Potential of imputation for cost-efficient genomic selection for resistance to *Flavobacterium columnare* in Rainbow trout

C. Fraslin[1*], A. Nousiainen[2] , H. Koskinen[2], A. Kause[3], and R.D. Houston[1]

[1] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, United-Kingdom; [*]clemence.fraslin@roslin.ed.ac.uk
[2] Natural Resources Institute Finland (Luke), Neulaniementie 5, FI-70210, Kuopio, Finland
[3] Natural Resources Institute Finland (Luke), Myllytie 1, FI-31600, Jokioinen, Finland

Abstract
Columnaris disease (CD) is a major emerging disease affecting rainbow trout aquaculture. Selective breeding through genomic selection (GS) has potential to improve host resistance. However, GS is expensive partly due to the cost of genotyping high numbers of animals using high-density (HD) SNP arrays. The objective of this study was to assess the efficiency of GS for resistance to CD using ten *in silico* LD-panels combined with imputation. After a natural outbreak of CD, 2,874 challenged fish and 469 fish from the parental generation (n=81 parents) were genotyped with 27,907 SNPs. Using LD-panels (3,000 makers) alone or lower density panels (300 markers) along with imputation resulted in comparable accuracy to the 28K HD-panel, and 11% higher accuracy than pedigree-based predictions. Compared to using the commercial HD-panel, LD-panels with imputation may provide a more affordable route to genomic prediction of breeding values, supporting wider adoption of GS in aquaculture breeding programmes.

Introduction
Columnaris disease (CD) is an emerging problem for various aquaculture species worldwide. *Flavobacterium columnare*, the pathogen agent of CD, causes high levels of mortality in rainbow trout fry, causing significant annual economic losses to the industry. Improvement of disease resistance through selective breeding is a promising approach to reduce mortality, and could reduce disease incidence and reduce antibiotic use. Genomic selection (GS) uses genome-wide marker information gathered in a reference population to predict the breeding value for traits of economic interest in a target broodstock. GS is particularly applicable for traits difficult to measure in breeding candidates (such as disease resistance traits) and has been shown to be more accurate, and thus more efficient, than pedigree-based selection. However, the cost of genotyping large numbers of individuals required to perform GS is still prohibitive for many breeding programmes. Using low-density (LD) SNP panels could reduce the cost of genotyping (since LD panels are typically cheaper to genotype) while still achieving higher prediction accuracy than standard pedigree-based selection. Genotype imputation from low-density to high-density genotypes is therefore a promising approach to achieve a more affordable genomic selection. The aim of this study was to assess the efficiency of GS using low-density SNP panels along with imputation for this disease resistance trait.

Materials & Methods
Fish rearing and disease outbreak management. In May 2019, 81 rainbow trout (33 females and 48 males) were selected among 567 fish from the Finnish national breeding programme maintained by Luke, based on their relationship and genetic contribution to maintain a predetermined inbreeding coefficient of less than 1% per generation. The 33 dams and 48 sires were mated to create 105 full-sib families. In June 2019, around 30,000 fry were

separated into three fingerling tanks, about 100 fish per family per tank, at the farm of Hanka-Taimen Oy (Finland). From arrival at this new farm (considered as day 0 of the experiment), the fish mortality and any disease signs were monitored twice a day. On day 11 of the experiment, fish in all three tanks started to show signs of CD, some dead fish were sampled and sent to a veterinarian to confirm the CD diagnosis. From day 15 to 24, a piece of tail was taken, for later DNA extraction, from 510 fish per tanks, randomly chosen amongst the dead or dying fish with clear CD signs (considered as susceptible). At day 26, the three tanks were treated following the veterinarian guidelines against *F. columnare* with an approved treatment until day 32. On the last day of the experiment, day 99, a piece of tail was collected, for latter DNA extraction, on about 506 fish per tanks, randomly sampled among the fish still alive at that time (considered as resistant). In addition, the biggest fish among the sampled fish from each tank (n=164, 167, 168 for tanks 1, 2 and 3, respectively) were PIT-tagged and are considered as breeding candidates for the future broodstock

Genotyping, in silico low-density panels, and imputation. A total of 3,054 challenged fish (1,538 susceptible and 1,519 resistant) and 570 fish from the parental generation (including the 81 parents) were genotyped using the 57K SNP Axiom^{TM} Trout Genotyping Array (Palti et al., 2015). After quality controls, 2,874 challenged fish and 469 fish from the parental generation (n=78 parents of the challenged fish) genotyped with 27,907 informative SNPs were retained. Those 28K SNPs were considered as the high-density (HD) panel. The impact of reducing the SNP density on genomic prediction was tested with ten low-density (LD) SNP panels, created *in silico* (300; 500; 700; 1,000; 3,000; 5,000; 7,000; 10,000; 15,000 and 20,000 SNPs). For each LD-panel, SNPs were sampled from the 28K HD-panel using the CVrepGPAcalc package (Tsairidou et al., 2020). Two sampling approaches were used to create LD-panels, i) SNPs were randomly sampled within each chromosome, with the number of SNP sampled from a given chromosome being proportional to the physical length of the chromosome in the *O. mykiss* reference genome (Omyk_0.1, Gao et al., 2018). ii) SNPs were selected so that they would be equally spaced within each chromosome. Each LD-panel was replicated 10 times, with replicates allowed to overlap by chance and the final number of SNPs within each panel being allowed to slightly vary from the target density. FImpute3 (Sargolzaei et al., 2014) was used to impute the *in silico* LD- genotyped offspring back to the full 28K SNPs using a combined population and pedigree based imputation method with the HD-genotyped parents (n=469) as the reference population.

Genetic parameter estimates and genomic selection. Disease resistance was analysed as a binary trait (0 = alive; 1 = dead) with the rearing tank as a fixed effect in the statistical model. Variance components and heritabilities were estimated using ASReml (v.4.1, Gilmour et al., 2014) with two different approaches, a mixed linear BLUP animal model and a logistic regression model to assess the trait on the observed and underlying scale, respectively.
The (genomic) estimated breeding values {(G)EBV} of fish were computed using mixed linear BLUP animal model based on pedigree (PBLUP) or genomic (GBLUP) information using the BLUPF90 software (Misztal et al., 2002). The efficiency of selection using pedigree-based or genomic information was estimated using Monte-Carlo "leave-one-group-out" method by removing the known phenotype from 20% of the fish, and then using the phenotype and genotype data of the remaining 80% fish to predict the (G)EBVs of the 20% validation fish group. This procedure was repeated 20 times for the PBLUP, the HD-GBLUP, and for each of the 10 replicates of the LD-panels, pre and post-imputation. Accuracy of prediction was computed, for all SNP panels, as the mean over the 20 replicates of the correlation between the (G)EBV and the true phenotype of fish in the validation group, divided by the square root of the genomic-based heritability.

Results

Resistance to *F. columnare* was moderately heritable in this population with pedigree-based heritability estimated to be 0.18 (± 0.038 se) and 0.35 (± 0.046) on the observed and underlying logit scales, respectively. Genomic heritability was estimated to be 0.21 (± 0.029 se) and 0.43 (± 0.042 se) on the observed and underlying logit scales respectively. Pedigree-based prediction accuracy was estimated as 0.59 (± 0.080 sd) and the use of genomic evaluation increased the prediction accuracy by 14% for the HD-GBLUP (0.68 ± 0.076 sd). Decreasing the number of SNPs tended to decrease the accuracy of genomic predictions (Figure 1), and no significant difference was observed between the two methods of sampling.
Prediction accuracies obtained with 300 to 500 SNPs were close to the accuracy obtained with the pedigree-based analysis (blue dotted line in Figure 1). Encouragingly, prediction accuracies obtained with LD-panels from 7,000 SNPs and above were close to the accuracy obtained with the HD-panel (-1% in accuracy compared with the HD-GBLUP). Accuracies obtained with 1,000 SNPs were only 4% higher than those obtained with the pedigree, whereas the accuracy obtained with only 3,000 SNPs was 3% lower than the accuracies obtained with the HD-panel and thus 11% higher than the accuracy obtained with the pedigree only.

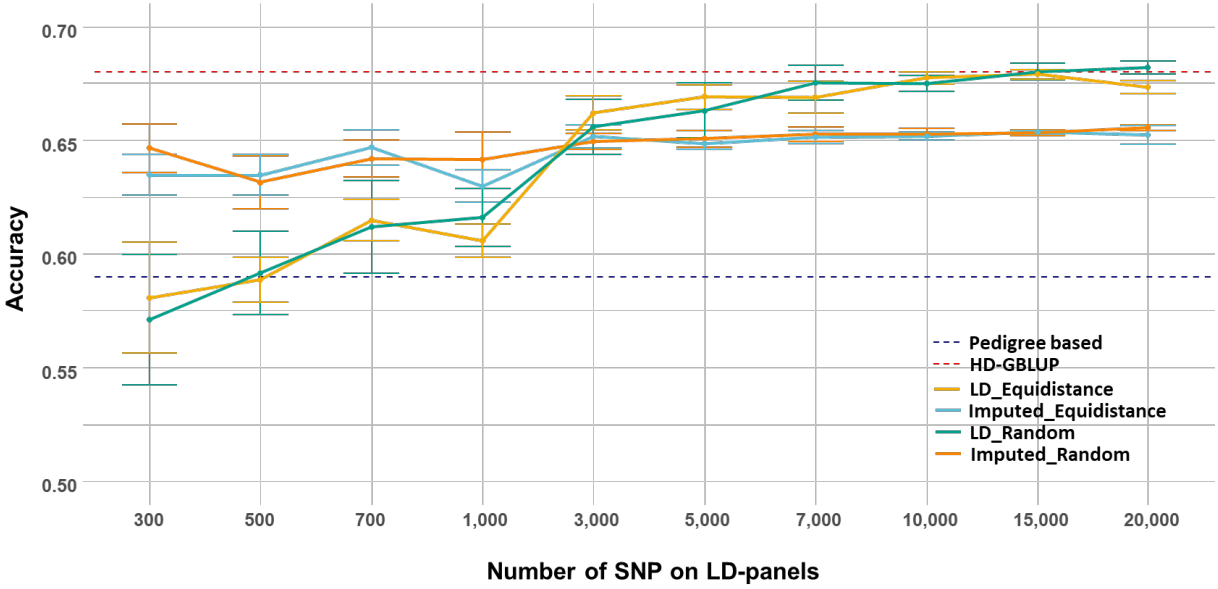

Figure 1. Accuracy of genomic prediction for resistance to *F. columnare* in rainbow trout, obtained with different SNP panels density, before and after imputation.

At the lowest densities (< 3,000 SNPs, Figure 1) imputation (with FImpute3) had a positive impact on the accuracy of genomic prediction with the accuracy of genomic prediction obtained after imputation similar to the accuracy obtained with 3,000 SNPs.
Surprisingly, the accuracy of genomic prediction obtained with 3,000 SNPs was not significantly different before and after imputation and for 5,000 SNPs or higher, the accuracy of genomic prediction obtained after imputation was slightly lower than without imputation.

Discussion

The results presented here suggest that resistance to *F. columnare*, is moderately heritable in this population and could be improved through selective breeding. The results on genomic prediction are consistent with those from various aquaculture species that relatively LD SNP panels are, at least more efficient than pedigree-based selection and sometimes as efficient as

full HD panels to accurately predict breeding values (see Houston et al., 2020 for review). More specifically, using LD SNP panels (e.g. 3,000 to 7,000 markers) would results in comparable selection accuracy to the full 28K HD SNP panel. Imputation of very low density SNPs panels (e.g. 300 – 700 SNPs) resulted in an increase in genomic prediction accuracies compared to the low density panel alone, but this was not observed with imputation from higher densities (e.g. > 5,000 SNPs).

For rainbow trout, the high-density SNP array (57K SNPs) costs approximately 20€ per sample when genotyping a breeding population of 8,000 offspring and 200 parents. This is a total cost of about 164K €, which can be highly prohibitive for most breeding programmes. To implement genomic selection in their breeding programme, our results imply that the use of a cheaper low-density SNP array might be beneficial. For example, under the assumption a 3K SNP panel or array could be typed for all 8,200 fish at a cost of 15€ per sample, this would mean a 25% reduction in genotyping cost for a 3% decrease in accuracy compared to the full HD panel. Another possible scenario is that all offspring were typed for a very low density SNP panel (300 SNPs) at a cost of 7.5€ per sample, with parents typed for the existing 57K array at a higher per sample cost of 40€ per sample (the price is highly dependent of the number of samples genotyped), genotyping would cost 68K€. This reduces the genotyping cost by 59% compared to the price of the 57K SNP for only a 4% decrease in accuracy using the imputation approach. Moreover, *F. columnare* infects small fish, well before they can be individually tagged and identified. Therefore, even for an accurate pedigree-based evaluation, the offspring and parents needs to be genotyped in order to recover the pedigree, meaning that the 300 SNP panel could be combined purpose for both parentage assignment and imputation-based genomic selection.

In conclusion, genomic selection using low-density SNP panels may reduce costs of genomic selection in rainbow trout breeding without major reduction in breeding value prediction accuracy. Using low-density SNP panels (about 3,000 markers) or very low-density SNP panels (about 300 markers) combined with imputation may be a cost-effective genomic prediction of breeding values to improve resistance to *F. columnare*, in this rainbow trout population.

References

Gao, G., Nome, T., Pearse, D.E., Moen, et al., (2018). Front. Genet. 9. https://doi.org/10.3389/fgene.2018.00147

Gilmour, A.R., Gogel, B.J., Cullis, B.R., et al., (2015). ASReml User Guide Release 4.1 Functional Specification. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Houston R.D., Bean T.P., Macqueen D.J., Kumar Gundappa M. et al., (2020). Nat. Rev. Genet. 211,389-409. https://doi.org/10.1038/s41576-020-0227-y

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., et al., (2002) Proc. Of the 7th WCGALP, Montpellier, France.

Palti Y., Gao G., Liu S., Kent M.P., et al., (2015) Mol. Ecol. Resour. 15,662-672. https://doi.org/10.111/1755-0998.12337

Sargolzaei, M., Chesnais, J.P. and Schenkel, F.S. (2014) BMC Genom. 15, 478. https://doi.org/10.1186/1471-2164-15-478

Tsairidou, S., Hamilton, A., Robledo, D., Bron, J.E., Houston, R.D. (2020). G3. 10, 581–590. https://doi.org/10.1534/g3.119.400800