ancestors having high-density genotypes (Table 1). For all cases, imputation accuracy decreased with reduced marker density going from 6K to 0.5K (Table 1). The largest increase in imputation accuracy occurred when increasing SNP density from 0.5K to 3K (rather than from 3K to 6K), indicating that the 3K panel would provide highly accurate imputed genotypes, with similar imputation accuracies than the 6K panel.

The lower number of animals in Group B reference set resulted in lower imputation accuracies compared to Group A, with the difference being more evident when fewer number of ancestors were used as reference set (Scenarios A1 *vs* A5 and B1 *vs* B5). In addition, the largest changes in accuracy between genotyping scenarios were observed for the SNPs chips with the lower density, LD0.5K. In the case where genotyping both parents is not possible, we observed that the sire's genotype information (B2) is more important to obtain better imputations accuracies than just the dam's genotype information (B3) for Group B (Table 1). For both Groups A and B, the HD genotype information of both parents is more critical to achieve high imputation accuracy than only sibs' information, especially for LD0.5K panels, but this relevance decreased with increasing SNP density of the LD panel.

Figures 1 and 2 show imputation accuracy in chromosome 1 for Group A and B, respectively, for Scenario 1 and 5, (for all chromosomes see Supplementary material). Imputation accuracy was not consistent across the chromosomes and depended on physical position of imputed SNP and location of low-density SNP. The largest differences were observed for LD0.5K panel in both Group A and B (Figure 1 and 2, respectively). The imputation accuracy decreased greatly at chromosomal ends, especially for LD0.5K panel. Increasing SNP density of the LD panel form 0.5K to 3K or 6K substantially improved imputation accuracy at chromosome ends.

3.2 Accuracy of genomic predictions using imputed genotypes

The genomic prediction accuracy using imputed genotypes, were identical or very similar among the scenarios and low-density panel tested compared to the use of the real 50K SNP genotypes (Figure 3). As expected, the lowest genomic prediction accuracy was observed for the scenario and SNP panel with lowest imputation accuracy (i.e. Scenario B5 and LD0.5K panel), which resulted in an accuracy slightly lower compared to the use of the real 50K SNP panel (0.71 *vs* 0.73, respectively). All other SNP panel densities from Scenario B1 and B5 resulted in genomic prediction accuracy higher than pedigree-based method (0.66) and identical to the use of a 50K SNP panel. The prediction accuracy improved 11% and 8% for the best and worst Scenario (Scenario B1 and Scenario B5/LD0.5K, respectively) compared to the pedigree-based method.

4. Discussion

The imputation method used by FImpute is based on the concept that close relatives share long haplotypes and the imputation is carried out using overlapping sliding windows starting with long haplotypes and moving towards short haplotypes (Sargolzaei et al., 2014). According to previous studies, the method results in high imputation accuracy when close relatives of targeted individuals are present in the reference group and computing requirements are considerably lower than other software used for imputation (Carvalheiro et al., 2014; Larmer et al., 2014; Sargolzaei et al., 2014).

Here we found that, in general, the imputation accuracy decreased in a non-linear manner from 6K to 0.5K, which is in accordance with the results obtained by

Habier et al. (2009) and Hickey et al. (2012), who also observed that the higher the proportion of genotypes to be imputed, the lower is the imputation accuracy. This can be due to the fact that panels with few SNPs could present low linkage and linkage disequilibrium between the markers, increasing imputation errors. The similarly high imputation accuracy between 3 and 6K SNP panels are in agreement with studies carried out with pigs (Duarte et al., 2013; Cleveland and Hickey, 2014), cattle (Druet et al., 2010; Zhang and Druet, 2010; Carvalheiro et al., 2014), sheep (Hayes et al., 2012) and Atlantic salmon (Kijas et al., 2016; Tsai et al., 2017). The 500 SNP panel showed the lowest imputation accuracies for the different scenarios tested; however, is most likely to be considerably much cheaper than any 3K and 6K SNP panel, thus cost-effectiveness must be carefully evaluated, taking genotyping cost and imputation and genomic prediction accuracies into account.

Based on the results from different scenarios for Group A and B, the effect of the number of genotyped individuals in the reference SNP panel seems to be smaller than the influence of the number of SNPs in LD panel. The size of the reference population should be large enough and not be a major factor influencing imputation accuracy using small panels for a portion of the population. The main benefit of increasing number of reference individuals will be obtained through increasing genotyping of highly related individuals between target and reference individuals (Zhang and Druet, 2010).

Here we also found that imputation accuracy was higher by using HD genotypes from sires instead of HD genotypes from dams, when HD genotypes from only one parent were available. These differences are more evident for Group B. The large paternal full-sib sample structure (~56 offspring per sire) could contribute for a better haplotype reconstruction than just using maternal full-sib samples (~31 offspring per dam). In addition, Gilbey et al. (2004); Lien et al. (2011) and Moen et al. (2004)

suggested that there are large differences in recombination rates between sexes in *Salmo salar*, in a ratio ranging from 1.38:1 to 8.26:1 (female:male). A slow decay in linkage disequilibrium could be a consequence of low recombination rates in males which resulted in higher accuracy of imputation when compared to females.

The lower accuracy found around the beginning and end of the chromosomes could be due to the fact the recombination is known to be higher around the telomeres, which would decrease the precision of haplotype reconstruction and imputation accuracy (Chowdhury et al., 2009; Tortereau et al., 2012). Low imputation accuracies in centromere regions might be attributed to incorrect order of markers on the reference genome in regions difficult to assemble.

In some chromosome regions, a notably low imputation accuracy is evident (e.g. chromosome 8 and 17 in Scenario 4, Supplementary material 1D and 2D). This suggest errors in the SNP position given by an incorrect anchoring of these markers to the genome or errors in the current reference genome assembly. These regions had markers with very low levels of linkage disequilibrium with neighboring markers, resulting in very low imputation accuracies (Carvalheiro et al., 2014; Druet et al., 2010). Sun et al. (2012) observed that imputation accuracy was positively associated with chromosome size due to the fact that longer chromosomes harbour more markers, and hence providing more information for inferring unknown haplotypes and imputing missing genotypes. In longer chromosomes, the problem of low imputation accuracy at the beginning and end of the chromosomes are relatively less important than in shorter chromosomes.

The imputation accuracy improved when Erbe et al. (2012) remapped the SNPs with high errors rates using linkage disequilibrium. However, they still found poorly imputed SNPs after remapping, suggested that recombination hot spots or regions on the

panel with lower SNP density, could result the high imputation error rates for some SNPs (Hozé et al., 2013). In our study, some of the markers with low imputation accuracy were removed before imputation (~3% of all SNPs in Scenario A5 and B5) to try to improve the accuracy. Only a marginal gain was observed, ranging, for example, from 0.83 to 0.86 and 0.74 to 0.77 for LD0.5K panel in Scenario A5 and B5, respectively, which were the scenarios with the highest accuracy gain. This result is most likely due to the small proportion of discarded markers. However, for genome wide association studies could be preferable to treat these markers with high error rates with caution, to avoid the negative impact of imputation errors in the QTL detection.

To test the impact of genotype imputation errors in genome-enabled selection methods, we estimated the accuracy of genomic predictions for body weight using the worst and the best scenarios for Group B, based on imputation accuracy. The present results are in accordance to previous studies carried out in aquaculture (Tsai et al., 2017) and livestock species (Berry and Kearney, 2011; Erbe et al., 2012), in which genomic prediction accuracies using imputed genotypes were always higher than those obtained using pedigree-based BLUP and not much lower than using HD genotypes.

The present study showed that the genomic prediction accuracy using imputed genotypes, for all densities of LD SNP panels, outperformed the pedigree-based method in the best and the worst scenarios of imputation accuracy (B1 and B5). In addition, the genomic prediction accuracies when genotyping both parents and a proportion of the progeny (10%) with the HD panel (Scenario B1), along the three low density panels used in the validation population, were identical to the accuracies obtained by using the 50K SNP panel. The same result was observed for LD panels of 3K and 6K when both parents were genotyped with the HD panel (Scenario B5). As expected, the lowest genomic prediction accuracy was observed for the scenario and SNP panel with lowest

imputation accuracy (i.e. Scenario B5 and LD0.5K panel), which resulted in an accuracy slightly lower compared to the use of the real 50K SNP panel (0.71 vs 0.73, respectively). These results indicate that the use of an appropriate genotyping strategy combining a genotyping of both parents and a percentage (10%) of the total progeny using a HD panel and a greater proportion (90%) of the progeny with panel Low density (500 SNPs) represents an alternative to reach similar accuracy precision to that achieved by genotyping all animals with a HD panel. These results may be used to plan genotyping strategies to reduce the costs for the practical implementation of genomic selection in Atlantic salmon.

References

Badke, Y.M., Bates, R.O., Ernst, C.W., Schwab, C., Fix, J., Van Tassell, C.P., Steibel, J.P., 2013. Methods of tagSNP selection and other variables affecting imputation accuracy in swine. BMC Genet. 14, 8. doi:10.1186/1471-2156-14-8

Bangera, R., Correa, K., Lhorente, J.P., Figueroa, R., Yáñez, J.M., 2017. Genomic predictions can accelerate selection for resistance against Piscirickettsia salmonis in Atlantic salmon (*Salmo salar*). BMC Genomics 18, 121. doi:10.1186/s12864-017-3487-y

Berry, D.P., Kearney, J.F., 2011. Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. Animal 5, 1162–1169. doi:10.1017/S1751731111000309

Carvalheiro, R., Boison, S.A., Neves, H.H.R., Sargolzaei, M., Schenkel, F.S., Utsunomiya, Y.T., Maria, A., Brien, P.O., Sölkner, J., Mcewan, J.C., Tassell, C.P. Van, Sonstegard, T.S., Garcia, J.F., 2014. Accuracy of genotype imputation in Nelore cattle. Genet. Sel. Evol. 46, 1–11.

Chowdhury, R., Bois, P.R.J., Feingold, E., Sherman, S.L., Cheung, V.G., Broman, K., Murray, J., Sheffield, V., White, R., Weber, J., Cheung, V., Burdick, J., Hirschmann, D., Morley, M., Coop, G., Wen, X., Ober, C., Pritchard, J., Przeworski, M., Shifman, S., Bell, J., Copley, R., Taylor, M., Williams, R., Petkov, P., Broman, K., Szatkiewicz, J., Paigen, K., Hassold, T., Hall, H., Hunt, P., Lamb, N., Sherman, S., Hassold, T., Henderson, K., Kee, K., Maleki, S., Santini, P., Keeney, S., Keeney, S., Neale, M., Stahl, F., Foss, H., Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Kong, A., Thorleifsson, G., Stefansson, H., Masson, G., Helgason, A., Weiss, L., Shen, Y., Korn, J., Arking, D., Miller, D., Dawber, T., Meadors, G., Jr, F.M., Kong, A., Gudbjartsson, D., Sainz, J., Jonsdottir, G., Gudjonsson, S., Su, A., Cooke, M., Ching, K., Hakak, Y., Walker, J., Barrett, T., Troup, D., Wilhite, S., Ledoux, P., Rudnev, D., Lamprecht, G., Seidler, U., Lassalle, B., Bastos, H., Louis, J., Riou, L., Testart, J., Bastos, H., Lassalle, B., Chicheportiche, A., Riou, L., Testart, J., Purcell, S., Neale, B., Todd-

Brown, K., Thomas, L., Ferreira, M., Li, Y., Abecasis, G., 2009. Genetic Analysis of Variation in Human Meiotic Recombination. PLoS Genet. 5, e1000648. doi:10.1371/journal.pgen.1000648

Cleveland, M.A., Hickey, J.M., 2014. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J. Anim. Sci. *91* 3583–3592. doi:10.2527/jas2013-6270

Correa, K., Bangera, R., Figueroa, R., Lhorente, J.P., Yáñez, J.M., 2017. The use of genomic information increases the accuracy of breeding value predictions for sea louse (*Caligus rogercresseyi*) resistance in Atlantic salmon (*Salmo salar*). Genet. Sel. Evol. 49, 15. doi:10.1186/s12711-017-0291-8

Correa, K., Lhorente, J.P., Bassini, L., López, M.E., Di Génova, A., Maass, A., Davidson, W.S., Yáñez, J.M., 2016. Genome wide association study for resistance to Caligus rogercresseyi in Atlantic salmon (*Salmo salar L.*) using a 50K SNP genotyping array. Aquaculture. doi:10.1016/j.aquaculture.2016.04.008

Correa, K., Lhorente, J.P., López, M.E., Bassini, L., Naswa, S., Deeb, N., Di Genova, A., Maass, A., Davidson, W.S., Yáñez, J.M., 2015. Genome-wide association analysis reveals loci associated with resistance against Piscirickettsia salmonis in two Atlantic salmon (*Salmo salar L.*) chromosomes. BMC Genomics 16, 854. doi:10.1186/s12864-015-2038-7

Druet, T., Schrooten, C., de Roos, A.P.W., 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J. Dairy Sci. 93, 5443–5454. doi:10.3168/jds.2010-3255

Duarte, J.L.G., Bates, R.O., Ernst, C.W., Raney, N.E., Cantet, R.J.C., Steibel, J.P., 2013. Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. BMC Genetics 14, 38

Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., Reich, C.M., Mason, B.A., Goddard, M.E., 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95, 4114–4129. doi:10.3168/jds.2011-5019

Gilbey, J., Verspoor, E., McLay, A., Houlihan, D., 2004. A microsatellite linkage map for Atlantic salmon (*Salmo salar*). Anim. Genet. 35, 98–105.

Habier, D., Fernando, R., Dekkers, J., 2009. Genomic selection using low-density marker panels. Genetics 182, 343-353. doi:10.1534/genetics.108.100289.

Hayes, B.J., Bowman, P.J., Daetwyler, H.D., Kijas, J.W., van der Werf, J.H.J., 2012. Accuracy of genotype imputation in sheep breeds. Anim. Genet. 43, 72–80. doi:10.1111/j.1365-2052.2011.02208.x

Hickey, J.M., Crossa, J., Babu, R., de los Campos, G., 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52, 654. doi:10.2135/cropsci2011.07.0358

Hozé, C., Fouilloux, M.-N., Venot, E., Guillaume, F., Dassonneville, R., Fritz, S., Ducrocq, V., Phocas, F., Boichard, D., Croiseau, P., 2013. High-density marker imputation accuracy in sixteen French cattle breeds. Genet. Sel. Evol. 45, 33. doi:10.1186/1297-9686-45-33

Kijas, J., Elliot, N., Kube, P., Evans, B., Botwright, N., King, H., Primmer, C.R., Verbyla, K., 2016. Diversity and linkage disequilibrium in farmed Tasmanian Atlantic salmon. Anim. Genet. doi:10.1111/age.12513

Larmer, S.G., Sargolzaei, M., Schenkel, F.S., 2014. Extent of linkage disequilibrium, consistency of gametic phase, and imputation accuracy within and across Canadian dairy breeds. J. Dairy Sci. 97, 3128–3141. doi:10.3168/JDS.2013-6826

Lien, S., Gidskehaug, L., Moen, T., Hayes, B.J., Berg, P.R., Davidson, W.S., Omholt,

S.W., Kent, M.P., 2011. A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. BMC Genomics 12, 615.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics 157*, 1819-1829.

Misztal, I., Tsuruta, S., Lourenco, D., Masuda, Y., Aguilar, I., Legarra, A., 2016. Manual for BLUPF90 family of programs. Athens: University of Georgia.

Moen, T., Hoyheim, B., Munck, H., Gomez-Raya, L., 2004. A linkage map of Atlantic salmon (*Salmo salar*) reveals an uncommonly large difference in recombination rate between the sexes. Anim. Genet. 35, 81–92.

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

Sargolzaei, M., Chesnais, J.P., Schenkel, F.S., 2009. Accuracy of a family-based genotype imputation algorithm. In: GEB Open Industry Session, Saint-Hyacynthe Quebec, Canada. Accessed Jun. 1, 2017. www.cdn.ca/Articles/GEBAPR2010/Mehdi%20-%20Accuracy%20of%20Family-Based%20Imputing.pdf.

Sonesson, A.K., Meuwissen, T.H.E., 2009. Testing strategies for genomic selection in aquaculture breeding programs. Genet. Sel. Evol. 41, 37. doi:10.1186/1297-9686-41-37

Sun, C., Wu, X.-L., Weigel, K.A., Rosa, G.J.M., Bauck, S., Woodward, B.W., Schnabel, R.D., Taylor, J.F., Gianola, D., 2012. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genet. Res. (Camb). 94, 133–150.

doi:10.1017/S001667231200033X

Tortereau, F., Servin, B., Frantz, L., Megens, H.-J., Milan, D., Rohrer, G., Wiedmann, R., Beever, J., Archibald, A.L., Schook, L.B., Groenen, M.A., 2012. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. BMC Genomics 13, 586. doi:10.1186/1471-2164-13-586

Tsai, H.-Y., Hamilton, A., Tinch, A.E., Guy, D.R., Bron, J.E., Taggart, J.B., Gharbi, K., Stear, M., Matika, O., Pong-Wong, R., Bishop, S.C., Houston, R.D., 2016. Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations. Genet. Sel. Evol. 48, 47. doi:10.1186/s12711-016-0226-9

Tsai, H.-Y., Matika, O., Edwards, S.M., Antolín–Sánchez, 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 Genes, Genomes, Genet. 7.

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 1–9. doi:10.1186/s12864-015-2117-9

VanRaden, P.M., O'Connell, J.R., Wiggans, G.R., Weigel, K.A., Calus, M., Meuwissen, T., Roose, A., Veerkamp, R., Solberg, T., Sonesson, A., Woolliams, J., VanRaden, P., Tassell, C. Van, Wiggans, G., Sonstegard, T., Schnabel, R., Taylor, J., Schenkel, F., Wiggans, G., VanRaden, P., Bacheller, L., Tooker, M., Hutchison, J., Cooper, T., Sonstegard, T., Weigel, K., Campos, G. de los, González-Recio, O., Naya, H., Wu, X., Long, N., Rosa, G., Gianola, D., Weigel, K., Campos, G. de los, Vazquez, A., Rosa, G., Gianola, D., Tassell, C. Van,

VanRaden, P., Wiggans, G., Tassell, C. Van, Sonstegard, T., Schenkel, F., Harris, B., Johnson, D., Li, Y., Willer, C., Sanna, S., Abecasis, G., Druet, T., Schrooten, C., Roos, A. de, Lund, M., Roos, A. de, Vries, A. de, Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, M., Su, G., Burdick, J., Chen, W., Abecasis, G., Cheung, V., Habier, D., Fernando, R., Dekkers, J., Zhang, Z., Druet, T., Villumsen, T., Janss, L., Wiggans, G., Sonstegard, T., VanRaden, P., Matukumalli, L., Schnabel, R., Taylor, J., Schenkel, F., Tassell, C. Van, Meuwissen, T., Hayes, B., Goddard, M., VanRaden, P., Sargolzaei, M., Schenkel, F., Toosi, A., Fernando, R., Dekkers, J., Yang, J., Benyamin, B., McEvoy, B., Gordon, S., Henders, A., Nyholt, D., Madden, P., Heath, A., Martin, N., Montgomery, G., Henderson, C., Cole, J., VanRaden, P., VanRaden, P., Sullivan, P., Liu, Z., FSeefried, null, Reinhardt, F., Reents, R., Flaquer, A., Fischer, C., Wienker, T., Cole, J., VanRaden, P., O'Connell, J., Tassell, C. Van, Sonstegard, T., Schnabel, R., Taylor, J., Wiggans, G., VanRaden, P., O'Connell, J., Wiggans, G., Weigel, K., Weigel, K., Campos, G. de los, Vazquez, A., Tassell, C. Van, Rosa, G., Gianola, D., O'Connell, J., VanRaden, P., Wiggans, G., Vanraden, P., Taylor, J., Bean, B., Marshall, C., Sullivan, J., Macciotta, N., Gaspa, G., Steri, R., Nicolazzi, E., Dimauro, C., Pieramati, C., Cappio-Borlino, A., Villa-Angulo, R., Matukumalli, L., Gill, C., Choi, J., Tassell, C. Van, Grefenstette, J., 2011. Genomic evaluations with many more genotypes. Genet. Sel. Evol. 43, 10. doi:10.1186/1297-9686-43-10

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423. doi:10.3168/jds.2007-0980

Vallejo, R.L., Leeds, T.D., Fragomeni, B.O., Gao, G., Hernandez, A.G., Misztal, I., Welch, T.J., Wiens, G.D., Palti, Y., 2016. Evaluation of genome-enabled selection

for bacterial cold water disease resistance using progeny performance data in rainbow trout: insights on genotyping methods and genomic prediction models. Front. Genet. 7, 1–13. doi:10.3389/fgene.2016.00096

Vallejo, R.L., Leeds, T.D., Gao, G., Parsons, J.E., Martin, K.E., Evenhuis, J.P., Fragomeni, B.O., Wiens, G.D., Palti, Y., 2017. Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree-based model in rainbow trout aquaculture. Genet. Sel. Evol. 49, 17. doi:10.1186/s12711-017-0293-6

Yañez, J.M., Bangera, R., Lhorente, J.P., Oyarzún, M., Neira, R., 2013. Quantitative genetic variation of resistance against Piscirickettsia salmonis in Atlantic salmon (*Salmo salar*). Aquaculture 414–415, 155–159.

Yáñez, Newman, S., Houston, R.D., 2015. Genomics in aquaculture to better understand species biology and accelerate genetic progress. Front. Genet. 6, 1–3. doi:10.3389/fgene.2015.00128

Yáñez, J.M., Houston, R.D., Newman, S., 2014a. Genetics and genomics of disease resistance in salmonid species. Front. Genet. 5, 1–13. doi:10.3389/fgene.2014.00415

Yáñez, J.M., Lhorente, J.P., Bassini, L.N., Oyarzún, M., Neira, R., Newman, S., 2014b. Genetic co-variation between resistance against both *Caligus rogercresseyi* and Piscirickettsia salmonis, and body weight in Atlantic salmon (*Salmo salar*). Aquaculture 433, 295–298. doi:10.1016/j.aquaculture.2014.06.026

Yañez, J.M., Martinez, V., 2010. Genetic factors involved in resistance to infectious diseases in salmonids and their application in breeding programmes. Arch. Med. Vet. 42, 1–13.

Yáñez, J.M., Naswa, S., López, 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. Genome-wide single nucleotide polymorphism (SNP) discovery in Atlantic salmon (*Salmo salar*): validation in wild and farmed American and European populations. Mol. Ecol. Resour. n/a-n/a. doi:10.1111/1755-0998.12503

Yáñez, J.M., Newman, S., Houston, R.D., 2015. Genomics in aquaculture to better understand species biology and accelerate genetic progress. Front. Genet. 6, 1–3. doi:10.3389/fgene.2015.00128

Yoshida, G.M., Bangera, R., Carvalheiro, R., Correa, K., Figueroa, R., Lhorente, J.P., Yáñez, J.M., 2017. Genomic prediction accuracy for resistance against piscirickettsia salmonis in farmed rainbow trout. G3 Genes|Genomes|Genetics; Genes|Genomes|Genetics g3.300499.2017. doi:10.1534/g3.117.300499

Yoshida, G.M., Lhorente, J.P., Carvalheiro, R., Yáñez, J.M., 2017. Bayesian genome-wide association analysis for body weight in farmed Atlantic salmon (*Salmo salar* L.). Anim. Genet. doi:10.1111/age.12621

Zhang, Z., Druet, T., 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. J. Dairy Sci. 93, 5487–5494. doi:10.3168/jds.2010-3501

Table 1. Imputation accuracy from low-density (LD) to high-density (HD) panel for Atlantic salmon using groups with different numbers of animals in reference and validation set and different scenarios of available genotypic information.

S cenario	Reference				Vali dation[b]	L D0.5K M ean	] D3K l ean	] D6K l ean
	ires	ams		ize				
	]	] ibs[a]						
	= 19	= 34						
Group A								
A 1	] D	] D	D	,015	0 LD	0 .948	(.980	(.983
A 2	] D		D	81	LD	0 .863	(.967	(.973
A 3		] D	D	96	LD	0 .864	(.962	(.971
A 4		D	62		LD	0 .851	(.976	(.982
A 5	] D	] D		3	LD	0 .829	(.942	(.951
Group B								
B 1	] D	] D	D	59	LD	0 .940	(.974	(.977
B 2	] D		D	25	LD	0 .820	(.944	(.955

						0	(	(
B		]			LD			
3		D	D	40		.786	.932	.948
B					LD	0	(	(
4			D	06		.737	.936	.950
B		]	]		LD	0	(	(
5	D	D		3		.736	.943	.952

[a] 963 and 106 offsprings with HD in reference set for Group A and B, respectively.

[b] 106 and 963 offsprings with LD in the validation set for Group A and B, respectively.

Figure 1. Correlations between observed and imputed genotypes for each SNP, for Group A and Scenarios 1 (A1), using the low-density panel 0.5 (A), 3 (B) and 6K (C), and Scenario 5 (A5), using the low-density panel 0.5 (D), 3 (E) and 6K (F), for chromosome 1. The red dots indicates the SNPs physical position for low-density panel.

Figure 2. Correlations between observed and imputed genotypes for each SNP, for Group B and Scenarios 1 (B1), using the low-density panel 0.5 (A), 3 (B) and 6K (C), and Scenario 5 (B5), using the low-density panel 0.5 (D), 3 (E) and 6K (F), for chromosome 1. The red dots indicates the SNPs physical position for low-density panel.

Figure 3. Accuracy of genomic prediction for body weight using the pedigree-based method (PBLUP), true 50 K genotypes (50 K) and imputed genotypes to Scenarios 1B and 5B, under different low-density panels (0.5, 3 and 6K) for *Salmon salar*.

Additional file 1. Correlations between observed and imputed genotypes for each SNP, for Group A and Scenarios 1 (A1), using the low-density panel 0.5 (A), 3 (B) and 6K (C), and Scenario 5 (A5), using the low-density panel 0.5 (D), 3 (E) and 6K (F), for all chromosomes. The red line indicates the mean imputation accuracy for each scenario.
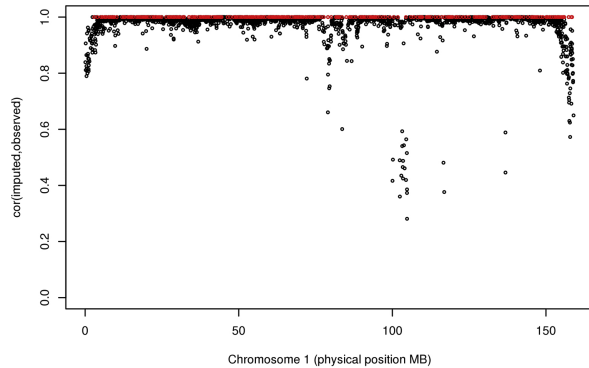
Additional file 2. Correlations between observed and imputed genotypes for each SNP, for Group B and Scenarios 1 (B1), using the low-density panel 0.5 (A), 3 (B) and 6K (C), and Scenario 5 (B5), using the low-density panel 0.5 (D), 3 (E) and 6K (F), for all chromosomes. The red line indicates the mean imputation accuracy for each scenario.

Highlights

- We assessed genotype imputation accuracy in farmed Atlantic salmon using low-density (LD) SNP panels
- Accuracy varied from 0.74 to 0.98 according to the LD SNP panel used
- Accuracy of genomic prediction for imputed genotypes were very similar to true genotypes
- Imputation represents a cost-effective approach for genomic selection and GWAs

Figure 1

Figure 2

Figure 3