Salmon goes allelotyping: fast search for diagnostic SNPs in Atlantic salmon - SNP-arrays and DNA pooling

OZEROV MIKHAIL Yu.[1], **VASEMÄGI A.**[1], **KENT M.P.**[2], **WENNEVIK V.**[3], **NIEMELÄ E.**[4], **SVENNING M.**[5], **PRUSOV S.V.**[6], **VÄHÄ J.-P.**[1]

[1]University of Turku, Finland; [2]Centre for Integrative Genetics, Norway; [3]Institute of Marine Research, Norway; [4]Finnish Game and Fisheries Research Institute, Finland; [5]Norwegian Institute for Nature Research, Norway; [6]Knipovich Polar Research Institute of Marine Fisheries and Oceanography (PINRO), Russia

Introduction

- Single nucleotide polymorphisms (SNPs) has become marker of choice for many purposes
- High density assays consisting of thousands of SNPs are becoming available in non-model organisms to answer various ecological and evolutionary questions
- The cost of individual genotyping of large number of samples is still relatively high
- DNA pooling - a cost effective alternative to individual genotyping

Material and methods

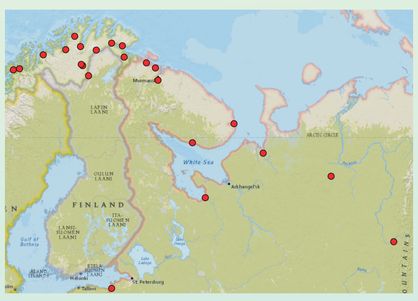


Fig. 1. Studied Atlantic salmon populations

- 23 Atlantic salmon populations (Fig. 1)
- 7K SNP-chip (3928 bi-allelic SNP markers)
- 2-3 technical replicates per pool
- Quality control (QC): call rate, array and pool construction variation of theta, cluster separation, spherical vs. linear filter of theta variation

Results

- 31 loci were excluded due to low call rate of individual genotypes (< 95%)
- 246 loci failed to pass cluster separation QC filter (Fig. 2)

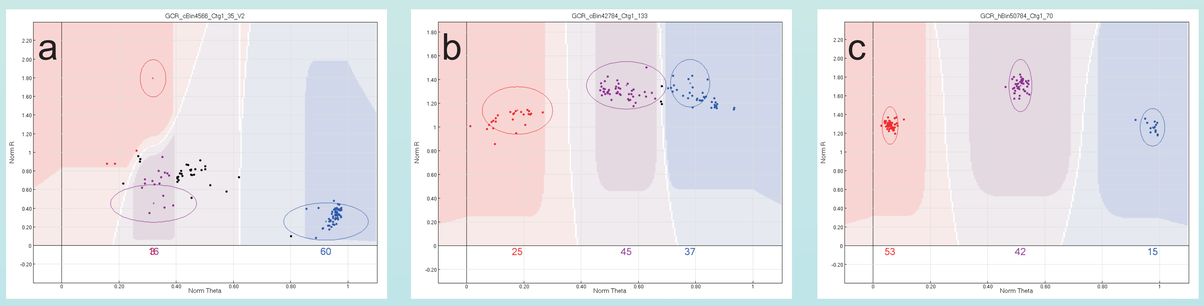
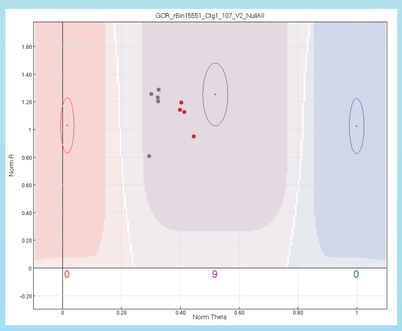


Fig. 2. Example of SNP loci failed to pass QC: a) call rate < 95%; b) cluster separation < 0.40; and c) "good" SNP locus (call rate 100%, cluster separation = 1.00)

- Array variation (i.e. technical replicates of the same pool allelotyped on different chips) per SNP was 20% higher than pool construction variation (i.e. technical replicates of the independently constructed, but identical pools, containing same individuals, allelotyped on the same chip) (Fig. 3, 4)

Fig. 3. Example of variation of theta (transformed normalized intensity of A and B alleles) among technical pool replicates of the same population. Technical replicates of Varzuga pool (n=70) are presented as brown circles, Lakselva pool (n=67) - gray circles



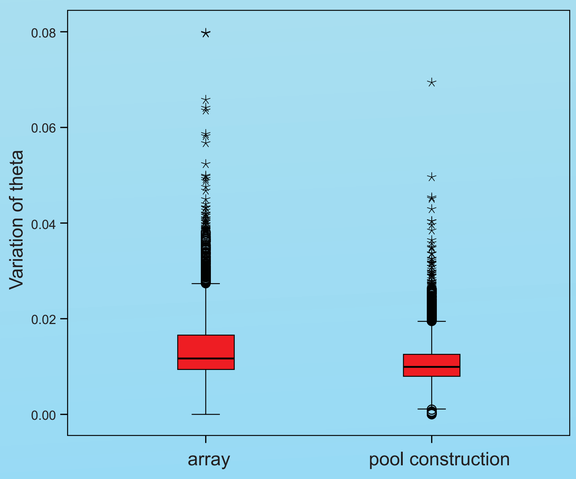


Fig. 4. Box-plot showing estimated array and pool construction variation of theta (Mann-Whitney U-test, $P < 0.0001$). Horizontal line, grey square, whiskers, open circles, and stars indicate median, 25th and 75th quartile, non-outlier range, outliers and extreme outliers, respectively

- Correlation between allele frequencies derived from individual genotyping and DNA pools was very high ($r = 0.991 - 0.992$) (Fig. 5)
- Mean error between true and estimated allele frequencies for all 3 pool sizes was small and similar (median error 0.023 - 0.025)
- Accurate allele frequency estimates in small DNA pools (Fig. 5)

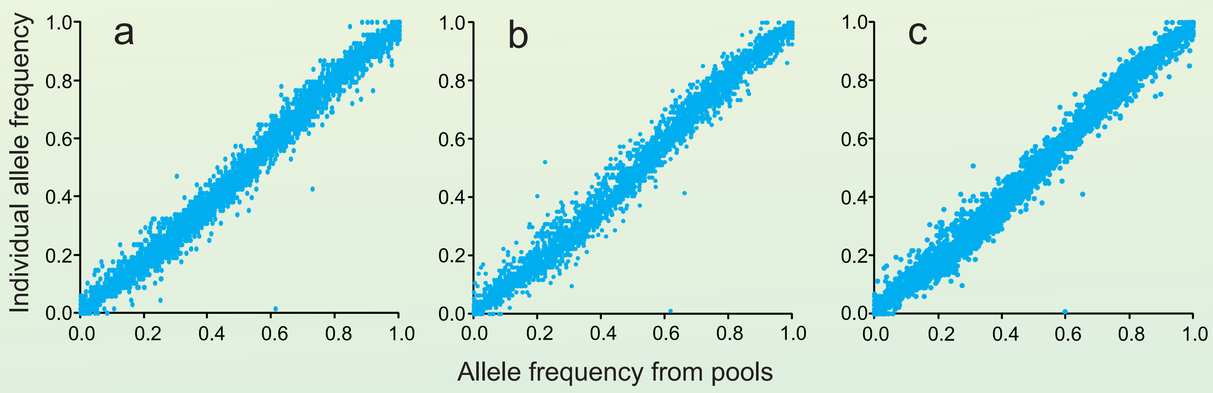


Fig. 5. Scatter plot of estimated allele frequencies from individual genotyping *vs.* pooled DNA. True allele frequencies from individual genotyping for Kola population were compared with estimated allele frequencies for three different pool sizes: a) 35 ($r = 0.992$), b) 50 ($r = 0.991$) and c) 70 ($r = 0.992$)

- Linear QC filter increases the proportion of non-informative loci
- Spherical QC filter retains larger proportion of polymorphic loci (Fig. 6)

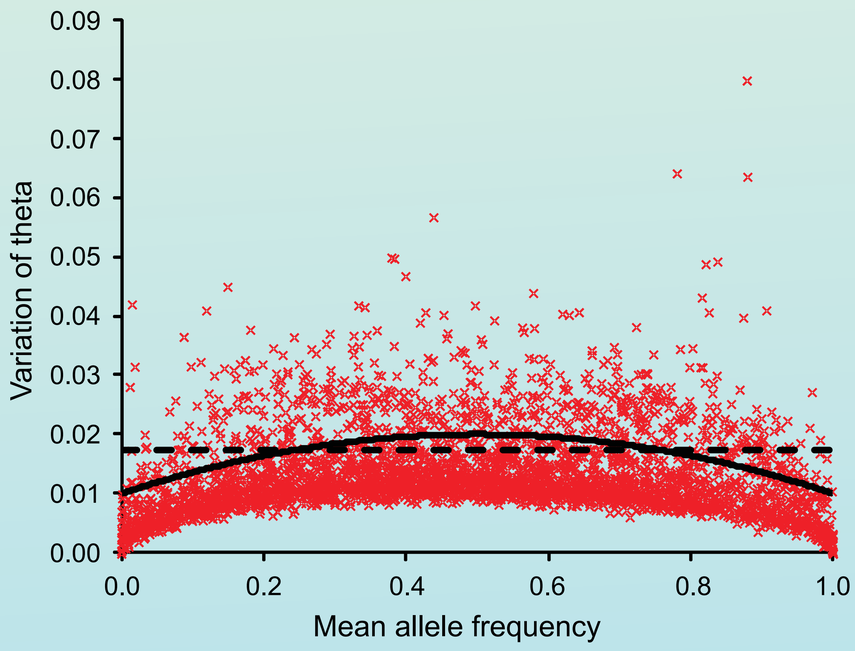


Fig. 6. A plot of mean estimated allele frequencies across 14 populations against array variation of theta. Solid and dashed lines indicate the boundaries of spherical and linear cut-offs, respectively

- Accurate and fast screening of allele frequency variation between populations in large number of SNPs allows to identify about 250 population-informative markers ($F_{ST} > 0.15$)
- Identified SNPs allowed >98% correct assignments of salmon individuals to population of origin in simulation studies

Conclusions

- DNA pooling provides a reliable, efficient and cost-effective means for obtaining genome-wide allele frequency estimates for multiple populations of Atlantic salmon
- DNA pool sizes of > 35 individuals are enough to provide relatively accurate allele frequency estimates
- QC filter based on spherical cut-off enables to exclude loci with relatively high error rate compared to the information content (minor allele frequency close to 0)
- DNA pooling enables cost-efficient identification of diagnostic markers e.g. for: a)detection of various ecologically and economically important traits (case-control studies); b) discrimination of salmon populations at various geographical scales; c)parentage determination etc.

$ Practically, the processing of 20 Atlantic salmon populations with 70 specimens in each will turn into 1400 SNP assays for individual genotyping or 60 assays for DNA pools (i.e. 20 pools x 3 technical replicates per pool). Therefore, in this case, an application of DNA pooling approach allows to reduce the costs in ~ 23 times

This project is funded by the European Union, Finnish Academy of Sciences, Norwegian Directorate of Nature Management and Norwegian Research Council