UNIVERSITÀ DEGLI STUDI DI SASSARI

SCUOLA DI DOTTORATO DI RICERCA
Scienze e Biotecnologie
dei Sistemi Agrari e Forestali
e delle Produzioni Alimentari

Indirizzo Scienze e Tecnologie Zootecniche

Ciclo XXVIII

Use of genetic, genomic and phenomic approaches to improve livestock performances

dr. Antonio Francesco Puledda

Direttore della Scuola	prof. Antonello Cannas
Referente di Indirizzo	prof. Gianni Battacone
Docente Guida	prof. Nicolò P.P. Macciotta
Tutor	dr. Giustino Gaspa

Anno Accademico 2014 – 2015

Alla mia Famiglia

Ringraziamenti

Desidero ringraziare tutti coloro che hanno reso possibile la realizzazione di questa tesi, professori, ricercatori e colleghi della scuola di dottorato di ricerca della Sezione di Scienze Zootecniche nonché il personale della stazione sperimentale Ifremer di Palavas les Flots.

Acknowledgements

I would like to acknowledge all people who contributed to this PhD thesis, professors, researches and colleagues of the Section of Animal Science as well as all people I met at the Ifremer experimental station of Palavas les Flots.

Index

CHAPTER 2

*Estimates of heritability and genetic correlation*s *for milk coagulation properties and individual laboratory cheese yield in Sarda ewes*.

CHAPTER 3

Effect of different Data Reduction on the Accuracy of Genomic Prediction on an Imputed High Density SNP panel in Italian Holstein cattle.

CHAPTER 4

Echography and 2D morphological predictors for indirect selective breeding of fillet and carcass percentage in European sea bas (Dicentrarchus labrax).

CHAPTER 5

Early landmark-based sexing of European Sea bass (Dicentrarchus labrax) using image analysis

Abstract of the thesis

Traditional methods to improve animal productions are derived from the quantitative genetics theory. The estimation of variance components, genetic parameters and genetic gains are essential steps of the selection process based on quantitative methods. The use of linear mixed models, algorithms for their solutions and computer to solve complex system of equations has made possible to estimate reliable genetic merits also from complex structures of relationships. More recently the high throughput technologies has provided a great increase of information both at genomic than phenomic level. This high informativeness has led two main consequences: improved accuracies and faster genetic gains from one hand, big data handling and unbalancing from the other hand. The present thesis will deal with these aspects. Two experimental contributions will focus on the estimation of genetic parameters for new criteria of selection on species of particular interest, with traditional and innovative phenotyping methods. A third contribution will deal with the research of new morphometric indicators that could have an economic impact if applied for the selection of the species of interest. Finally, a fourth contribution will address the issue of the data unbalancing derived from the current information gap generated by the speed of genotyping and the capacity of the phenotyping practices. All these contributions will have the same aim, the improvement of livestock performances.

CHAPTER 1

General Introduction

Animal breeding

Animal breeding is aimed at improving the traits of economic interest in livestock. The genetic merit of animals is estimated on the basis of phenotypes and the relationships among individuals. Objective of the selection is to identify the candidates with the highest genetic merit and to transmit this improvement to their offspring through programmed matings. The genetic merit is called "*Breeding Value*" (BV) and represents the most important information achievable for the improvement of animal productions. By choosing the highest BVs, the possibility to fix the genetic improvement in the following generations is maximised. The measurement of the selection response is the way to estimate the genetic progress. A selection on specific alleles and an alteration in the gene frequencies of the population is obtained by mating animals on the basis of their BVs, resulting in a shift in the population mean. The speed in getting the genetic goals rely on the nature of the trait under selection and, in particular, from its heritability. The heritability is the genetic parameter that measure the proportion of the trait variance which is genetically determined and, hence the ability of the trait to be transmitted through the generations. The heritability, together with the knowledge of the relatedness among individuals is the basic component of any breeding programme. Indeed, the knowledge of the degree of resemblance between the individuals provides, through the decomposition of the phenotypic variances and covariances, the genetic merit of an individual, its Breeding Value (Falconer and Mackay, 1996).

The opportune combination of phenotypic and pedigree information into a selection index (Hazel and Lush 1942; Hazel, 1943) allows the evaluation of the individual genetic merit. In practice the index assumed the form of a simple or multiple regression of the breeding value on multiple sources of information. In the simplest case of a single trait, the index is obtained by multiplying the individual phenotypic deviation from the population mean by the heritability of the trait (h^2) that represents the coefficient of regression of the breeding value on the phenotype.

$$I = h^2 P$$

The BV estimation obtained by the simultaneous consideration of multiple sources of information is more accurate. In this case, the index is the result of a multiple regression and the best combination of partial regression coefficients is found by minimizing the sum of squares of the regression of the index on the breeding value, in other words, by maximizing their correlation. The form of the index in the case of multiple sources of information is:

$$I = b_1 P_1 + \ldots + b_n P_n$$

Several environmental factors are able to affect the estimation of breeding values. The main problem is the absence of the *a priori* knowledge of the magnitude of these effects. Knowing these factors, observations can be adjusted and estimates properly corrected. In practice this knowledge was not available in advance and it was the reason that hampered genetic evaluation before the introduction of statistical procedures able to estimate fixed and random effects simultaneously. This was done thanks to the main contribution of Charles Roy Henderson that developed the methodology called Best Linear Unbiased Prediction (BLUP). BLUP assumes the knowledge of genetic and residual variances that are normally assessed by different kind of iterative methods.

The Best Linear Unbiased Prediction

The statistical procedure introduced by Henderson, known as Best Linear Unbiased Prediction (BLUP) is worldwide used to estimate breeding values through models including both fixed and random effects simultaneously. In matrix notation, the general form of a mixed model is as following:

$$y = X\beta + Za + e$$

where y is a vector of *n by 1* observable variables used in the prediction, β is a fixed, unknown vector of length p, a is the random vector including breeding values with q elements, null means and variance $V(a) = A\sigma_a^2 = \boldsymbol{G}$. Moreover, X and Z are known matrices composed by 0 and 1, relating records to the relative fixed and random effects, respectively. Finally e is the random vector of errors with n elements, null

means and variance V(e) = $I\sigma_e^2$ = $\boldsymbol{R}$. Assuming the absence of covariance between *a* and *e*, Cov(*a*, *e*) = Cov(*e*, *a*) = 0

$$V\begin{bmatrix} a \\ e \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}$$

and the variance of y

$$V(y) = ZGZ' + R$$

the *estimation* of β (BLU*E*) and the *prediction* of *a* (BLU*P*) can be obtained from:

$$\hat{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} y$$

$$\hat{a} = G Z' V^{-1} (y - X \hat{\beta})$$

where $G Z' V^{-1}$ is the covariance between *a* and *y* ($cov_{a,y}$) divided by *V* (var_y), that is the regression coefficient of *y* on *a,* with *y* corrected for the fixed effects. The solution of $\hat{a}$ and $\hat{\beta}$ require the inversion of *V*, not always computationally feasible. To avoid this problem, Henderson proposed the following equation:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

and factoring out the terms R^{-1} from both sides of the equation and transforming G = $A^{-1}\lambda$, the equation becomes:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

where *A* is the Numerator Relationship Matrix (NRM) reporting the additive genetic relationships among individuals and λ is the ratio σ_e^2/σ_a^2. This is the canonical form of the Mixed Model Equation (MME). Properties of BLUP are:

- the mean square error $E(\hat{a}_j - a_j)^2$ is minimized (the Best),
- $\hat{a}$ is a linear function of $\boldsymbol{y}$ (Linear),
- $\hat{a}_j$ is an unbiased estimator of a_j (Unbiased).

The main use of the mixed models is the estimation of genetic additive effects ($\boldsymbol{a}$) by adjusting the data for the known fixed effects. The accuracy of these predictions is

improved by considering relationship between all evaluated animals. However, when the number of animals is large, the computation of *A* and its inversions are not feasible through conventional routines. Henderson (1976) proposed a rapid method to compute the inverse of NRM bypassing the computation of *A*. Since its introduction, BLUP has been developed and extended to obtain a variety of genetic evaluation in animal breeding, especially for dairy animals (Quaas and Pollack, 1980).

BLUP was for the first time implemented in the genetic evaluation of bulls through the Sire Model. This kind of model made possible the estimation of genetic merit of candidates using the comparison of the performance of contemporary progeny groups. The evaluation of dams was considered secondary and obtained using the sire merit. With the expansion of the artificial insemination (AI), most of the hypothesis on which the earlier models were based on, became invalid. In particular, the consequence of the intense use of the best bulls for AI, was the creation of genetic subpopulations in the same breed, thus the assumption of sampling from a homogeneous population was not longer acceptable. The consideration of the existence of subpopulations in the estimation of the breeding value started in the 1972 by the introduction of the North East Artificial Insemination Sire Comparison model (NEAISC) with the aim to account for the genetic trends deriving from these subpopulations. Despite the advantage in terms of accuracy by using the relationship among sires, the use genealogical information was not possible for technical reason until the discovery of the Henderson's rules (1976) for the direct inversion of the numerator relationship matrix (NRM).

With the use of genealogical information, the genetic merit started to be corrected by the genetic level of females. In 1979 the Sire-Maternal Grand Sire model was introduced to account for the genetic level of the daughters of the maternal grand sire. This model was used until 1989 when the progress in computer machinery made feasible the advent of the Animal model (AM). The concept of the individual AM was announced by Henderson in the 1949 but this name was firstly used in the paper of Quaas and Pollack (1980). Due to the excessive computational requirements needed to obtain routine evaluations, a number of computational simplifications have been proposed to reduce the number of equations derived from full AM. An example

was the Reduced Animal Model (RAM) proposed by Quaas and Pollack (1980) and implemented by Henderson (1985), Westell (1984, 1988) and Van Vleck (1990). With these methods, the only breeding values included in the equations are of animals having one or more tested progeny. Through a back approach, appropriate linear functions, derived from the vector of solutions were used to evaluate related individuals with no progeny. When the RAM models were not solvable due to computational reason, other procedures such as the Gauss-Seidel iterative methods and the ordination of animals by traits were used to obtain single and multiple traits evaluations, respectively. The computational adequacy to solve the great number of equations generated by AM was reached only in 1989. Since that moment, the AM became the reference methods for animal selection. AM allows the contemporary evaluation of all individuals included in the relationship matrix by using the own and all available performances from known relatives (ascendants and descendants), enhancing the accuracy of predictions, overcoming the problem of the comparisons between animals coming from different environment and periods and having different sources of parental and phenotypic information.

Early dairy breed evaluation systems had been based on lactation model describing milk, fat and protein produced in standardized lactation in term of genetic and environmental variation without considering factors affecting productions in different lactation stages. In other words, the intra-lactation variance was assumed to be absorbed by the residual variance. This limitation was overcame when the computational power allowed to use daily individual performances recorded during the entire lactation and the implementation of the test day (TD) models (Stanton et al., 1992). The TD represents the amount of milk/fat/protein produced and measured on a specific day during the lactation. The number of tests depends on the length of lactation and the frequency of measurements. The frequency of testing varies according to different recording plans and it is mainly imposed by economic reasons. The advantage of using TD model is the possibility to correct observations for environmental effects specifically affecting the records during the lactation. Different mathematical functions have been proposed by many authors to draw the lactation curves. On the basis of these functions, different models have been applied to calculate genetic merit of the sires. Among these, Random Regression Models

(RRM) assumes that the shape of a lactation curve is different for each cow according to individual values of coefficients and giving the possibility to change milk recording programs in accordance with the expected level of accuracy.

The Variance Component Estimation

The total variance of the phenotypic value (σ_P^2) can be subdivided in the following components:

$$\sigma_P^2 = \sigma_a^2 + \sigma_d^2 + \sigma_i^2 + \sigma_e^2$$

where subscripts *a*, *d* and *i* refers to the additive (average effects of individual alleles), dominance (interaction between alleles at each locus) and epistatic (interaction between genotypes at different loci) effects in which the total genetic variance is decomposed (Walsh B., 2001). Subscript *e* refers to the environmental source of variance, an important source of variability able to reduce the precision of the genetic estimations (Falconer and Mackay, 1996). Variance components estimation for standard designs (balanced and with random sampling of parents) can be obtained by using least square analysis (analysis of variance and regression). In more realistic cases, when more complicated population structures, several generations or selected groups are considered, other statistical tools as the maximum likelihood (ML) procedures, are used. Based on iterative algorithms, these procedures allow to exploit all the available information deriving from any structure of genetic relationships. By assuming some starting values (the priors) the procedure search, through iterations, values of variance at which the likelihood is maximized. Common fixed effect such as season, year and herd need to be taken into account because they could determine an underestimation of the variance components. A modification of the ML, known as restricted maximum likelihood (REML), is able to adjust observations by weighting all possible fixed effects. Although these procedures require more computational power than simple ML, they become the methods of choice for the estimation of genetic parameters and are currently widespread in the most common software packages for animal breeding.

Genetic parameters: heritability and genetic correlations.

Heritability is the population parameter that measure the genetic proportion of the variation existing among individuals. It is formally defined as the quota of the phenotypic variability genetically determined and potentially transmittable to the progeny. In a broad sense, heritability derives from the ratio between the genetic (σ_G^2) and the phenotypic (σ_P^2) varince. However, being dominance and epistatic effects (*d* and *i*, respectively) not responding to selection and assuming an absence of genotype by environment interaction (G x E = 0), heritability is more practically referred to the ratio between the heritable additive genetic variance (σ_a^2) and the phenotypic variance (σ_p^2) and given by:

$$h^2 = \frac{\sigma_a^2}{\sigma_p^2}$$

In this form heritability assumes a narrow sense and accounts for the differences and resemblances among individuals due to their additive genetic effects. Heritability is one of the main important biological concepts in animal breeding having a central role in the prediction of the response to selection. In breeding programmes, the heritability suggests the efficiency of the prediction of the breeding value from phenotypic information allowing comparisons of the same trait on different populations or different traits on the same population. Heritability is a dimensionless value that ranges from 0 to 1, where 0 means the same average effect of allele on the character (variation is completely due to environmental reason) and 1, indicating that all variation is due to the average effect of alleles. The estimation of heritability relies on the use of phenotypes collected on a huge number of individuals. When phenotypes derive from population with a mixture of relationships, multiple generations or unbalanced designs, estimates of genetic parameters are obtained through the use of complex iterative methods based on linear mixed models (Visscher et al., 2008). The development of mixed models improved the reliability of estimates by keeping into account all existing relationship between relatives. Lynch and Walsh (1998) and Thompson, (2005) reviewed methods employed for the genetic parameter estimation occurred in the last half century.

The knowledge of the relationships existing between traits is of great interest since the improvement of one character by selection could cause simultaneous changes in other correlated characters. The correlation between traits derives from genetic (pleiotropy and linkage) and environmental causes. Pleiotropy is the property of a gene to affect more characters simultaneously and its influence can affect correlated traits either positively or negatively. Moreover, genes closely located on the chromosomes are more likely to segregate together and cause interactions (correlations) on the expression of more than one characteristic. The environment operates the same influence on all characters simultaneously. The phenotypic correlations between traits can be dissected into the genetic and environmental causes (Falconer and Mackay, 1996):

$$r_{P(XY)} = h_X h_Y r_{A(XY)} + e_X e_Y r_{E(XY)}$$

where r_A and r_E are the genetic and environmental correlations, h_X and h_Y are the square root of the heritability of both traits, e_X and e_y are the square root of the environmental portion of the phenotypic variance $\sqrt{(1 - h^2)}$. The terms $h_X h_Y r_{A(XY)}$ is also known as the *coheritability* (Falconer and Mackay, 1996) and it allows the estimation of the correlated response, a particular kind of selection based on the use of correlated traits. The estimation of the genetic correlation (r_A) is based on the use of variance and covariance between traits:

$$r_A = \frac{cov_{XY}}{\sqrt{(var_X \, var_Y)}} = \frac{cov_{XY}}{\sigma_X \, \sigma_Y}$$

or considering the offspring–parent relationship through the cross-variance estimation:

$$r_A = \frac{cov_{XY}}{\sqrt{(cov_{XX} \, cov_{YY})}}$$

where cov_{XY} is the covariance between traits X in the parent and Y in the offspring (or *vice versa*), cov_{XX} and cov_{YY} are the between-relatives (parent-offspring) covariances of trait X and Y, respectively. If both cross-variances are available (trait X in parent and Y in offspring and X in offspring and Y in parent), the r_A is the mean of the two estimations (Falconer and Mackay, 1996).

More accurate estimates of variance and covariance between traits are obtained by using mixed linear model based on single and multi-traits evaluations rather than the parent-offspring models. In a series of studies (Kruuk, 2004; Åkesson et al. 2008), genetic parameters obtained from parent-offspring regression are compared with those obtained through animal models. In particular more accurate values (lower standard error) of genetic parameters were obtained when using animal models, suggesting a lesser sensitivity of these methods against possible sources of bias such as selection of non-additive covariances between relatives. On the basis of these results the authors suggested particular attentions when comparing genetic parameters arising from different methods of estimation.

The response to selection

The *response to selection* (R) is the change of the population mean of a phenotypic value from the population from which parents are chosen and the offspring of the selected parents. The difference between the mean of the phenotypic value of the selected parents and the whole population before their selection, is the *selection differential* (S). Interest of the breeders is to know what improvement in a character can be obtained in the offspring by selecting some parents instead of others. Knowing the selection differential, the prediction of the response can be calculated considering the heritability of the character by the following equation, also known as "*Breeder's equation*" (Lush, 1937):

$$R = h^2 S$$

Assuming that the information about the mean value of the character in the offspring is already unknown, the response to selection depends on the heritability of the trait estimated on the parental generation. Due to the change in gene frequencies and the reduction of variability obtained through selection, heritability, selection differential and consequently response to selection have to be updated after few generations (Falconer and Mackay, 1996). Moreover, assuming that selection is performed by truncation (all individuals beyond the truncation point are selected as parents) and

that character of interest follow a normal distribution, the selection differential should be expressed in terms of standard deviation and symbolised by:

$$S = i\,\sigma_P$$

where *i* is the *intensity of selection* that is the degree of the pressure of the selection applied depending on the proportion of the population selected as parents and σ_P is the standard deviation of the trait. The intensity of selection is given by:

$$i = S/\sigma_P$$

and suggests how many times, in standard deviations unit, the mean of the selected group exceeds the mean of the whole population. If the trait under selection follows the normal distribution, the value of the intensity of selection can be easily determined. The standardized selection differential allows to compare the responses obtained for traits having different units of measures. By replacing the selection differential with the intensity of selection the equation of response become:

$$R = i\,h^2\sigma_P$$

that is the classical form of the response to selection. Different strategy of selection can be adopted to get a response. Individual or *Mass Selection* (MS), *Family* (sib selection and progeny testing) and *Within-Family* (WF) selection, are different examples, each one adopted under specific circumstances.

For each type of selection, the individual phenotypic value can be thought as the sum of two different components: the deviation of the family from the population mean and the within-family deviation of the individual from its family mean. Depending on the kind of selection applied, different weights are given to these two forms of deviation. In particular, in the case of the individual selection, neither deviations are considered, being the selection based on the deviation of the individual value from the population mean. In the case of family or within-family selection, one of the two sources of deviation is set to 0. However, the maximum rate of response is obtained by applying the selection knows as *combined selection*. This kind of method uses the family and within-family heritabilities as regression coefficients and gives at the same time, the right weight to each component of the above mentioned forms of

phenotypic deviation. For each type of selection the form of the expected response varies taking into account the provenience of the sources of variation as following:

$$R_I = i\, h_i^2 \sigma_i$$

$$R_{FS} = i\, h_f^2 \sigma_f$$

$$R_{WF} = i\, h_w^2 \sigma_w$$

where subscripts *i*, *f* and *w* refer to the individual, family and within-family selection. To allow comparison between expected responses obtained between individual and family, within-family and combined selection, different comparison factors are conveniently inserted in the expected response equations that according to Falconer and Mackay (1996), becomes:

$$R_{FS} = i\, h^2 \sigma_P \frac{1 + (n-1)r}{\sqrt{n\,\{\,1 + (n-1)t}}$$

$$R_{WF} = i\, h^2 \sigma_P\, (1 - r) \sqrt{\left[\frac{n-1}{n(1-t)}\right]}$$

$$R_C = i\, h^2 \sigma_P \sqrt{\left[1 + \frac{(r-t)^2}{(1-t)} \times \frac{(n-1)}{1 + (n-1)t}\right]}$$

where *n* is the number of individuals in the family (family size), *r* is the coefficient of relationship between relatives (that changes considering the method of selection) and *t* is the *phenotypic intraclass correlation.*

Two different characters can show a correlation if traits are related for some reasons. The sum of genetic and environmental influences can produce a covariance between related traits that can be directly observable through the phenotypes of a number of individuals in the population. As the same for the variance of one trait, the covariance between two traits can be divided in these two components, genetic (additive genetic) and environmental (non additive genetic). The partitioning on these two components should be treated in terms of correlations, being the correlation a ratio of the covariance with the standard deviation of the two traits. The genetic and environmental correlation combine together to give the phenotypic

correlation. Once obtained the observational components for each trait by using dedicated software, selection of animals can be performed indirectly by using the correlation between characters. This kind of selection is named *Indirect Selection* (IS) and is based on the use of secondary characters, correlated with the trait that is desired to improve, obtainable by applying the Correlated Response:

$$CR_Y = i\ h_X h_Y r_A \sigma_{PY}$$

where the *coheritability* ($h_X h_Y r_A$) is simply multiplied, in the same manner of the heritability under the individual selection, by the intensity of selection and by the phenotypic standard deviation of the trait of interest, that is directly observable from individuals in the population. Indirect selection is normally preferable when the heritability of the character of interest is lower of the secondary one and when the correlation between traits is relatively high. Indirect selection is preferable to direct selection when it is difficult or impossible to measure the direct character on individuals. For example, when the measurement of a character is affected by not negligible errors, heritability is reduced by the predominance of the non additive and environmental components and the direct response is also reduced. In this case, the use of a related trait easier to measure could result in the indirect improvement of the character of interest.

Indirect selection is also preferred instead of direct selection when the second method results costly or even lethal for individual under selection. In these cases, the choice of characters whose measurements are cheaper or that avoid the sacrifice of individuals could be more ethical, economic and time savings. The indirect selection applied directly on candidate by using indirect criteria, avoiding the use of sibs in replacement of candidates, being the response not affected by the relationship between individuals, could result in faster genetic gain.

So, the response to selection can be improved by increasing the heritability of a character through more accurate rearing or measuring techniques, by the increase of the intensity of selection or by the reduction of the generation interval. The response of selection that take into account of the generation interval become

$$dR = \frac{i\, h^2 \sigma_P}{T}$$

where T is the generation interval, or equivalently the time that elapses between the same event in two successive generations, as for example the interval occurring between the mating of parents and the mating of their offspring (Falconer and Mackay, 1996). By dividing the equation of the response to selection by the generation interval, the response is expressed in terms of speed of genetic gain or more specifically in gain per unit of time.

From quantitative genetics to genomic selection

Minimizing the denominator of the previous formula, the response to selection tends to be maximized for the considered unit of time. There is an inverse relationship between the generation interval and the speed of the genetic gain, since the shorter the time between two following generations, the greater the response to selection achievable within the interval. If considered in terms of accuracy, the above form of response can be rewritten as:

$$dR = \frac{i\, r_{IA}\, \sigma_A}{T}$$

where r_{IA} is the accuracy of the breeding value expressed as correlation between the selection index and real genetic merit ($r_{IA} = Cov_{IA}/\sigma_I \sigma_A$). In this form, the genetic gain is directly proportional to the intensity of selection, to the accuracy of the breeding value and to the genetic standard deviation of the character under selection, remaining inversely related to the generation interval. The accuracy of the breeding value depends on the number and kind of information available. From this point of view, a longer generation interval allows to obtain a greater number of information from which to estimate more accurate breeding values but at the same time, the increase of the generation interval tends to reduce the response. The optimal combination of these three entities (i, r_{IA}, T) is at the basis of a selection program. Since 1980, the advancements in DNA technologies made it possible to obtain

information at genotype level for different livestock species. More recently, with the substantial reduction of the costs of the genotyping, the availability of genomic information increased dramatically. The recent availability of High Throughput Technologies (HTT) and of dense molecular marker maps, composed by many thousands of Single Nucleotide Polymorphisms (SNPs), made feasible since 2001, the development of the genomic selection technology (Meuwissen et al., 2001). Since the initial simulation studies to the current real practice, this new approach to the animal selection, showed the ability to nearly double the genetic responses simultaneously reducing the generation interval and improving the breeding values accuracy, if compared with the classic selection tools.

Genomic Selection (GS) relies on the estimation of direct genomic breeding values (DGV) of animals by assessing the effect of the total number of segment in which the genome is divided into. Molecular Markers are the boundaries of these segments and their number determines the density of the marker maps. Generally, genomic selection is a two-step procedure (Hayes et al., 2009a): at first, marker effect estimation is performed on a reference population with known genotypes and phenotypes and later, the effects previously estimated are used for the prediction of DGV of young animals without phenotypes. GS has made the selection process faster and cheaper than traditional selection schemes, particularly for low heritability (behaviour), sex-related (milk production), lethal (carcass quality and meat content) or costly to measure (feed efficiency) traits. Animals are chosen on the basis of their genomic merit and genomic profiles are rapidly available early in the life resulting in a substantial reduction of the generation interval and improvement of the genetic response.

The accuracy of the genomic predictions, measured by the correlation between the DGV and phenotypes available, is also increased and generally intermediate between that obtained with the pedigree index (evenly characterised by absence of phenotypes but using the average genetic merit of parents) and progeny testing, that is the most accurate but expensive system of assessment of genetic merit. The accuracy of GS is dependent from multiple causes such as size and structure of the reference population, number of markers, the statistical model and different genetic aspects related to the evaluated trait. Among the others, the aspect that mainly affects the

DGV estimations is the great disproportion between the number of animals having phenotypic records (*n*) and the number of markers used as predictors (*p*), the so called $n << p$ problem. Apart from automatic data recording systems, generally the phenotypes are commonly obtained from field recording. Thus, expensive to measure traits may limit the total number of animals that can be included in the reference population. Conversely, the unitary cost of the single markers is progressively decreasing. To address this issue different strategies have been undertaken. One possibility is the reduction of the number of genetic markers by keeping constant the quantity of individuals. Variable reduction was obtained through the pre-selection of molecular markers with different regression procedures (Solberg et al., 2009; Moser et al., 2010; Dimauro et al., 2013), by ranking and choosing markers on the basis of their effect (Weigel et al., 2009) or statistical association (Schulze et al., 2004, Pattaro et al., 2008) with the trait of interest or by substituting molecular markers with a reduced number of predictors obtained applying multivariate statistical techniques as the Principal Component Analysis (Solberg et al., 2009; Macciotta et al., 2010; Moser et al. 2010). Another way is to increase the number of observations by sharing phenotypes and genotypes of animals belonging to the same breed (Lund et al., 2011; Jorjani et al., 2012) or using mixed reference population by *multi-breed approach* (Hayes et al., 2009b). Despite the benefits of the enlargement of the reference population, this approach showed side effects derived from the genetic structures of the admixed population (Pryce et al., 2011). Alternatively, the increase of the number of animals can be achieved by reducing the costs associated with genotyping, via *imputation* (VanRaden et al. 2013). The imputation technique is a *in silico* reconstruction of missing genotypes, allowing to obtain with reliable accuracies, higher marker densities starting from cheapest low-density panels (3–7K). In this manner a larger number of animals can be genotyped and dataset with different marker densities uniformed and compared. Nowadays, new technologies are revolutionizing the field of the data collection through the use of phenotyping procedure able to reduce the costs of data recording. The introduction of automated systems is allowing the real time phenotyping of a larger number of animals. The next two paragraphs will deal with some of the high throughput technologies that are revolutionizing the animal breeding.

The large number of SNPs variants discovered by genome sequencing of different livestock species, offers today the opportunity to use these genetic differences for implementing studies at different genetic scales. Different technologies in the field of the SNP genotyping have been developed until now, ranging from single to multiplexed assays, and most of these are commercially available. Ragoussis (2006) reviewed the main technologies, describing pros and cons for each of them. Large scale genomic studies are based on the analysis of the whole genome of many individuals, then systems capable to detect simultaneously tens or hundreds of thousands of genotypes for each animal, on thousands of individuals, are needed. Nowadays the whole genome genotyping technology (WGG) proposed by Illumina, is one of the most common high-throughput technologies able to investigate DNA at the genomic level. WGG is well appreciated because it is able to reduce working time, operation errors and unitary cost per single SNP analyzed (Shen et al., 2005).

High-throughput phenotyping technologies

The success of the selection for quantitative traits is largely dependent on the availability of reliable phenotypes. However the phenotyping process has constituted for many years a *bottleneck* for genetic analysis and genomic predictions (Furbank et al., 2011; Cabrera-Bosquet et al., 2012; Rahaman et al., 2015) particularly in the last decades, with the advancements obtained in the molecular genomics framework. High dimensional phenotypic data are needed both to discover genomic variants affecting more than one phenotype at the same time and to explain their pleiotropic effects. Moreover, the large amount of phenotypic information nowadays achievable by the implementation of high-throughput phenotyping platforms (HTPP) associated with the improvement of the accuracies of the traits measured, results in better heritability estimates (Houle et al., 2010), and improved breeding outcomes. HTPPs are offering the opportunity to measure a great amount of detailed traits with real-time systems, then providing the solution to overcome the limitation imposed by the number of records in the genomic selection.

The first phenotyping systems have been developed in the field of plant breeding where many thousands of individuals are often involved in breeding programs

(Cabrera-Bosquet et al., 2012). Phenotyping systems make the process of trait measurement faster and cost-effective, increasing the precision of the phenotypic measurements at all biological level and avoiding, when possible, the destructivity associated with traditional phenotyping procedures. The image analysis technologies are very popular in plant breeding. Rahaman et al. (2015) described the imaging devices for high throughput plant phenotyping based on the interaction between organisms and light. Visible light on 2D and 3D dimension, fluorescence and spectroscopy imaging can provide information regarding growth and biomass (morphological), nutrition (physiological) and health status for disease detection or stress diagnosing. More recently functional imaging based on PET (Positron emission tomography), MRI (magnetic resonance imaging) and FRET (Förster resonance energy transfer) were integrated to screen at molecular level, the functional and structural plant dynamics. After image acquisition, raw data are processed and once integrated with molecular and environmental metadata, statistical models are performed to describe factors affecting development and to predict performances. Despite the well consolidated availability of dedicated high-throughput phenotyping facilities, including image processing and analysis software in plant breeding (Lobet et al., 2013), phenomics represents a relatively new approach of trait measurement in farmed animals, especially if compared with huge progresses made in animal genomics.

Livestock phenomics represents the new trend of the individual animal measuring. It relies on the real-time high-dimensional phenotypic data acquisition, aimed at obtaining an improvement of animal performances through the enhancement of accuracy and reproducibility of conventional traits or the introduction of novel traits of economic interest, otherwise unachievable. All phenotypes are measurable from the whole organism to the sub-cellular components by implementing adequate measuring systems (Greenwood et al., 2016). The importance of an universal recognition of economic relevant traits and of the methods for their measurements has led to the development of a worldwide uniformed reference ontology for phenotypic traits of farm animals, the "*Animal Trait Ontology of Livestock*" (ATOL) (Hocquette et al., 2012), and the "*Environment Ontology for Livestock"* (EOL) for

the standardization of breeding conditions which may influence the expression of these phenotypes.

In contrast with the plant breeding, where organism are motionless, animal phenotyping often requires tracking or alternatively the introduction of technologies that do not prevent movements, particularly for animals fed at pasture. The real-time monitoring of the behaviour of ruminants for example, is important to make effective decisions about pasture management and feed supplementation, allowing the optimization of animal performances and enabling a more rational environmental management and improving economic outcomes. Until now, several techniques have been used to improve the management of pasture by discriminating grazing and ruminating behaviours in different species (Andriamasinoro et al., 2016). Sensors of movement such as accelerometers, magnetometers, gyroscopes coupled with audio-visual recording devices and satellite tracking systems are suitable to obtain real-time information about feeding behaviour, directly from field. Systems for recording the behaviour of grazing animals at pastures have been recently implemented on cattle (Bishop-Hurley et al. 2014; Williams et al., 2014; Dutta et al., 2015;) and on dairy sheep (Sneddon et al., 2014; Alvarenga et al., 2016; Decandia et al., 2016). The knowledge of the pasture location and timing of grazing can be fundamental to understand, and eventually to model, the changing on milk yield and composition since these changes are able to influence derived products. When animals are fed at pasture in fact, additional aspect such as phenological stage and the specific contribution (SC) of the main botanical families grazed, other than animal-related variables, are able to influence milk properties (Gorlier et al., 2012). Traditional methods of mapping vegetations directly on the field are generally time consuming, expensive and often not practical, especially when referring to large extensions. The introduction of *hyperspectral sensors* in agriculture is increasing the amount of information achievable by analysing the entire electromagnetic spectrum, beyond the visible light, both on large spatial and laboratory scale.

Estimates of pasture production (biomass and growth rate) based on *Remote Sensing Imagery* is allowing for example to measure the *foliage greenness* from satellite images. An example is the "*Pastures from Space*" program, developed in Australia by CSIRO Livestock Industries in collaboration with the Department of Agriculture

of Western Australia and the Western Australian Land Information Authority (Landgate) that is allowing farmers to make decision regarding an efficient use of feed resources. This technology is based on the normalized difference vegetation index (NDVI), through which seasonal changes in leaf areas are quantified by analyzing the red and near-infrared (NIR) solar radiation reflectance through satellite image processing (see: http://www.pasturesfromspace.csiro.au/).

At laboratory level, vibrational spectroscopy represents a promising approach for the rapid prediction of a large variety of phenotypes, in replacement of more laborious "gold standard" tests (Egger-Danner et al., 2015). Rapid phenotype predictions are essential in both production and processing industry when payment are based on quality or when, frequent monitoring, accuracy and precision of measurements are fundamentals. The study of the absorbance of the matter in the region of the electromagnetic spectrum ranging between 4000 – 400 cm^{-1}, called medium infrared (MIR) region, represents an appealing method to predict a large variety of traits (Boichard and Brochard, 2012). MIR-based predictors have been evaluated for the prediction of methane production (Dehareng et al., 2012), the estimation of energy intake and efficiency in lactating cows (McParland et al., 2014), the quantification of metabolites such as lactoferrin and major minerals (Soyeurt et al., 2009 and 2012) as possible indicator of udder status (for the mastitis prediction) in bovine milk. Moreover, the fatty acid composition of the milk was spectroscopically determined to find physiological indicator of fertility (Bastin et al., 2012) and adaptation to climate changes (De Rensis and Scaramuzzi, 2003). MIR spectroscopy has been also applied to predict the fine composition of milk, representing a valuable method to determine fatty acids (Soyeurt et al., 2006), also on a large scale (Soyeurt et al., 2011).

Among milk composition, novel traits generally defined as milk coagulation properties (MCPs) are currently investigated for their technological impact on cheese making process (Bittante et et al., 2012; Pazzola et al., 2014; Manca et al., 2016), but also regarding their correlations with milk quality and sanitary status of the mammary gland (Rovai et al., 2015). Rennet coagulation time (RCT), curd firmness at 30 min after rennet addition (a30), and curd-firming time (k20) are MCPs with undoubted importance in the cheese making process. These traits are commonly

measured by using the *Formagraph* instrument (Foss Electric A/S, Hillerød, Denmark). The output of the analysis is an individual lactodinamographic path able to summarize simultaneously the three coagulation properties in one unique run per animal. Despite the informativeness of this analysis, the laboratory procedure still remains the limiting factor to the integration of these parameters in a large scale breeding programs. However, the collection of these phenotypes on a large sample of animals, jointly with the integration of their genomic information, could results in accurate prediction equations, that, once integrated in automated milking system (AMS), may lead to predict these new traits on large scale, drastically reducing the time and the costs of analysis (Cecchinato et al., 2009). After appropriate calibrations of instruments and standardization of the methods, these predicting equations may be finally extended across countries taking into account for genetic and environmental variability (De Marchi et al., 2014). With the project "*Latte ovino della Sardegna*" supported by the RAS, Regional Law n 7, August 2007 (Grant n. CRP 61608) and conducted in collaboration with different breeders farmer associations, MCPs from a representative sample of sheep of the Sarda breed has been genetically investigated to find correlation between milk and cheese yield, accounting for significant environmental factors (Puledda et al., 2016). The collection of reference measures on a large scale would be the first step in the process of the MCPs modelling for the Sarda breed. The next step will be the comparison of these measures with MIR spectra to obtain a powerful prediction equation suitable for integrating the in-line milk phenotyping system with no additional cost for milk laboratories, allowing the possibility to routinely introduce MCPs as selection criteria in the breeding program of the Sarda breed for the genetic improvement of the individual cheese making ability.

With the increase of the global demand for safe foods, quality-based payment systems have been developed to induce producers in enhancing the quality of their products. The evaluation of the quality, whose concept implies the satisfaction of standard criteria generally accepted by consumers, still represents a drawback, particularly when involving the destruction of the samples analyzed. Non-destructive methods for quality evaluation of products are for example the abovementioned infrared technologies adopted in the milk quality evaluation at laboratory level. The

current trend is to move surveys directly in the production environment by integrating instruments directly into production lines. An example is the assessment of milk quality and udder health through automated milking system that shifted the level of the analysis from the bulk milk to the individual (daily) sampling.

Computer vision (CV) technologies jointly with pattern recognition techniques are often used at industrial and farming level to find appearance features useful for the quality inspection of agricultural products in manner to allow consumer-oriented production choices. In this context size, shape, form and colour are important external factors contributing in describing the overall aspects of products and image analysis is a valuable, not-subjective and automatable instrument for their evaluation (Costa et al., 2011). An example is the use of the CSB-ImageMeater® (CSB-System International) for the on-line grading of pig carcasses on the basis of the visual determination of the lean meat percentage (LMP), in substitution of traditional invasive photodiode-equipped instruments based on light reflectance. The visual system has several advantages if compared with methods based on hand-held instruments. Being characterized by a contactless procedure and a higher degree of automation, it results more hygienic, allowing the (S)EUROP classification (European standard classification for meat percentage) of up to 1.500 carcass per hour. Due to the high speed of return of investments (ROI), great interest is addressed towards this new high throughput recording system. However, the initial cost of installation, around 200.000 Euros (Rossi, 2013; unpublished data), is the main factor that limits the diffusion of this system to slaughterhouses with the highest processing levels. Such an high throughput technology is able to increase the number of phenotypes sampled for pig analyzed, allowing the transition of the characterization of LMP, EUROP class and predicted weight from the level of the whole carcass to that of each single cut. The use of indexes accounting for the overall quality will provide a standardization of the given products. The increase of animal automatically processed per unit of time will results in faster and objective evaluation of the quality resulting in fairer payment for breeders. Current methods of payments (and those in development) are based on the use of grids in which the indicators of quality are economically quantified on the basis of criteria defined by local markets and in respect to local and global legislation. In Italy the system of

quality evaluation is already based on the use the Fat-O-Meater (Carometec A/S, Herlev, Denmark) and the Hennessy Grading Probe (Hennessy Grading System Ltd., Auckland, New Zealand) but the interest is moving towards the automation of the quality assessment and the creation of a payment system able to account for the economical differences that each cut has in each own local market (Project "Pagamento dei suini a peso morto", Centro Ricerche Produzioni Animali CRPA, Reggio Emilia - Italy, 2013).

A classical computer vision apparatus is generally composed by two parts, the first one including the hardware components, the second one regarding the software implementation, with the first evolving technologically faster than the second (Du and Sun, 2006). Hardware components are lighting, optical (high resolution digital cameras) and other mechanical devices that once integrated into the production chain are able to reproduce the human evaluation process in a coordinated manner through an artificial brain (the computer). The complex of rules needed to coordinate these devices and classify samples through the image analysis, represents the software part of the system. Image analysis is the process of producing quantitative information starting from the analysis of representation of whole objects or their regions of particular interest. It is divided in a low (acquisition and pre-processing), intermediate (segmentation and description) and high (recognition and interpretation using classifiers) levels of processing complexity. The result of these steps is the construction of a base knowledge among the object, the use of this knowledge to create rules and to use these rules to set an artificial decision making (Brosnan et al., 2004). Various algorithms have been proposed to emulate the human-thinking and the decision-making process and to obtain qualitative characterizations of products through computer vision. That is the field of the *machine learning technology*.

The processes of machine learning is based on specific algorithm creating decisional rules from training samples and the application of these rules to make decision on unknown samples. Among these algorithms, artificial neural network (ANN), statistical learning (SL), fuzzy logic, decision tree and genetic algorithms are playing an important role on quality evaluation, sort and grading of food products such as meat, fish, pizza, cheese, and bread. Extensive reviews were done on many food products by Brosnan and Sun (2002), Brosnan et al. (2004), Du and Sun (2006) and,

particularly for a shape-discrimination of fruits and vegetables by Costa et al. (2011). In animal production visual inspection has been successfully employed for the evaluation of tenderness, colour, marbling and textural features of the meat (Tan, 2004) for lean meat evaluation using tissue thickness on pig carcass sections (Engel et al., 2012) for the meat quality assessment in chickens (Barbin et al., 2016) for quality grading in fish processing (Misimi, 2007) but also to monitor the behaviour and welfare of animals (Banhazi et al., 2016) to detect phenotypic appearance of species and morphological trait from shape (Kühl et al., 2013). Recently, the study of biological shape variations has taken advantage from the development of geometric morphometrics methods (Monteiro et al., 2002). Traditionally based on the analysis of morphometric distances, angles and ratios (Marcus, 1990), geometric morphometric has progressively moved toward more modern methods (Rohlf and Marcus, 1993) based on the study of spatial coordinates of morphological points having biological meaning, called "*landmarks*". These techniques are more often integrated in breeding programs for assessing the phenotypic variability of farmed animals through image analysis. An example will be given in the present thesis.

Objectives of the Thesis

The thesis is structured in five main chapters and includes the outlines of four experimental contributions, two of those conducted at the Section of Animal Science of the Department of Agriculture of the University of Sassari and two at Ifremer, the French Research Institute for the Exploitation of the Sea.

The *Chapter 2* contains results of the first large scale investigation on *milk coagulation properties* (MCPs) measured on Sarda breed ewes whose results have been recently published by an international peer reviewed journal. Covariance components were estimated for Rennet Coagulation Time (RCT), curd firming time (k_{20}), curd firmness (a_{30}) and individual laboratory cheese yield (ILCY) by fitting both single and multiple-trait animal models. Objective of this work was to provide genetic parameters for coagulation traits as well as their relationship with milk production and composition traits as first step in their integration as potential selection criteria.

The *Chapter 3* reports the second experimental contribution that deals with issues of reduction of high dimensional genotypic data aimed at estimating genomic prediction equation in cattle. With the advent of the microarray technology, the amount of molecular information obtainable per unit of time has increased dramatically in contrast to the decrease of the associated costs. With the increase of information available, a huge unbalancing between the number of records and the number of variables has been created. Tools able to reduce this unbalancing avoiding the loose of information are preferred. The objective of this work was to evaluate the effect of the genomic data reduction through the application of the Principal Component Analysis (PCA) on the accuracies of Direct Genomic Values (DGV) obtained under different conditions of marker density and repartition.

The *Chapter 4* reports the third experimental contribution. This work was aimed at estimating the genetic parameters and genetic gains obtainable by using both traditional and innovative traits for the improvement of the fillet yield in European sea bass, one of the most important species for the Mediterranean aquaculture. Data derived from the Project BAR3D, whose activities were financed by France Agrimer (convention SIVAL NL: 2012-0302). Measurements were obtained from digital

images and ultrasound echographies and were used to derive morphological predictors (indirect criteria) of processing yields, directly on live candidates. The objective of this work was to integrate these traits in a hypothetical selection scheme that, avoiding the sacrifice of candidates and the use of relatives under family selection, may be able to speed up their genetic gains. This work provides genetic parameters and the potential genetic gain for these traits under different breeding strategies.

The *Chapter 5* reports preliminary results obtained through image analysis of juveniles of European sea bass. The aim of this study was to evaluate the possibility to obtain an early sorting of fingerlings based on their sex. Gender determination is an open issue for this species, with great economic implications due to the differences in growth existing among sexes at the commercial size, with females on average 30% heavier than males. The method of evaluation tested was based on the use of several morphometric distances, called *inter-landmark distances,* obtained by connecting all landmarks manually positioned on fish pictures using *Fiji* software. The objective of this work was to obtain an early sexual sorting of juveniles in manner to obtain commercial populations with sex-ratios skewed toward the breeder's interests.

REFERENCES

Åkesson, M., Bensch, S., Hasselquist, D., Tarka, M., Hansson, B., 2008. Estimating heritabilities and genetic correlations: Comparing the "animal model" with parent-offspring regression using data from a natural population. PLoS One 3. doi:10.1371/journal.pone.0001739

Alvarenga, F.A.P., Borges, I., Palkovič, L., Rodina, J., Oddy, V.H., Dobos, R.C., 2016. Using a three-axis accelerometer to identify and classify sheep behaviour at pasture. Appl. Anim. Behav. Sci. 181, 91–99. doi:10.1016/j.applanim.2016.05.026

Andriamasinoro, A.L.A., Bindelle, J., Mercatoris, B., Lebeau, F., 2016. A review on the use of sensors to monitor cattle jaw movements and behavior when grazing. Biotechnol. Agron. Soc. Environ. 20, 273–286.

Banhazi, T.M., Tscharke, M., 2016. A brief review of the application of machine vision in livestock behaviour analysis. J. Agric. Informatics 7, 23–42. doi:10.17700/jai.2016.7.1.279

Barbin, D.F., Mastelini, S.M., Barbon, S., Campos, G.F.C., Barbon, A.P.A.C., Shimokomaki, M., 2016. Digital image analyses as an alternative tool for chicken quality assessment. Biosyst. Eng. 144, 85–93. doi:10.1016/j.biosystemseng.2016.01.015

Bastin, C., Berry, D.P., Soyeurt, H., Gengler, N., 2012. Genetic correlations of days open with production traits and contents in milk of major fatty acids predicted by mid-infrared spectrometry. J. Dairy Sci. 95, 6113–6121. doi:10.3168/jds.2012-5361

Bishop-Hurley, G., Henry, D., Smith, D., Dutta, R., Hills, J., Rawnsley, R., Hellicar, A., Tims, G., Morshead, A., Rahman, A., D'Este, C., Shu, Y. 2014 An investigation of cow feeding behaviour using motion sensors. In 'Proceedings of the 2014 IEEE instrument and measurement technology conference (I2MTC)'. pp. 1285–1290. (IEEE: Montevideo, NY)

Bittante, G., Penasa, M., and Cecchinato, A. 2012. Invited review: Genetics and modeling of milk coagulation properties. Journal of Dairy Science 95, 6843-6870

Boichard, D., Brochard, M., 2012. New phenotypes for new breeding goals in dairy cattle. Animal 6, 544–550. doi:10.1017/S1751731112000018

Brosnan, T., Sun, D.-W., 2002. Inspection and grading of agricultural and food products by computer vision systems - A review. Comput. Electron. Agric. 36, 193–213. doi:10.1016/S0168-1699(02)00101-1

Brosnan, T., Sun, D.W., 2004. Improving quality inspection of food products by computer vision - A review. J. Food Eng. 61, 3–16. doi:10.1016/S0260-8774(03)00183-3

Cabrera-Bosquet, L., Crossa, J., von Zitzewitz, J., Serret, M.D., Luis Araus, J., 2012. High-throughput Phenotyping and Genomic Selection: The Frontiers of Crop Breeding Converge. J. Integr. Plant Biol. 54, 312–320. doi:10.1111/j.1744-7909.2012.01116.x

Cecchinato, A., De Marchi, M., Gallo, L., Bittante, G., Carnier, P., 2009. Mid-infrared spectroscopy predictions as indicator traits in breeding programs for enhanced coagulation properties of milk. J. Dairy Sci. 92, 5304–5313. doi:10.3168/jds.2009-2246

Costa, C., Antonucci, F., Pallottino, F., Aguzzi, J., Sun, D.-W., Menesatti, P., 2011. Shape Analysis of Agricultural Products: A Review of Recent Research Advances and Potential Application to Computer Vision. Food Bioprocess Technol. 4, 673–692. doi:10.1007/s11947-011-0556-0

Decandia, M., Giovanetti,V., Acciaro, M., Mameli, M., Molle, G., Cabiddu, A., Manca, C., Cossu, R., Serra, M.G., Dimauro, C, and Rassu, S.P.G., 2016. BEHARUM project : use of ICT to The multiple roles of grassland in the European bioeconomy. Proceedings of the 26th General Meeting of the European Grassland Federation Trondheim, Norway, 4-8 September 2016. ISBN 978-82-17-01677-9

Dehareng, F., Delfosse, C., Froidmont, E., Soyeurt, H., Martin, C., Gengler, N., Vanlierde, A., Dardenne, P., 2012. Potential use of milk mid-infrared spectra to predict individual methane emission of dairy cows. Animal 6, 1694–1701. doi:10.1017/S1751731112000456

De Marchi, M., Toffanin, V., Cassandro, M., Penasa, M., 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J. Dairy Sci. 97, 1171–1186. doi:10.3168/jds.2013-6799

De Rensis, F., Scaramuzzi, R.J., 2003. Heat stress and seasonal effects on reproduction in the dairy cow - a review. 2014. Theriogenology 60, 1139–1151. doi:10.1016/S0093-691X(03)00126-2

Dimauro, C., Cellesi, M., Gaspa, G., Ajmone-Marsan, P., Steri, R., Marras, G., Macciotta, N.P., 2013. Use of partial least squares regression to impute SNP

genotypes in Italian Cattle breeds. Genet. Sel. Evol. 45, 15. doi:10.1186/1297-9686-45-15

Du, C.J., Sun, D.W., 2006. Learning techniques used in computer vision for food quality evaluation: A review. J. Food Eng. 72, 39–55. doi:10.1016/j.jfoodeng.2004.11.017

Dutta, R., Smith, D., Rawnsley, R., Bishop-Hurley, G., Hills, J., Timms, G., Henry, D., 2015. Dynamic cattle behavioural classification using supervised ensemble classifiers. Comput. Electron. Agric. 111, 18–28. doi:10.1016/j.compag.2014.12.002

Egger-Danner, C., Cole, J.B., Pryce, J.E., Gengler, N., Heringstad, B., Bradley, A., Stock, K.F., 2015. Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. Anim. an Int. J. Anim. Biosci. 9, 191–207. doi:10.1017/S1751731114002614

Engel, B., Lambooij, E., Buist, W.G., Vereijken, P., 2012. Lean meat prediction with HGP, CGM and CSB-Image-Meater, with prediction accuracy evaluated for different proportions of gilts, boars and castrated boars in the pig population. Meat Sci. 90, 338–344. doi:10.1016/j.meatsci.2011.07.020

Falconer, D.S., MacKay, T.F.C., 1996. Introduction to Quantitative Genetics, 4th ed. Longman Scientific & Technical, Harlow, UK.

Furbank, R.T., Tester, M., 2011. Phenomics - technologies to relieve the phenotyping bottleneck. Trends Plant Sci. 16, 635–644. doi:10.1016/j.tplants.2011.09.005

Gorlier, A., Lonati, M., Renna, M., Lussiana, C., Lombardi, G., Battaglini, L.M., 2012. Changes in pasture and cow milk compositions during a summer transhumance in the western Italian Alps. J. Appl. Bot. Food Qual. 85, 216–223.

Greenwood, P.L., Bishop-Hurley, G.J., González, L.A., Ingham, A.B., 2016. Development and application of a livestock phenomics platform to enhance productivity and efficiency at pasture. Anim. Prod. Sci. 56, 1299–1311. doi:10.1071/AN15400

Hayes, B.J., Bowman, P.J., Chamberlain, A.J., Goddard, M.E., 2009a. Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92, 433–443.

Hayes, B.J., Bowman, P.J., Chamberlain, A.C., Verbyla, K., Goddard, M.E., 2009b. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41, 51. doi:10.1186/1297-9686-41-51.

Hazel, L.N., Lush, J.L., 1942. The Efficiency of Three Methods of Selection. J. Hered. 33, 393–399.

Hazel, L.N., 1943. The Genetic Basis for Constructing Selection Indexes. Genetics 28, 476–490.

Henderson, C.R., 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biomctrics 32:69.

Henderson, C.R., 1985. Equivalent Linear Models to Reduce Computations. J. Dairy Sci. 68, 2267–2277. doi:10.3168/jds.S0022-0302(85)81099-7

Hocquette, J.F., Capel, C., David, V., Guémené, D., Bidanel, J., Ponsart, C., Gastinel, P.L., Bail, P.Y. Le, Monget, P., Mormède, P., Barbezant, M., Guillou, F., Peyraud, J.L., 2012. Objectives and applications of phenotyping network set-up for livestock. Anim. Sci. J. 83, 517–528. doi:10.1111/j.1740-0929.2012.01015.x

Houle, D., Govindaraju, D.R., Omholt, S., 2010. Phenomics: the next challenge. Nat. Rev. Genet. 11, 855–866. doi:10.1038/nrg2897

Jorjani, H., Jakobsen, J., Hjerpe, E., Palucci, V. and Dürr, J. 2012. Status of genomic evaluation in the brown swiss populations. Interbull Bulletin 46, 46–54.

Kruuk, L.E., 2004. Estimating genetic parameters in natural populations using the "animal model." Philos. Trans. R. Soc. London B Biol. Sci. 359, 873–890. doi:10.1098/rstb.2003.1437

Kühl, H.S., Burghardt, T., 2013. Animal biometrics: quantifying and detecting phenotypic appearance. Trends Ecol. Evol. 28, 432–441. doi:10.1016/j.tree.2013.02.013

Lobet, G., Draye, X., Périlleux, C., 2013. An online database for plant image analysis software tools. Plant Methods 9, 1–7. doi:10.1186/1746-4811-9-38

Lynch, M. and Walsh, B., 1998. Genetics and Analysis of Quantitative Traits. Book 980pp. ISBN: 978-0878934812

Lund, M.S., Roos, A.P., Vries, A.G., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F. and Su, G. 2011. A common reference population from four european holstein populations increases reliability of genomic predictions. Genetics Selection Evolution 43, 43

Lush, J. L., 1937. Animal Breeding Plans, Collegiate Press Inc, Ames, IA.

Macciotta, N.P.P., Gaspa, G., Steri, R., Nicolazzi, E.L., Dimauro, C., Pieramati, C., Cappio-Borlino, A., 2010. Using eigenvalues as variance priors in the prediction

of genomic breeding values by principal component analysis. J. Dairy Sci. 93, 2765–74. doi:10.3168/jds.2009-3029

McParland, S., Lewis, E., Kennedy, E., Moore, S.G., McCarthy, B., O'Donovan, M., Butler, S.T., Pryce, J.E., Berry, D.P., 2014. Mid-infrared spectrometry of milk as a predictor of energy intake and efficiency in lactating dairy cows. J. Dairy Sci. 97, 5863–5871. doi:10.3168/jds.2014-8214

Manca, M.G., Serdino, J., Gaspa, G., Urgeghe, P., Ibba, I., Contu, M., Fresi, P., Macciotta, N.P.P., 2016. Derivation of multivariate indices of milk composition, coagulation properties, and individual cheese yield in dairy sheep. J. Dairy Sci. 99, 4547–4557. doi:10.3168/jds.2015-10589

Marcus, L.F., 1990. Traditional morphometrics. Pp. 77–122 in Rohlf, F.J., Bookstein, F.L., 1990. Proc. Michigan Morphometrics Work. ISBN: 0-9628499-0-1

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. doi:11290733

Misimi, E., 2007. Computer vision for quality grading in fish processing (Doctoral dissertation). Faculty of Information Technology, Mathematics and Electrical Engineering, Department of Engineering Cybernetics. ISBN: 978-82-471-54448-9.

Monteiro, L.R., Diniz-Filho, J.A.F., Dos Reis, S.F., Araujo, E.D., 2002. Geometric Estimates of Heritability in Biological Shape. Evolution (N. Y). 56, 563–572. doi:10.1111/j.0014-3820.2002.tb01367.x

Moser, G., Khatkar, M.S., Hayes, B.J., Raadsma, H.W., 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol. 42, 37. doi:10.1186/1297-9686-42-37

Pattaro, C., Ruczinski, I., Fallin, D.M., Parmigiani, G., 2008. Haplotype block partitioning as a tool for dimensionality reduction in SNP association studies. BMC Genomics 9, 405. doi:10.1186/1471-2164-9-405

Pazzola, M., Dettori, M.L., Cipolat-Gotet, C., Cecchinato, A., Bittante, G., and Vacca, G.M. 2014. Phenotypic factors affecting coagulation properties of milk from Sarda ewes. Journal of Dairy Science 97, 7247-7257

Pryce, J.E., Gredler, B., Bolormaa, S., Bowman, P.J., Egger-Danner, C., Fuerst, C., Emmerling, R., Solkner, J., Goddard, M.E., and Hayes. B.J. 2011. Short communication: genomic selection using a multi-breed, across-country reference population. Journal of Dairy Science 94, 2625–2630.

Puledda, A., Gaspa, G., Manca, M.G., Serdino, J., Urgeghe, P.P., Dimauro, C., Negrini, R., Macciotta, N.P.P., 2016. Estimates of heritability and genetic correlations for milk coagulation properties and individual laboratory cheese yield in Sarda ewes. Animal 1–9. doi:10.1017/S1751731116002147

Quaas, R.L., and Pollak., E. J. 1980. Mixed Model Methodology for Farm and Ranch Beef Cattle Testing Programs. J. Anim. Sci. 51:1277-1287. doi:10.2527/jas1981.5161277

Ragoussis, J., 2006. Genotyping technologies for all. Drug Discov. Today Technol. 3, 115–122. doi:10.1016/j.ddtec.2006.06.013

Rahaman, M.M., Chen, D., Gillani, Z., Klukas, C., Chen, M., 2015. Advanced phenotyping and phenotype data analysis for the study of plant growth and development. Front. Plant Sci. 6, 619. doi:10.3389/fpls.2015.00619

Rossi, A., Bertolini, A., Gorlani, E., 2013. Valutazione carcasse suine: verso metodiche aggiornate. Agricoltura. Mensile della Regione Emilia-Romagna; unpublished data.

Rohlf, F.J., Marcus, L.F., 1993. A revolution morphometrics. Trends Ecol. Evol. 8, 129–132. doi:10.1016/0169-5347(93)90024-J

Rovai M, Natalia R, Gerardo C, Jordi S and Gabriel L 2015. Effect of subclinical intrammamay infection on milk quality in dairy sheep: I. Fresh-soft cheese produced from milk of uninfected and infected glands and from their blends. Small Ruminant Research 125, 127-136

Schulze, T.G., Zhang, K., Chen, Y.S., Akula, N., Sun, F., McMahon, F.J., 2004. Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. Hum. Mol. Genet. 13, 335–342. doi:10.1093/hmg/ddh035

Shen, R., Fan, J.B., Campbell, D., Chang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Garcia, E.W., McBride, C., Steemers, F., Garcia, F., Kermani, B.G., Gunderson, K., Oliphant, A., 2005. High-throughput SNP genotyping on universal bead arrays. Mutat. Res. - Fundam. Mol. Mech. Mutagen. 573, 70–82. doi:10.1016/j.mrfmmm.2004.07.022

Sneddon, J., Mason, A., 2014. Automated Monitoring of Foraging Behaviour in Free Ranging Sheep Grazing a Bio-diverse Pasture using Audio and Video Information. 8th Int. Conf. Sens. Technol. 2–4. doi:10.1109/ICSensT.2013.6727614

Solberg, T.R., Sonesson, A.K., Woolliams, J. a, Meuwissen, T.H.E., 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol. 41, 29. doi:10.1186/1297-9686-41-29

Stanton, T.L., Jones, L.R., Everett, R.W., Kachman, S.D., 1992. Estimating Milk, Fat, and Protein Lactation Curves with a Test Day Model. J. Dairy Sci. 75, 1691–1700. doi:10.3168/jds.S0022-0302(92)77926-0

Soyeurt, H., Dardenne, P., Dehareng, F., Lognay, G., Veselko, D., Marlier, M., Bertozzi, C., Mayeres, P., Gengler, N., 2006. Estimating Fatty Acid Content in Cow Milk Using Mid-Infrared Spectrometry. J. Dairy Sci. 89, 3690–3695. doi:10.3168/jds.S0022-0302(06)72409-2

Soyeurt, H., Bruwier, D., Romnee, J.-M., Gengler, N., Bertozzi, C., Veselko, D., Dardenne, P. 2009. Potential estimation of major mineral contents in cow milk using mid-infrared spectrometry. J. Dairy Sci. 92, 2444–2454. doi:10.3168/jds.2008-1734

Soyeurt, H., Dehareng, F., Gengler, N., McParland, S., Wall, E., Berry, D.P., Coffey, M., Dardenne, P. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J. Dairy Sci. 94, 1657–1667. doi:10.3168/jds.2010-3408

Soyeurt, H., Bastin, C., Colinet, F.G., Arnould, V.M.-R., Berry, D.P., Wall, E., Dehareng, F., Nguyen, H.N., Dardenne, P., Schefers, J., Vandenplas, J., Weigel, K., Coffey, M., Théron, L., Detilleux, J., Reding, E., Gengler, N., McParland, S., 2012. Mid-infrared prediction of lactoferrin content in bovine milk: potential indicator of mastitis. animal 6, 1830–1838. doi:10.1017/S1751731112000791

Tan, J., 2004. Meat quality evaluation by computer vision. J. Food Eng. 61, 27–35. doi:10.1016/S0260-8774(03)00185-7

Thompson, R., Brotherstone, S., White, I.M.S., 2005. Estimation of quantitative genetic parameters. Philos. Trans. R. Soc. B-Biological Sci. 360, 1469–1477. doi:10.1098/rspb.2007.1417

Visscher, P.M., Hill, W.G., Wray, N.R., 2008. Heritability in the genomics era - concepts and misconceptions. Nat. Rev. Genet. 9, 255–266. doi:10.1038/nrg2322.

VanRaden, P.M., Null, D.J., Sargolzaei, M., Wiggans, G.R., Tooker, M.E., Cole, J.B., Sonstegard, T.S., Connor, E.E., Winters, M., van Kaam, J.B.C.H.M., Valentini, a, Van Doormaal, B.J., Faust, M. a, Doak, G. a, 2013. Genomic imputation and evaluation using high-density Holstein genotypes. J. Dairy Sci. 96, 668–78. doi:10.3168/jds.2012-5702

Van Vleck, L.D., 1990. Absorption of equations for non-parents for an animal model with maternal effects and genetic groups. J. Anim. Sci. 68, 4014–4025.

Walsh, B., 2001. Quantitative Genetics 1–7. doi:10.1038/npg.els.0005408

Weigel, K., de los Campos, G., González-Recio, O., Naya, H., Wu, X.L., Long, N., Rosa, G.J.M., Gianola, D., 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92, 5248–5257. doi:10.3168/jds.2009-2092

Westell, R.A., Quaas, R.L., Van Vleck, L.D., 1984. Genetic groups in an animal model. J. Anim. Sci. 59 (Suppl. 1):175 (Abstr.).

Westell, R.A., Quaas, R.L., Van Vleck, L.D., 1988. Genetic Groups in an Animal Model. J. Dairy Sci. 71, 1310–1318. doi:10.3168/jds.S0022-0302(88)79688-5.

Williams, M.L., Mac Parthaláin, N., Brewer, P., James, W.P.J., Rose, M.T., 2014. A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques. J. Dairy Sci. 99, 2063–2075. doi:10.3168/jds.2015-10254

CHAPTER 2

Estimates of heritability and genetic correlations for milk coagulation properties and individual laboratory cheese yield in Sarda ewes

Published as

A. Puledda, G. Gaspa, M. G. Manca, J. Serdino, P. P. Urgeghe, C. Dimauro, R. Negrini and N. P. P. Macciotta. 2016. Estimates of heritability and genetic correlations for milk coagulation properties and individual laboratory cheese yield in Sarda ewes. *Animal* doi: 10.1017/S1751731116002147

(Received 2 May 2016; Accepted 10 September 2016)

Short title: Genetic parameters for clotting properties of milk

Abstract

Objective of this study was to estimate genetic parameters of milk coagulation properties (MCPs) and individual laboratory cheese yield (ILCY) in a sample of 1.018 Sarda breed ewes farmed in 47 flocks. Rennet coagulation time (RCT), curd firming time (k_{20}) and curd firmness (a_{30}) were measured using Formagraph instrument whereas ILCY were determined by a micro manufacturing protocol. About 10% of the milk samples did not coagulate within 30 minutes and 13% had zero value for k_{20}. The average ILCY was 36%. (Co)variance components of considered traits were estimated by fitting both single and multiple-trait animal models. Flock-test date explained from 13% to 28% of the phenotypic variance for MCPs and 26% for ILCY, respectively. The largest value of heritability was estimated for RCT (0.23 ± 0.10) whereas it was about 0.15 for the other traits. Negative genetic correlations between RCT and a_{30} (–0.80 ± 0.12), a_{30} and k_{20} (–0.91 ± 0.09), and a_{30} and ILCY (–0.67 ± 0.08) were observed. Interesting genetic correlations between MCPs and milk composition ($r_G > 0.40$) were estimated for pH, NaCl and Casein. Results of the present study suggest to use only one out of three MCPs to measure milk renneting ability, due to the highly genetic correlations among them. Moreover, negative correlations between ILCY and MCPs suggest that great care should be taken when using these methods to estimate cheese yield from small milk samples.

Keywords: clotting properties, rennet, dairy sheep, genetic parameters, variance components.

Implication

The estimation of genetic parameters is the first and essential step to select for coagulation traits and cheese yield in dairy sheep. The aim of this paper is to fill a gap in comparison to what happens in dairy cattle; indeed heritability estimates for clotting properties are missing for small ruminants. As the sheep milk is almost totally destined for cheese making, the estimation of heritability of coagulation traits may enable future scenarios of selection for such traits, with potential implications on selection schemes and milk payment tables.

Introduction

The Italian dairy sheep stock consists of about 5.5 million ewes. The largest breed is the Sarda with more than 3 million sheep (BDN, 2015). The Sarda breed accounts for about 43% of the national total ovine stock and about 4% of EU sheep stock (Eurostat, 2015). The total milk production of the Sarda breed is about 300.000 t of milk per year (about 4% of total world production, FAOSTAT, 2014). The breeding programme currently involves a breeding nucleus of 212.941 milk recorded sheep farmed in 1.032 flocks, and the commercial population (ICAR, 2014). Since the beginning of the program, the total milk yield (MY) per lactation has been the main selective goal of Sarda sheep (Sanna et al., 1997; Carta et al., 2009). Recording for fat percentage (FP) and protein percentage (PP) on first lambing ewes started in 1998. The milk is totally transformed into cheese with a production of 50-60.000 tons per year of three Protected Designation Origin products, mostly destined for export (Furesi et al., 2013). Thus the milk cheese making ability could be a breeding goal of great interest for this breed.

Pecorino Romano cheese yield could be predicted from bulk milk composition using suitable equations (Pirisi et al., 1994). However, predictions on individual milk are less accurate due to the variability of milk solids, thus individual laboratory cheese yield (ILCY) has been proposed as an indicator of potential cheese yield in individual ovine milk samples (Othmane et al., 2002). ILCY shows low heritability and a positive genetic correlation with milk composition and negative genetic correlation with milk yield in Spanish Churra sheep (Othmane et al., 2002).

Other indicators of milk cheese making ability, extensively studied in cattle, are milk coagulation properties (MCPs) (Aleandri et al., 1989; Ikonen et al., 1999; De Marchi et al., 2008; Bonfatti et al., 2014). They are usually defined by three parameters: rennet coagulation time (RCT, min), curd firming time (k_{20}, min) and curd firmness (a_{30}, mm), commonly measured by using either mechanical or optical devices (Bittante et al., 2012). Several studies have reported that an appreciable proportion of the MCPs variation in cow milk is of additive genetic nature. A recent review by Bittante et al., (2012) reported moderate values of heritability (about 0.26) for RCT and a_{30}, whilst few study report reliable estimates for k_{20}. In general, MCPs exhibit

moderate to high genetic correlations with pH and somatic cell count, and very low to null with milk traits, respectively (Bittante et al., 2012).

Few studies have been carried out on MCPs in small ruminants, especially in sheep. In particular effects of environmental factors, feeding, breed, parity and lactation stage on MCPs, milk composition and laboratory cheese yield, have been investigated (Jaramillo et al., 2008; Bittante et al., 2014; Pazzola et al., 2014). In some researches, novel milk coagulation and syneresis parameters, estimated by a nonlinear modelling of the entire curd-firming process were used (Vacca et al., 2015). Finally, relationships between MCPs, sanitary status of the mammary gland were also investigated in sheep (Rovai et al., 2015). Analysis of environmental factors affecting MCPs and ILCY and the estimation of their genetic parameters for MCPs and ILCY are essential steps for planning their improvement by means of selection. Aim of this study was to estimate heritability of MCPs, ILCY and their phenotypic and genetic correlations with milk yield and composition in Sarda dairy ewes.

Material and methods

Animals, milk sample collection and laboratory analysis

The study involved 1.018 Sarda ewes from 47 flocks located in the four historical provinces of Sardinia. Pedigree and milk recording information were supplied by the national association of small ruminant breeders (ASSONAPA, Rome, Italy). The pedigree file included more than 1.8 million animal records. Dairy ewes were offspring of 499 rams; other details about the structure of the population are given in Table 1.

Table 1 Descriptive statistics for animals and flocks structure of Sarda population involved in this study.

Items	n.[1]	average	sd	min	max
Flock size	-	95.5	57.1	13	233
Number of Lactation	1.016	4.0	2.3	1	12
Age (Month)	1.016	48.1	27.2	12	133
Ewes per ram	499	2.1	1.7	1	15
Rams per flock	47	16.9	5.8	6	33

[1] For the number of lactation and age n. is the number of records; for the last two items, n. is the number of rams and flocks, respectively.

Individual milk samples (100 ml) were collected in the mid-late lactation (from 45 to 249 days in milk (DIM), average = 156 ± 37.4 days) from April to July 2014 by the provincial association of breeders (APA). Preservatives (bronopol, 62,5 µl/100 ml) were added to the milk samples. Analyses were carried out within the 24 h after sampling and the milk samples were kept refrigerated during transportation from the farms to the laboratory. The milk samples were split into two subsamples of 50 ml each and analysed in order to determine composition and cheese making attitude of milk, respectively. Standard milk analysis were performed at the milk lab of the Regional Association of Animal Breeds of Sardinia (ARA, Oristano, Italy). Milk composition was spectroscopically determined by MilkoScan^{TM} (Foss Electric, Denmark). Somatic cell count (SCC) was also determined using the Fossomatic^{TM} (Foss Electric, Denmark). MCPs were measured by using a Formagraph Instrument (Foss Electric A/S, Hillerød, Denmark). In brief, 10 ml of each individual sample were heated to 35° C before the addition of 200 µL of rennet solution (Hansen Naturen 215, with 80% ± 5% and chymosin 20% ± 5 % pepsin, Pacovis Amrein AG, Bern, Switzerland) diluted to 0.8% in distilled water. This analysis ended within 30 min after rennet addition and produced a lactodinamographic path as reported by Bittante et al., (2012). RCT is the time between rennet addition and the start of the milk coagulation, k_{20} the time at which the typical oscillation graph reaches the width of 20 mm, and a_{30} the width of the graph at 30 min after rennet addition. ILCY was determined according to a modified method of Othmane et al., (2002), further details of the methodology used are provided in Manca et al., (2016). The predicted pecorino cheese yield (PPCY) was also calculated using the equation proposed by Pirisi et al., (1994): PPCY = 1.747 x protein% + 1.272 x fat%.

Statistical analysis

Thirteen traits were analysed: RCT, k_{20}, a_{30}, ILCY, PPCY, MY, FP, PP, casein percentage (CSN), conjugated linolenic acid percentage (CLA), pH, NaCl, Somatic cell score (SCS). Non-coagulating samples were eliminated, as well as the missing records for the other traits. As the k_{20} parameter presented a skewed distribution, a log transformation was also applied on this trait. (Co)variance components were estimated by using Restricted Maximum Likelihood (REML) methodology

implemented in VCE v. 6.0 software (Groeneveld et al., 2010). Both a Single- (ST) and multiple-trait (MT) animal models were fitted. The raw data included 1.018 animals with phenotypes. Ancestors were extracted from the pedigree file considering up to the fourth previous generation. A total of 5234 animals were included in the numerator relationship matrix (A). ST and MT analysis were carried out on MCPs using the following linear mixed model

$$y_{ijklmno} = \mu + PAR_J + LM_k + DIM_l + RP_m + ftd_n + a_o + e_{ijklmno} \quad [1]$$

In the ST model, for each trait, $y_{ijklmno}$ is a single measure for the i_{th} individual; μ the overall mean; *PAR* the fixed effect of j_{th} parity with 6 levels (j = 1 to 5 and PAR ≥ 6); *LM* is the fixed effect of k_{th} lambing month with 6 levels (k = October to March); *DIM* is the fixed effect l_{th} class of days in milk with 6 levels (l = 1 : from 45 - 80 d, l = 2 : 81 – 120 d, l = 3 : 121 – 150 d, l = 4 : 151 – 180 d, l=5: 181-210, l=6: >210d); *RP* is the fixed effect of the m_{th} position of the milk samples in the rack of Formagraph, (m = 1 to 10); *ftd* is the cross-classified random effect of the n_{th} combination flock-test date (n = 1 to 70 levels) with $ftd_n \sim N(0, I\sigma^2_{ftd})$, where I and σ^2_{ftd} are the identity matrix and the variance associated with the flock-test date, respectively; a_o is the random genetic effect for the o_{th} animal (o = 1 to 5234 levels) with $a_o \sim N(0, A\sigma^2_a)$ and $e_{ijklmno}$ is the random residual term with $e_{ijklmno} \sim N(0, I\sigma^2_e)$ where σ^2_a and σ^2_e are the additive genetic and residual variance, respectively. Genetic parameters of ILCY, MY and composition traits were estimated with a mixed linear model that had the same structure of eq. [1], but that did not included the effect of the rack position.

In the MT animal model y_i represented the vector of dependent variables for the i_{th} individual, whereas the fixed and random effects were the same as eq. [1]. Two different MT analyses were carried out: i) a five-traits animal model, including the 3 MCPs, ILCY and PPCY, that was aimed at estimating genetic correlations among coagulation properties and cheese yields; ii) a series of bi-variate analyses for ILCY and each of the MCPs with all the remaining variables was performed to evaluate the genetic correlations among abovementioned properties and milk yield and

composition. For both MT animal models the (co)variances for random effects were assumed to follow a multivariate normal distribution and they were modelled as

$$[\boldsymbol{a}_1 \dots \boldsymbol{a}_2]' \sim N(0, \boldsymbol{A} \otimes \boldsymbol{G})$$

$$[\boldsymbol{ftd}_1 \dots \boldsymbol{ftn}_2]' \sim N(0, \boldsymbol{I} \otimes \boldsymbol{F})$$

$$[\boldsymbol{e}_1 \dots \boldsymbol{e}_2]' \sim N(0, \boldsymbol{I} \otimes \boldsymbol{R})$$

where: **A** and **I** have previously been defined; *n* was the number of traits analysed; **G**, **F** and **R** were the *n* by *n* genetic additive, flock-test-date and residual covariance matrices, respectively (the element $\sigma^2_{i,j(i=j)}$ in the diagonal are the variance and the off-diagonal $\sigma_{i,j(i\neq j)}$ covariance between trait *i* and trait *j*). For each trait, both for ST and MT analyses, heritability was computed as

$$h^2 = \frac{\sigma^2_a}{\sigma^2_a + \sigma^2_{ftd} + \sigma^2_e}$$

for each analysed trait. Moreover, phenotypic and genetic correlations among MCPs and all the other traits have been computed.

Results

Descriptive statistics

FP, PP and CSN (Table 2) were similar to those observed in individual data either in Sarda (Pazzola et al., 2014; Manca et al., 2016) and in other Italian breeds, such as Valle del Belice (Cappio-Borlino et al., 1997), Massese (Martini et al., 2008) and alpine breeds (Bittante et al., 2014). Milk solid were lower than those reported for Spanish breeds (Othmane et al., 2002; Jaramillo et al., 2008). The average values of the test-day records for CLA (1.26 ± 0.57) was lower than values reported for Sarda and Massese breeds (Nudda et al., 2005; Martini et al., 2008). The pH exhibited the smallest variability whereas SCC varied accordingly to the breed average (Pazzola et al., 2014).

Table 2
Number of records, overall mean, standard deviation, minimum (Min), Maximum (Max) and coefficient of variation (CV) for sheep milk yield and composition, milk coagulation properties and individual laboratory cheese yield.

Trait[1]	n.[2]	Mean	SD	Min	Max	CV (%)
Milk yield and composition						
MY (L/day)	1.005	1.72	0.42	0.61	3.30	24.4
FP (g/100 mL)	998	6.06	1.35	2.55	12.00	22.3
PP (g/100 mL)	998	5.47	0.61	3.86	8.76	11.2
CSN (g/100 mL)	998	4.25	0.50	2.91	6.89	11.8
CLA (g/ 100g FAME)	908	1.26	0.57	0.00	3.04	45.2
SCS	994	4.68	2.33	0.06	11.13	49.8
pH (U)	1.002	6.58	0.14	5.65	7.36	2.1
NaCl (mg/100 mL)	998	146.88	45.16	64.30	551.70	30.7
Cheese-related traits						
RCT (min)	1.008	15.18	4.29	2.37	30.07	28.3
k20 (min)	879	1.75	0.74	0.50	7.00	42.3
logk20	879	1.99	0.16	1.48	2.62	8.1
a30 (mm)	990	52.63	16.08	0.98	107.80	30.6
ILCY (%,w/v)	1.017	36.24	9.33	4.67	80.14	25.7
PPCY (%,w/v)	998	17.28	2.43	11.71	29.55	14.1

[1]MY = test day milk yield; FP = fat; PP = protein; CSN = casein; CLA = Conjugated linoleic Acid, FAME=fatty acid methyl esters; SCS = somatic cell score, $\log_2[(SCC\mu l^{-1}/100)+3]$; RCT = rennet coagulation time; k_{20} = curd firming time; $Logk_{20} = \log_{10}$ of curd firming time in sec; a_{30} = curd firmness; ILCY= individual laboratory cheese yield; PPCY=Predicted Pecorino Cheese Yield, U=pH unit. [2] Number of samples used to compute descriptive statistics.

About 6% (n = 64) of the samples did not coagulate within 30 min (**NC** samples thereafter), and at the same time they did not present any values for a_{30} and k_{20}. The samples with RCT > 29 min and a_{30} < 1 mm, about 9% (n=90) and 10% (n=101) respectively, were discarded. In addition 129 samples (about 13%) that did not reach a curd firmness of 20 mm were excluded from the analysis (Figure 1). The percentage of NC samples within each DIM class tended to increase along the lactation. However, when the NC samples are referred to the total number of samples, largest values were observed in the central DIM classes. The same trend is observed for the percentage of missing k_{20} records (Figure 2).

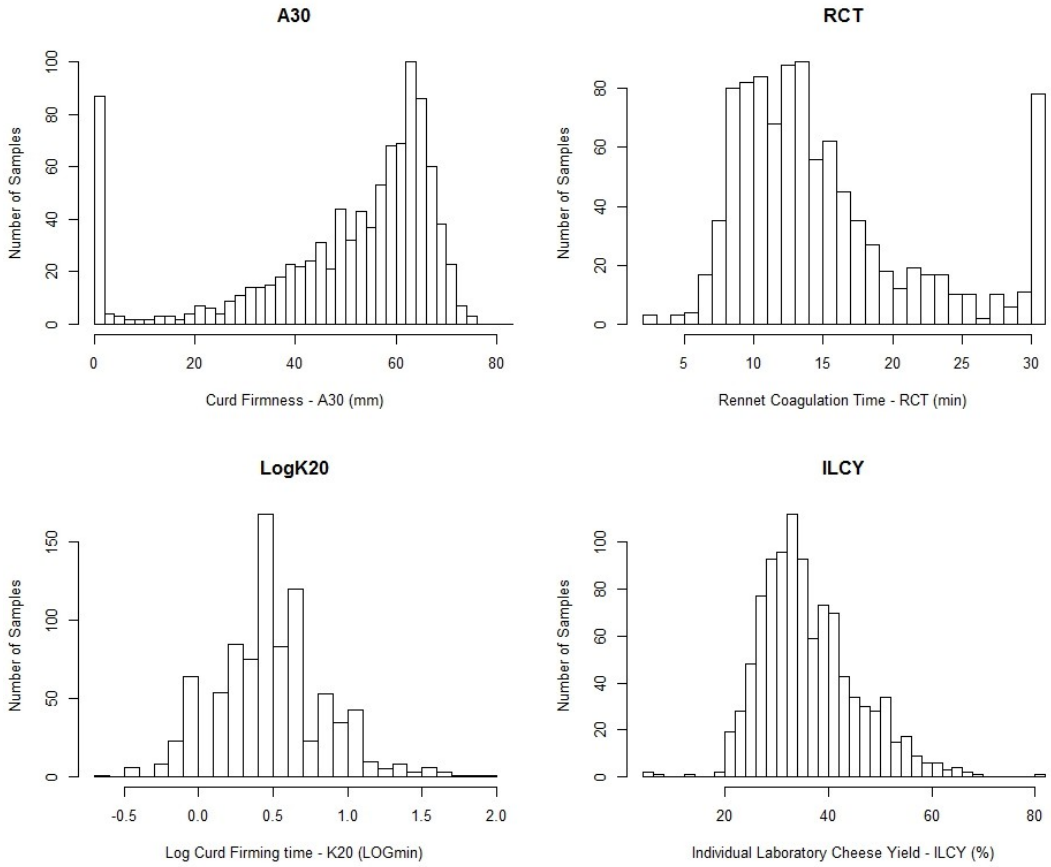

Figure 1 Frequency distribution of the three milk coagulation properties and the Individual laboratory cheese yield (ILCY) in Sarda sheep, for the raw data (before data editing). The very first (a_{30}) e the last two bars (RCT) represent those samples that have been discarded from the analysis.

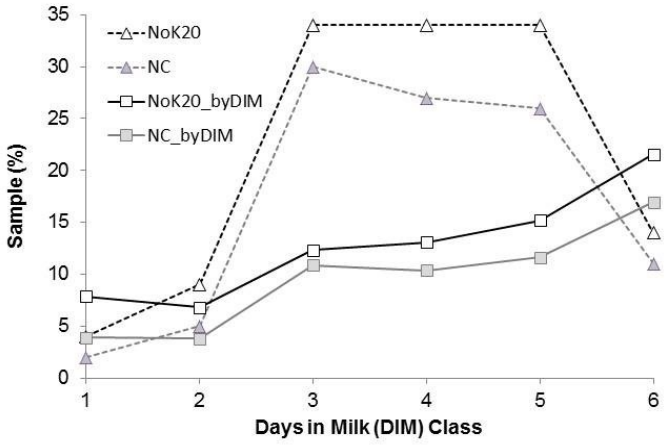

Figure 2 Percentage of not coagulating (NC) or missing k_{20} samples (NoK20), percentage of not coagulating samples (NC_byDIM) and missing k_{20} (NoK20_byDIM) within each class of DIM.

The a_{30} and k_{20} parameters presented a more skewed distributions compared to RCT and ILCY (Figure 1). The average values for RCT were lower than those reported by Pazzola et *al.* (2014) and Vacca et *al.* (2015), whereas k_{20} and a_{30} varied accordingly to the values reported in literature for dairy sheep (Table 2). Average ILCY measured in the present study was 36.2 ± 9.3%, whereas predicted cheese yield using the equation was equal 17.3 ± 2.4% (Table 2).

A not negligible fraction of the phenotypic variance can be ascribed to the differences between flocks. The flock-test date effect explained a quota of phenotypic variation ranging from 13% to 33% for the three milk coagulation parameters and cheese yield (Table 3). The largest percentage of variance explained by flock-test date was recorded for MY (55%), followed by FP (47%), pH (30%), PP and CSN (20%). Moreover, for the last two traits the flock variance matched those of MCPs.

Table 3

Estimates of genetic (σ^2_a) and environmental variance ($\sigma^2_{ftd,}\sigma^2_e$), heritability ($h^2$) and percentage of variance explained by flock-test-day (r^2_{ftd}) and standard errors (SE) for sheep milk coagulation properties, individual cheese yield, milk production and composition traits.

Trait[1]	n.[2]	σ^2_a	$\sigma^2_{ftd,}$	σ^2_e	$h^2_{(SE)}$	$r^2_{ftd\ (SE)}$
Cheese-related traits						
RCT (min)	908	5.83	3.32	16.17	$0.23_{(0.10)}$	$0.13_{(0.03)}$
logk20	879	0.004	0.006	0.016	$0.15_{(0.11)}$	$0.23_{(0.04)}$
a30 (mm)	907	27.74	54.92	110.49	$0.14_{(0.10)}$	$0.28_{(0.05)}$
ILCY (%,w/v)	1.017	12.45	20.40	46.46	$0.16_{(0.09)}$	$0.26_{(0.05)}$
PPCY (%,w/v)	998	0.66	1.55	2.46	$0.14_{(0.09)}$	$0.33_{(0.05)}$
Milk yield and Composition						
MY (L/day)	1.005	0.013	0.09	0.06	$0.08_{(0.05)}$	$0.55_{(0.04)}$
FP (g/100 mL)	998	0.16	0.76	0.71	$0.10_{(0.07)}$	$0.47_{(0.05)}$
PP (g/100 mL)	998	0.04	0.06	0.20	$0.13_{(0.10)}$	$0.20_{(0.04)}$
CSN (g/100 mL)	998	0.03	0.04	0.13	$0.15_{(0.11)}$	$0.20_{(0.04)}$
CLA (g/100g FAME)	908	0.02	0.14	0.12	$0.09_{(0.06)}$	$0.50_{(0.05)}$
SCS	994	0.16	0.72	4.18	$0.03_{(0.07)}$	$0.14_{(0.03)}$
pH (U)	1.002	0.003	0.006	0.01	$0.16_{(0.08)}$	$0.30_{(0.05)}$
NaCl (mg/100 mL)	998	543.1	289.1	1104.7	$0.28_{(0.13)}$	$0.15_{(0.03)}$

[1] RCT = rennet coagulation time; $logk_{20}$ = log_{10} of curd firming rate; a_{30} = curd firmness; ILCY = individual laboratory cheese yield; PPCY = Predicted Pecorino Cheese Yield; MY = test day milk yield; FP = fat; PP = protein; CSN = casein; CLA = Conjugated linoleic Acid, FAME = fatty acid methyl esters; SCS = somatic cell score $log_2(SCC\mu l^{-1}/100+3)$, U = pH unit.

[2] Number of samples used to estimate variance components.

Heritability, phenotypic and genetic correlations of coagulation traits

Heritability estimates (Table 3) were moderate for RCT ($h^2 = 0.23$) and lower for a_{30} and k_{20} ($h^2 = 0.15$). The estimate for ILCY ($h^2 = 0.16$) was as twice as the values by Othmane et al., (2002) heritability for PPCY was of the same magnitude of ILCY. Values for other milk traits ranged from 0.03 (SCS) to 0.28 (NaCl content). Intermediate values were observed for MY (0.08), FP (0.10), PP (0.13) and CSN (0.15) (Table 3). Moderate to high phenotypic correlations were observed among coagulation traits (Table 4). The curd firmness at 30 min showed a negative correlation with RCT and k_{20}. Conversely, k_{20} was positively associated with RCT ($r_P = 0.84$). ILCY and PPCY presented moderate to low positive correlations with RCT-k_{20} and negative with a_{30}, respectively. Genetic correlations between MCPs were large and negative, those involving a_{30}, positive (between RCT and k_{20}) respectively. The largest value was for the correlation between k_{20} and a_{30} ($r_A = -0.91$). Genetic correlations for a_{30} and RCT are in the range of variability of MCPs as recently reviewed by Bittante et al., (2012). Unexpected results were the negative genetic correlation between ILCY and a_{30} as well as the positive correlations between ILCY, RCT ($r_A = 0.55$) and k_{20} ($r_A = 0.64$). Conversely, PPCY was positively associated with a_{30} ($r_A = 0.22$) and negatively with k_{20} ($r_A = -0.19$) even if the magnitude of these estimates were lower than those involving ILCY and with larger standard errors. PPCY and ILCY were moderately associated ($r_A = 0.47$). Heritabilities of MCPs and ILCY estimated by the MT animal model were very close to those obtained with the ST model, with lower standard errors though (data not shown).

Table 4
Phenotypic correlation (below the diagonal) and genetic correlation (above the diagonal) between sheep milk traits and milk coagulation properties estimated with a 5-trait animal model (standard error in brackets).

Trait [1]	RCT	logk20	a30	ILCY	PPCY
RCT (min)	-	$0.84_{(0.09)}$	$-0.80_{(0.12)}$	$0.55_{(0.15)}$	$0.08_{(0.21)}$
logk20 (min)	0.79	-	$-0.91_{(0.09)}$	$0.64_{(0.11)}$	$-0.19_{(0.16)}$
a30 (mm)	–0.60	–0.76	-	$-0.67_{(0.08)}$	$0.22_{(0.17)}$
ILCY (%,w/v)	0.41	0.32	–0.34	-	$0.47_{(0.18)}$
PPCY (%,w/v)	0.23	0.07	–0.13	0.51	-

[1]RCT = rennet coagulation time; k_{20} = curd firming rate; a_{30} = curd firmness; ILCY = individual laboratory cheese yield; PPCY = Predicted Pecorino Cheese Yield.

Phenotypic and genetic correlations between milk coagulation and milk quality traits

Phenotypic correlations among the three MCPs and milk traits were negligible apart from those involving SCS, pH and NaCl (Table 5). In particular, pH was negatively and moderately correlated with a_{30} and pH ($r_P = -0.42$) and positively and strongly with RCT, respectively. NaCl and SCS were moderately correlated with three MCPs. Moreover, RCT only was weakly associated with PP and CSN. Phenotypic correlation of ILCY (PPCY) with fat and protein percentage were r_P=0.46 ($r_P = 0.91$) and $r_P = 0.37$ ($r_P = 0.72$), respectively.

The majority of genetic correlations among MCPs and milk production and composition variables were close to zero (MY, FP with RCT) or presented very large standard errors (FP, PP, CS, SCS with a_{30}). Of interest are those between pH and RCT ($r_A = 0.68$) and pH and a_{30} ($r_A = -0.83$). RCT was also moderately correlated to casein ($r_A = 0.44$), but unexpectedly close to zero genetic association (with large standard errors though) were found among protein and casein and a_{30}. Moreover, both RCT and k_{20} showed a positive genetic association with NaCl, whereas, no reliable associations were found with functional compound like CLA. As expected, ILCY was positively correlated with milk composition (FP, PP and CSN) and negatively correlated with MY. High trivial genetic correlation were observed among fat, protein and casein and PPCY, and although of reduced magnitude when compared to ILCY, a negative correlation between PPCY and MY was observed.

Table 5
Phenotypic (r_P) and genetic correlation (r_G) among coagulation traits analyzed with bi-variate animal model in combination with sheep milk yield and composition

Trait[1]	r_P					r_G				
	RCT	logk20	a30	ILCY	PPCY	RCT	logk20	a30	ILCY	PPCY
MY (L/day)	–0.09	0.07	–0.04	–0.09	–0.15	$0.03_{(0.39)}$	$0.04_{(0.48)}$	$0.27_{(0.46)}$	$–0.88_{(0.42)}$	$–0.60_{(0.46)}$
FP (g/100mL)	0.09	0.03	–0.12	0.46	0.91	$–0.02_{(0.38)}$	$–0.34_{(0.42)}$	$0.32_{(0.18)}$	$0.45_{(0.31)}$	$0.93_{(0.06)}$
PP (g/100mL)	0.30	–0.04	0.03	0.37	0.72	$0.41_{(0.32)}$	$–0.42_{(0.42)}$	$0.09_{(0.37)}$	$0.75_{(0.27)}$	$0.85_{(0.13)}$
CSN (g/100mL)	0.29	–0.06	0.05	0.38	0.73	$0.44_{(0.11)}$	$–0.43_{(0.45)}$	$0.00_{(0.48)}$	$0.65_{(0.27)}$	$0.84_{(0.13)}$
CLA (g/100g FAME)	–0.01	–0.05	0.14	–0.25	–0.37	$–0.27_{(0.38)}$	$–0.46_{(0.46)}$	$0.27_{(0.40)}$	$–0.32_{(0.34)}$	$–0.33_{(0.36)}$
SCS (U)	0.45	0.35	–0.30	0.35	0.29	$–0.14_{(0.92)}$	$–0.72_{(1.35)}$	$0.11_{(0.77)}$	$0.58_{(0.61)}$	$0.29_{(0.93)}$
pH (U)	0.70	0.55	–0.42	0.18	–0.28	$0.68_{(0.19)}$	$0.44_{(0.34)}$	$–0.83_{(0.21)}$	$0.58_{(0.42)}$	$–0.21_{(0.53)}$
NaCl (mg/100mL)	0.45	0.44	0.35	0.08	0.09	$0.52_{(0.27)}$	$0.68_{(0.32)}$	$0.05_{(0.51)}$	$0.87_{(0.65)}$	$0.24_{(0.51)}$

MY = test day milk yield; FP = test day fat percentage; PP = test day protein percentage; CSN = test day casein percentage; CLA = Conjugated linoleic Acid, FAME=fatty acid methyl esters; SCS = somatic cell score $\log_2[(SCC\mu l^{-1}/100)+3]$, RCT = rennet coagulation time; k_{20} = curd firming rate; a_{30} = curd firmness; ILCY= individual laboratory cheese yield; PPCY=Predicted Pecorino Cheese Yield.

Discussions

In general, studies on milk rheological properties are characterised by a relevant variability of results. Moreover, several variables affecting clotting properties have been identified so far (Bittante et al., 2012). In the present study, some milk samples did not coagulate within the reference time of 30 minutes. The percentage of non-coagulating milks was larger than in previous studies on Sarda and Alpine breeds (Bittante et al., 2014; Pazzola et al., 2014), but smaller than the value (24%) observed on Sarda bulk milk (Giangolini et al., 2004). The result of the present study are similar to those observed in dairy cattle where the proportion of samples that did not coagulate and those with missing k_{20} are on average 19% and 33% across studies, respectively (Bittante et al., 2012). According to some authors, the milk coagulation process should be faster in ovine than bovine milk (Bittante et al., 2012; Pazzola et al., 2014). The average RCT found in the present study does not confirm this hypothesis. The values measured in Sarda and Alpine were twofold (Bittante et al., 2014; Pazzola et al., 2014). On the other hand, it is in agreement with results obtained in other studies on Sarda (Pirisi et al., 2000; Mele et al., 2006) Massese (Pugliese et al., 2000; Martini et al., 2008) and Spanish (Jaramillo et al., 2008; Rovai et al., 2015) sheep breeds. Average values and distributions of k_{20} and a_{30} are in agreement with a previous report on Sarda ewes (Pazzola et al., 2014). Average ILCY measured in the present study was similar to those estimated by Jaramillo et al., (2008) but 10% higher than previous finding on Churra sheep (Othmane et al., 2002). Anyhow, the actual cheese yield is clearly overestimated by the use of ILCY. Whereas the PPCY were in accordance to the average Pecorino Romano cheese yield (Pirisi et al., 2002) and it was moderately correlated with ILCY. The overestimation of cheese yield could be ascribed to the method of micro-manufacturing used (Othmane et al., 2002; Bonfatti et al., 2014) [explained later in the discussion]. Differently from what is observed in dairy cattle, the flock environment exerted a significant role. Compared to previous works on Sarda (Pazzola et al., 2014; Vacca et al., 2015) the proportion of variance explained by flock-test day was similar for k_{20}, but slightly lower for RCT and a_{30}. The fraction of variance explained by flock for MCPs was dramatically higher in comparison with studies on cattle (Ikonen et al., 2004; Tyrisevä et al., 2004; Vallas et al., 2010), probably due to the peculiarities of sheep farming.

Heritability of milk coagulation, composition and cheese yield traits

For some traits a significant quota of phenotypic variance was additive genetic, in other cases the majority of the variation was of environmental nature. Estimates of heritability for RCT were moderate and just in few cases presented small standard errors. The comparison can be made only with dairy cattle data due to the lack of information for sheep in the literature. Values obtained in the present study confirmed reports for dairy cattle (Ikonen et al., 1999; Tyrisevä et al., 2004; Cassandro et al., 2008). However, RCT heritability was below the findings of Ikonen et al., (2004) and Vallas et al., (2010). In the case of a_{30} the heritability was of the same extent of other studies (Cassandro et al., 2008; Cecchinato et al., 2011) but sensibly lower than values obtained by Ikonen et al., (1999; 2004) and Tyrisevä et al., (2004). The k_{20} parameter had a similar heritability of a_{30} but few reports have been found on heritability of k_{20} in literature (Bittante et al., 2012).The heritability estimate of ILCY was double in magnitude compared to values reported by Othmane et al., (2002) whose estimates derived from a sample of similar size, even if with ~7,500 test-day records of sheep milk over two generations. The cheese yield equation-predicted on the basis of fat and protein percentage had a heritability of the same magnitude of ILCY. Heritabilities for milk composition traits were from low to intermediate. Values for fat and protein were markedly lower than those reported for Sarda sheep (Sanna et al., 1997). The use of one test day per animal and the reduced sample size in comparison to other works, may at least partially justify these differences. However, values observed in the present study were close to those reported by Othmane et al., (2002) and they were in the range of variability observed for dairy sheep (Oravcová et al., 2005).

Phenotypic and genetic correlations between milk coagulation traits and cheese yield

The knowledge of genetic associations among coagulation, milk yield and quality traits is essential when exploring the possibility to select in favour of one of the MCPs traits. The overall phenotypic correlation pattern of MCPs confirmed what observed in Sarda and Churra breeds (Nudda et al., 2001; Jaramillo et al., 2008), whereas partially disagree with the results of Pazzola et al., (2014). The latter authors

found a low negative correlation between RCT and a_{30} (–0.15) indicating a substantial phenotypic independency between these two traits. In the present paper, moderate phenotypic and high negative genetic correlations were obtained between these two traits, respectively. This result is similar to previous reports in dairy cattle (Cassandro et al., 2008, Ikonen et al., 2004; Bittante et al., 2012). Indeed, if milk takes less time to coagulate, then more time is available for the process of curd firming. As the repeatability of RCT is quite higher (Bittante et al., 2012), this means that even in sheep one measure of RCT is enough to predict both traits. Furthermore, the strong negative phenotypic correlation between k_{20} and curd firmness at 30 min was expected, due to the positive association of k_{20} with RCT. Similar relationships were previously found in Sarda sheep (Pazzola et al., 2014). Additive genetic variances that is common between these two traits (k_{20}, a_{30}) have been scarcely investigated in dairy cattle, due to the higher percentage of missing values for k_{20} parameter. Negative correlation between ILCY and a_{30} (either phenotypic or genetic) and positive correlations between ILCY and RCT-k_{20} were unexpected. Conversely, PPCY presented a weak positive genetic association with a_{30} and negative with k_{20}. The possible explanation for this correlation pattern can be formulated considering two conflicting aspects. The first is the interaction between predicted cheese yield and a_{30} as function of the fat percentage (Aleandri et al., 1989). These authors found that the predicted cheese yield was positively associated with a_{30} with low fat milk, and negatively associated to a_{30} with high fat milk. Hence, the high fat level of ovine milk, compared to cow milk, could partially explain our results. A second issue is represented to the method used for measuring cheese yield. Indeed micro-cheese factoring can produce biased estimation of actual cheese yield, due to the small amount of milk used. This fact is also confirmed by the overestimation of cheese yield, found also in other works (Othmane et al., 2002; Jaramillo et al., 2008). Moreover, Bonfatti et al., (2014) found that cow milk with short RCT and high a_{30} did not exhibit higher cheese yield in model cheeses, being the cheese yield variation in their experiment more likely associated to variation in milk fat and protein percentages. The modest genetic correlation between the cheese yield predicted by regression (PPCY) and a_{30} seem to suggest this second hypothesis, even if further investigations are needed to clarify the relationship between ILCY and MCPs in sheep milk.

Phenotypic and genetic correlation among milk coagulation, milk yield and composition

The study of the genetic associations between MCPs and milk yield and chemical composition is crucial for evaluate proper selection strategies. The phenotypic correlations between RCT and protein and casein percentages found in the present study were in agreement with results on sheep (Jaramillo et al., 2008; Nudda et al., 2001) and in cattle (Bittante et al., 2012). The worsening of the coagulation properties of sheep milk (>RCT and $<a_{30}$) with increased somatic cell count is documented in literature (Pirisi et al., 2000; Nudda et al., 2001; Raynal-Ljutovac et al., 2007). An increased somatic cell count can also be the result of intramammary inflammatory process (Rovai et al., 2005). However in sheep, high somatic cell count in milk can be often unrelated with pathological conditions, differently from cow. Several factors (breed, parity, stage of lactation, type of birth, estrus, diurnal) affect SCC variation in sheep milk (Raynal-Ljutovac et al., 2007). Increased RCT and k_{20} and reduced a_{30} with increasing pH were previously reported in sheep milk (Pirisi et al., 2000; Bencini et al., 2002). Finally, low to moderate phenotypic correlation was observed among individual cheese yield and milk traits. The highest association was between ILCY and fat percentage, and it was half of the correlation found by Jaramillo et al., (2008) but agreed with the results of Othmane et al., (2002).

Interesting genetic correlations have been estimated between pH, casein, NaCl and RCT-a_{30}. For the pH, similar values have been reported for dairy cattle (Ikonen et al., 2004; Cassandro et al., 2008; Vallas et al., 2010; Cecchinato et al., 2011). Moderate correlations between RCT and casein percentage and no association among protein, casein percentages and a_{30} were reported by Ikonen et al., (2004). On the other hand, results of the present study were opposite to what was found by some other authors (Cassandro et al., 2008; Cecchinato et al., 2011). A suggestive negative association between CLA with rennet properties might confirm what was found by Bittante et al. (2014) in milk of sheep supplemented with rumen-protected conjugated fatty acid source. Moderate to high positive genetic correlation were found between NaCl, a_{30}-k_{20}. Conversely unreliable negative genetic correlation were found between SCS and RCT, differently from what observed by other authors in cow milk (Ikonen et al., 2004; Cassandro et al., 2008; Cecchinato et al., 2011). A very high genetic

correlation among NaCl and SCS (0.98 ± 0.31) was found in the present study. This is an interesting result and it suggests the possibility to use indirect indicators of udder status different from SCS. As far as cheese yield is concerned, there is a positive genetic correlation between milk composition and ILCY and negative with MY, which confirm what was found by Othmane et al., (2002).

Conclusions

This study provided estimates of genetic parameters for milk coagulation properties of sheep milk of Sarda Breed. From the selective point of view, a not negligible proportion of phenotypic variance was additive genetic, and the heritability estimated for MCPs were in agreement with those found in cow milk for MCPs. Genetic correlations found in the present study suggest the chance to use only one of the rennet parameter, as they are highly genetic correlated, however negative correlation between ILCY and favourable rennet properties suggests to be careful in the use of this methods to predict cheese yield from small milk samples.

Acknowledgements

This research was supported by the RAS, Regional Law n 7, August 2007, Grant n. CRP 61608 "Il latte ovino della Sardegna". The author wish to thank the provincial breeders farmer association (APA) of Cagliari, Nuoro, Sassari and Oristano for their support in sample collection; the Laboratory of Associazione Regionale allevatori della Sardegna (Oristano, Italy) for performing milk composition analysis, the Italian association of animal breeders (AIA) and of sheep breeders (ASSONAPA) for providing pedigree information.

REFERENCES

Aleandri, R., Schneider, J.C, and, Buttazzoni, L.G., 1989. Evaluation of milk for cheese production based on milk characteristics and formagraph measures. Journal of Dairy Science 72, 1967-1975

BDN. Banca dati Nazionale dell'anagrafe zootecnica, 2015. Consistenza allevamenti Ovini e Caprini. Retrieved on 11 December 2015 from http://statistiche.izs.it/portal/page?_pageid=73,12918and_dad=portal

Bencini, R. 2002. Factors affecting the clotting properties of sheep milk. Journal of the Science of Food and Agriculture 82, 705-719

Bittante, G., Pellattiero, E., Malchiodi, F., Cipolat-Gotet, C., Pazzola, M., Vacca, G.M., Schiavon, S., and Cecchinato, A. 2014. Quality traits and modeling of coagulation, curd firming, and syneresis of sheep milk of Alpine breeds fed diets supplemented with rumen-protected conjugated fatty acid. Journal of Dairy Science 97, 4018-4028

Bittante, G., Penasa, M., and Cecchinato, A., 2012. Invited review: Genetics and modeling of milk coagulation properties. Journal of Dairy Science 95, 6843-6870

Bonfatti, V., Tuzzato, M., Chiarot, G., and Carnier, P., 2014. Variation in milk coagulation properties does not affect cheese yield and composition of model cheese. International Dairy Journal 39, 139-145

Cappio-Borlino, A., Portolano, B., Todaro, M., Macciotta, N.P.P., Giaccone, P., and Pulina, G., 1997. Lactation curves of Valle del Belice dairy ewes for yields of milk, fat, and protein estimated with test day models. Journal of Dairy Science 80, 3023-3029

Carta, A., Casu, S., and Salaris, S., 2009. Invited review: Current state of genetic improvement in dairy sheep. Journal of Dairy Science 92, 5814-5833

Cassandro, M., Comin, A., Ojala, M., Dal Zotto, R., De Marchi, M., Gallo, L., Carnier, P., and Bittante, G., 2008. Genetic parameters of milk coagulation properties and their relationships with milk yield and quality traits in Italian Holstein cows. Journal of Dairy Science 91, 371-376

Cecchinato, A., Penasa, M., De Marchi, M., Gallo, L., Bittante, G., and Carnier, P., 2011. Genetic parameters of coagulation properties, milk yield, quality, and

acidity estimated using coagulating and noncoagulating milk information in Brown Swiss and Holstein-Friesian cows. Journal of Dairy Science 94, 4205-4213

De Marchi, M., Bittante, G., Dal Zotto, R., Dalvit, C., and Cassandro, M., 2008. Effect of Holstein Friesian and Brown Swiss breeds on quality of milk and cheese. Journal of Dairy Science 91, 4092-4102

Eurostat 2015. Number of Sheep. Retrieved on 12 February 2016 from http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&init=1&plugin=1&language=en&pcode=tag00017.

Food and Agriculture Organization of the United Nation Statistics Division 2014. Statistical Database of the Food and Agriculture Organization of the United Nation. Retrieved on 06 December 2015 from http://faostat3/fao.org/home/E

Furesi, R., Madau, F.A., and Pulina, P., 2013. Technical efficiency in the sheep dairy industry: an application on the Sardinian (Italy) sector. Agricultural and Food Economics 1, 1-11

Giangolini, G., Amatiste, S., Battisti, S., Boselli, C., Filipetti, F., Rosati, R., Fagiolo, A., 2004. Clotting properties of Sarda ewe's bulk milk coming from flocks raised in two province of the latium region, in proceeding XVI congresso nazionale della società italiana di allevamento degli ovini e dei caprini (SIPAOC), 29 September 2004,Siena, Italy.

Groeneveld, E., Kovac, M., Mielenz, N., 2010. VCE User's Guide and Reference Manual Version 6.0. Retrieved on 10 September 2015 from ftp://ftp.tzv.fal.de/pub/vce6/doc/vce6-manual-3.1-A4.pdf

Ikonen, T., Ahlfors, K., Kempe, R., Ojala, M., and Ruottinen, O., 1999. Genetic parameters for the milk coagulation properties and prevalence of noncoagulating milk in Finnish dairy cows. Journal of Dairy Science 82, 205-214

Ikonen, T., Morri, S., Tyriseva, A.M., Ruottinen, O., and Ojala, M., 2004. Genetic and phenotypic correlations between milk coagulation properties, milk production traits, somatic cell count, casein content, and pH of milk. Journal of Dairy Science 87, 458-467

International Commitee for animal Recording, 2014. Dairy sheep milk survey online database. Retrieved on 8 February 2016 from http://www.icar.org/survey/pages/tables.php.

Jaramillo, D.P., Zamora, A., Guamis, B., Rodriguez, M., and Trujillo, A.J., 2008. Cheesemaking aptitude of two Spanish dairy ewe breeds: Changes during lactation and relationship between physico-chemical and technological properties. Small Ruminant Research 78, 48-55

Manca, M.G., Serdino, J., Gaspa, G., Urgeghe, P., Ibba, I., Contu, M., Fresi, P. and N. P. P. Macciotta, 2016. Derivation of multivariate indices of milk composition, coagulation properties, and individual cheese yield in dairy sheep. Journal of Dairy Science, http://dx.doi.org/10.3168/jds.2015-10589, Published online 6 April 2016

Martini, M., Mele, M., Scolozzi, C., and Salari, F., 2008. Cheese making aptitude and the chemical and nutritional characteristics of milk from Massese ewes. Italian Journal of Animal Science 7, 419-437

Mele, M., Buccioni, A., Petacchi, F., Serra, A., Banni, S., Antongiovanni, M., and Secchiari, P., 2006. Effect of forage/concentrate ratio and soybean oil supplementation on milk yield, and composition from Sarda ewes. Animal Research 55, 273-285

Nudda, A., Feligni, G., Battacone, G., Murgia, P., and Pulina, G., 2001. Relationship between somatic cells count, whey protein and coagulation properties in sheep milk. in Proceeding of 14th ASPA congress, 12-15 June 2001, Firenze, Italy, pp 511-513

Nudda, A., McGuire, M.A., Battacone, G., and Pulina, G., 2005. Seasonal variation in conjugated linoleic acid and vaccenic acid in milk fat of sheep and its transfer to cheese and ricotta. Journal of Dairy Science 88, 1311-1319

Oravcová, M., Groeneveld, E., Kovac, N., Peškovicová, D., Margetin, M., 2005. Estimation of genetic and environmental parameters of milk production traits in Slovak purebred sheep using test-day model. Small Ruminant Research 56, 113-120

Othmane, M.H., De La Fuente, L.F., Carriedo, J.A., and San Primitivo, F., 2002. Heritability and genetic correlations of test day milk yield and composition, individual laboratory cheese yield, and somatic cell count for dairy ewes. Journal of Dairy Science 85, 2692-2698

Pazzola, M., Dettori, M.L., Cipolat-Gotet, C., Cecchinato, A., Bittante, G. and Vacca, G.M., 2014. Phenotypic factors affecting coagulation properties of milk from Sarda ewes. Journal of Dairy Science 97, 7247-7257

Pirisi, A., Murgia, A., Scintu, M.F., 1994. Estimate of Pecorino Romano and Pecorino Sardo cheese yield from the protein and fato contents in sheep milk. Scienza e Tecnica Lattiero Casearia 45, 476-483

Pirisi, A., Piredda, M., Corona, M., Pes, S., Pintus, S., Ledda, A. 2000. Influence of somatic cell count on ewe's milk composition, cheese yield and cheese quality. in Proceeding Great Lakes Dairy Sheep Symposium, 2-4 November 2000, Guelph, Canada, Pages 47-59

Pirisi, A., Piredda, G., and Carta, A., 2002. Resa di trasformazione del latte ovino: applicabilità alla produzione industriale di un'equazione ricavata mediante microcaseificazioni. In proceeding of XV Congresso nazionale della società italiana di allevamento degli ovini e dei caprini (SIPAOC), 11-14 September 2002, Cagliari, Italy

Pugliese, C., Acciaioli, A., Rapaccini, S., Parisi, G., and Franci, O., 2000. Evolution of chemical composition, somatic cell count and renneting properties of the milk of Massese ewes. Small Ruminant Research 35, 71-80

Raynal-Ljutovac, K., Pirisi, A., de Crémoux, R., and Gonzalo, C., 2007. Somatic cells of goat and sheep milk: Analytical, sanitary, productive and technological aspects. Small Ruminant Research 68, 126-144

Rovai, M., Natalia, R., Gerardo, C., Jordi, S., and Gabriel, L., 2015. Effect of subclinical intrammamay infection on milk quality in dairy sheep: I. Fresh-soft cheese produced from milk of uninfected and infected glands and from their blends. Small Ruminant Research 125, 127-136

Sanna, S.R., Carta, A., and Casu, S., 1997. (Co)variance component estimates for milk composition traits in Sarda dairy sheep using a bivariate animal model. Small Ruminant Research 25, 77-82

Tyriseva, A.M., Vahlsten, T., Ruottinen, O., and Ojala, M., 2004. Noncoagulation of milk in Finnish Ayrshire and Holstein-Friesian cows and effect of herds on milk coagulation ability. Journal of Dairy Science 87, 3958-3966

Vacca, G.M., Pazzola, M., Dettori, M.L., Pira, E., Malchiodi, F., Cipolat-Gotet, C., Cecchinato, A., and Bittante, G., 2015. Modeling of coagulation, curd firming, and syneresis of milk from Sarda ewes. Journal of Dairy Science 98, 2245-2259

Vallas, M., Bovenhuis, H., Kaart, T., Paerna, K., Kiiman, H., and Paerna, E., 2010. Genetic parameters for milk coagulation properties in Estonian Holstein cows. Journal of Dairy Science 93, 3789-3796

CHAPTER 3

Research paper

Effect of different Data Reduction on the Accuracy of Genomic Prediction on an Imputed High Density SNP panel in Italian Holstein cattle.

Abstract

In this study the effect of principal component analysis (PCA) on the accuracy of Direct Genomic Values (DGV) for 32 traits of 2.822 Italian Holstein bulls was assessed under different conditions of predictor dimensionality and repartition. A sample of 2.924 Italian Holstein bulls were genotyped using different marker density: 916 were genotyped using the Illumina High Density (800K) beadchip, the remaining 2.008, using the Illumina Medium Density beadchip. The latter bulls were imputed from medium to high density through *FImpute* program. A number of 610.415 markers were retained after data editing and from these, three panels with different marker density were used: one high-density (HD) containing all markers which passed the Quality Control (QC) and two medium-density (MD) panels that were simulated from the HD in two ways: 1) by merging HD with the BovineSNP50v2 (39.884 SNPs) and 2) by selecting starting marker from haplotype blocks (64.647 SNPs). From these panels, based on marker partitioning, four scenarios of markers reduction using Principal Component Analysis (PCA) were tested. PCA was carried out: 1) on the entire genome and 2) by chromosome in the MD obtained by merging, 3) by chromosome in MD obtained from haplotypes blocks and 4) by (sub)-chromosome on HD panel for computational reasons. A number of PCs explaining 80, 90 and 99 percent of the total variance was retained for each scenario. The impact of using PC scores as predictors instead of SNP markers was assessed comparing accuracies of DGVs estimations (*r*DGV). The effect of predictors were firstly estimated on 2.301 training bulls and subsequently used to estimate DGVs on 521 younger validation bulls. Accuracy of predictions was computed as correlation between DGVs and phenotypes in validation animals. Average *r*DGVs across 32 traits were compared to demonstrate the effect of the repartition and reduction (through PCA, haplotype pre-selection and both of them) on the DGV estimation. An increase in accuracy was generally observed with the increase of the number of PCs retained and with the reduction of the asymmetry between the number of predictors and observations. The shrinkage of the total variance of different SNPs panel size in a quite similar number of PCs (caused by the limited population size) occurring during PCs extraction, could be seen as possible explanation of differences in rDGVs obtained across scenarios. The better repartition

of variance in the MD resulted in higher rDGV in comparison with HD. Finally rDGVs obtained through PCA were compared to those obtained using the full panel of SNP markers with effects estimated using a commonly available program for Genomic Selection.

Keywords: genomic selection, direct genomic value, accuracy, principal component analysis.

Introduction

Genomic selection (GS) is based on the integration of single nucleotide polymorphism (SNP) and phenotypic information available in a reference population with the aim to estimate genomic enhanced breeding values on younger animals that have only genotypes (Meuwissen et al., 2001). Thanks to their high coverage along the genome, it is assumed that at least one SNP is in linkage disequilibrium with a chromosomal region affecting the trait (Goddard and Hayes, 2009). The simultaneous estimation of the effects of all markers on the trait of interest in the reference population and their use to obtain direct genomic values (DGV) on young candidates allows an early use of bulls in the selection programmes (VanRaden et al., 2009) leading to a reduction in the generation interval. Moreover the use of the genomic instead of traditional pedigree-based relationship matrix, is able to increase the accuracies of the predicted breeding values (Hayes et al. 2009)

However, the accuracy of the DGV relies on the size of the reference population, the degree of relatedness among individuals, the density of the marker panels used and the estimation methods (Goddard and Hayes 2009.). Actually, one way to increase the number of animals in the reference population, overcoming the increase of the costs associated to the GS, is the imputation of the missing genotypes when dealing with bead-chips of different density. Imputation consists in the use of the known haplotype of a set of reference animals (ancestors) genotyped with a high density SNP panel or with whole sequence data, to reconstruct the missing genotypes of a population genotyped with a lower density SNP chip. The inclusion of a larger number of genotyped animals allows the enlargement of the reference population (including also elite cows) in a cost-effective manner with an acceptable degree of accuracy. Different strategies of genotype imputation have been proposed: albeit not compulsory, some requires pedigree information (Sargolzaei et al., 2011; Hickey et al., 2012; VanRaden et al., 2013) whereas others, developed for human genetics, do not such as Beagle (Browning and Browning, 2009) and IMPUTE2 (Howie et al., 2009). Alternative imputation methods based on partial least square regression have been also proposed (Dimauro et al., 2013). The huge increase of molecular information deriving from real or *in silico* genotyping process generates a larger

asymmetry between variables (SNP markers) and observations (animals) especially when dealing with HD panels of markers or whole genome sequence data.

Strategies for the reduction of the data asymmetry are different varying from the manipulation of the number of animals includes in the reference populations and/or to the modification of the number of variables (generally exceeding the number of observations). The enlargement of the population size is generally obtained by combining genotyped animals belonging from the same (VanRaden et al., 2013) or different breeds (Gaspa et al., 2015; Lund et al., 2016). The reduction of the number of predictors is obtainable through the application of sometime time-consuming variable selection methods (Gianola et al., 2009; Verbyla et al., 2009) or different dimensionality reduction tools.

Reduction could be reached for example either using algorithm for block partitioning or applying multivariate methods (Solberg et al., 2009, Macciotta et al., 2010). In the first case the reduction consists in the passage from the marker to the haplotype level, by reducing the sequence of markers in a fewer number of sequence (the blocks) including markers characterized by the same level of linkage disequilibrium (LD) (Pattaro et al., 2008). In the second case, multivariate methods are applied to obtain a small number of predictors linearly combined with original variables (Solberg et al., 2009). In the last case the asymmetry of data matrix, resulting in the known *S rank* issue was identified as the cause of the not reliable distribution of the original variance into the replacement variables and the reason of spurious results obtained by applying multivariate approaches on the DGV accuracy.

Splitting the genome by chromosomes, on the basis of their biological ortogonality and performing the predictor's reduction at this "chromosome wide" level, avoiding the not full rank of the variance-covariance square matrices characterizing deriving from whole genome datasets, is a way to obtain reliable results without loss of information (Dimauro et al., 2011). The increase of number of SNP markers included in the HD panels has reduced the benefits obtained by the application of the chromosome wide approach. Under an assumption of an equal repartition of markers along the genome, the subdivision of chromosomes in equally-spaced blocks of markers could be seen as possible way to overcome the rank problem.

Aim of this study was to investigate the effects of different ways of marker repartition obtained with the application of a multivariate predictor reduction technique to obtain genomic evaluation of Italian Holstein bulls genotyped with HD SNP panels.

Materials and methods

Data

A high density panel including autosomal genotypes distributed on 29 chromosomes (no sexual chromosomes were available) of 2.924 Italian Holstein bulls was used in this study. Genotypes were partially obtained through imputation. A training population (TP) composed by 916 bulls was genotyped with the BovineHD Beadchip (Illumina Inc., San Diego, CA). Genotypes of 2.008 bulls of the prediction population (PP) were initially obtained with the BovineSNP50 Illumina Beadchip and then imputed to High Density using *FImpute* program (Sargolzaei et al., 2014). Editing on genotypes was performed using Plink v1.90 (Purcell et al., 2015) on the basis of minor allele frequency (MAF<0.01), deviance from Hardy-Weinberg equilibrium (p<0.001), absence of heterozygous and monomorphic loci. No missing data were present in the dataset due to imputation. After quality control (QC), SNPs were coded in the '012' format, according to the number of copies of SNP allele, where 0 and 2 was used to indicate homozygous, 1 for heterozygous. Finally 610.415 markers passed the QC and were retained in the HD panel.

The phenotypes used in the analysis were de-regressed proofs (DRPF) of 32 traits provided by the Italian Holstein Breeder Association (ANAFI). In particular DGVs accuracies (*r*DGVs) were calculated for production, conformation and functional traits. Reliability of DRPF and heritability (h^2) of all traits were also provided by ANAFI (table 1)

Table 1
List of traits,values of heritability (h^2) and reliability (r^2_{EBV}) of traits used in the study. Data provided by ANAFI.

Trait	**h^2**	**r^2_{EBV}**
1. Milk yield	0.31	0.91
2. Fat Kg	0.29	0.93
3. Protein Kg	0.30	0.91
4. Fat (%)	0.50	0.96
5. Protein (%)	0.50	0.95
6. SCC	0.21	0.87
7. Conformation	0.24	0.75
8. Stature	0.40	0.83
9. Strength and vigor	0.25	0.80
10. Body Depth	0.30	0.81
11. Angularity	0.23	0.79
12. Rump Angle	0.23	0.78
13.Rump width	0.22	0.78
14.Rearleg side view	0.16	0.74
15. Foot angle	0.10	0.68
16. Rear leg rear view	0.06	0.60
17. Feet and Legs functionality	0.10	0.68
18. Front Udder Attachment	0.20	0.78
19. Rear udder height	0.21	0.77
20. Rear udder width	0.21	0.75
21. Udder support	0.16	0.74
22. Udder depth	0.30	0.81
23. Front teat position	0.19	0.78
24. Teat length	0.19	0.80
25. Rear teat position	0.14	0.81
26. Calving Interval	0.04	0.55
27. Days first insemination	0.06	0.58
28. Direct calving ease	0.08	0.77
29. Maternal calving ease	0.04	0.36
30. Direct Longevity	0.11	0.60
31. Combined Longevity	0.10	0.69
32. Fertility	0.05	0.54

Experimental design

Three panel densities and five markers partitioning were generated for this study: one high-density panel (HD) with two marker partitions, and two different medium-density panels (MD) with one and two marker partitions, respectively.

The HD panel was composed by all markers who passed QC (610.415 SNPs). The multivariate predictor reduction on HD, for computational reasons, was applied after its repartition in a *"chromosome wide"* level (HD_CHR). Moreover to see how the rank problem generated by the disproportion between variables and records may be still able to inflate estimates in the HD_CHR, a sub-chromosomal repartition of markers was performed on HD (HD_SPLIT). In this last case each chromosome was split into 10 parts having a number of markers evenly distributed, creating a total of 290 datasets. Marker reduction was finally applied on each of them.

The first type of MD was obtained by merging the whole set of markers in HD with the BovineSNP50v2 marker's map and extracting only the markers sheared between the two panels. A total of 39.884 overlapping markers were retained. In this case the predictor reduction was applied on two level of markers repartition: 1) on the whole set of markers in a *"genome wide"* level (MD_GW) and 2) after the repartition at the chromosome level (MD_CHR).

The second type MD panel, named MD_BLOCK, was generated through the haplotype block partitioning of the HD panel. More specifically, chromosomes of HD were split into blocks of markers having the same level of linkage disequilibrium (LD). For each block the first SNP was selected and pooled (within chromosome) with markers preselected at the same manner. Finally the panel containing the SNPs preselected by block partitioning was subjected to the chromosome wide multivariate marker reduction.

Animals with all available phenotypic records were kept in the analysis. Finally, 2.822 resulting bulls were sorted by year and divided into Reference (REF) and Validation (VAL) subsets. In particular, oldest bulls born between 1964 and 2003 and youngest born between 2004 and 2011 were included in the REF and VAL populations, respectively. Accuracies of DGVs obtained after the abovementioned strategies of data reduction were finally compared to those obtained on the same traits with the same model but using the whole set of SNP markers (HD_BLUP).

Principal Component Analysis

Principal Component analysis (PCA) was the statistical tool used for obtaining the multivariate reduction of the (*p)* number of markers (*X*) by producing a set of uncorrelated (*k*) predictors, re-oriented along new spatial directions on the basis of their variance distribution (Krzaznowsky, 2000). Predictors are named Principal Components (PCs) and are composed by linear combination of the original variables as following:

$$PC_j = a_{1j} X_1 + a_{2j} X_2 + \ \ldots \ + \ a_{pj} X_p$$

where a_{pj} are the coefficients with j = 1,...,p (weights) allowing the conversion of the original variables into the j_{th} principal component. PCA starts from the generation of the correlation matrix of the original variables (R). Weights are elements of the vectors (*eigenvectors*) associated to the *eigenvalues* of the R matrix, symbolized by λ and representing the variance of the relative PCs. Eigenvalues are sorted in decreasing order:

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p \geq 0$$

in manner that their cumulative sum corresponds to the total variance of the system, corresponding to the trace of *R* (*trR*) and symbolized by:

$$\sum_{j}^{p} \lambda_j = trR$$

and the variance explained by the j_{th} principal component as the ratio of its eigenvalue on the total variance:

$$\sigma^2_{PC_j} = \frac{\lambda_j}{trS}$$

The reduction of the system is obtained by retaining a number of principal components lesser than the original number of variables ($k << p$).

There are three main criteria are commonly available to impose the reduction limit:

1) Kaiser's criterion: only PCs whose eigenvalues are greater than 1 are kept.
2) Cumulative proportion of the variance explained by the PCs.
3) Point of truncation in the *Scree Plot.*

Once determined the number of PCs to be retained on the basis of one of the abovementioned criteria, the scores of PCs are fitted into models to estimate genomic breeding values instead of the use of SNP markers.

In this study the cumulative variance was the criterium adopted for the selection of the number of PCs to be retained. The data reduction was obtained on each scenario by selecting a number of PCs able explain the 80%, 90% and 99% of the total original variance, respectively.

PCA was applied both on MD and HD panels under the five scenarios of marker repartition previously described and here summarized:

i) MD panel genome-wide (MD_GW);

ii) MD panel chromosome-wide (MD_CHR);

iii) MD panel with SNP pre-selected on the basis of LD (MD_BLOCK);

iv) HD panel chromosome-wide (HD_CHR);

v) HD panel at sub-chromosomal level (HD_SPLIT).

Principal Component extraction was performed by applying the "*prcomp*" function implemented in R software. This procedure is based on the Singular Value Decomposition (SVD) of the matrix. That procedure is particularly recommended when the numbers of variable (in the present case the markers) strongly exceed the number of observations (animals), a common problem when using genomic data. With the SVD the original data matrix is decomposed into three sub-matrices: an orthogonal (n,n) matrix U, a diagonal (n,p) matrix Σ (rectangular with the same

dimensions of the original matrix) and a (*p*,*p*) matrix V' corresponding to the transpose of the orthogonal matrix. The diagonal entries of Σ (positive values in decreasing order) are called *"singular values"*, whereas the columns of U and V are called *"left"* and *"right singular vectors"*. The Principal Component extraction is obtained by the product of the three matrices in which the original data matrix is factorized ($A = U \cdot \Sigma \cdot V'$). Once performed the Principal Component extraction, in manner to determine the number of PCs that had to be retained, the cumulative proportion of PCs, calculated as the cumulative variance explained by their eigenvalues, was accounted for until reaching the threshold of variance imposed.

Statistical models

PCA BLUP method

Scores of the PCs were subsequently used as new variables to estimate DGV and relative accuracies for the PC (PCA-BLUP). The effects of PCA scores predictors on phenotypes were estimated on the REF population using the following mixed linear model:

$$y = 1\mu + Zg + e \qquad [1]$$

where *y* is the vector of DRPF, *1* is a vector of ones, μ is the overall mean, *Z* is the design matrix of predictors (PCs scores or SNP genotypes), *g* is the vector of their random effects and *e* is the vector of the random residuals. Covariance matrices of PC scores (*G*) and residual (*R*) effects were assumed to be normally distributed with mean zero and (co)variances $I\sigma^2_g$ and $I\sigma^2_e$, respectively, with $\sigma^2_g = \sigma^2_a / p$ (*p* = number of predictors) assuming an equal contribution of each predictors to the total additive variance. Variance components (σ^2_g and σ^2_e) were provided by ANAFI. BLUP solutions were predicted solving the mixed model equation by Gauss-Seidel Residual Update (GSRU) iterative algorithm.

SNP-BLUP method

Results obtained using PC scores in the PCA-BLUP method were compared with those obtained with a more traditional approach, named SNP-BLUP method, that uses SNP genotypes as predictor in model [1]. This approach was carried out using the GS3 program (Legarra et al., 2010).

A three step procedure was implemented including: 1) estimation of the variance components associated to the random effects through the VCE procedure with flat prior, 2) use of the estimated variance components to set the solution of the BLUP model in REF, 3) use of estimated effects in REF to obtain DGVs in VAL through the 'PREDICT' option. In particular variance components estimated in step 1 through the VCE estimator were used to solve the mixed model equations through a GIBBS Sampling approach assuming known variances for all random effects. Estimates were done using a Linux server with 4 x 4 Quad Core processors and 128 Gb RAM.

DGV and accuracy estimation

For both the abovementioned methods the vectors of DGV for the individuals of the validation set were calculated as:

$$DGV = 1\mu + Z\hat{g}$$

where μ is the overall mean of the trait of interest, Z is matrix of PCs scores or SNP genotypes (depending on the method used), and ĝ is the vector of the effects estimated with BLUP in the REF population. Accuracies of prediction were calculated as Pearson correlations between DGV and DRPF for each trait in the VAL population. The average *r*DGV across traits was used to compare the effect of the different method of predictors partitioning and different level of variable reduction (threshold of variance explained by PCs) obtained applying PCA.

Results

Principal Components

In total, 2.301 and 521 bulls were included in the reference and validation populations (~80:20%), respectively. The number of markers and PCs retained for each scenario and the degree of data reduction (difference percentage between PCs retained after PCA and initial number of SNPs markers) are reported in table 1. MD_GW showed the lowest number of component extracted. As already reported by Gaspa et al. (2013), the asymmetry between the number of (***p***) variables and (***n***) observations in the genome wide approach, was the reason of the limited number of PCs retained. In this case the total number of PCs was equal to the maximum number of animals analyzed (2.822) and the lowest compared to the other scenarios.

Table 1.
Methods of predictor reduction for scenarios characterized by different marker density and partitioning. Type and number of predictors used in the analysis after the variance threshold imposition and degree of reduction (in brackets) compared to the original number of SNP markers.

Model	Scenario	$N._{SNP}$	Datasets[1]	PC retained (%)		
				80%	90%	99%
				(% of data reduction)		
	MD_GW	39.884	1	915 (98%)	1.430 (96%)	2.502 (94%)
	MD_CHR	39.884	29	2.572 (94%)	4.748 (88%)	15.561 (61%)
PCA-BLUP [2]	MD_BLOCK	64.647	29	2.548 (96%)	4.878 (92%)	18.186 (72%)
	HD_CHR	610.415	29	2.747 (99.5%)	5.263 (99%)	20.958 (96%)
	HD_SPLIT	610.415	290	5.335 (99%)	9.217 (98%)	35.229 (94%)
SNP-BLUP [3]	HD_BLUP	610.415	29	-	-	-

[1] Number of dataset used to extract PCs; [2] Predictors = PC scores; [3] Predictors = SNPs.

In the other scenarios the total number of PCs was composed by the sum of PCs extracted and being in particular equal to the rank of the covariance matrix [min(*n*,*p*)-1] where *n* and *p* are the number of rows and columns of the data matrix, respectively. MD_CHR and MD_BLOCK had a total number of PCs extracted comparable to the number of initials predictors. In HD_CHR *p* was in all cases bigger than *n* and the total number of PCs retained was equal to *n* (limiting factor) multiplied by the number of chromosomes (2.822*29). The maximum number of components extracted was obtained in HD_SPLIT which was characterized by the maximum number of datasets (290), each one having the same number of observations but a lesser number of variables and consequently, the minimum disproportion between *n* and *p*.

Degree of dimensionality reduction

The application of the PCA has highlighted, through the comparison of different scenarios, the problem associated with the rank of the *S* matrix. The degree of data reduction ranged from a minimum of 61% (MD_CHR) to a maximum of 99% (HD_SPLIT). Among the medium density panels, the data reduction for MD_GW (from 94% to 98%) was higher than other MD (61% – 94% for MD_CHR and 72% – 96% for MD_BLOCK) because a minimum number of PCs were extracted due to the strong limiting effect generated by the relatively lower number of observations. In MD_GW a total of 2.822 PCs were obtained starting from a set of 39.884. Although the number of markers in MD_GW was the same of MD_CHR, the repartition of markers in a greater number of datasets (the chromosomes) and consequently in the greater number of PCs extracted, was the cause of the smaller degree of reduction of MD_CHR. MD_BLOCK had a quite higher degree of data reduction compared to MD_CHR because, despite the similar number of PCs extracted among scenarios, the initial number of SNPs included in the two panels was different with the first being around the double of the second. Also in this case, the higher disproportion between markers and observations occurred in MD_BLOCK was the reason of the higher degree of data reduction. As expected, the abovementioned trend was amplified in the HD where the degree of reduction obtained in HD_SPLIT was outperformed by

the reduction obtained in HD_CHR. In such conditions, the number of PCs having eigenvalues greater than zero is heavily influenced by the number of observations, resulting in possible biased estimations. The extraction of PCs by chromosome allowed to overcame this problem by reducing the disproportion between observations and the number of markers on each dataset (chromosomes). However, with the recently release of High Density panels as the BovineHD Beadchip (Illumina Inc., San Diego, CA) featuring more than 700K markers, the rank issue has been raised again, also when performing analysis with the chromosome wide approach.

Computation time

In scenarios in which number of variables approximated the number of observations, in particular for MD_CHR, MD_BLOCK, and HD_SPLIT, the total computation time needed to obtain PC extraction for all chromosomes were 60 min for MD_CHR, 50 min for MD_BLOCK and 6 h 30 min in the HD_SPLIT (average computation time was 1.3 min per dataset) respectively. For scenarios in which the number of markers exceeded excessively the number of observations (MD_GW and HD_CHR) PCs extraction had to be performed using a Linux server (4 x 4 Quad Core processors and 128 Gb RAM) and computational time was 1 h 15 min for MD_GW and approximately 22 h 50 min for HD_CHR (47 min per chromosome). DGV computation was performed by personal computer for all scenarios. Computation time required to reach the convergence by GSRU algorithm varied between scenarios and within them, across threshold of PCs retention. Time required for DGV computation on the PCA-BLUP method, ranged from a maximum of 1 hour and 85 minutes for Udder depth on HD_SPLIT at 99% of variance explained to a minimum of 2.17 seconds (trait 29 of MD_GW at 80% of variance explained). Average computation time needed by GS3 to obtain estimation for single traits implied 98 hours for the variance components estimation with a number of iteration set to 50.000, 4 hours (from 37 minutes to 8 hours) to solve the BLUP models on the REF population with priors of variance components estimated in the previous step and 8 minutes to predict DGVs and accuracies in the validation set.

Accuracies (rDGVs)

Accuracies for the different scenarios analyzed and the average rDGVs across traits are shown in table 2, heritabilities and reliabilities in table 1. In figure 1 the same values are divided by groups of traits and different degrees of variance accounted for by the principal components. Among different scenarios of PCA reduction, **production traits** as milk yield, protein (Kg) and somatic cell count, showed the highest values of accuracies. Milk yield accuracy was moderate (0.47 – 0.54) as well as somatic cell count (0.31 – 0.63). Accuracies observed for protein Kg (0.60 – 0.63) and fat Kg (0.20 – 0.28) were generally higher compared to protein % (0.36 – 0.41) and fat % (0.16 – 0.24). **Conformation traits** showed moderate values of *r*DGVs (0.22– 0.56). Overall conformation traits (overall conformation, stature, strength and vigor, body depth, angularity and rump) generally had moderate accuracies (0.22 – 0.56). Feet and legs related traits (rear leg rear view, foot angle, feet and legs functionality) showed on average low to moderate rDGVs (0.14 – 0.37) apart for the rear leg side view that showed lower values of accuracy particularly in HD_CHR (0.07). Udder conformations traits (udder attachment, height width, support and depth, teat length and position) showed on overall moderate values of accuracies (0.15 – 0.48). **Functional traits** were characterized by low DGV accuracies (0.03 – 0.36). The lowest value were obtained for Maternal Calving Ease (0.01 – 0.06), the highest for fertility (0.15 – 0.36). Other reproduction related traits (calving interval, days at first insemination, direct calving ease, direct and combined longevity) showed values of *r*DGVs generally intermediate in this range (0.07 – 0.31). Values of accuracies obtained through GS3 were comparable with those obtained through PCA particularly for MD_CHR at 99% of variance explained by principal component retained. MD_CHR sometimes outperformed the SNP-BLUP method. Within each scenario of marker repartition, excepted for MD_GW, accuracies generally showed an increase with the number of principal component retained. Considering the average *r*DGVs across traits between scenarios, values of accuracies generally decreased passing from MD_CHR to MD_BLOCK, HD_SPLIT, HD_CHR indicating an overall inverse relation between the numbers of SNP markers includes in the initial panel from which PCs extractions have been performed and the accuracies of the predictions obtained with the PCA-BLUP method for DGV estimation.

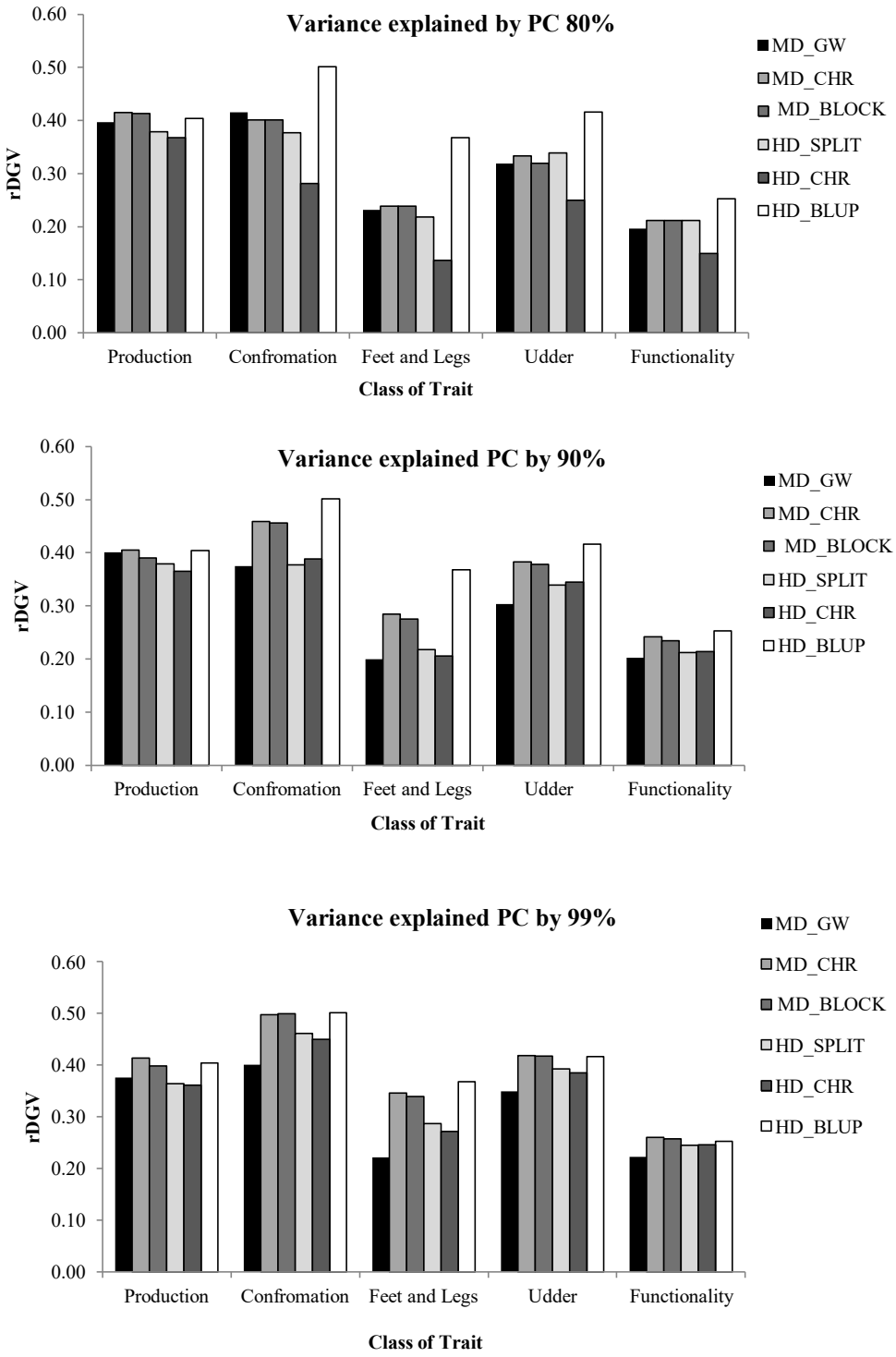

Figure 1. Levels of accuracy for groups of traits under different level of variance accounted for by PCs.

Table 2. Accuracies of DGVs estimated for 32 traits analyzed under different conditions of markers repartition and predictors reduction. Average rDGV across traits with standard deviations.

Traits	MD_GW			MD_CHR			MD_BLOCK			HD_CHR			HD_SPLIT			HD_BLUP
	80%	90%	99%	80%	90%	99%	80%	90%	99%	80%	90%	99%	80%	90%	99%	
1. Milk yield	0.50	0.51	0.48	0.52	0.52	0.54	0.51	0.51	0.52	0.47	0.47	0.47	0.49	0.49	0.48	0.55
2. Fat Kg	0.28	0.27	0.22	0.28	0.27	0.28	0.28	0.24	0.25	0.21	0.21	0.20	0.23	0.23	0.21	0.26
3. Protein Kg	0.62	0.60	0.60	0.63	0.63	0.62	0.63	0.62	0.62	0.60	0.61	0.61	0.63	0.63	0.61	0.62
4. Fat (%)	0.21	0.24	0.20	0.24	0.22	0.22	0.24	0.20	0.20	0.19	0.17	0.16	0.18	0.18	0.16	0.19
5. Protein (%)	0.38	0.38	0.37	0.40	0.39	0.41	0.40	0.39	0.40	0.37	0.36	0.36	0.37	0.37	0.36	0.40
6. SCC	0.55	0.55	0.58	0.55	0.61	0.63	0.55	0.58	0.61	0.31	0.54	0.60	0.55	0.55	0.61	0.62
7. Conformation	0.39	0.35	0.37	0.38	0.44	0.48	0.38	0.43	0.47	0.25	0.36	0.41	0.35	0.35	0.43	0.48
8. Stature	0.49	0.46	0.47	0.47	0.53	0.56	0.47	0.52	0.56	0.34	0.44	0.51	0.42	0.42	0.52	0.55
9. Strength and vigor	0.42	0.40	0.44	0.43	0.49	0.53	0.43	0.48	0.53	0.32	0.43	0.49	0.41	0.41	0.50	0.53
10. Body Depth	0.41	0.37	0.42	0.41	0.46	0.49	0.41	0.46	0.49	0.30	0.40	0.46	0.39	0.39	0.47	0.50
11. Angularity	0.36	0.28	0.33	0.32	0.40	0.45	0.32	0.44	0.47	0.22	0.35	0.40	0.32	0.32	0.42	0.47
12. Rump Angle	0.38	0.38	0.39	0.40	0.43	0.47	0.40	0.42	0.47	0.26	0.36	0.43	0.37	0.37	0.44	0.48
13.Rump width	0.46	0.38	0.39	0.40	0.46	0.50	0.40	0.45	0.50	0.27	0.38	0.45	0.37	0.37	0.46	0.50
14.Rearleg side view	0.23	0.14	0.20	0.22	0.26	0.34	0.22	0.24	0.34	0.07	0.15	0.24	0.17	0.17	0.26	0.37
15. Foot angle	0.20	0.23	0.23	0.25	0.30	0.34	0.25	0.28	0.33	0.14	0.22	0.28	0.24	0.24	0.30	0.35
16. Rear leg rear view	0.22	0.20	0.22	0.23	0.27	0.33	0.23	0.27	0.32	0.16	0.21	0.27	0.22	0.22	0.28	0.36
17. Feet and legs functionality	0.27	0.23	0.24	0.26	0.31	0.37	0.26	0.31	0.36	0.17	0.24	0.30	0.24	0.24	0.31	0.39
18. Front udder Attachment	0.34	0.37	0.43	0.44	0.45	0.47	0.44	0.45	0.48	0.30	0.41	0.44	0.42	0.42	0.45	0.46
19. Rear udder height	0.37	0.38	0.42	0.39	0.44	0.45	0.39	0.42	0.45	0.30	0.40	0.43	0.38	0.38	0.44	0.44
20. Rear udder width	0.31	0.33	0.33	0.33	0.37	0.41	0.33	0.37	0.41	0.30	0.34	0.40	0.36	0.36	0.40	0.42
21. Udder support	0.31	0.31	0.34	0.32	0.36	0.37	0.32	0.34	0.37	0.26	0.33	0.35	0.33	0.33	0.36	0.38
22. Udder depth	0.37	0.37	0.43	0.42	0.46	0.47	0.38	0.45	0.47	0.29	0.42	0.44	0.42	0.42	0.45	0.45
23. Front teat position	0.22	0.23	0.32	0.27	0.33	0.38	0.23	0.31	0.37	0.22	0.28	0.35	0.27	0.27	0.35	0.37
24. Teat length	0.34	0.26	0.33	0.28	0.38	0.44	0.26	0.38	0.43	0.18	0.33	0.39	0.32	0.32	0.40	0.43
25. Rear teat position	0.29	0.18	0.21	0.22	0.27	0.36	0.21	0.29	0.36	0.15	0.24	0.29	0.22	0.22	0.30	0.38
26. Calving Interval	0.07	0.10	0.09	0.10	0.10	0.11	0.10	0.08	0.10	0.11	0.08	0.09	0.08	0.08	0.09	0.12
27. Days first insemination	0.14	0.21	0.20	0.18	0.22	0.27	0.18	0.22	0.27	0.14	0.17	0.22	0.18	0.18	0.22	0.27
28. Direct calving ease	0.19	0.16	0.19	0.20	0.22	0.23	0.20	0.24	0.24	0.13	0.22	0.24	0.18	0.18	0.24	0.17
29. Maternal calving ease	0.02	0.06	0.01	0.03	0.03	0.02	0.03	0.04	0.03	0.05	0.06	0.04	0.04	0.04	0.04	0.03
30. Direct Longevity	0.11	0.13	0.15	0.13	0.15	0.14	0.13	0.13	0.13	0.15	0.15	0.16	0.15	0.15	0.15	0.12
31. Combined Longevity	0.28	0.20	0.27	0.26	0.30	0.31	0.26	0.29	0.31	0.17	0.25	0.30	0.29	0.29	0.30	0.31
32. Fertility	0.22	0.22	0.30	0.23	0.30	0.36	0.23	0.29	0.36	0.15	0.23	0.32	0.23	0.23	0.32	0.38
Average accuracy (SD)	**0.31** (0.14)	**0.30** (0.13)	**0.32** (0.13)	**0.32** (0.13)	**0.36** (0.14)	**0.39** (0.14)	**0.31** (0.13)	**0.35** (0.14)	**0.38** (0.14)	**0.24** (0.11)	**0.31** (0.13)	**0.35** (0.14)	**0.31** (0.13)	**0.31** (0.13)	**0.35** (0.14)	**0.39** (0.14)

Discussion

An overall increase in the accuracy of the estimates was generally observed with the increase of the number of PCs retained into each panel density. Accuracies progressively increased by retaining a greater number of PCs and specifically passing from the 80, 90 to 99% of variance explained by the components. A possible explanation of this behaviour should be that by increasing the number of PCs retained all the variance of the initial SNP panel tends to be accounted for by components. Despite the apparent contradiction of increasing the amount of PCs to obtain better results, from table 1 is possible to notice that, also retaining the greatest number of components as possible, in never cases (except for MD_GW) their total amount exceeded the half of the total number of original variables. Then while having results comparable to those obtained through the SNP-BLUP method (used as control), the number of predictors used in the estimations consistently changed, with a great reduction particularly in favour of the PCA-BLUP method.

Comparing the changes in accuracies obtained between panels with different marker densities, results suggested the existence of an inverse relationship between the original number of SNPs and the accuracies achieved after PC extraction and DGV estimations. More specifically, a possible explanation of this behaviour should be founded in the shrinkage of the total variance in a quite similar number of PCs obtained between panels with different marker densities. For all panels the population size was the limiting factor during PCs extractions. What was different between panels was the disproportion existing between the number of animals and the number of variables, being this disproportion effectively greater for HD panels. In example, the total quantity of PCs extracted at the 99% of the variance explained, was relatively higher for MD than for HD panels. More specifically, a total of 15.561 and 18.186 PCs were respectively extracted for MD_CHR and for MD_BLOCK starting from an initial number of 39.884 and 64.647 SNP markers respectively. A quite similar number of PCs were obtained for HD_CHR (20.958) and HD_SPLIT (35.229) but starting from a panel of markers of 610.415 SNPs. It results evident as the total variance deriving from each panel was differently distributed among a quite similar number of PCs extracted. The difference in magnitude observed between

eigenvalues derived from different PC extractions (data not shown) supported this hypothesis. Changes observed in *r*DGVs between panels having the same marker densities but different ways of marker repartitioning (for example the MD_GW and MD_CHR, the HD_CHR and the HD_SPLIT), were associated to the data asymmetry. The same markers were in fact manipulated in different ways in manner to have a progressive reduction of the unbalancing between the number of variable (by splitting markers at chromosome and sub-chromosome level) and the (fixed) number of animals. In practise what happened at the upper level (abovementioned) between different panel densities, occurred also within scenario depending on the way of data manipulation. In particular, when dealing with a MD_GW (all markers kept in a unique dataset) the difference between the number of markers and the number of observations is maximum. All the variance of this dataset will be compressed into a number of PCs not greater than the number of observations. At the contrary, by dividing the markers on chromosome or sub-chromosome level, a better repartition of the variance occurred at each PC extraction, resulting in more accurate DGV estimations. Finally among the PCA-BLUP approach, MD panel generally outperformed accuracies carried out by the HD panels, comparing results obtained through the SNP-BLUP method but implying lesser computational effort and a reduced time of work. The MD_GW and the HD_CHR required the use of the server for the PCs extraction and a higher computation time. The repartition of markers in a sub-chromosome level (HD_SPLIT) allowed for the first time, both the reduction of the data unbalancing and the computation demand making possible to perform all the analysis directly on the HD panel both avoiding a great computational power and information loses.

Conclusions

Since the dramatic increase of markers due to the introduction of high density SNP panels, the use of PCA applied to reduce the asymmetry of the genomic datasets, seemed no longer able to provide accurate breeding values. With the release of the BovineHD beadchip a dramatic increase on the number of markers has been achieved. Passing from 50.609 SNPs of the Illumina Bovine50k technology to the actual 777.962 SNP markers, this technology has exacerbated the data unbalancing. Considering that the population size of the main cattle breeds has not increased at the same rate of the technology to them applied, the reduction of predictors has become mandatory in manner to obtain reliable breeding values. Population size for the Italian Holstein breed accounts for less than 3 thousand progeny tested genomic bulls. With a similar degree of unbalancing generated by the technology, tools able performs estimates avoiding loss of information are preferred. Pre-selection of markers based on their Linkage Disequilibrium has been extensively applied in GWAS analysis to reduce markers on the basis of their association with traits of interest and avoiding the presence of not associated markers. In this work the pre-selection of markers through the haplotype block partitioning implied a potential loss of informative markers. The use of PCA, relying on the linear combination of all available markers, have a double positive effect. From one hand it allows to prevent the markers rejections, on the other hand, by retaining only PCs carrying the greater quota of variability associated to the original dataset, it allows to discard the nuisance variability. From results of this study three main trends for the values of accuracies were observed: 1) accuracy increased within scenario with the increase of the number of component retained; 2) accuracy increased between scenarios according to the reduction of the initial marker density; 3) at the same level of marker density, accuracy increased with the different kind of markers repartition, with better results obtained when applying the chromosome than the genome wide approach for markers repartition.

Acknowledgements This research was supported by Italian Ministry of Agriculture, grant INNOVAGEN. I would like to thank Dr. Giustino Gaspa for the support given in the creation of the experimental design, editing and processing of the data. I wish to acknowledge Prof. Macciotta for the helpful suggestions and the final revision.

REFERENCES

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, 210–223. doi:10.1016/j.ajhg.2009.01.005

Dimauro, C., Cellesi, M., Pintus, M. A., and Macciotta, N. P. P., 2011. The impact of the rank of marker variance–covariance matrix in principal component evaluation for genomic selection applications. *Journal of Animal Breeding and Genetics*, *128*(6), 440–445. http://doi.org/10.1111/j.1439-0388.2011.00957.x

Dimauro, C., Cellesi, M., Gaspa, G., Ajmone-Marsan, P., Steri, R., Marras, G., and Macciotta, N. P., 2013. Use of partial least squares regression to impute SNP genotypes in Italian Cattle breeds. *Genetics Selection Evolution*, *45*(1), 15. http://doi.org/10.1186/1297-9686-45-15

Gaspa, G., Pintus, M. a., Nicolazzi, E. L., Vicario, D., Valentini, a., Dimauro, C., and Macciotta, N. P. P., 2013. Use of principal component approach to predict direct genomic breeding values for beef traits in Italian Simmental cattle. *Journal of Animal Science*, *91*(1), 29–37. http://doi.org/10.2527/jas.2011-5061

Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., Fernando, R., 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics 183, 347 LP – 363.

Goddard, M.E., Hayes, B.J., 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10, 381–391. doi: 10.1038/nrg2575.

Hayes, B.J., Visscher, P.M., Goddard, M.E., 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. (Camb). 91, 47–60. doi:10.1017/S0016672308009981

Hickey, J.M., Kinghorn, B.P., Tier, B., van der Werf, J.H.J., Cleveland, M.A., 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. 44, 9. doi:10.1186/1297-9686-44-9

Howie, B.N., Donnelly, P., Marchini, J., 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5. doi:10.1371/journal.pgen.1000529

Krzanowski W J. Principle of Multivariate analysis: a user's perspective. Oxford Statistical Science Series (21), 2000.

Legarra, A., Ricard, A., and Filangi, O., 2010. GS3-Genomic Selection, Gibbs Sampling, Gauss Seidel and BayesCπ. Manual, 1–16.

Lund, M.S., van den Berg, I., Ma, P., Brøndum, R.F., Su, G., 2016. Review: How to improve genomic predictions in small dairy cattle populations. animal 10, 1042–1049. doi:10.1017/S1751731115003031

Macciotta, N. P. P., Gaspa, G., Steri, R., Nicolazzi, E. L., Dimauro, C., Pieramati, C., and Cappio-Borlino, A., 2010. Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. *Journal of Dairy Science*, *93*(6), 2765–74. http://doi.org/10.3168/jds.2009-3029

Meuwissen, T.H., E., Hayes, B.J., and, Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), 1819–1829. http://doi.org/11290733

Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S. 2014. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*, *15*(1), 478. http://doi.org/10.1186/1471-2164-15-478

Solberg, T. R., Sonesson, A. K., Woolliams, J. a, and Meuwissen, T. H. E., 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genetics, Selection, Evolution : GSE, 41, 29. http://doi.org/10.1186/1297-9686-41-29

Pattaro, C., Ruczinski, I., Fallin, D. M., and Parmigiani, G., 2008. Haplotype block partitioning as a tool for dimensionality reduction in SNP association studies. BMC Genomics, 9, 405. http://doi.org/10.1186/1471-2164-9-405

Purcell S., Chang C., 2015. PLINK [Build (v1.90b3.29 64-bit)]. Electronic resource (URL): https://www.cog-genomics.org/plink2

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., and Schenkel, F. S. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science, 92(1), 16–24. http://doi.org/10.3168/jds.2008-1514

Verbyla, K.L., Hayes, B.E.N.J., Bowman, P.J., Goddard, M.E., 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet. Res. (Camb). 91, 307–311. doi:10.1017/S0016672309990

CHAPTER 4

Research paper

Echography and 2D morphological predictors for indirect selective breeding of fillet and carcass percentage in European sea bass (*Dicentrarchus labrax*).

Submitted to Aquaculture as

Marc Vandeputte, Antonio Puledda, Anne Sophie Tyran, Anastasia Bestin, Céline Coulombet, Aline Bajek, Gwénaëlle Baldit, Alain Vergnet, François Allal, Jérôme Bugeon, Pierrick Haffray. Echography and 2D morphological predictors for indirect selective breeding of fillet and carcass percentage in European sea bass (*Dicentrarchus labrax*).

Abstract

Genetic parameters for carcass and fillet percentage were estimated in 795 European sea bass reared under commercial conditions and slaughtered at 573 days post fertilization (average length of 31.5 cm). Fishes were offspring of 45 sires and 20 dams crossed in a factorial mating design. Pedigrees were re-constructed with 90.7% success using microsatellites. The heritability of fillet yield was moderately low (0.21), while it was high for carcass yield (0.57). Different predictors were derived from 2D measurement of surfaces on digital pictures and echography measurements at various points of the body. The accuracy of the phenotypic prediction was rather low for fillet yield (r^2 = 0.02 - 0.16), but higher for carcass yield (r^2 = 0.27 – 0.41). However, genetic correlations with the traits to predict were reasonably high (up to 0.67 for fillet yield and 0.95 for carcass yield), thus allowing to consider them for performing indirect selection instead of sib selection. However, it was difficult to design a predictor that would simultaneously increase fillet yield and carcass yield because of contradicting effects of relative head size, an important component of the predictors which was positively correlated to carcass yield but not to fillet yield.

Introduction

European sea bass (*Dicentrarchus labrax* L., 1758) is a member of the family Moronidae and one of the most important economic species reared in the Mediterranean Basin since the 1980's when the development of techniques for the mass production of juveniles became feasible (Chatain and Chavanne, 2009). Sea bass is a eurythermic (10 - 28°C) and euryhaline (3 - 35‰ salinity) species, widely distributed in the Mediterranean, the Black Sea and the Eastern coast of the Atlantic Ocean (from Norway to Morocco). The global production of European sea bass accounted for 161,059 tons in 2013 (FAO, 2011 - 2016 ©). To date, aquaculture is the agricultural sector with the highest rate of growth and as any other growing industry the optimisation of production is becoming a necessity and genetic improvement, a major way to optimize productivity. Several previous studies on the genetics of sea bass have focused on the improvement of economic traits such as growth rate (Saillant et al., 2006; Dupont-Nivet et al., 2008; Vandeputte et al., 2009;), body composition and processing traits (Gorshkov et al., 2004; Haffray et al., 2007; Saillant et al., 2009), survival and control of sex (Saillant et al., 2002; Vandeputte et al., 2007). For growth rate, which has been the main objective of selection for most fish species, sea bass showed a large within population genetic variability (Dupont-Nivet et al., 2008; Vandeputte et al., 2009). Quantitative genetic variation was evidenced for most of the production traits investigated (Saillant et al., 2002, 2006, 2009; Vandeputte et al., 2007, 2009, 2014; Dupont-Nivet et al., 2008, 2010; Bardon et al., 2009; Karahan-Nomm et al., 2013). Despite the progress obtained in the knowledge of these traits, new characters such as carcass quality, processing efficiency (improvement of the edible fraction) and impact of production (environmental footprint) are becoming interesting for aquaculture. Texture, colour, abdominal and flesh fat content are qualitative traits able to influence the final products and consumer preferences (Neira et al., 2004; García-Celdrán et al., 2015;). Processing traits are more interesting for fish producers, directly influencing their profitability (Saillant et al., 2009). An important processing trait, the fillet percentage, is able to affect the interest of producers, consumers and citizens. A higher fillet percentage involves a higher production of edible portions, a more favourable global feed conversion (amount of feed needed to produce a given

amount of edible flesh), the reduction of waste and an improvement of the ecological efficiency when the functional unit considered is the quantity of edible flesh (Haffray et al., 2012; Acosta-Alba et al., 2015). As happened in France for rainbow trout (*Oncorhynchus mykiss*), where consumer demand for fresh and smoked fillet induced the increase of slaughtering weight from 0.3 to 3.5 kg (Bisault, 2009) and a change towards a salmon-like elongated shape (Haffray et al., 2013), the increase of harvest weight is getting into practice for European sea bass. In 2005, 20% of the sea bass production was already constituted by fish bigger than 0.8 kg (Chatain and Chavanne, 2009). From the genetic point of view, the introduction in the selection index of traits correlated with fillet percentage, easier to score and more heritable (e.g. headless gutted carcass percentage – Haffray et al., 2012) as well as the adoption of individual scoring of candidates through indirect and non-lethal selection criteria (Haffray et al., 2013) are seen as methods of choice to increase fillet percentage in the French trout breeding industry. The advent of new trait measurement systems such as digital or ultrasounds imagery, could allow to automate fish grading by computer vision for size, morphology and sex during the rearing cycle (Costa et al., 2013). At the same time, the use of these new sources of morphological information (external and internal measurement with ultrasound) combined in different regression models to predict the processing yields is becoming, particularly for traits requiring the sacrifice of individuals, a reliable practice in fish breeding as reported in rainbow trout. Exploiting the within family variability, Individual Selection (IS) based on Indirect Criteria (IC) is potentially more efficient than the traditional sib selection, due to the possibility to obtain higher selection intensities. However, the high rate of genetic gain that may be obtained in farmed fish for the increase of specific body compartments may hide undesirable side effects and requires specific studies to investigate interactions between correlated traits, especially for those negatively correlated (Haffray et al., 2012). An accurate evaluation of the genetic component of the phenotypic variance (heritability) and covariance between traits (genetic correlations), is also needed to allow an accurate prediction of selection response. In the last decades, estimates for direct production or processing traits (growth and carcass) were produced for Mediterranean fish species, in particular for Gilthead sea bream (*Sparus aurata*) (Navarro et al., 2009; García-Celdrán et al., 2015) and European sea bass (Haffray et al., 2007; Saillant et

al., 2009; Bestin et al., 2014), but in the latter species without reporting indirect selection traits that can be measured on the live breeding candidate.

The aim of this study was to provide estimates of genetic parameters for body and carcass traits for a population of European sea bass reared under commercial conditions, determine morphological indicators able to predict processing yields (indirect criteria) and predict the genetic progress achievable with their application in a hypothetical selection scheme.

Materials and methods

Experimental design

A total of 423 families of European sea bass were produced in September 2012 at the "Ecloserie Marine de Gravelines" (EMG) breeding station through artificial fertilization of gametes deriving from 47 sires and 27 dams crossed in two 16x9 and one 15x9 full factorial mating plans. The parents were broodstock from EMG's ongoing fifth generation of sea bass breeding program. Each individual spawn was separated in 15 or 16 aliquots, each aliquot being fertilized by the sperm from one sire. After fertilization, all aliquots from each female were grouped and put into an individual incubation jar. After 24 h, a 45 g aliquot of live (floating) eggs was collected from each of the 31 jars, and all aliquots were pooled to obtain a single batch containing an equal quantity of eggs from each of the dams. All fish were kept in a single tank and fed with the same diet until April 2014 when they reached an average Fork Length of 31.5 ± 2.8 cm (range 23.9 – 38.9) and average Body Weight of 395 ± 127.1 g (range 115.0 – 783.0).

Slaughter chain and processing data collection

The fish traits measurements and the rearing environments were indexed according to the ATOL (Animal Trait Ontology for Livestock) and EOL (Environment Ontology for Livestock) ontologies, available on the ATOL (http://www.atol-

ontology.com/index.php/en/les-ontologies-en/visualisation-en). At 573 days post fertilization (April, 2014), 1002 sea bass were randomly sampled from a total of 2000 fish. Fish were first lightly anaesthetized in the tank (0.1 ml/l phenoxyethanol), then separated in groups of 5-6 individuals and killed with a lethal dose of phenoxyethanol (0.5 ml/l). This terminal anaesthesia delays the effects of *rigor mortis* that might compromise the quality of digital photos and the filleting operation. Each fish was identified with electronic transponders (ISO 11784/11785 BiologID, France) previously implanted in the dorsal musculature. The use of Passive Integrated Transponder (PIT) tags allowed to trace back the identity of each fish during the following phenotyping steps. Fish was initially photographed (Canon EOS 1000D) with the dorsal fin deployed in order to distinctly show all morphometric landmarks (figure 1). A sample of caudal fin was then individually cut and stored in ethanol for later DNA extraction and parentage assignment. For each fish, fork length was recorded to the nearest 0.1 mm with an in-house electronic ruler, total body weight was measured to the nearest 0.5g, and 10 echography measurements were collected using an ultrasound scanner (Hospimedi, LC100, 7.5 MHz – see below). After biometry, each fish was portioned into the following body compartments: left fillet, head, viscera, gonads and half carcass (right part). Compartments were individually weighed to the nearest 0.5 g, except viscera and gonads which were weighed to the nearest 0.1 g. Sex was determined by visual inspection of the gonads and the binomial variable (1 = male, 2 = female) was added to the phenotype file. The following direct measurement were available for each fish: body length (BL, ATOL:0001658) and body weight (BW, ATOL: 0000351), viscera weight (ViscW, ATOL:0002258), head weight (HeadW, ATOL:0001545), skin-on untrimmed left fillet weight (leftFilW, ATOL:0002304) and half carcass weight (HCarcW). A new set of derived variables was created combining the previous traits: Condition Factor or K (ATOL:0001653), and 5 composite weight traits: carcass weight (ATOL:0001057), headless gutted carcass weight (ATOL:0002260), fillet weight (ATOL:0002304), vertebral axis weight and bone weight (table 1). In addition, six yield percentages were calculated by dividing the weight of the different body compartments by total body weight: fillet (ATOL:0002305), head, headless gutted carcass (ATOL:0002261), carcass (ATOL:0000548), vertebral axis and viscera (table 1).

Moreover, considering the possible presence of scale effects, log-transformed variables were calculated and an allometric correction performed. Natural logarithm was computed for the weight of all body compartments. The log-transformed body compartments were regressed on the logarithm of BW to obtain BW-independent allometric residuals (table1).

Table 1.
Traits recorded and calculated for the different body compartements.

body compartment	weight trait	combination of variables	Yield trait (traitW/BW)	Log residual trait (log_trait ~ log_BW)
Fillet	FilW	LeftFilW x 2	Fil%	logr_Fil
Head	HeadW	HeadW	Head%	logr_ Head
Headless Gutted Carcass	HGCarcW	HalfCarcW + LeftFilW	DGcarc%	logr_ Dgcarc
Total Carcass	CarcW	HGCarcW + HeadW	Carc%	logr_ Carc
Axis	AxisW	HGCarcW - (2 x leftFillW)	Axis%	logr_ Axis
Viscera	Viscera W	Viscera W	Viscera%	logr_ Viscera

Ultrasound tomography data collection

A series of internal measurements were performed on each fish through ultrasound imagery (Hospimedi, LC100, 7.5 MHz) using Haffray et al., (2013) positions reported in figure 2. The scans obtained resulted in 9 muscular thicknesses (from Echo1 to Echo9) and the internal depth of the ventral cavity (Echo23). Thicknesses were individually collected from anterior (E1, E4, E7), intermediate (E2, E5, E8) and posterior (E3, E6, E9) muscles in the dorsal, ventral and abdominal position respect to the lateral line (Figure 2). In order to compare echography measurements among fish, ratios of these variables were computed dividing ultrasound thickness by individual BL according to the following formula:

$$Ratio_{echo_i} = {}^{Echo_i}/_{BL}$$

where $Echo_i$ is the echography at the i_{th} survey point and BL is the individual body length.

Image analysis

Thirty coordinates of morphological points with biological meaning (landmarks and semi-landmarks) were obtained by semi-automated processing of the digital pictures on Visilog® software (FEI Visualization Sciences Group). Photo processing was realized at the INRA Laboratory of Fish Physiology and Genomics (Rennes, France). Briefly, after graphical positioning of the head and tail extremity of the fish, a macro-command adapts the scale to the selected size so that all fish can be comparable having the same perspective. Several blue lines divided the image in equidistant section in manner to facilitate the manual imposition of landmarks and semi-landmarks on the surface of the fish (Figure 2). Each point was characterized by two coordinates (X, Y) and the distance between two landmarks A and B was calculated by applying the formula:

$$d = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}.$$

A set of these segments was used to delimit 28 triangles (Figure 3a) whose areas were calculated using the following formula:

$$area = {}^{1}/_{2} \left|(x_B - x_A)(y_C - y_A) - (x_C - x_A)(y_B - y_A)\right|$$

where the subscripts A, B and C refer to the three sides of a triangle. The sum of different triangle combinations gave 12 representative surfaces covering the entire shape of each fish (Figure 3b). Considering the fact that surfaces are proportional to the square of body size, in order to assess shape rather than size, proportion of the surfaces were preferred to obtain a better comparison among fish. Proportions (*P*) were computed as the following ratio:

$$P_i = {}^{S_i}/_{S_{tot}}$$

where for each fish S_i is the i_{th} surface and S_{tot} is the individual sum of all surfaces. Moreover, the distance between landmarks 1 and 30, representing the 2D Fork Length, due to its more accurate evaluation, were used as total Body Length (BL) in substitution of the on-site measurement.

Phenotype data editing

The sum of the weights of the body compartments was compared with the total body weight to find discrepancies. Fish with a sum of body compartments greater than the original body weight or with a loss on weight bigger than 2% were discarded from the analysis. Moreover, linear regression among traits (BW against BL and all body compartments against BW) was used to identify values not consistent with the prediction (bivariate errors). Animals with residuals (in absolute value) greater than three times the standard error of the regression were discarded. Values of proportions of surfaces and ratios of echography measurements outside 3 standard deviations around the mean were treated as missing values. Animals with undetermined sex were not included in the analysis.

Prediction models

The residuals of the log-log regressions between weights of compartments and total BW were used as a surrogate for percent yields and treated as dependent variables in a series of multiple regressions having different combination of independent variables (table 6). Types of independent variables combinations tested were: I) only proportion of surfaces, II) only ratios of echography measurements, III) a combination of proportion and ratios, IV) a combination of the most significant variables in the model III and IV_{bis}) characterized by the substitution of the proportion of the head surface with the "square head proportion" (SQ_Head). The formula to calculate SQ_Head was:

$$\mathrm{SQ_{Head}} = \frac{(d_{1\_6}.d_{4\_11})}{d_{1\_30}{}^{2}}$$

where d_{1_6} and d_{4_11} are the distances between the landmarks 1 - 6 and 4 - 11 and $d_{1\_30}^2$ is the squared BL. This measurement was tested as a surrogate to relative head surface as it can be measured directly on the field with a caliper. For each model, significant variables were selected through a stepwise selection. The adjusted R^2 was used as indicator of goodness of fit, and its maximization as the criterion to choose the best combination of variables. Prediction models were tested for Fil% and Carc%, both considered the major objectives for the sea bass selection on processing yields. For each model, the linear predictor was computed for each fish and added to the phenotype file. Models were fitted with the REG procedure of SAS® software.

DNA analysis and genotype data editing

The samples of caudal fin were stored in 95% ethanol and sent to LABOGENA-DNA, the French laboratory for livestock genotyping (ISO 17025 accredited, Jouy-en-Josas, France), that provided the DNA analysis using 12 microsatellite markers. Animals with more than 2 allelic mismatches for a sire, dam, offspring trio or not comprised in the experimental design were considered not adequately assigned and discarded from further analysis.

Quantitative genetic analysis

Descriptive statistics (mean, standard deviation and coefficient of variation) and significance of effects were estimated using SAS® software. Mean and standard deviation were calculated for all fish and for both sexes. Data were analysed with the following animal model using VCE 6.0 software (Groeneveld et al., 2008):

$$Y = X\beta + Z\alpha + \varepsilon$$

where Y is the vector of observations (weights, proportion of surfaces, echography ratios, yields, log-log residuals of weights), X and Z are the incidence matrices of β and α, respectively the vectors of fixed effects (sex with 2 levels and mating plan with 3 levels corresponding to the 3 independent full-factorial blocks) and the vector

of random additive genetic effects (860 levels corresponding to the records of the pedigree file), and ε is the vector of random residual effects. Heritabilities were estimated with single trait models and genetic correlations through a series of bivariate models using all phenotypic data available from offspring. Parental fish had no own phenotypic records.

The response to selection was estimated using the formulas of Falconer and McKay (1996) both for fillet and carcass percentage under a mass (MS), full sib (FS) and indirect (IS) selection. Response to selection for the theoretical MS based on direct (lethal) criteria was estimated applying the breeder's equation:

$$R_{MS} = i\ h^2\sigma_P$$

where *i* is the selection intensity, h2 and σ_P are the heritability and phenotypic standard deviation of the trait for which genetic gain is estimated. Expected genetic gain for FS was estimated by:

$$R_{FS} = i\,\sigma_P\,h^2\frac{n\,r}{\sqrt{n\,(1+(n-1)\,t)}}$$

where n is the number of sibs sampled for family, *r* is the genetic additive correlation between sibs (*r* = 0.5 for full sibs) and *t* is the phenotypic intra class correlation. The genetic progress obtainable through IS was estimated applying the Correlated Response (CR) equation:

$$CR_Y = i\,h_{P_X}h_{P_Y}r_A\,\sigma_{PY}$$

where h_{PX} and h_{PY} are respectively the square roots of heritability of the predictor X (on which selection is applied) and of the target trait Y, r_A is the genetic correlation existing between traits and σ_{PY} is the phenotypic standard deviation of the trait of interest.

The relative efficacy of selection (RES) related to each predictor and their components were estimated to evaluate their interest for performing indirect selection for the trait to predict. RES was computed as the variable part of the equation of the response to indirect selection for a given trait of interest: $RES = h_{P_1}|r_A|$, where $h_{P_1}$is the square root of the heritability of the predictors or their simple variables, and $|r_A|$ is the absolute value of the genetic correlation existing between the substituted variables and the character to predict. Response to selection was evaluated both for fillet and carcass percentage under MS and FS with 10% and 30% of, with 10 sibs per family in FS. For *IS* the correlated response generated by predictors and variables with high RES were tested only with a proportion of the population of 10%.

Results

Genotyping and parental assignment

A total of 915 offspring only gave some amplification, due to an unidentified DNA quality problem. From these, 830 were assigned to a single parental pair, with up to 2 mismatches allowed (90.7%), 67 were assigned to multiple parent pairs (7.4%) while 18 (2.0 %) were not assigned to any parent pair. Among the 830 fish correctly assigned, 34 that were not included in the declared mating plan and 1 with undetermined sex, were removed from the dataset. After editing, 795 offspring, 45 sires and 20 dams (2 sires and 1 dam were not represented in the offspring) were used to create the pedigree file that finally included 860 animals, from 45 sire half-sib families, 20 dam half-sib families, and 254 full-sib families. After raw data quality control, phenotypic data from 760 fish (378 males, 382 females) with 60 variables were available.

Distribution of traits

Mean and standard deviation for weight and percentage traits for the overall population and among sexes are shown in table 2. The average Carc% and DGCarc% were 88.9% and 69.7%, respectively. Fil%, represented 57.4% of the whole body weight whereas weights of head, vertebral axis and viscera accounted for 19.1%, 12.3% and 10.0% of the total BW. A residual of 1.1% could be ascribed to the gonads removal and body fluids losses occurred during the slaughter chain. Fish appertaining to each mating plan were: 195 for Mating Plan 1 (MP1), 222 for Mating Plan 2 (MP2) and 343 for Mating Plan 3 (MP3). Distribution of sexes was homogeneous for the three mating plans with a little predominance of females in MP2 (55.5%) and males in MP3 (53.3%). Sexes were divided as follow: 378 males and 382 females in the overall population and 97 and 98 in MP1, 98 and 124 in MP2, 183 and 160 in MP3. The effect of sex and mating plan (MP) was significant for all weight traits, and for most percent traits, so sex and mating plan were kept as fixed effects in all models.

Table 2.
Mean and standard deviation for processing traits (weights and percent yield) in males and females of European sea bass (*Dicentrarchus labrax L.*).

Trait	Mean ± S.D.	Males	Females	Min.	Max.
BW	395.4 ± 127.1	356.4 ± 113.4	433.9 ± 128.4	115.0	783.0
CarcW	350.2 ± 108.8	316.1 ± 96.8	384.1 ± 109.6	104.5	690.0
HGCarcW	275.5 ± 87.8	249.5 ± 79.2	301.3 ± 88.3	78.5	545.5
FilW	227.6 ± 74.7	205.8 ± 67.6	249.2 ± 75.1	62.0	453.0
HeadW	74.7 ± 21.7	66.6 ± 18.1	82.8 ± 21.9	26.0	144.5
AxisW	47.9 ± 13.9	43.7 ± 12.3	52.1 ± 14.2	14.5	99.0
Viscera W	41.0 ± 18.4	37.3 ± 17.0	44.6 ± 19.1	8.2	120.6
Carc%	88.9 ± 1.9	89.0 ± 2.0	88.8 ± 1.9	81.3	94.0
HGCarc%	69.7 ± 1.2	69.9 ± 1.3	69.5 ± 1.2	65.7	73.9
Fil%	57.4 ± 1.6	57.5 ± 1.7	57.3 ± 1.6	50.1	62.9
Head%	19.1 ± 1.6	19.0 ± 1.7	19.3 ± 1.6	14.7	24.9
Axis%	12.3 ± 1.3	12.4 ± 1.4	12.1 ± 1.2	9.2	16.8
Viscera%	10.0 ± 1.9	10.1 ± 1.9	10.0 ± 1.9	5.4	17.4

Prediction equations

A total of 10 models were validated for the prediction of fillet and carcass percentage. Predicted yields were represented by the residual of the log-log regression of the trait on body weight (table 6). For the prediction of fillet percentage model F1 (proportion of surfaces) was obtained including 8 variables with a R^2 of 16%. Model F2, constituted by 3 ratios of echography measurements, with a R^2 of 2% was the least explicative of all models. Model F3, composed by 7 significant variables (5 surfaces and 2 echography measurements), gave the best combinations in terms of phenotypic variance explained (R^2 = 18%) for fillet percentage. Model F4 and $F4_{bis}$ had respectively a R^2 of 14% and 10%. Although maintaining the same pattern of relative values, R^2 of the models for the prediction of the carcass percentage were systematically higher than those obtained in the prediction of fillet percentage. In particular model C3 (7 surfaces and 3 echography measurements) showed the best predicting ability (R^2 = 41%) while model C2 (4 echography measurements) had the lowest efficiency (R^2 = 27%). Unlike models F4 and F4bis, the models C4 and C4bis gave comparable R2.

Heritabilities

Heritability was generally moderate for all variables. Heritability for BW at slaughter was 0.41 ± 0.09, whereas heritability of BL and K were 0.38 ± 0.09 and 0.50 ± 0.09, respectively (table 3). Estimates obtained for all other carcass traits (weights of body parts) ranged from 0.36 to 0.41 (table 3). Heritabilities for the residuals of the log-log regression (table 8) were comparable to those obtained for the percent yields (table 4) ranging from 0.16 (logr_Axis) to 0.57 (logr_Carc), with a relatively low value (0.21) for logr_Fil. Heritabilities of prediction models F_1, F_2, F_3, C_1, C_2 and C_3 (table 9), were systematically lower than those obtained from the traits to predict (fillet and carcass percent). Moreover, models F_2 and C_2 showed the lowest heritability compared to the other prediction models, for both traits to predict. Heritabilities of the models F_4 and F_{4bis} were higher than these obtained for fillet percentage. Models C_4 and C_{4bis} showed lower values of heritability compared to values obtained for carcass percentage.

Table.3
Heritability and standard error (diagonal), genetic correlations ± standard error (above the diagonal) and phenotypic correlations (below the diagonal) for growth and carcass traits in European sea bass (*Dicentrarchus labrax* L.).

	Lfork_2D	**BW**	**K**	**CarcW**	**HGCarcW**	**FilW**	**HeadW**	**BoneW**	**AxisW**	**VisceraW**
Lfork_2D	**0.38 ± 0.09**	0.96 ± 0.01	0.58 ± 0.11	0.98 ± 0.01	0.97 ± 0.01	0.96 ± 0.01	0.97 ± 0.01	0.98 ± 0.01	0.99 ± 0.01	0.79 ± 0.06
BW	0.95	**0.41 ± 0.09**	0.75 ± 0.07	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.001	0.95 ± 0.01	0.96 ± 0.01	0.97 ± 0.01	0.90 ± 0.03
K	0.55	0.72	**0.50 ± 0.09**	0.71 ± 0.08	0.73 ± 0.08	0.74 ± 0.07	0.62 ± 0.10	0.62 ± 0.11	0.61 ± 0.11	0.85 ± 0.05
CarcW	0.96	0.99	0.69	**0.40 ± 0.09**	0.99 ± 0.01	0.99 ± 0.001	0.97 ± 0.01	0.98 ± 0.01	0.98 ± 0.01	0.86 ± 0.04
HGCarcW	0.96	0.99	0.71	0.99	**0.40 ± 0.09**	0.99 ± 0.001	0.96 ± 0.01	0.97 ± 0.01	0.97 ± 0.01	0.88 ± 0.04
FilW	0.95	0.99	0.72	0.99	0.99	**0.40 ± 0.09**	0.95 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.88 ± 0.04
HeadW	0.96	0.95	0.58	0.97	0.95	0.95	**0.39 ± 0.08**	0.99 ± 0.001	0.98 ± 0.01	0.75 ± 0.07
BoneW	0.96	0.96	0.59	0.97	0.96	0.95	0.98	**0.39 ± 0.09**	0.99 ± 0.003	0.77 ± 0.07
AxisW	0.92	0.93	0.59	0.94	0.94	0.91	0.93	0.97	**0.36 ± 0.08**	0.80 ± 0.06
VisceraW	0.80	0.90	0.81	0.87	0.88	0.89	0.78	0.80	0.78	**0.41 ± 0.09**

Table 4.
Heritability and standard error (diagonal), genetic correlation and standard error (above the diagonal), phenotypic correlations (below the diagonal) for processing traits in European sea bass (*Dicentrarchus labrax* L.)

	BW	**Carc%**	**HGCarc%**	**Fil%**	**Head%**	**Bone%**	**Axis%**	**Viscera%**
BW	**0.41 ± 0.09**	− 0.50 ± 0.13	− 0.14 ± 0.19	0.42 ± 0.17	− 0.61 ± 0.11	− 0.67 ± 0.10	− 0.70 ± 0.12	0.47 ± 0.14
Carc%	− 0.48	**0.57 ± 0.10**	0.72 ± 0.11	− 0.01 ± 0.20	0.90 ± 0.03	0.89 ± 0.04	0.82 ± 0.09	− 0.99 ± 0.001
HGCarc%	− 0.01	0.55	**0.32 ± 0.08**	0.64 ± 0.13	0.36 ± 0.18	0.35 ± 0.18	0.35 ± 0.20	− 0.75 ± 0.09
Fil%	0.33	− 0.01	0.61	**0.20 ± 0.07**	− 0.41 ± 0.18	− 0.44 ± 0.18	− 0.47 ± 0.19	− 0.02 ± 0.21
Head%	− 0.57	0.78	− 0.07	− 0.47	**0.46 ± 0.10**	0.98 ± 0.01	0.87 ± 0.08	− 0.88 ± 0.04
Bone%	− 0.59	0.77	0.03	− 0.64	0.9	**0.43 ± 0.10**	0.94 ± 0.03	− 0.88 ± 0.04
Axis%	− 0.44	0.54	0.16	− 0.67	0.52	0.84	**0.21 ± 0.07**	− 0.82 ± 0.09
Viscera%	0.48	− 0.98	− 0.55	0.01	− 0.76	− 0.76	− 0.54	**0.52 ± 0.10**

Genetic correlations

All body part weights showed a highly positive genetic (0.75 to 0.99) and phenotypic (0.78 to 0.99) correlation (table 3). K also showed a positive, but lower genetic (0.61 to 0.85) and phenotypic (0.55 to 0.81) correlation with the weights of body parts. Genetic correlations between percent traits and log-residual traits were close to unity (table 5). Genetic correlations between yields estimated by log residuals were generally positive (table 8), except for correlations between viscera and the other parts, which were negative (- 0.25 to - 0.99), and correlations between fillet and head or axis (- 0.19 and - 0.31, respectively). Carcass and fillet yields were genetically independent (r_A = - 0.01 ± 0.20), while the genetic correlation between logr_Fil and logr_Carc was positive but rather low and imprecise (0.28 ± 0.20). The genetic correlation of logr_Fil with logr_HGCarc was much higher (0.79 ± 0.09). Genetic correlations obtained between log−residuals of the trait to predict and predictors (table 9) were lower for those predicting fillet yield (0.17 to 0.67) than for those predicting carcass yield (0.91 to 0.95). Predictors of logr_Carc showed higher correlations between them (0.77 to 0.98) than predictors of logr_Fil (0.003 to 0.87). The correlation between models F2 and $F4_{bis}$, was the lowest (0.003). Although the genetic correlation between logr_Fil and logr_Carc was weakly positive (see above), the genetic correlation between predictors of logr_Fil and predictors of logr_Carc was always (and sometimes strongly) negative (- 0.13 to - 0.77). The genetic correlations between predictors of logr_Fil (F1 F4bis) and logr_Carc was also always negative (- 0.20 to -0.56) while the correlation between predictors of logr_Carc (C1 to C4bis) and logr_Fil was weakly positive (0.03 to 0.26).

Table 5.
Genetic correlations ± standard errors between processing traits expressed as percentage (%) and log-residual traits (logr)

model	r_g	S.E.
Carc% ~ logr_Carc	0.99	0.00002
Carc% ~ logr_HGCarc	1	0
Fil% ~ logr_Fil	1	0
Head% ~ logr_Head	0.99	0.00043
Bone% ~ logr_Bone	0.99	0.00032
Axis% ~ logr_Axis	0.99	0.000607
Viscera% ~ logr_Viscera	0.99	0.00116

Table 6.
Multiple regression models for the prediction of fillet and carcass percentage with type and name of models, linear combination of variables used to obtain predictors and relative R^2. Models F1-F4bis relate to prediction of fillet percentage, models C1-C4bis to prediction of carcass percentage.

Model		Linear combination of coefficients and variables	Adjusted R^2
Type	**Name**		
I	**F1**	0.93 − 1.91 P_Head − 1.41 P_1D − 0.99 P_1V − 1.17 P_2V − 1.25 P_3V − 0.68 P_4D − 0.54 P_4V − 1.71 P_5D	0.16
II	**F2**	− 0.07 + 0.52 Recho2 + 0.70 Recho7 + 1.84 Recho9	0.02
III	**F3**	0.227 − 1.25 P_Head − 0.29 P_1V + 0.49 P_2D − 0.18 P_2V + 0.29 P_4V + 1.30 Recho8 − 0.48 Recho23	0.18
IV	**F4**	0.26 − 1.17 P_Head − 0.60 Recho23	0.14
IVbis	**F4bis**	0.17 − 4.06 SQ_Head − 0.003 Recho23	0.10
I	**C1**	−0.03 + 0.58 P_Head − 1.15 P_1D + 1.39 P_2D − 0.66 P_2V − 0.78 P_5D + 0.39 P_5V	0.33
II	**C2**	0.05 − 0.31 Recho6 + 0.54 Recho7 + 1.95 Recho8 − 0.70 Recho23	0.27
III	**C3**	0.08 + 0.35 P_Head − 0.90 P_1D + 1.17 P_2D − 0.57· P_2V − 1.10 P_5D + 0.36 P_5V − 0.75 P_Tail + 0.42 Recho7 + 1.56 Recho8 − 0.53· Recho23	0.41
IV	**C4**	− 0.01 + 0.41 P_Head + 0.79 P_2D − 0.33 P_2V − 0.45 Recho23	0.36
IVbis	**C4bis**	0.007 + 0.78 P_2D − 0.33 P_2V + 1.74 SQ_Head − 0.65 Recho23	0.36

Relative efficiency of indirect predictors

The relative efficacy (RES) of predictors and the simple variables composing the predictors was computed individually. Values of heritabilities, genetic correlations with logr_Fil and logr_Carc and RES of the proportion of surfaces, ratios of echography measurements and composite predictors are shown in table 7. Values of RES were generally higher for variables and predictors involved in the prediction of carcass than in that of fillet. Logr_Carc predictors showed the best results of RES (0.52 to 0.70). Two simple variables, Prop_Head and R_Echo23, showed values of RES (0.54 and 0.51, respectively) comparable to logr_Carc predictors motivating their use in the selection on carcass yield (table 7). Ranging from 0.08 (F2) to 0.28 (F4), RES of fillet yield predictors were in all cases lower than RES of carcass yield predictors. The highest RES were obtained with F3 and F4 predictors (0.28 and 0.27) only marginally higher than with the simple R_Echo9 predictor (0.23). No predictor or single trait showed both a high efficiency for increasing fillet and carcass yield, the best single traits being R_Echo23 and P_Head.

Table 7
Relative efficacy comparison for selection on fillet and carcass yield using elements of predictors or predictors themselves. Heritability and ± standard error of components of predictors, genetic correlations component of predictors and trait to predict and relative efficiency of selection for fillet yield or carcass yield are shown.

Trait	h^2	r_{A1} *(trait,logr_FilW)*	r_{A2} *(trait, logr_CarcW)*	RES Fillet $\sqrt{h^2}.\lvert r_{A1}\rvert$	RES Carcass $\sqrt{h^2}.\lvert r_{A2}\rvert$
P_Head	0.40 ± 0.09	− 0.26 ± 0.20	0.85 ± 0.06	0.16	0.54
SQ_Head	0.36 ± 0.08	− 0.12 ± 0.22	0.54 ± 0.14	0.07	0.32
P_1D	0.30 ± 0.08	0.20 ± 0.22	0.52 ± 0.14	0.11	0.28
P_1V	0.28 ± 0.07	− 0.33 ± 0.22	− 0.62 ± 0.13	0.17	0.33
P_2D	0.23 ± 0.06	0.12 ± 0.24	0.76 ± 0.10	0.06	0.36
P_2V	0.31 ± 0.08	− 0.14 ± 0.22	− 0.08 ± 0.19	0.08	0.04
P_3D	0.27 ± 0.07	0.05 ± 0.24	− 0.57 ± 0.15	0.03	0.30
P_3V	0.27 ± 0.07	0.03 ± 0.23	− 0.66 ± 0.13	0.02	0.34
P_4D	0.22 ± 0.06	0.09 ± 0.23	0.23 ± 0.20	0.04	0.11
P_4V	0.24 ± 0.07	0.005 ± 0.23	− 0.33 ± 0.17	0.002	0.16
P_5D	0.19 ± 0.06	0.32 ± 0.25	0.02 ± 0.22	0.14	0.01
P_5V	0.16 ± 0.06	0.29 ± 0.25	0.05 ± 0.22	0.12	0.02
P_Tail	0.29 ± 0.07	− 0.10 ± 0.23	0.40 ± 0.16	0.05	0.22
R_Echo1	0.07 ± 0.04	− 0.67 ± 0.24	*	0.18	*
R_Echo2	0.10 ± 0.04	− 0.42 ± 0.27	− 0.97 ± 0.14	0.13	0.31
R_Echo3	0.06 ± 0.04	− 0.49 ± 0.30	*	0.12	*
R_Echo4	0.21 ± 0.07	− 0.04 ± 0.24	− 0.28 ± 0.20	0.02	0.13
R_Echo5	0.24 ± 0.06	− 0.21 ± 0.24	− 0.57 ± 0.15	0.10	0.28
R_Echo6	0.17 ± 0.05	− 0.03 ± 0.24	− 0.47 ± 0.17	0.01	0.19
R_Echo7	*	*	*	*	*
R_Echo8	0.05 ± 0.04	− 0.04 ± 0.37	0.84 ± 0.26	0.01	0.19
R_Echo9	0.23 ± 0.07	0.49 ± 0.21	0.01 ± 0.20	0.23	0.004
R_Echo23	0.38 ± 0.08	− 0.30 ± 0.21	− 0.83 ± 0.06	0.18	0.51
F1	0.15 ± 0.05	0.45 ± 0.22	−	0.17	−
F2	0.14 ± 0.05	0.22 ± 0.26	−	0.08	−
F3	0.18 ± 0.07	0.67 ± 0.17	−	0.28	−
F4	0.27 ± 0.08	0.52 ± 0.20	−	0.27	−
F4bis	0.38 ± 0.09	0.17 ± 0.22	−	0.10	−
C1	0.51 ± 0.09	−	0.91 ± 0.03	−	0.65
C2	0.33 ± 0.08	−	0.91 ± 0.04	−	0.52
C3	0.50 ± 0.09	−	0.95 ± 0.02	−	0.67
C4	0.55 ± 0.10	−	0.94 ± 0.03	−	0.70
C4bis	0.50 ± 0.09	−	0.94 ± 0.03	−	0.66

* = lack of convergence

Table 8.
Heritability and standard error (diagonal), genetic correlation and standard error (above the diagonal), phenotypic correlations (below the diagonal) for processing traits obtained as residual of the linear regression between logarithm of body compartment and logarithm of body weight in European sea bass (*Dicentrarchus labrax* L.).

	BW	**logr_Carc**	**logr_HGCarc**	**logr_Fil**	**logr_Head**	**logr_Axis**	**logr_Viscera**
BW	**0.41 ± 0.09**	− 0.12 ± 0.17	− 0.04 ± 0.20	0.01 ± 0.21	− 0.18 ± 0.18	− 0.11 ± 0.22	0.12 ± 0.18
logr_Carc	− 0.03	**0.57 ± 0.11**	0.77 ± 0.08	0.28 ± 0.20	0.87 ± 0.05	0.72 ± 0.13	− 0.99 ± 0.002
logr_HGCarc	0.06	0.64	**0.32 ± 0.08**	0.79 ± 0.09	0.36 ± 0.18	0.33 ± 0.23	− 0.76 ± 0.09
logr_Fil	0.05	0.21	0.66	**0.21 ± 0.08**	− 0.19 ± 0.23	− 0.31 ± 0.24	− 0.25 ± 0.20
logr_Head	− 0.11	0.71	− 0.06	− 0.32	**0.44 ± 0.10**	0.78 ± 0.12	− 0.88 ± 0.05
logr_Axis	− 0.003	0.41	0.2	− 0.59	0.35	**0.16 ± 0.06**	− 0.74 ± 0.12
logr_Viscera	0.06	− 0.96	− 0.62	− 0.17	− 0.69	− 0.42	**0.56 ± 0.11**

Table 9.
Heritability and standard error (diagonal), genetic correlations ± standard error (above the diagonal) and phenotypic correlations (below the diagonal) for growth and carcass traits in European sea bass (*Dicentrarchus labrax* L.). Predictors F1-F4bis relate to prediction of fillet percentage, models C1-C4bis to prediction of carcass percentage

	logr_Fil	**logr_Carc**	**F1**	**F2**	**F3**	**F4**	**F4bis**	**C1**	**C2**	**C3**	**C4**	**C4bis**
logr_Fil	**0.20 ± 0.07**	0.28 ± 0.20	0.45 ± 0.22	0.22 ± 0.26	0.67 ± 0.17	0.52 ± 0.20	0.17 ± 0.22	0.03 ± 0.22	0.26 ± 0.22	0.14 ± 0.21	0.14 ± 0.21	0.22 ± 0.21
logr_Carc	0.21	**0.57 ± 0.10**	− 0.40 ± 0.19	− 0.20 ± 0.23	− 0.22 ± 0.21	− 0.56 ± 0.15	− 0.52 ± 0.13	0.91 ± 0.03	0.91 ± 0.04	0.95 ± 0.02	0.94 ± 0.03	0.94 ± 0.03
F1	0.38	− 0.18	**0.15 ± 0.05**	0.62 ± 0.20	0.86 ± 0.06	0.74 ± 0.11	0.47 ± 0.17	− 0.56 ± 0.15	− 0.46 ± 0.19	− 0.51 ± 0.17	− 0.56 ± 0.16	− 0.54 ± 0.17
F2	0.13	0.02	0.13	**0.14 ± 0.05**	0.49 ± 0.22	0.39 ± 0.22	0.003 ± 0.22	− 0.36 ± 0.20	− 0.35 ± 0.22	− 0.29 ± 0.23	− 0.40 ± 0.21	− 0.30 ± 0.21
F3	0.39	− 0.09	0.87	0.18	**0.18 ± 0.07**	0.87 ± 0.05	0.65 ± 0.12	− 0.46 ± 0.17	− 0.12 ± 0.21	− 0.30 ± 0.19	− 0.35 ± 0.19	− 0.34 ± 0.18
F4	0.34	− 0.25	0.74	0.08	0.88	**0.27 ± 0.08**	0.81 ± 0.06	− 0.77 ± 0.09	− 0.35 ± 0.18	− 0.61 ± 0.13	− 0.62 ± 0.12	− 0.63 ± 0.14
F4bis	0.26	− 0.25	0.62	− 0.06	0.68	0.76	**0.38 ± 0.09**	− 0.58 ± 0.11	− 0.13 ± 0.18	− 0.46 ± 0.14	− 0.40 ± 0.15	− 0.49 ± 0.14
C1	− 0.04	0.67	− 0.26	− 0.04	− 0.29	− 0.56	− 0.44	**0.51 ± 0.09**	0.77 ± 0.08	0.95 ± 0.01	0.95 ± 0.01	0.95 ± 0.01
C2	0.02	0.68	− 0.28	0.06	0.002	− 0.07	− 0.08	0.54	**0.33 ± 0.08**	0.87 ± 0.04	0.91 ± 0.03	0.88 ± 0.04
C3	0.02	0.73	− 0.23	0.08	− 0.08	− 0.32	− 0.31	0.89	0.80	**0.50 ± 0.09**	0.97 ± 0.01	0.98 ± 0.009
C4	− 0.03	0.74	− 0.3	− 0.08	− 0.13	− 0.37	− 0.30	0.91	0.75	0.92	**0.55 ± 0.10**	0.98 ± 0.004
C4bis	− 0.01	0.70	− 0.27	− 0.04	− 0.18	− 0.3	− 0.37	0.87	0.75	0.91	0.97	**0.50 ± 0.09**

Expected Genetic Gain

Estimation of genetic gains was performed using the genetic and phenotypic parameters of the different traits and predictors. Gains obtained for selection on carcass yield were higher than those obtained for selection on fillet yield in all cases of direct selection. Under the hypothetical mass selection (not feasible for selection purposes involving the sacrifice of the individuals), genetic gains achievable per generation for the two intensities of selection tested ranged from 1.13 to 1.71% per generation for carcass percentage and from 0.38 to 0.57% for fillet percentage (table 10). Genetic gain for FS with 10 sibs selected per family and 30% selection pressure was 0.95% for carcass yield and 0.43% for fillet yield. Genetic gains obtained by indirect selection (IS) were higher for carcass yield than for fillet yield. Genetic gain achieved applying carcass predictors ranged from 1.18% to 1.58%, whereas fillet predictors resulted in values between 0.10% and 0.36%. Indirect selection applied with simple variables (R_Echo23 and Prop_Head) showed better gains for carcass (1.17% and 1.22%) than for fillet yield (0.14% and 0.16%). Correlated response in carcass percent when selecting on fillet predictors was negative, while this was not the case for fillet when selection was performed on carcass predictors. No method could achieve both a high gain in fillet percent and a high gain in carcass percent

Table 10.
Expected genetic gain (in percent body weight units)for a generation of selection based on Massal (MS), Full Sib (FS), and Indirect (IS) selection for fillet and carcass yield, with different selection intensities (% selected).

Trait selected	Type of selection	Selection intensity	Genetic gain per generation	
			Fil%	Carc%
F1	IS	10%	0.22%	- 0.35%
F2	IS	10%	0.10%	- 0.17%
F3	IS	10%	0.36%	0.21%
F4	IS	10%	0.34%	- 0.66%
F4bis	IS	10%	0.13%	- 0.73%
C1	IS	10%	0.03%	1.47%
C2	IS	10%	0.19%	1.18%
C3	IS	10%	0.12%	1.52%
C4	IS	10%	0.13%	1.58%
C4bis	IS	10%	0.20%	1.51%
R_Echo23	IS	10%	0.16%	1.17%
Prop_head	IS	10%	0.14%	1.22%
Logr_Fil	MS	10%	0.57%	0.29%
Logr_Fil	MS	30%	0.38%	0.19%
Logr_Carc	MS	10%	0.19%	1.71%
Logr_Carc	MS	30%	0.12%	1.13%
Logr_Fil	FS	10%	0.65%	0.33%
Logr_Fil	FS	30%	0.43%	0.22%
Logr_Carc	FS	10%	0.16%	1.43%
Logr_Carc	FS	30%	0.10%	0.95%

Discussions

This study provides for the first time a complete description of growth and processing traits in European sea bass, including the knowledge of their genetic parameters (heritability and genetic correlations) and the potential genetic gain for processing traits in different breeding strategies. Results were however only obtained for one strain and in one environment and extension to other strains or environments should then be considered with caution. Previous studies had estimated genetic parameters for fillet percent in European sea bass either on a small family setting (27 families from 9 sires and 3 dams - Saillant et al., 2009) from wild parents from West Mediterranean origin, on different populations derived from wild parents (Vandeputte et al., 2014) or on the same population (Bestin et al., 2014). Here, we also predicted processing yields using morphological traits obtained from external (digital photo) and internal (ultrasound imagery) measurements. This was done with the aim to replace selection methods implicating the sacrifice of the individuals (mass selection not feasible in practice on lethal traits) or their relatives (sib selection), by direct selection of candidates based on the use of predictors as indirect criteria. A similar approach has already been done in other species such as catfish (*Ictalurus spp*) (Bosworth et al., 2001) where the phenotypic relationship between body shape and ultrasound images of muscle sections and meat yield was used to identify high meat-producing individuals. Moreover, Rutten et al. (2004) perfprmed a regression model based on strain-specific coefficients to predict both fillet weight and fillet yield in Nile tilapia (*Oreochromis niloticus*). Sang et al. (2009) estimated phenotypic fillet weight and fillet yield using six body measurements (including volume) and four circumferences integrated in a multiple regression model in a reference population composed by 200 river catfish (*Pangasianodon hypophthalmus)*. Haffray et al. (2013) predicted processing yields through multiple linear regression models comprising 2D and ultrasound measures providing genetic parameters and theoretical gains achievable applying predictors for the direct selection of live candidates on rainbow trout. Recently Perazza et al. (2015) also estimated phenotypically loin-eye area in tambaqui (*Colossoma macropomum*) using ultrasounds. To our knowledge this study is the first large scale analysis on fillet

yield prediction for a Mediterranean fish species reared under commercial conditions, integrating ultrasound imagery and body shape measurements.

Fixed effects

Effects of sex and mating plan were seen on all weight traits and some of the yields. The effect of mating plan can be explained by heterogeneity in the composition of the female groups used, and this heterogeneity can be of genetic (different subsets of the selection nucleus) or environmental (egg quality in separately reared female groups) origin. However, as we used mating plan as fixed effect, this is not expected to bias genetic parameter estimates as long as these subgroups are a representative sample of the selection nucleus. Females tended to have a lower fillet yield compared to males (- 0.2%), but this difference was not significant contrary to that reported by Saillant et al. (2009), who found a difference among sexes of 2.4%. A size effect may be hypothesised to explain this difference as fish from Saillant et al. (2009) were almost twice as big as the fish processed in the present study (741 g vs 395 g). This difference is most likely due to the sexual maturation in the older fish recorded by Saillant et al. (2009) while here gonads were on average smaller than 1% of the body weight. The only significant sexual dimorphisms reported here were for head yield (F>M), headless gutted carcass yield (F<M) and axis yield (F<M) showed in table 2.

Genetic parameters of carcass and processing traits

The heritabilities estimated for BW (0.40) and K (0.50) were generally intermediate or high when compared with previous experiments in sea bass that varied from 0.19 to 0.63 for BW and 0.10 to 0.49 for K (Bardon et al., 2009; Dupont-Nivet et al., 2008, 2010; Haffray et al., 2007; Karahan et al., 2013; Saillant et al., 2006, 2009; Vandeputte et al., 2014). They are in the upper range for Carc% (0.57 *vs* 0.48 - 0.74), equivalent (0.20 *vs* 0.20 - 0.25) for Fil% (Haffray et al., 2007; Saillant et al., 2009; Vandeputte et al., 2014) and lower (0.46 vs 0.87) for Head% (Saillant et al., 2009). Compared to another year class of the same strain estimated in the previous generation of selection (Bestin et al., 2014), heritability was higher in the present

study for BW (0.40 *vs* 0.19), K (0.50 *vs* 0.37) and Head% (0.46 *vs* 0.32) and similar for Carc% (0.57 *vs* 0.54), HGCarc% (0.32 *vs* 0.35) and Fil% (0.20 *vs* 0.20). The last results confirmed also that after 4 generations of selection, the EMG line exhibited high genetic variation for growth and processing traits.

The weights of body parts were highly correlated both genetically and phenotypically, and had heritability close to that of body weight, as generally seen in this type of studies – there is much more phenotypic variation in body size and weight than in proportions of the different body parts. Fillet and carcass percent were approached both by the ratio trait and by the residual of the log-log regression of the body part weight on body weight, as earlier proposed (Haffray et al., 2013; Vandeputte et al., 2014). Log-residuals and yields had similar heritabilities (0.20 - 0.21 for fillet yield, 0.57 for carcass yield) and were very highly genetically correlated (>0.99) showing they describe the same trait.

The present study showed a low genetic correlation between carcass yield and fillet yield (- 0.01 ± 0.20 with the ratio trait or 0.28 ± 0.20 with the log residuals,) as previously reported in another marine species such as sea bream (- 0.05 ± 0.19, Navarro et al., 2009). This limited genetic correlation may be due to the fact that head yield has a strongly positive genetic correlation with carcass yield (0.87 - 0.90), so that sea bass with high carcass percentage have also a larger head. In addition, the genetic correlations between head yield and fillet yield were negative (- 0.19 to - 0.41) as reported earlier in sea bass (- 0.73) by Saillant et al. (2009), rainbow trout (- 0.50 to - 0.53, Haffray et al., 2012) and tilapia (- 0.94, Rutten et al., 2005). Head yield has a very high mean value (19.1 %) when compared with rainbow trout (11.1 %, Haffray et al., 2012). This difference in relative importance of head, fillet and viscera between species may lead to different types of genetic and phenotypic relationships between body traits among different species. In sea bass for example, Head% had a coefficient of variation (8.4 %) higher than that of other body compartments (2.1 % for Carc%; 2.8 % for Fil%), questioning on the unexpected consequences that selection for smaller head, already applied on other fish species for the improvement of fillet yield, could have when applied to sea bass.

Predictors

We evaluated the use of headless gutted carcass yield as a possible lethal predictor of fillet yield not requiring the error-prone filleting operation, as reported in rainbow trout (Haffray et al., 2013). This trait had a heritability higher than that of fillet yield (0.32 vs. 0.20 - 0.21) and a reasonably high genetic correlation with fillet yield (0.72 - 0.79), although the figures were less favourable than in rainbow trout (h^2 = 0.54 - 0.55, r_A = 0.97 - 0.98, Haffray et al. (2013)). The use of headless gutted carcass yield as a surrogate for fillet yield in slaughtered sibs (faster measurement, less subject to errors) in this population of sea bass is thus not as straightforward as it is in rainbow trout.

Through a multiple linear regression approach we selected ten models (five types per character to predict), to predict indirectly fillet and carcass percentage. Models of type I (using proportion of surfaces as predictors) and II (using ratios of echography measurements), gave medium to low values of R^2. The lower R^2 and lower heritability of type II models could be related to the low heritability values obtained from ratios of echography measurements. These poorer results may be linked to measurement errors, which are easily done through inexact positioning of the probe (for Recho 1, 2, 3, $h^2 \leq 0.10$ – table 7) or measurement of small distances (for Recho 7 and 8 – table 7). It is symptomatic that the highest heritabilities are obtained for the echographies measuring the largest distance (Recho23, depth of the body cavity h^2=0.38 – table 7) or measuring fillet thickness at well-defined points on the lateral line (Recho 4 to 6, $0.17 \leq h^2 \leq 0.24$ – table 7). We also have to mention that in this case, all fish in the dataset were used to set up the regression equations, and not only the 10% higher or lower performers as in Haffray et al. (2013), which may also contributed to the apparently lower precision of phenotypic prediction.

In order to simplify phenotyping, especially in fish farms and/or in a context of real-time measurement to integrate the predictors in an hypothetical automated sorting system, we tested models of type IV and IV_{bis}, with less, easier to measure variables such as the relative head surface (type IV) or the relative surface of a square containing the head (SQ_Head, in type IV_{bis}) together with the deepness of the abdominal cavity, that was shown (table 7) to have a high potential for predicting the

genetic variation of both fillet and carcass yield. In the case of carcass yield, the type III model including both surfaces and echography measurements (C3) performed better than models C1 and C2 (R^2=0.41 vs 0.33 and 0.27). The simplified models (C4 and $C4_{bis}$) were equivalent (R^2=0.36 for both), justifying the application of the SQ_Head. Models C4 and $C4_{bis}$ were characterized by the presence of dorsal (P_2D) and ventral (P_2V) proportion of the body surfaces below the first dorsal fin. The coefficients of these two body surfaces were positively (dorsal) and negatively (ventral) related with the carcass yield.

Moreover the regression coefficients for the head were positive in all five models confirming as seen before that fish with large heads also have a high phenotypic carcass percentage. In the end P_Head (or alternatively SQ_Head), P_T2D and P_T2V, jointly to R_echo23, were the measurements with the highest explanatory power on carcass yield. Heritabilities of predictors of carcass yield, ranging from 0.33 to 0.55, were generally close to the heritability of the trait to predict logr_Carc (0.57). Genetic and phenotypic correlations between predictors and carcass yield were high (0.91 to 0.95 – table 9), thus making them suitable for indirect selection on live candidates (+1.18 to 1.58% carcass per generation with 10% pressure – table 10), better than sib selection on slaughtered sibs (0.95% gain with 30% pressure, 1.43% with 10%). Carcass predictors were also positively correlated with fillet yield, and gave positive (albeit very low) responses in fillet percentage (0.03 to 0.20%).

The prediction capability of fillet predictors was generally lower than that of carcass predictors. The best model including only proportion of surfaces (F1) gave a R^2 of 16%, and F3 was the best model with a R^2 of 18%. The simplified models F4 and $F4_{bis}$ exhibited lower values of R^2 (0.14 and 0.10, respectively). In terms of potential genetic gains, the best models were F3 and F4 which gave 0.36% and 0.34% of fillet gains, respectively. They were outperformed by sib selection (0.65% with the same selection intensity), although when a more realistic selection intensity (30% of the families selected) was applied to sib selection on fillet, the results were rather similar (0.43% per generation). Interestingly, and in contradiction with the limited but positive effect on fillet yield of selection on carcass predictors reported before, selection on all fillet predictors provided negative gains on carcass yield (- 0.17 to - 0.73%). This means that the two types of predictors did not act on the same parts of

the sea bass body. This is especially due to the importance of head surface in these models, which, contrary to what is seen in carcass predictors, has negative coefficients in fillet predictors and thus opposite results.

Conclusions

Then, selecting for improved fillet percentage in sea bass cannot follow the general and easy picture in fish of increasing carcass percentage (Haffray et al., 2012; Kause et al., 2007; Kocour et al., 2007; Rutten et al., 2005) and decreasing head size (Haffray et al., 2012; Kocour et al., 2007; Rutten et al., 2005), as head size has opposed effects on carcass yield and fillet yield. Sib selection on both carcass and fillet yield could be an opportunity and should create genetic progress on the two traits, as they are weakly if at all genetically correlated. However, selection using predictors need to be considered with caution to not deteriorate fillet yield when selecting to improve carcass yield or vice versa. These unusual results may however be linked to the population or the rearing conditions used, and should be confirmed with similar studies in other sea bass populations or ongrowing environments, also considering factors that may interact with fat deposition in the belly, such as feed composition. More generally, this study highlighted that the relative development of different body compartments might be rather species specific, and requires further studies to be able to draw a more general picture.

Acknowledgements

This work was financed by France Agrimer under the BAR-3D project (convention SIVAL NL: 2012-0302). I would like to thank the Erasmus-plus mobility programme and the Section of Animal Science of the Department of Agriculture of Sassari that financed the experience abroad, Dr. Béatrice Chatain for allowing my participation in the project and Dr. Marc Vandeputte for the welcome, all suggestions and the assistance given on data analysis during the traineeship. I wish to acknowledge Alain Vergnet, Clémence Fraslin, Doan Quoc Khanh, Sébastien Ferrari and François Allal, all researchers, technicians and employees that I met at the Ifremer Station of Palavas les Flots.

REFERENCES

Acosta-Alba, I., Aubin, J., Cariou, S., Haffray, P., Quittet, B., Vandeputte, M., 2015. Le programme FilDor: associer sélection génétique et anlayse d'impact environnemental des élevages de poissons. Courr. Envir. INRA 65, 49-57.

Bardon, A., Vandeputte, M., Dupont-Nivet, M., Chavanne, H., Haffray, P., Vergnet, A., Chatain, B., 2009. What is the heritable component of spinal deformities in the European sea bass (*Dicentrarchus labrax*)? Aquaculture 294, 194–201. doi:10.1016/j.aquaculture.2009.06.018.

Bestin, A., Dupont-Nivet, M., Haffray, P., Médale, F., Quillet, E., Vandeputte, M., Cariou, S., Desgranges, A., Laureau, S., Ricoux, R., Beutin, C., 2014. Genotype by diet interactions on growth and processing traits in rainbow trout (*Oncorhynchus mykiss*), European sea bass (*Dicentrarchus labrax*), gilthead sea bream (*Sparus aurata*) and meagre (*Argyrosomus regius*) fed diets with almost complete substitution of both fish meal and fish oil by vegetal ingredients. Proc.10th WCGALP, Vancouver, Canada.

Bisault, L., 2009. La salmoniculture française à l'étiage. Agreste Primeur 227, 4.

Bosworth, B. G., Holland, M., Brazil, B. L., 2001. Evaluation of ultrasound imagery and body shape to predict carcass and fillet yield in farm-raised catfish. J Anim. Sci. 79, 1483-1490.

Chatain, B., Chavanne, H., 2009. Genetics of European sea bass (*Dicentrarchus labrax* L.). *Cah. Agricult.* 18, 249-255. http://dx.doi.org/10.1684/agr.2009.0296.

Costa, C., Antonucci, F., Boglione, C., Menesatti, P., Vandeputte, M., Chatain, B., 2013. Automated sorting for size, sex and skeletal anomalies of cultured seabass using external shape analysis. Aquac. Eng. 52, 58–64. doi:10.1016/j.aquaeng.2012.09.001.

Dupont-Nivet, M., Vandeputte, M., Vergnet, A., Merdy, O., Haffray, P., Chavanne, H., Chatain, B., 2008. Heritabilities and GxE interactions for growth in the European sea bass (*Dicentrarchus labrax L.)* using a marker-based pedigree. Aquaculture 275, 81–87 doi:10.1016/j.aquaculture.2007.12.032.

Dupont-Nivet, M., Karahan-Nomm, B., Vergnet, A., Merdy, O., Haffray, P., Chavanne, H., Chatain, B., Vandeputte, M., 2010. Genotype by environment interactions for growth in European seabass (*Dicentrarchus labrax*) are large when growth rate rather than weight is considered, Aquaculture. doi:10.1016/j.aquaculture.2010.05.025.

Falconer, D.S., MacKay, T.F.C., 1996. Introduction to Quantitative Genetics, 4th ed. Longman Scientific and Technical, Harlow, UK.

FAO 2011-2016 © . Fisheries and aquaculture software. FishStatJ - software for fishery statistical time series. In: *FAO Fisheries and Aquaculture Department* [online]. Rome. Updated 21 July 2016.

García-Celdrán, M., Ramis, G., Manchado, M., Estévez, A., Afonso, J.M., Armero, E., 2015. Estimates of heritabilities and genetic correlations of carcass quality traits in a reared gilthead sea bream (*Sparus aurata L.*) population sourced from three broodstocks along the Spanish coasts. Aquaculture 446, 175–180. doi:10.1016/j.aquaculture.2015.04.028.

Gorshkov, S., Gorshkova, G., Meiri, I., Gordin, H., 2004. Culture performance of different strains and crosses of the European sea bass (*Dicentrarchus labrax*) reared under controlled conditions at Eilat, Israel. J. Appl. Ichthyol. 20, 194–203. doi:10.1111/j.1439-0426.2004.00513.x.

Groeneveld, E., Kovač, M., Mielenz., N., 2008. VCE - User's Guide and Reference Manual - Version 6.0. Institute of Farm Animal Genetics, Neustadt, Germany.

Haffray P., Pincent C., Dupont-Nivet, M., Vandeputte, M., Merdy, O., Chavanne, H., Chatain, B., 2007. Heritabilities and G × E interactions for quality traits in the European sea bass (*Dicentrarchus labrax L.*). Aquaculture, 272, S265. doi:10.1016/j.aquaculture.2007.07.080.

Haffray, P., Bugeon, J., Pincent, C., Chapuis, H., Mazeiraud, E., Rossignol, M.-N., Chatain, B., Vandeputte, M., Dupont-Nivet, M., 2012. Negative genetic correlations between production traits and head or bony tissues in large all-female rainbow trout (*Oncorhynchus mykiss*). Aquaculture 368-369, 145–152. doi:10.1016/j.aquaculture.2012.09.023.

Haffray, P., Bugeon, J., Rivard, Q., Quittet, B., Puyo, S., Allamelou, J.M., Vandeputte, M., Dupont-Nivet, M., 2013. Genetic parameters of in-vivo prediction of carcass, head and fillet yields by internal ultrasound and 2D external imagery in large rainbow trout (*Oncorhynchus mykiss*). Aquaculture 410-411, 236–244. doi:10.1016/j.aquaculture.2013.06.016.

Karahan, B., Chatain, B., Chavanne, H., Vergnet, A., Bardon, A., Haffray, P., Dupont-Nivet, M., Vandeputte, M., 2013. Heritabilities and correlations of deformities and growth-related traits in the European sea bass (*Dicentrarchus labrax* , L) in four different sites. Aquac. Res. 44, 289–299. doi:10.1111/j.1365-2109.2011.03082.x.

Kause, A., Paananen, T., Ritola, O., Koskinen, H., 2007. Direct and indirect selection of visceral lipid weight, fillet weight, and fillet percentage in a rainbow trout breeding program. J. Anim. Sci. 85, 3218–3227. doi:10.2527/jas.2007-0332

Kocour, M., Mauger, S., Rodina, M., Gela, D., Linhart, O., Vandeputte, M., 2007. Heritability estimates for processing and quality traits in common carp (Cyprinus carpio L.) using a molecular pedigree. Aquaculture 270, 43–50. doi:10.1016/j.aquaculture.2007.03.001

Navarro, A., Zamorano, M.J., Hildebrandt, S., Ginés, R., Aguilera, C., Afonso, J.M., 2009. Estimates of heritabilities and genetic correlations for growth and carcass traits in gilthead seabream (*Sparus auratus* L.), under industrial conditions. Aquaculture 289, 225–230.

Neira, R., Lhorente, J.P., Araneda, C., Díaz, N., Bustos, E., Alert, A., 2004. Studies on carcass quality traits in two populations of Coho salmon (*Oncorhynchus kisutch*): phenotypic and genetic parameters. Aquaculture 241, 117–131. doi:10.1016/j.aquaculture.2004.08.009.

Perazza, C. A., Pinaffi, F. L. V., Silva, L. A., Hilsdorf, A. W. S., 2015. Evaluation of ultrasound imaging to predict loin eye area in tambaqui. Bol. Inst. Pesca, São Paulo, 41, 803 – 809.

Rutten, M. J. M., Bovenhuis, H., Komen, H., 2004. Modeling fillet traits based on body measurements in three Nile tilapia strains (*Oreochromis niloticus* L.). Aquaculture 231, 113–122.

Rutten, M.J.M., Bovenhuis, H., Komen, H., 2005. Genetic parameters for fillet traits and body measurements in Nile tilapia (*Oreochromis niloticus L.*). Aquaculture 246, 125–132.

Saillant, E., Fostier, A., Haffray, P., Menu, B., Thimonier, J., Chatain, B., 2002. Temperature effects and genotype-temperature interactions on sex determination in the European sea bass (*Dicentrarchus labrax L.*). J. Exp. Zool. 292, 494–505. doi:10.1002/jez.10071.

Saillant, E., Dupont-Nivet, M., Haffray, P., Chatain, B., 2006. Estimates of heritability and genotype–environment interactions for body weight in sea bass (*Dicentrarchus labrax L.*) raised under communal rearing conditions. Aquaculture 254, 139–147. doi:10.1016/j.aquaculture.2005.10.018.

Saillant, E., Dupont-Nivet, M., Sabourault, M., Haffray, P., Laureau, S., Vidal, M.-O., Chatain, B., 2009. Genetic variation for carcass quality traits in cultured sea bass (*Dicentrarchus labrax*). Aquat. Living Resour. 22, 105–112. doi:10.1051/alr/2009010.

Sang, N. Van, Thomassen, M., Klemetsdal, G., Gjøen, H.M., 2009. Prediction of fillet weight, fillet yield, and fillet fat for live river catfish (*Pangasianodon hypophthalmus*). Aquaculture 288, 166–171. doi:10.1016/j.aquaculture.2008.11.030.

Vandeputte, M., Dupont-Nivet, M., Chavanne, H., Chatain, B., 2007. A polygenic hypothesis for sex determination in the European sea bass *Dicentrarchus labrax*. Genetics 176, 1049–1057.

Vandeputte, M., Dupont-Nivet, M., Haffray, P., Chavanne, H., Cenadelli, S., Parati, K., Vidal, M.-O., Vergnet, A., Chatain, B., 2009. Response to domestication and selection for growth in the European sea bass (*Dicentrarchus labrax*) in separate and mixed tanks. Aquaculture 286, 20–27. doi:10.1016/j.aquaculture.2008.09.008.

Vandeputte, M., Garouste, R., Dupont-Nivet, M., Haffray, P., Vergnet, A., Chavanne, H., Laureau, S., Ron, T.B., Pagelson, G., Mazorra, C., Ricoux, R., Marques, P., Gameiro, M., Chatain, B., 2014a. Multi-site evaluation of the rearing performances of 5 wild populations of European sea bass (*Dicentrarchus labrax*). Aquaculture 424-425, 239–248. doi:10.1016/j.aquaculture.2014.01.005.

Figures

Figure 1

Scheme of the different echography measurements E1: anterior-dorsal; E2: intermediate-dorsal; E3: posterior-dorsal; E4: anterior-ventral; E5: intermediate-ventral; E6: posterior-ventral; E7: anterior-abdominal; E8: intermediate-abdominal; E9: posterior-abdominal; E23: depth of the belly cavity.

Figure 2

Landmarks (bold) and semi-landmarks (italics) placed on each sea bass picture

1: head extremity, **2**: top of the eye; **3**: bottom of the eye; **4**: top end of the head and beginning of the dorsal fillet; 5: lateral line and head intersection; **6**: operculum extremity; **7**: left top extremity of pectoral fin; **8**: left bottom extremity of pectoral fin; **9**: right extremity of pectoral fin; **10**: pelvic fin insertion; **11**: bottom end of the head and beginning of the ventral fillet; **12**: jaw extremity; **13**: beginning of the first dorsal fin; *14*: intersection of the lateral line with the vertical of point 13; *15*: intersection of the ventral wall with the vertical of point 13; **16**: end of the first dorsal fin; *17*: intersection of the lateral line with the vertical of point 16; *18*: intersection of the ventral wall with the vertical of point 16; **19**: beginning of the second dorsal fin; *20*: intersection of the lateral line with the vertical of point 19; *21*: intersection of the ventral wall with the vertical of point 19; **22**: beginning of the anal fin; **23**: end of the second dorsal fin; *24*: intersection of the lateral line with the vertical of point 23; *25*: intersection of the ventral wall with the vertical of point 23; **26**: end of the anal fin; **27**: top of the caudal peduncle; *28*: intersection of the lateral line with the vertical of point 27; *29*: intersection of the ventral wall with the vertical of point 27; **30**: fork of the caudal fin.

Figure 3

a) subdivision of the fish in 28 triangles b) triangles grouped in 12 representative surfaces

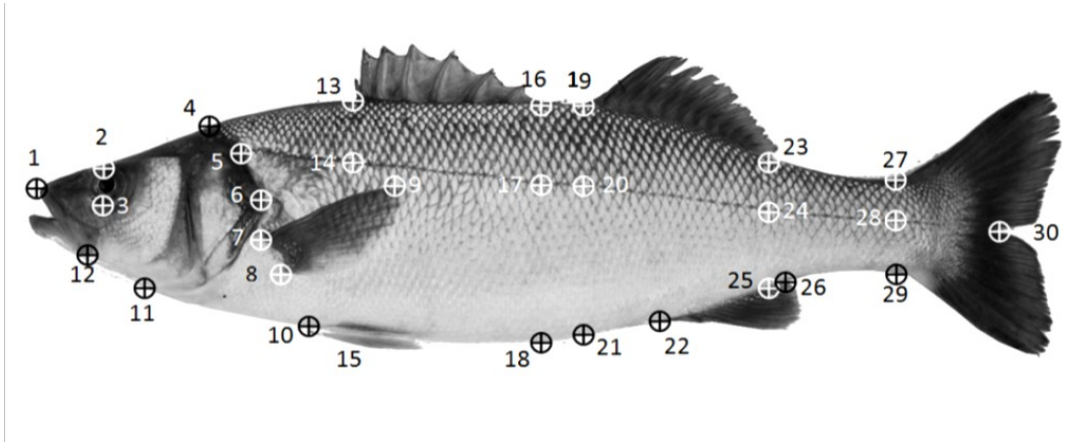

Figure 1

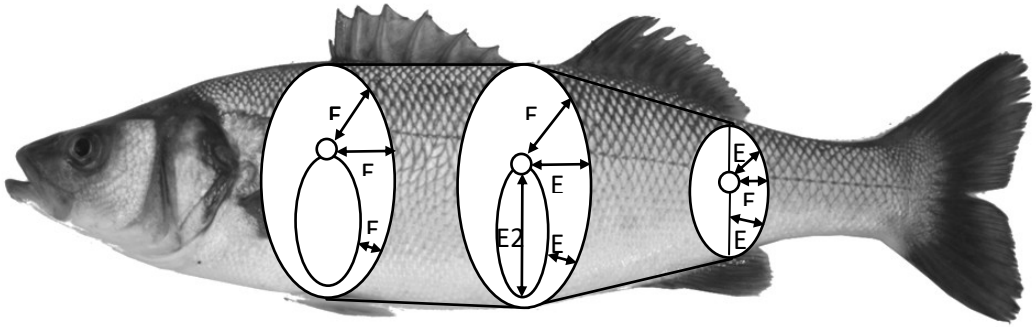

Figure 2

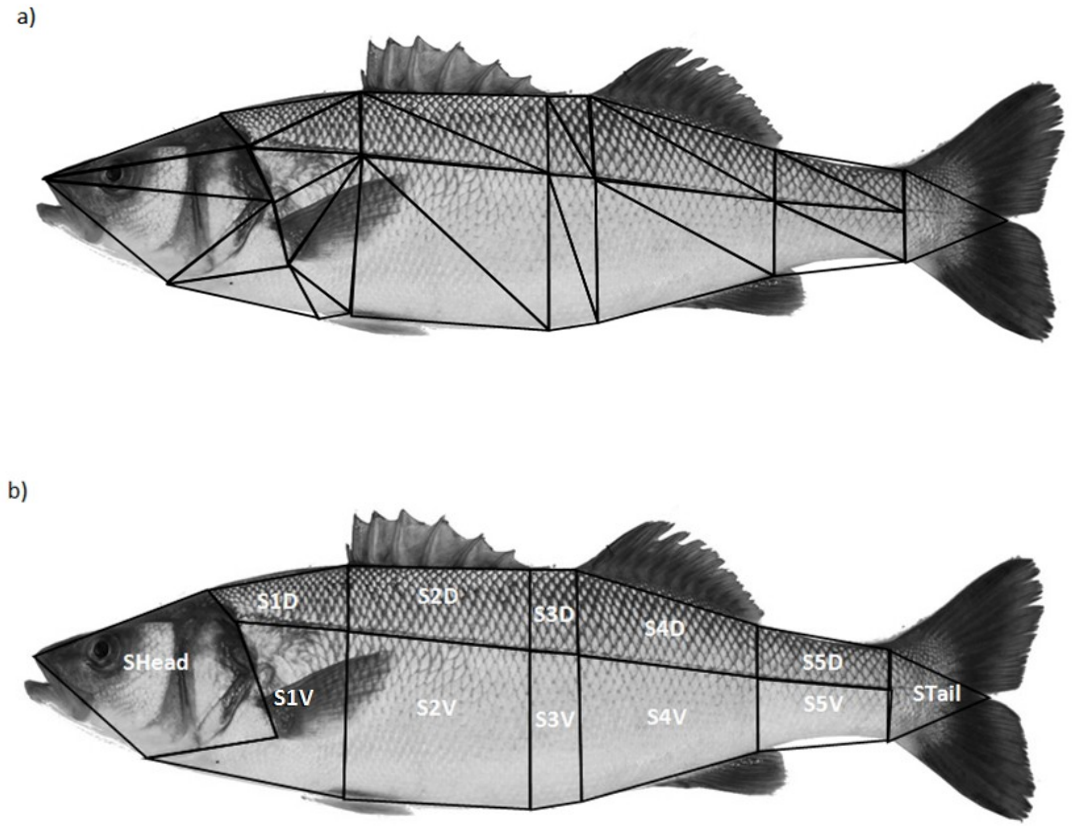

Figure 3

List of Authors

Marc Vandeputte[1,2],* Antonio Puledda, Anne Sophie Tyran[5], Anastasia Bestin[5], Céline Coulombet[6], Aline Bajek[6], Gwénaëlle Baldit[1,2], Alain Vergnet[2], François Allal[3], Jérôme Bugeon[7], Pierrick Haffray[5]

[1]GABI, INRA, AgroParisTech, Université Paris-Saclay, F-78350 Jouy-en-Josas, France

[2]Ifremer, Chemin de Maguelone, F-34250, Palavas-les-Flots, France

[3]Ifremer, UMR9190 MARBEC, Chemin de Maguelone, F-34250, Palavas-les-Flots, France

[4] Section of Animal Science, Department of Agriculture, University of Sassari, Sassari, Italy

[5] SYSAAF Section Aquacole, Campus de Beaulieu, F-35000 Rennes, France

[6] Ecloserie Marine de Gravelines -Ichtus, Voie des Enrochements, F-59820 Gravelines, France

[7] LPGP, INRA, F-35000 Rennes, France

CHAPTER 5

Research

Early landmark-based sexing of European Sea bass (*Dicentrarchus labrax*) using image analysis

---

Abstract

This study reports preliminary results obtained from an image analysis performed on juveniles of European sea bass with the aim to obtain an early sexual sorting by using several morphometric linear measurements, called *inter-landmark distances*. A previous result, based on the earliest individual performance monitoring obtained through the application of the microtag technology on sea bass fingerlings, has highlighted the existence of sex-related biometric differences in early life stages. In this study biometric measurements of experimental fish were collected periodically (~ every 20 days) since the beginning of the trial when fish were photographed (95 dph), microtagged and finclipped for a further parentage assignment. *ImageJ* software was used to manually place 19 landmarks on the surface of 866 photographed fingerlings whose sex was macroscopically determined at the end of the experiment. To assess the possibility to assign fingerlings to their (known) sex using image analysis, several segments were obtained by connecting all pairs of landmarks manually positioned on the surface of photographed fish. To remove some effects able to affect the comparison among individuals, lengths were scaled to the individual standard length obtaining body proportions that were used instead of linear measurements. A discriminant analysis was performed on interlandmark distances after their standardization. The 84% of fish were assigned to the correct sex by using this method. These preliminary results suggested the need to better investigate the use of morphometric differences as possible indicator of the sex. Considering the economic interests that lie behind the obtaining of sea bass sex-skewed female populations, the finding of sexual indicators joined with the application of the computer vision technologies may results in potential economic benefits.

Introduction

The European sea bass (*Dicentrarchus labrax L*. 1758) is a gonochoristic teleost, member of the family Moronidae, characterised by the lacking of heteromorphic sex chromosome (Cataudella et al., 1973; Sola et al., 1993). Primordial germ cells of the gonads appear in sea bass at 10.6 mm of Standard Length (SL) while differentiated gonads are not recognizable until 90 – 120 mm of SL (Roblin and Bruslé, 1983; Blázquez et al., 1999). Although adults of this species are sexually differentiated, gonads remain undifferentiated until the first year of life, with precocious males that in cultured condition generally reach the puberty earlier than normal males (two years of age) and than females (three year of age) (Blázquez et al., 1995). Due to the delay in their gonad development, sea bass females generally tend to grow faster and to be larger than males at the marketable size (Haffray et al., 2007; Navarro-Martin et al., 2009). Therefore most research in the field of sea bass aquaculture in the last years has been focused on obtaining mono-sex population composed only by females. To date, it is well known that sex ratio of cultured sea bass is consistently skewed towards males (Blazquez et al., 1995, 1998; Pavlidis et al., 2000, Koumoundouros et al., 2002; Piferrer et al., 2005; Navarro-Martin et al., 2009; Çoban et al., 2011; Vandeputte et al., 2012) as opposed to natural population characterised by a predominance of females (Barnabé, 1976; Arias, 1980; Pavlidis et al., 2000; Vandeputte et al., 2012). Although the biased predominance of females in natural conditions could be due to the different longevity between sexes (females older than males) or to a sampling bias related to the sex distribution respect to the capture areas, with large males living offshore (Vandeputte et al., 2012), there are strong scientific evidences that the sex ratio of European sea bass, as in many other fish species, is affected by environmental conditions.

Starting from the work of Blázquez et al. (1998), temperature and photoperiods were investigated as potential causes of the sex differentiation during the labile period, from 57 and 137 days post fertilization (dpf), when exogenous steroids are able to modify the sex in this species (Blázquez et al., 1995, 1998). A predominance of males was found using low temperatures, results apparently in contrast with those obtained from further experiments. Sea bass female-skewed populations were obtained by the exposure of the very early developmental stages to low rearing

temperatures (13 or 15°C) by Pavlidis et al. (2000). Thanks to this research, the thermo-sensitive period (TSP), period in which sex of fish is subjected to the temperature of the environment, was localized before the completion of the metamorphosis (total length of 17-18 mm). The high percent of males obtained by Blázquez et al. (1998) was probably due not to the thermal treatment applied on juveniles, but to the environmental conditions applied during the eggs and larval rearing (20 – 24 °C from 0 to 43 dpf).

Koumoundouros et al. (2002) found a significant increase in the incidence of female with the increase of the exposure of fish to low temperatures (15°C). From these results sex determination in sea bass seems to be sensibly affected by thermal treatments and sex ratio significantly correlated with the growth rate of the fish well before the gonads differentiation. Koumoundouros et al. (2002), concluded that temperature is able to influence the sex determination even before the histological differentiation of the first germ cells (9 mm TL; Roblin and Bruslé, (1983) and that presumptive gonads are committed to a certain sex phenotype before their differentiation under the influence of genes differently expressed depending on temperatures and involved in the release of sexual hormones able to regulate sex differentiation. Piferrer et al. (2005), presented a review on the knowledge of sex determination and differentiation in the European sea bass providing an extensive description of the interaction between genetic, endocrine and environment aspects related to size, age and sex of fish. However, despite the different approaches undertaken to understand the process of the sex determination of sea bass, authors concluded that the real mechanism of sex determination was still not clearly understood and that temperature is the only factor able to clearly influence sex ratios in this species.

To date, the existence of early tagging protocols (Cousin et al., 2012; Ferrari et al., 2014) coupled with the availability of different software for image analysis, are allowing a longitudinal monitoring of individual traits with biometric information obtained from digital pictures (Haffray et al., 2013; Chapter 4). The recent availability of miniaturized RFID (Radio Frequency Identification) tags, has permitted to follow sea bass individuals from the 3th month of life until the end of their lifecycle, providing for the first time, the early research of sex-related growth

differences among sexes (Ferrari et al., 2014). At the same time, computer vision technology in fish processing was aimed to find real time systems able to reduce operating costs, improve the quality of products and the profit of producers, solving particular issues as grading for size, prevention of skeletal anomalies and implementation of breeding programs based on morphologic recognizable traits (Costa et al., 2013). European sea bass is characterized by a not clear visible morphological sex dimorphism that has avoided till now, despite its importance in optimizing sex-based breeding schemes, efficient methods for the sorting of females from males in a rapid and automatic manner. Coban et al. (2011) discovered some biometric traits related to the sex by using 15 morphometric measures sampled by image analysis on 219 adults of sea bass reared under commercial conditions. Significant differences among sexes were found using particular indices (ratio between a specific measurements and the total length). However, despite this interesting result, the discriminating power of the linear traits used in the analysis was not evaluated.

Despite the existence of a morphologic dimorphism among sexes on adults of European sea bass, to our best knowledge, no studies have been already performed to predict sex in relatively early life stages on this species. Previous results in fact suggested that significant growth differences due to phenotypic sex could be detectable between males and females as from 105 dpf (Ferrari et al. 2015). Aim of this work was to evaluate the possibility to discriminate sex in European sea bass through image analysis at 95 dph, intermediately between the formation of the gonadal ridges (35 dph) and the earliest histological signs of ovarian and testicular differentiation (150 dpf), then earlier than previously made with similar methods. This work will show results obtained by applying a Canonical Discriminant Analysis on several linear measurements, called Inter Landmark Distances (ILDs), obtained by connecting landmarks drawn on pictures of juveniles of European sea bass.

Materials and Methods

Fish production and rearing condition

500 full-sib family were produced (March, 2013) throughout a full factorial mating plan involving a first generation of domesticated West Mediterranean broodstock. Gametes from 10 dams and 50 sires were *in vitro* fertilized according to the procedure described by Grima et al. (2010). In particular, after fertilization, eggs were pooled for 48 hours at 14°C (from 14,2 to 14,9°C) and salinity decreased from 36,6 to 34‰ until the complete hatching (day 0). After hatching, larvae were reared in common garden following the sea bass rearing standards (Chatain et al., 1994), with temperature and salinity kept until the 69th day post hatch (dph) at 15°C and 25‰, respectively. Water renewed was ensured throughout a flow rate that progressively varied from 75 l h^{-1} (0 dph) to a maximum of 400 l h^{-1} (68 dph). From 70 to 75 dph, until the end of the experiment (397 dph), temperature was progressively increased up to 25°C and salinity set up to natural condition (average salinity 35,3 ± 3,5‰, min = 24.8‰, max = 38.4‰) (Figure 1).

Phenotypic data

The total fish population accounted for 1174 fish of which 558 were males, 416 females (200 did not have a sex assignation). The Body Weight (BW) and the Fork Length (FL) were repeatedly collected during the experimental period. Biometries were made with an average interval of 23 days between two consecutive manipulations. In manner to obtain a longitudinal growth-monitoring during the early life stages, at the first biometry (95 dph) fish were tagged by using 1x6 mm microtags (Lutronics, Nonatec RFID, Lutronic International, Rodange, Luxembourg, mass10 mg; frequency: 13.56 MHz). Tagging procedure took place following the protocol described by Cousin et al., (2012). BW was collected (to the nearest mg) starting from 95 dph, until the last biometry that was performed at 325 dph. At 95 dph fish were photographed using a digital camera (Canon EOS 1100D, 12.2 Mpixel). Left side pictures of the offspring were obtained placing fish on a numbered glass slide over a light table, using a graduated ruler for scale adjusting. To prevent

identity losses during the trial possibly caused by the limited reading distances of microtags (1 - 2 cm), and to ensure a correct sex assignment at the end of the experiment, fish were re-tagged at 179 dph, using conventional PIT tags (Passive Integrated transponder). At the same time, biological samples (a portion of the caudal fin) were collected for the further parental assignment. From 179 dph the FL was individually recorded (to the nearest mm) at each biometry by using an in-house electronic ruler. At the end of the experiment (397 dph) fish were sacrificed for sex determination. Fish were first pre-anesthetized and then killed by a lethal dose of benzocaine (conc. 70 ppm, 400 ppm, respectively). After death, fish were dissected to remove the gonads. Sex was macroscopically determined by visual inspection of gonads, according to the method described by Barnabé (1976).

DNA analysis and parentage assignment

DNA extraction and genotyping for parentage assignment was performed using a panel of 12 microsatellite markers by *Labogena*, the French laboratory for livestock genotyping (ISO 17025 accredited, Jouy-en-Josas, France). Parental assignment was performed by using VITASSIGN (Vandeputte et al., 2006) with a tolerated threshold of two allelic mismatches for each kinship trio.

Experimental population

A total of 866 pictures of fingerlings having a valid sex assignation were used to create the experimental sample. Sexes were distributed almost uniformly in the sample with 476 males and 390 females. The kinship of 762 fish was known after parentage assignment and information on relatives (Sire and Dams) was provided by a pedigree. On the whole there were 10 Dams and 49 Sires (one sire was not represented in the sample) generating a total of 10 groups (families hereafter). Among these, three were widely, four averagely and three poorly represented with one family having only two individuals, respectively (table 1). A reduced dataset of 588 fish, including a balanced number of individuals of each sex, was created to perform further (univariate) statistical analyses.

Table 1. Number of offspring per family (overall) and separated by sex for fish having a correct parentage assignment (762 fish).

Family (Dams)	Overall	Males	Females
Dam 1	198	100	98
Dam 2	192	103	89
Dam 3	102	66	36
Dam 4	68	36	32
Dam 5	68	15	53
Dam 6	56	36	20
Dam 7	40	36	4
Dam 8	22	9	13
Dam 9	14	7	7
Dam 10	2	2	0

Image Analysis and Landmark Capture

Image analysis was performed using Fiji (*ImageJ*) software (Schindelin et al., 2012). The first operation was to set the scale of the pictures by revealing the length (in pixel) of a longitudinal bar drawn on a photographed ruler. After the conversion of pixels to millimeters, 19 landmarks were manually digitized on the surface of each fish with the same order (figure 2) and their coordinates were stored in a datafile.

Inter Landmark Distances

Coordinates were used to create a dataset of derived measures called *Inter Landmark Distances* (ILDs) by recursively measuring the length of segments connecting all landmarks on each picture. To create the set of ILDs, coordinates were firstly transposed using *R* software. Subsequently to estimate all possible ILDs, the Euclidean formula (used in Chapter 4 to estimate the sides of triangles drawn on fish surfaces) was recursively applied:

$$ILD = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$$

The whole set of ILDs was firstly statistically investigated and subsequently used to perform a Canonical Discriminant Analysis (CDA) with the aim to find morphometric variables possibly related to the sex. The complete list of ILDs as well

as their descriptive statistics are presented in table $S1_{(a,b,c,d,e)}$. Briefly, ILDs were named with the combination of the landmark numbers for example the Standard Length (SL), obtained by connecting landmarks 1 and 19, was named *L1_19*.

Data Quality Control

Data (FL and ILDs) were checked for the presence of errors before the analysis. For each variable, a correction for values exceeding three times the standard deviation of the intra-group (sex) distribution was performed. More specifically, values falling outside the fixed (upper or lower) limits were converted into the relative intra-group (sex) median. More specifically, for each variable, values for males and females were corrected separately considering the specific sex. The intra-group median was used for two reasons: 1) the median is less affected by the presence of possible outliers, 2) the use of the overall instead of the intra-group median, could affect the power of the discrimination by reducing the distance between groups (sex) for the trait. After QC, descriptive statistics were generated for each sex, using the proc MEANS function of SAS software.

Statistical analysis

After QC, all variables were tested for normality with the *Shapiro-Wilk's test* using the *shapito.test* function (R software). A log natural transformation of variables far from normality (p <0.05) was performed. The effect of the sex was investigated on FL and ILDs. In particular to evaluate the effect of the sex on FL an *Analysis of Variance with Repeated Measures* (RM-ANOVA) was performed by applying the GLM procedure with the repeated statement (SAS software). The model was performed by using all six FL measures repeatedly sampled over time from 179 to 325 dph, as explained above in "phenotypic data". The effect of the sex on each ILDs was evaluated with a series of two-sided *Student's t-test* using the sex as classifiers. The t-test function of R software was included in a loop and recursively performed in manner to investigate all ILDs. The Student's t-tests were performed after data standardization (see after ILDs standardization).

Multivariate analysis

The Canonical Discriminant Analysis (CDA) was the multivariate method adopted for investigating the sex-based discrimination of fingerlings using ILDs. CDA is a classification technique that allows, with the derivation of linear equations called *canonical functio*ns, the assignment of an object to a specific group (Dimauro et al., 2015). The general structure of the canonical function is:

$$can = d_1X_1 + d_2X_2 + \ldots\ldots + d_nX_n,$$

where d_i are the canonical coefficients that indicate the contribution of each single variable to the global function. If more than two groups are involved, *k-1* functions are extracted and, based on the *Mahalanobis distance*, and corresponding *Hotelling's T-squared test* (De Maesschalck et al., 2000) is evaluated to assess the effective separation among groups. In the present research, having a binomial class variable as the sex (male and female), only one canonical function was extracted. The final assignation of fish into groups was obtained using the cross-validated DISCRIM procedure of SAS. The discriminant analysis was firstly performed on all set of ILDs and secondly on the set of the most significant variables selected through the stepwise selection by applying the STEPDISC procedure (SAS) with a significance level of the *enter* and *stay* selection criteria set to 0.15.

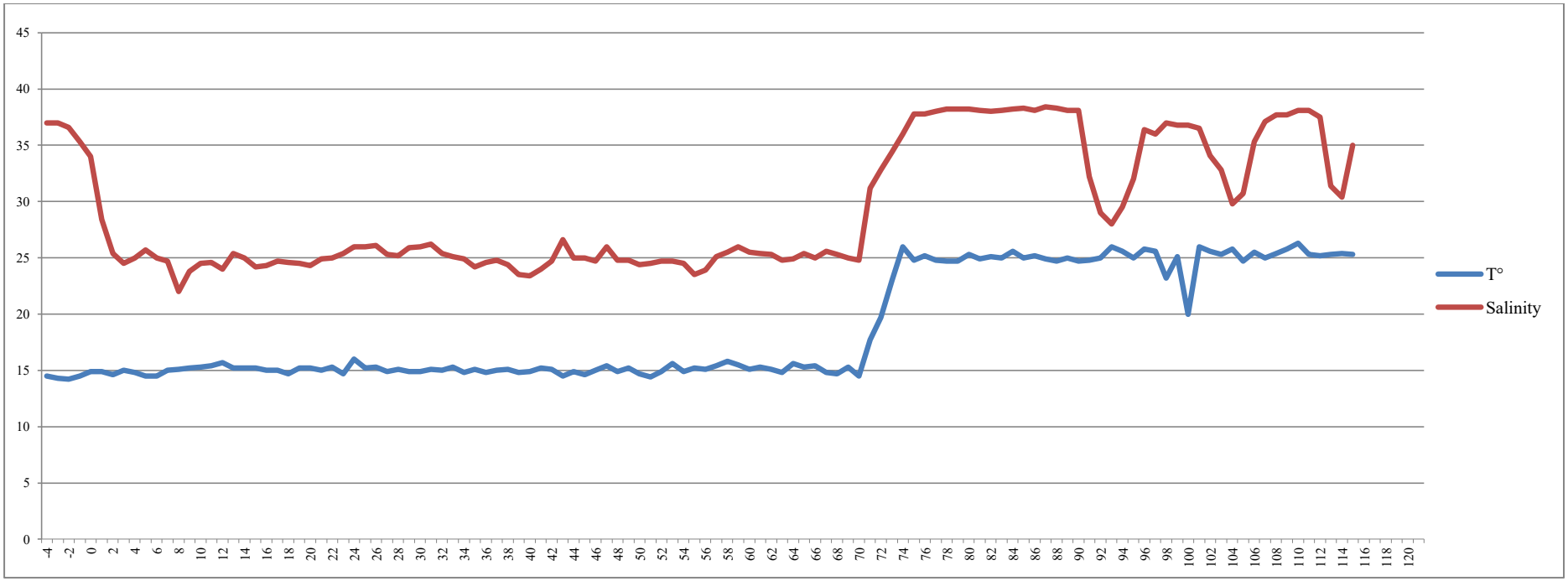

Figure1. Profiles of temperature and salinity recorded during the experimental period and adopted following the sea bass rearing standards (Chatain et al., 1994).

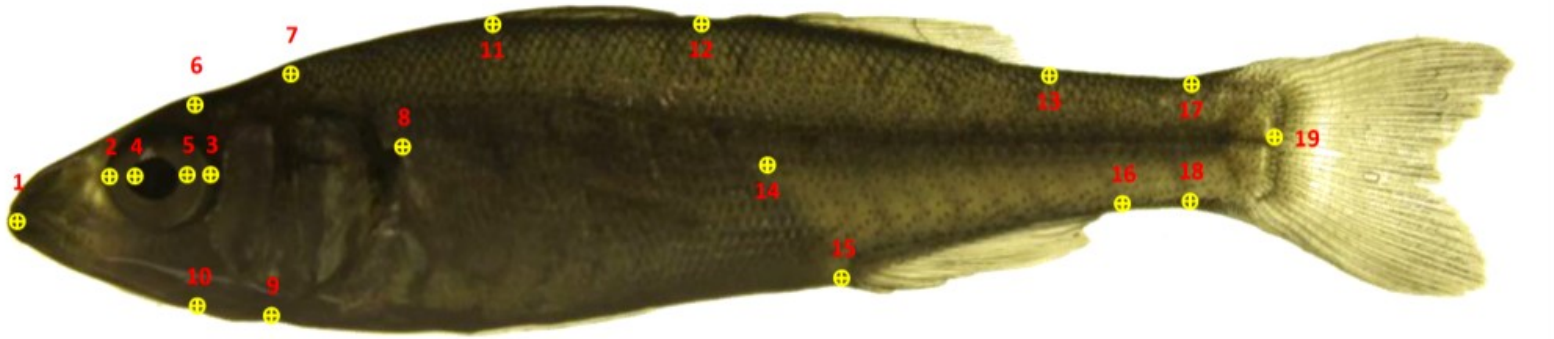

Figure 2. Position of landmarks on the surface of the fish.

Results and Discussions

Biometric measurements

The mean BW at 95 dph of males (0.52 ± 0.11 g) was significantly different from the mean BW of females (0.69 ± 0.12 g) (Levene's test, p-value ≤ 0.0016; Welch ANOVA, F = 422, p-value < 0.0001). Descriptive statistics for FL are showed in table 2. The first measurement of the table relies on the standard length (SL) obtained from image analysis. At 95 dph the mean SL of males (32.5 ± 2.2 mm) was statistically different from the mean SL of females (35.6 ± 2.0 mm) (t-test p-value < 0.001) with females 9% longer than males at this age. FL were collected on fish starting from 179 dph. The between-sex difference in FL reached a maximum at 200 dph (12%), remaining stable until 256 dph and decreasing from 305 dph until the end of the experiment (325 dph, 10%). Figure 3 highlights the quite constant difference in FL between sexes during the experimental period. The gap in length reach at 95 dph should be due to different growth rates occurred before then beginning of the trial with males growing more slowly than females. FL resulted normally distributed at the test for normality. The effect of the sex on FL was statistically significant (RM-ANOVA $F_{(1,861)} = 355.70$; p-value <0.001).

Inter Landmark Distances

A total of 171 inter landmark distances (ILDs) were obtained by linking all 19 landmarks. During the quality control (QC), 689 values were changed into their intragroup (sex) median as explained in materials and methods. The average CV among ILDs was 7.8% for males and 6.7% for females. These values were slightly lower than that obtained for the FL collected on field (8.5%). Two ILDs associated to the caudal peduncle (L16_18 and L13_17) showed the highest values of CV both on males (19% and 16%) and females (16 and 14%). The high variability of these ILDs was probably related to the high flexibility of the caudal part of the body. The moderate variability of L9_10 (CV = 11.45%) on both sexes was probably due to the difficult in identifying the bottom side of the operculum encountered in part of the photos due to the high contrast generated by the background. Data were tested for normality with the ShapiroWilk normality test. Only 3 ILDs were not normal distributed, however their log transformation guaranteed the normalization.

Table 2.
Descriptive statistics for FL (in millimeters) collected at different days post hatch (dph) for fish included in the study. Mean and standard deviation, minimum and maximum values and coefficient of variability (CV) are showed separated by sex, after quality control (QC).

Variable	dph	N.oss	N.eff	Mean ± sd	Min	Max	CV
Males							
SL	95		476	32.55 ± 2.23	26.00	38.45	6.84
FL5	179		476	103.82 ± 8.56	81.00	128.60	8.25
FL6	200		476	111.98 ± 10.06	84.50	141.90	8.98
FL7	228	476	476	128.23 ± 12.31	96.70	164.50	9.60
FL8	256		476	135.33 ± 13.35	100.30	175.00	9.86
FL9	305		476	155.87 ± 14.51	116.30	201.80	9.31
FL10	325		475	163.80 ± 15.42	114.50	213.80	9.42
Females							
SL	95		390	35.59 ± 2.04	30.06	41.82	5.72
FL5	179		390	115.73 ± 8.04	96.00	139.00	6.94
FL6	200		390	125.36 ± 9.51	102.60	152.00	7.58
FL7	228	390	390	143.37 ± 11.83	114.20	176.60	8.25
FL8	256		390	151.86 ± 13.74	114.50	192.80	9.05
FL9	305		389	173.18 ± 16.00	127.70	223.00	9.24
FL10	325		389	180.41 ± 18.10	128.70	232.20	10.03

dph = days post hatching; N.oss = total number of fish; N.eff = number of fish with available records;

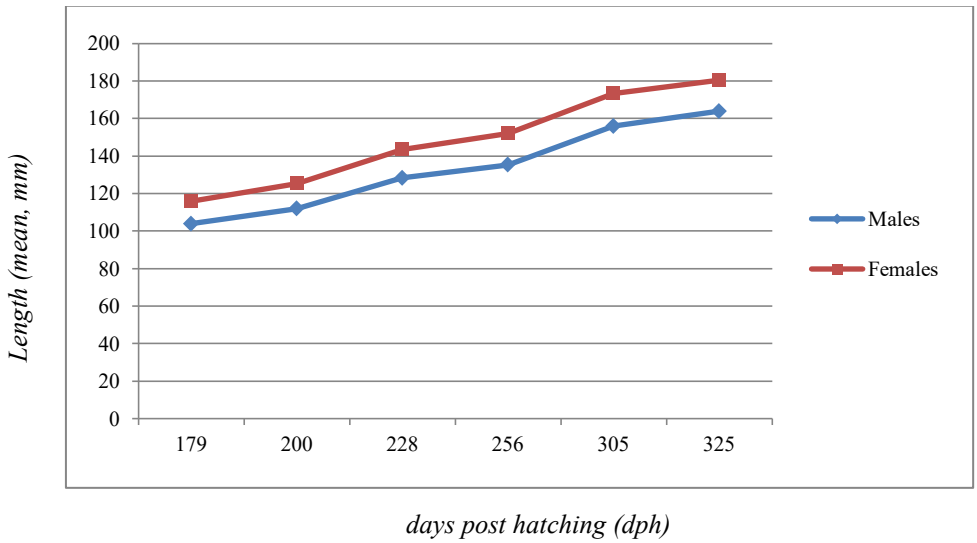

Figure 3. Growth differences in body length (mean, mm) between males and females over 6 biometries, from 179 to 325 dph.

Standardization

Data were used to perform univariate and multivariate analysis. The univariate approach included the analysis of mean differences of ILDs among sexes (table S2), whereas the multivariate approach implied the use of ILDs to find a discriminant function through Discriminant Analysis. Before performing analysis, data were standardized dividing each length by the individual Standard Length (L1_19). The data transformation allowed to move the interest from the length measures to the individual body proportions, removing possible size and scale effects. The influence of size effects and the effect of their removal by standardization can be graphically displayed by using the Principal Component Analysis. In biometric measurements involving body lengths, the first PC generally accounts for the most part of the variability associated to the size. In figure 4a (on the left) one of the ILDs (*L1_8*) is plotted against the first principal component (PC1) extracted from the whole set of ILDs (after QC). In figure 4b (on the right), the same PC1 is plotted against the same variable but after its standardization (*L1_8* / *L1_19*). The relation of the original variable from "size components" (summarized by the PC1) is clearly evidenced by the linear trend on figure *a*. On the contrary, in figure b, the random distribution of observations suggests the disappearance of this dependence.

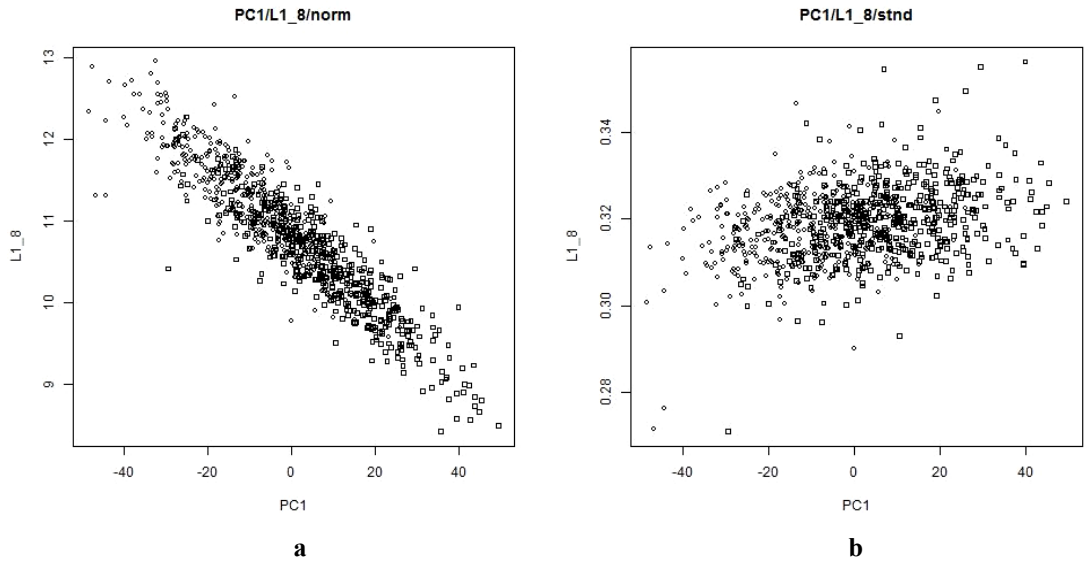

Figure 4. Dependance of the shape from the size on one of the ILD, the *L1_8* (from the tip of the snout to the posterior end of the opercolum), before (a) and after (b) data standardization.

Statistical analysis

The effect of the sex on each ILD was evaluated with a two sided Student's t-test on a dataset with a balanced number of males and females (balanced design) including standardized data of a total of 588 fish (294 for each sex). The dataset was created by randomly selecting an equal number of males and females from each family. For each family the maximum number of individuals was equal to the double of the number of individuals in the less represented sex. One family was discarded because having only two individuals (with the same sex). For each ILD, the t-test was performed verifying the homogeneity of variance. In case of heteroscedasticity (evaluated through Bartlett's test) a Welch approximation to the degrees of freedom was adopted. Statistical test were performed on the whole population and within families. Results are reported in tables S2, where p-values of t-tests are ranked in manner to highlight ILDs statistically significant both at population and family level. From table S2 is possible to notice that within the first ten ranked segments seven included the landmark 14. This landmark is located at the end of the abdominal cavity. In particular, the segment L8_14 can highlight a possible difference between sexes related to the site of gonad development.

Another significant difference was observed in the dimension and position of the eyes. In this case the most significant variables were: the intraocular (L2_3) and iris (L4_5) length, the distance between the tip of the snout and the posterior margin of the eye (L1_3), the distances between the anterior and posterior margin of the eye and the posterior and anterior margin of the iris (L2_5 and L3_4, respectively), as well as their relative distance with the abovementioned landmark 14 (L2_14, L3_14, L4_14, L5_14).

Despite L1_2 was not significant at all investigated levels (population and familes) the differences in eye dimension/position could be due to some functional reasons. It should be noted that in less represented families (Dam 4,5,6,7,8,9) the number of non significant variables was higher than in other families (Dam1,2,3). In smaller families, problems of sampling size could have affected the results of the tests.

Multivariate analysis

The full dataset (866 fish) of standardized ILDs was used to perform the Discriminant Analysis by using SAS software. The sex was set as classifiers and two conditions were tested: 1) the full set of ILDs and 2) the combination of ILDs selected through stepwise using the STEPDISC procedure. More specifically the proc DISCRIM with crossvalidation was performed on the full set of variables and on these selected through stepwise. Results obtained by fitting the full set of ILDs are showed in the table 3. In this case 77% of fish were correctly assigned to each sex (figure 5a). In table 4 are showed crossvalidated results of DA obtained from the subset of significant ILDs selected by the stepwise procedure. In the last case the discriminant analysis was able to correctly assign 83% of fish to the relative sex (figure 5b).

Table 3

Discriminant Analysis. Sex discrimination of individuals using the full set of ILDs with crossvalidation.

Groups	Males	Females	Total
Males	356	120	476
Females	85	305	390

Classification/misclassification:

77% : 23%

Table 4

Discriminant Analysis. Sex discrimination of individuals using variables selected through stepwise selection with crossvalidation.

Groups	Males	Females	Total
Males	386	90	476
Females	59	331	390

Classification/misclassification:

83% : 17%

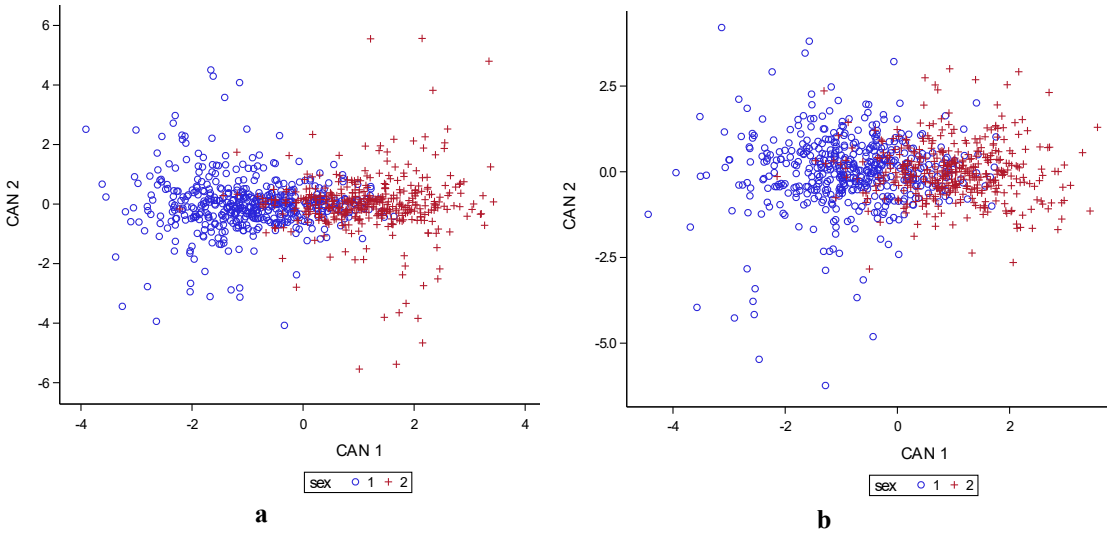

Figure 5. Scatterplots of the first two canonical variables for the full set of ILDs (a) and the set of ILDs selected by the stepwise (b) analyzed with the Discriminant Analysis.
Sex (1 = Males; 2 = Females)

In table 5 the list of ILDs selected by the stepwise as well as coefficients of the Linear Discriminant Function for sex are reported. The Mahalanobis distance between sex groups was significant (generalized quadratic distance for sex 3.55; p-value < 0.001). From table 5 is possible to observe that more than 50% of the discriminant variables are related to the head (from L1_15 to L10_12) as already observed in the univariate approach. The 32% of the remaining variables are related to the dorsal profile (from L12_13 to L13_19) and the remaining 12% to the abdominal body compartment. Four variables out of 25 selected are in common with those higlighted through the univariate statistical analysis (L1_5, L2_5, L2_14 and log_L13_14) and related to differences in the position/dimension of the eye and belly cavity, respectively.

Table 5. Linear Discriminant Function for sex

Variable	sex	
	Males	**Females**
Constant	-78415	-78509
L1_5	71487	71421
L1_10	20983	21070
L1_13	69372	69434
L2_5	-46780	-46874
L2_10	-218.9681	-150.70978
L2_12	1452	1391
L3_13	-1233	-1145
L4_13	-17981	-18064
L4_18	78826	78901
L6_17	-4.20835	-82.0461
L7_8	-1757	-1681
L7_18	-28037	-28135
L7_19	35186	35289
L10_12	18128	18252
L12_13	6619	6646
L12_14	-3920	-4073
L12_15	13605	13685
log_L13_14	21923	21368
L13_15	-31316	-31262
L13_16	5047	4987
L13_18	-27138	-27194
L13_19	54714	54828
L14_16	-1960	-1921
L15_18	-18702	-18600
L15_19	58659	58547

Head
Dorsal Profile
Abdomen/Anal Fin/Peduncle

In 2009 Blázquez et al., implemented a Canonical Discriminant Analysis (CDA) using the expression levels of different genes involved in sexual differentiation, as predictors of the phenotypic sex. The author found that *cyp19a1a*, the gene encoding for gonadic aromatase, was implicated in the ovarian differentiation. The author obtained a complete discrimination between sexes when fish reached 60 mm SL then before the comparison of histological signs of sex differentiation (79 - 95 mm SL, 150 – 200 dpf, Saillant et al., 2003). By predicting sex using the expression levels of *cyp11b*, a gene encoding for the enzyme 11-β-hydroxylase implicated in the synthesis of the androgen 11-ketotestosterone during the testicular differentiation, the complete separation between sexes was observed only after 87 mm SL. No sex-related differences were observed by using expression levels of other sex-related genes by applying CDA. The promising results obtained by Blázquez et al. (2009) suggested that an early discrimination of fingerlings it could be obtained even if the proposed method seems difficult to apply into the real practice.

More recently a 83% of correct assignation was obtained by Costa et al. (2013) in a test population of European sea bass at market size (250 g mean weight) by performing a Partial Least Square Regression Discriminant Analysis (PLSDA) on high resolution images, demonstrating the possibility to discriminate sex in adult of this species using image analysis.

In this work the assignation has been assessed on fingerlings of ~ 35 mm of SL (mean BW = 0.60 g) at 95 dph, about half the length reported by Blázquez et al. (2009) and age reported by Saillant et al. (2003) by using morphometric measurements derived from digital images. Results obtained from this work represents the first large scale assessment of sex assignation for juveniles of European Sea bass by using the image analysis.

Conclusions

Results of this study suggested the presence of significant differences among sexes mainly related to the head, the central part of the abdomen (in correspondence of the end of the belly cavity) and the dorsal profile of the body on juveniles of European sea bass at 95 days post hatching. An interesting level of discrimination between sexes (83%) was obtained by using a subset of significant linear measurements derived by the image analysis. Also in this context, the relevant effect of the head has been highlighted by both univariate and multivariate analysis. The presence of differences in this body compartment since the early life stages, also emerged from the study on processing traits (chapter 4) on adults of European sea bass, could suggest a relation of this part of the body with the sex. It could be interesting to evaluate if this difference, that seem to be persistent during the development, if related to the sex, could be able to affect genetic parameters when estimates are done in populations with mixture of sexes. Finally, a complete knowledge of the morphologic development of European sea bass may be achieved by repeatedly taking high-resolution pictures on the same individuals, starting from the first stages of their life, taking advantage of the microtagging technology, and throughout the entire lifecycle until the completion of the sexual maturation. If a discrimination between sexes could be obtained simply using morphometric measurements (obtained in this case without complex technological equipments) and if the obtaining of these measure could be automated by applying some algorithm for the automatic landmark detection, the process of sorting could be inexpensively introduced in the commercial practice, allowing a rapid sorting of sexes with undoubted benefits for the breeders.

Acknowledgements

This research project was supported by the European Commission under the 7th Framework Program FP7-KBBE-2010-4 Contract No. 265957 COPEWELL. I would like to thank Dr. Sébastien Ferrari, Dr. François Allal, Dr. Marc Vandeputte, Dr. Marie-Laure Bégout for providing suggestions, materials and data, in particular fish pictures and biometric measurements. I wish also to thank Prof. Nicolò Pietro Paolo Macciotta and Prof. Corrado Dimauro for the comments and the revision of the work.

REFERENCES

Arias, A. 1980. Crecimiento, régimen alimentario y reproducción de la dorade (Sparus aurata L.) y del róbalo (*Dicentrarchus labrax* L.) en los esteros de Cádiz. Inv Pesq 44:59–83.

Barnabé, G., 1976. Contribution à la connaissance de la biologie du loup, *Dicentrarchus labrax* (L.) (Poisson Serranidae). Thèse de Doctorat d'état, mention Sciences Université des Sciences et Techniques du Languedoc, Montpellier (426 pp.).

Blazquez, M., Piferrer, F., Zanuy, S., Carrillo, M., and Donaldson, E. M., 1995 Development of sex control techniques for European sea bass (*Dicentrarchus labrax* L.) aquaculture: effects of dietary 17amethyltestosterone prior to sex differentiation. *Aquaculture*, *135*, 329–342. http://doi.org/10.1016/0044-8486(95)01030-0

Blázquez, M., Zanuy, S., Carillo, M., and Piferrer, F., 1998. Effects of rearing temperature on sex differentiation in the European sea bass (*Dicentrarchus labrax* L.). *Journal of Experimental Zoology*, *281*(3), 207–216. http://doi.org/10.1002/(SICI)1097-010X(19980615)281:3<207::AID-JEZ6>3.0.CO;2-R

Blázquez, M., Carrillo, M., Zanuy, S., and Piferrer, F., 1999. Sex ratios in offspring of sex-reversed sea bass and the relationship between growth and phenotypic sex differentiation. Aquaculture, 55(July 2016), 916–930. http://doi.org/10.1111/j.1095-8649.1999.tb00730.x

Blázquez M., Navarro-Martín L., Piferrer F. 2009. Expression profiles of sex differentiation-related genes during ontogenesis in the European sea bass acclimated to two different temperatures. J. Exp. Zool. (Mol. Dev. Evol.) 312B:686–700.

Cataudella, S., Civitelli, M. V, and Campanna, E., 1973. The chromosomes of some Mediterranean teleosts: Scorpaenidae, Serranidae, Labridae, Blennidae, Gobiidae (pisces - Scorpaeniformes, Perciformes). *Boll. Zool.*, *40* (March 2014), 384–389. http://doi.org/10.1080/11250007309429255

Chatain, B., 1994. Estimation et amélioration des performances zootechniques de l'élevage larvaire de *Dicentrarchus labrax* et de Sparus auratus. Retrieved from http://archimer.ifremer.fr/doc/00000/1750/

Çoban, D., Yildirim, Ş., Kamaci, H. O., Suzer, C., Saka, Ş., and Firat, K., 2011. External morphology of European seabass (*Dicentrarchus labrax*) related to sexual dimorphism, *35*(2), 255–263. http://doi.org/10.3906/zoo-0810-3

Costa, C., Antonucci, F., Boglione, C., Menesatti, P., Vandeputte, M., and Chatain, B., 2013. Automated sorting for size, sex and skeletal anomalies of cultured seabass using external shape analysis. *Aquacultural Engineering*, *52*, 58–64. http://doi.org/10.1016/j.aquaeng.2012.09.001

Cousin, X., Daouk, T., Péan, S., Lyphout, L., Schwartz, M.-E., and Bégout, M.-L. 2012. Electronic individual identification of zebrafish using radio frequency identification (RFID) microtags. *The Journal of Experimental Biology*, *215*(Pt 16), 2729–34. http://doi.org/10.1242/jeb.071829

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L., 2000. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*(1), 1–18. http://doi.org/10.1016/S0169-7439(99)00047-7

Dimauro, C., Nicoloso, L., Cellesi, M., Macciotta, N. P. P., Ciani, E., Moioli, B., Crepaldi, P., 2015. Selection of discriminant SNP markers for breed and geographic assignment of Italian sheep. *Small Ruminant Research*, *128*(August 2016), 27–33. http://doi.org/10.1016/j.smallrumres.2015.05.001

Ferrari, S., Chatain, B., Cousin, X., Leguay, D., Vergnet, A., Vidal, M. O., Bégout, M. L., 2014. Early individual electronic identification of sea bass using RFID microtags: A first example of early phenotyping of sex-related growth. *Aquaculture*, *426-427*(April), 165–171. http://doi.org/10.1016/j.aquaculture.2014.01.033

Grima, L., Vandeputte, M., Ruelle, F., Vergnet, A., Mambrini, M., and Chatain, B., 2010. In search for indirect criteria to improve residual feed intake in sea bass (*Dicentrarchus labrax*). Part I: Phenotypic relationship between residual feed intake and body weight variations during feed deprivation and re-feeding periods. *Aquaculture*, *300*(1-4), 50–58. http://doi.org/10.1016/j.aquaculture.2010.01.003

Haffray, P., Tsigenopoulos, C.S., Bonhomme, F., Chatain, B., Magoulas, A., Rye, M., Triantafyllidis, A., Triantaphyllidis, C., 2007. European sea bass-*Dicentrarchus labrax*. Geninpact final scientific report (available at www.geninmpact.imr.no)

Haffray, P., Bugeon, J., Rivard, Q., Quittet, B., Puyo, S., Allamelou, J. M., Dupont-Nivet, M., 2013. Genetic parameters of in-vivo prediction of carcass, head and fillet yields by internal ultrasound and 2D external imagery in large rainbow trout (Oncorhynchus mykiss). *Aquaculture*, *410-411*, 236–244. http://doi.org/10.1016/j.aquaculture.2013.06.016

Koumoundouros, G., Pavlidis, M., Anezaki, L., Kokkari, C., Sterioti, A., Divanach, P., and Kentouri, M., 2002. Temperature sex determination in the European sea bass, *Dicentrarchus labrax* (L., 1758) (Teleostei, Perciformes, Moronidae): Critical sensitive ontogenetic phase. *Journal of Experimental Zoology*, *292*(6), 573–579. http://doi.org/10.1002/jez.10095

Navarro-Martín, L., Blázquez, M., Viñas, J., Joly, S., and Piferrer, F., 2009. Balancing the effects of rearing at low temperature during early development on sex ratios, growth and maturation in the European sea bass (*Dicentrarchus labrax*). Limitations and opportunities for the production of highly female-biased stocks. *Aquaculture*, *296*(3-4), 347–358. http://doi.org/10.1016/j.aquaculture.2009.07.022

Pavlidis, M., Koumoundouros, G., Sterioti, A., Somarakis, S., Divanach, P., and Kentouri, M., 2000. Evidence of temperature-dependent sex determination in the European sea bass (*Dicentrarchus labrax* L.). *Journal of Experimental Zoology*, *287*(3), 225–232. http://doi.org/10.1002/1097-010X(20000801)287:3<225::AID-JEZ4>3.0.CO;2-D

Piferrer, F., Blázquez, M., Navarro, L., and González, A., 2005. Genetic, endocrine, and environmental components of sex determination and differentiation in the European sea bass (*Dicentrarchus labrax* L.). *General and Comparative Endocrinology*, *142*(1-2 SPEC. ISS.), 102–110. http://doi.org/10.1016/j.ygcen.2005.02.011

Roblin, C., Bruslé, J., 1983. Ontogenèse gonadique et différenciation sexuelle du Loup *Dicentrarchus labrax*, en conditions d'élevage. Reproduction Nutrition Développement, 23 (1), pp.115- 127. <hal-00898033>

Saillant, E., Chatain, B., Menu, B., Fauvel, C., Vidal, M. O., and Fostier, A., 2003. Sexual differentiation and juvenile intersexuality in the European sea bass (*Dicentrarchus labrax*). *Journal of Zoology*, *260*(1), 53–63. http://doi.org/10.1017/S0952836903003467

Schindelin, J.; Arganda-Carreras, I. and Frise, E. et al., 2012, "Fiji: an open-source platform for biological-image analysis", Nature methods 9(7): 676-682, PMID 22743772 (on Google Scholar).

Sola, L., Bressanello, S., Rossi, A. ., Iaselli, V., Crosetti, D., and Cataudella, S., 1993. A karyotype analysis of the genus Dicentrarchus by different staining techinues. *Journal of Fish Biology*, *43*(April 2005), 329–337. http://doi.org/10.1111/j.1095-8649.1993.tb00567.x

Valentin, A.E., Penin, X., Chanut, J.P., Sévigny, J.M., Rohlf, F.J., 2008. Arching effect on fish body shape in geometric morphometric studies. J. Fish Biol. 73, 623–638. doi:10.1111/j.1095-8649.2008.01961.x

Vandeputte, M., Mauger, S., and Dupont-Nivet, M., 2006. An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion. *Molecular Ecology Notes*, *6*(1), 265–267. http://doi.org/10.1111/j.1471-8286.2005.01167.x

Vandeputte, M., Quillet, E., and Chatain, B., 2012. Are sex ratios in wild European sea bass (*Dicentrarchus labrax*) populations biased? *Aquatic Living Resources*, *25*(1), 77–81. http://doi.org/10.1051/alr/2012002

SUPPLEMENTAL MATERIALS

(CHAPTER 5)

Table S1(a).
Descriptive statistics for all ILDs (original data, after QC) separated by sex.

Males					Females				
ILD	**Mean ± sd**	**Min**	**Max**	**CV**	**ILD**	**Mean ± sd**	**Min**	**Max**	**CV**
L1_2	2.76 ± 0.25	2.07	3.44	9.00	L1_2	3 ± 0.25	2.41	3.64	8.17
L1_3	5.7 ± 0.37	4.67	6.77	6.49	L1_3	6.1 ± 0.35	5.17	7.05	5.77
L1_4	3.34 ± 0.27	2.6	4.02	8.04	L1_4	3.6 ± 0.26	2.95	4.32	7.29
L1_5	4.76 ± 0.33	3.83	5.74	6.89	L1_5	5.09 ± 0.31	4.19	5.95	6.06
L1_6	5.6 ± 0.44	4.39	6.96	7.85	L1_6	6.14 ± 0.42	5.00	7.28	6.77
L1_7	8.26 ± 0.57	6.62	9.69	6.87	L1_7	8.91 ± 0.52	7.43	10.35	5.84
L1_8	10.41 ± 0.68	8.42	12.28	6.56	L1_8	11.3 ± 0.63	9.58	12.95	5.60
L1_9	7.72 ± 0.61	5.99	9.39	7.85	L1_9	8.43 ± 0.57	6.92	10.04	6.73
L1_10	5.43 ± 0.4	4.26	6.44	7.39	L1_10	5.93 ± 0.39	4.96	7.05	6.65
L1_11	13.79 ± 0.93	11.37	16.38	6.74	L1_11	15 ± 0.87	12.53	17.64	5.79
L1_12	20.48 ± 1.5	16.44	24.93	7.34	L1_12	22.48 ± 1.37	18.66	26.48	6.10
L1_13	27.27 ± 1.85	21.91	32.62	6.78	L1_13	29.79 ± 1.7	25.11	34.96	5.71
L1_14	19.96 ± 1.57	15.42	24.67	7.85	L1_14	22.03 ± 1.43	17.92	25.95	6.49
L1_15	21.79 ± 1.54	17.24	26.4	7.07	L1_15	23.88 ± 1.43	19.95	28.03	5.98
L1_16	28.27 ± 1.94	22.8	33.79	6.87	L1_16	30.95 ± 1.82	25.78	36.06	5.87
L1_17	30.43 ± 2.14	24.46	36.34	7.03	L1_17	33.21 ± 1.93	28.01	39.01	5.83
L1_18	30.3 ± 2.17	24.25	36.46	7.15	L1_18	33.15 ± 1.94	27.91	38.84	5.86
L1_19	32.55 ± 2.23	26.00	38.45	6.84	L1_19	35.59 ± 2.04	30.06	41.82	5.72
L2_3	2.98 ± 0.18	2.5	3.44	6.01	L2_3	3.15 ± 0.16	2.56	3.57	5.09
L2_4	0.62 ± 0.06	0.44	0.8	9.53	L2_4	0.65 ± 0.06	0.46	0.83	9.28
L2_5	2.03 ± 0.12	1.66	2.36	5.91	L2_5	2.14 ± 0.1	1.86	2.43	4.56
L2_6	3.07 ± 0.28	2.24	3.87	9.07	L2_6	3.33 ± 0.25	2.58	4.04	7.60
L2_7	5.58 ± 0.4	4.47	6.75	7.24	L2_7	5.98 ± 0.32	5.00	6.85	5.42
L2_8	7.7 ± 0.51	6.24	9.17	6.57	L2_8	8.36 ± 0.46	7.14	9.68	5.47
L2_9	5.74 ± 0.48	4.53	7.17	8.32	L2_9	6.33 ± 0.44	5.17	7.57	7.02
L2_10	3.91 ± 0.31	2.98	4.74	8.04	L2_10	4.34 ± 0.29	3.54	5.17	6.79
L2_11	11.04 ± 0.77	8.75	13.2	6.98	L2_11	12.01 ± 0.71	9.91	14.08	5.87
L2_12	17.72 ± 1.36	13.94	21.49	7.66	L2_12	19.48 ± 1.18	16.05	22.54	6.06
L2_13	24.55 ± 1.7	19.66	29.3	6.92	L2_13	26.84 ± 1.54	22.56	31.37	5.74
L2_14	17.31 ± 1.42	13.21	21.3	8.20	L2_14	19.15 ± 1.27	15.41	22.79	6.64
L2_15	19.25 ± 1.42	15.12	23.14	7.36	L2_15	21.16 ± 1.29	17.6	25.02	6.11
L2_16	25.65 ± 1.8	20.6	30.45	7.02	L2_16	28.13 ± 1.68	23.29	33.12	5.97
L2_17	27.71 ± 2.01	21.95	32.98	7.26	L2_17	30.26 ± 1.76	25.52	35.31	5.80
L2_18	27.68 ± 2.03	22.04	33.11	7.32	L2_18	30.33 ± 1.81	25.44	35.72	5.96
L2_19	29.89 ± 2.08	23.99	35.06	6.96	L2_19	32.69 ± 1.85	27.54	38.05	5.67
L3_4	2.37 ± 0.15	1.93	2.83	6.40	L3_4	2.52 ± 0.13	2.21	2.92	5.30
L3_5	0.95 ± 0.09	0.64	1.21	9.97	L3_5	1.02 ± 0.09	0.77	1.27	9.23
L3_6	2.21 ± 0.2	1.63	2.8	9.10	L3_6	2.35 ± 0.23	1.82	3.00	9.61
L3_7	3.29 ± 0.29	2.45	4.14	8.74	L3_7	3.59 ± 0.29	2.7	4.41	8.14

Table S1(b).
Descriptive statistics for all ILDs (original data, after QC) separated by sex.

Males					Females				
ILD	**Mean ± sd**	**Min**	**Max**	**CV**	**ILD**	**Mean ± sd**	**Min**	**Max**	**CV**
L3_8	4.73 ± 0.37	3.60	5.9	7.79	L3_8	5.22 ± 0.33	4.31	6.18	6.36
L3_9	4.02 ± 0.38	3.04	5.22	9.54	L3_9	4.43 ± 0.37	3.45	5.43	8.39
L3_10	3.55 ± 0.31	2.69	4.45	8.73	L3_10	3.85 ± 0.28	3.14	4.66	7.34
L3_11	8.23 ± 0.66	6.17	9.92	7.97	L3_11	9.07 ± 0.62	7.33	10.91	6.83
L3_12	14.8 ± 1.23	11.51	18.17	8.34	L3_12	16.4 ± 1.08	13.3	19.22	6.56
L3_13	21.58 ± 1.59	16.88	25.99	7.36	L3_13	23.71 ± 1.43	19.85	28.01	6.05
L3_14	14.37 ± 1.3	10.6	18.09	9.02	L3_14	16.03 ± 1.16	12.57	19.47	7.26
L3_15	16.39 ± 1.3	12.56	20.12	7.95	L3_15	18.11 ± 1.16	14.87	21.6	6.39
L3_16	22.72 ± 1.68	18.01	27.29	7.39	L3_16	25.02 ± 1.56	20.63	29.72	6.25
L3_17	24.75 ± 1.88	19.33	29.78	7.61	L3_17	27.13 ± 1.65	22.72	32.1	6.08
L3_18	24.75 ± 1.9	19.44	29.93	7.69	L3_18	27.2 ± 1.68	22.49	32.08	6.17
L3_19	26.94 ± 1.96	21.38	31.82	7.26	L3_19	29.56 ± 1.74	24.65	34.84	5.90
L4_5	1.43 ± 0.09	1.16	1.72	6.55	L4_5	1.51 ± 0.08	1.31	1.71	5.12
L4_6	2.73 ± 0.25	2.01	3.48	9.23	L4_6	2.97 ± 0.24	2.33	3.58	8.08
L4_7	5.11 ± 0.37	4.17	6.23	7.34	L4_7	5.49 ± 0.31	4.52	6.36	5.72
L4_8	7.09 ± 0.49	5.67	8.39	6.89	L4_8	7.74 ± 0.44	6.52	9.05	5.63
L4_9	5.23 ± 0.46	3.98	6.54	8.85	L4_9	5.79 ± 0.43	4.60	7	7.35
L4_10	3.56 ± 0.31	2.70	4.32	8.81	L4_10	3.96 ± 0.28	3.24	4.78	6.97
L4_11	10.48 ± 0.75	8.25	12.55	7.19	L4_11	11.44 ± 0.7	9.37	13.44	6.08
L4_12	17.13 ± 1.34	13.54	20.95	7.80	L4_12	18.87 ± 1.17	15.47	21.91	6.20
L4_13	23.93 ± 1.69	18.94	28.65	7.08	L4_13	26.21 ± 1.53	22.07	30.7	5.82
L4_14	16.69 ± 1.4	12.63	20.77	8.38	L4_14	18.52 ± 1.26	14.79	22.13	6.80
L4_15	18.64 ± 1.4	14.53	22.65	7.50	L4_15	20.52 ± 1.28	16.94	24.36	6.24
L4_16	25.03 ± 1.78	20.02	29.93	7.12	L4_16	27.49 ± 1.66	22.79	32.45	6.05
L4_17	27.1 ± 1.99	21.40	32.38	7.34	L4_17	29.63 ± 1.74	24.93	34.7	5.88
L4_18	27.06 ± 2	21.46	32.59	7.40	L4_18	29.68 ± 1.77	24.81	34.66	5.97
L4_19	29.27 ± 2.06	23.42	34.53	7.03	L4_19	32.05 ± 1.84	26.92	37.43	5.74
L5_6	2.11 ± 0.2	1.57	2.65	9.32	L5_6	2.28 ± 0.20	1.71	2.87	8.77
L5_7	3.93 ± 0.32	3.10	4.81	8.15	L5_7	4.26 ± 0.28	3.50	5.03	6.57
L5_8	5.67 ± 0.42	4.44	6.94	7.38	L5_8	6.23 ± 0.38	5.23	7.39	6.08
L5_9	4.42 ± 0.4	3.29	5.59	9.12	L5_9	4.90 ± 0.40	3.78	6.05	8.12
L5_10	3.39 ± 0.3	2.6	4.27	8.79	L5_10	3.73 ± 0.28	2.97	4.55	7.53
L5_11	9.12 ± 0.69	7.01	11.03	7.57	L5_11	10.00 ± 0.64	8.21	11.92	6.40
L5_12	15.72 ± 1.27	12.32	19.25	8.06	L5_12	17.38 ± 1.12	14.25	20.26	6.44
L5_13	22.52 ± 1.61	17.93	27	7.15	L5_13	24.71 ± 1.48	20.74	29.07	5.99
L5_14	15.29 ± 1.33	11.47	19.11	8.71	L5_14	17.02 ± 1.21	13.55	20.49	7.11
L5_15	17.28 ± 1.33	13.39	21.07	7.71	L5_15	19.08 ± 1.24	15.58	22.83	6.48
L5_16	23.63 ± 1.71	18.86	28.29	7.26	L5_16	26.01 ± 1.62	21.49	30.88	6.21
L5_17	25.68 ± 1.92	20.12	30.84	7.49	L5_17	28.13 ± 1.69	23.66	33.15	6.03

Table S1(c).
Descriptive statistics for all ILDs (original data, after QC) separated by sex.

Males					Females				
ILD	**Mean ± sd**	**Min**	**Max**	**CV**	**ILD**	**Mean ± sd**	**Min**	**Max**	**CV**
L5_18	25.67 ± 1.94	20.3	30.94	7.54	L5_18	28.19 ± 1.73	23.57	33.11	6.12
L5_19	27.86 ± 1.99	22.17	32.86	7.15	L5_19	30.55 ± 1.79	25.68	35.88	5.87
L6_7	2.73 ± 0.3	1.83	3.57	11.01	L6_7	2.83 ± 0.24	2.11	3.51	8.48
L6_8	5.82 ± 0.37	4.71	6.88	6.35	L6_8	6.24 ± 0.35	5.14	7.28	5.67
L6_9	6.22 ± 0.42	5.12	7.38	6.68	L6_9	6.77 ± 0.41	5.59	7.96	6.05
L6_10	5.48 ± 0.37	4.42	6.53	6.66	L6_10	5.99 ± 0.34	5.04	6.91	5.68
L6_11	8.4 ± 0.62	6.61	10.13	7.38	L6_11	9.07 ± 0.58	7.42	10.78	6.43
L6_12	15.31 ± 1.18	12.04	18.59	7.72	L6_12	16.81 ± 1.07	13.94	19.88	6.35
L6_13	22.33 ± 1.55	17.74	26.67	6.93	L6_13	24.36 ± 1.44	20.06	28.53	5.92
L6_14	15.43 ± 1.27	11.64	19.03	8.23	L6_14	17.02 ± 1.17	13.82	20.27	6.86
L6_15	17.73 ± 1.27	14.03	21.32	7.16	L6_15	19.4 ± 1.19	16.02	22.92	6.11
L6_16	23.8 ± 1.65	19.04	28.39	6.94	L6_16	26.03 ± 1.55	21.56	30.55	5.95
L6_17	25.54 ± 1.85	20.23	30.99	7.24	L6_17	27.84 ± 1.66	23.48	32.83	5.98
L6_18	25.81 ± 1.87	20.45	30.86	7.25	L6_18	28.2 ± 1.68	23.73	33.1	5.95
L6_19	27.86 ± 1.92	22.25	33.13	6.90	L6_19	30.42 ± 1.76	25.64	35.72	5.80
L7_8	3.83 ± 0.27	3.02	4.45	7.02	L7_8	4.16 ± 0.26	3.52	4.86	6.28
L7_9	6.47 ± 0.42	5.25	7.59	6.43	L7_9	7.04 ± 0.42	5.92	8.19	5.89
L7_10	6.75 ± 0.44	5.47	8.05	6.50	L7_10	7.32 ± 0.42	6.13	8.46	5.68
L7_11	5.68 ± 0.51	4.3	6.92	9.04	L7_11	6.25 ± 0.48	5.00	7.59	7.72
L7_12	12.66 ± 1.06	9.9	15.36	8.35	L7_12	14.04 ± 0.95	11.55	16.89	6.80
L7_13	19.71 ± 1.44	15.47	23.49	7.30	L7_13	21.66 ± 1.31	18.33	25.36	6.07
L7_14	13.03 ± 1.17	10.08	16.11	8.95	L7_14	14.52 ± 1.06	11.63	17.43	7.28
L7_15	15.52 ± 1.2	11.96	18.66	7.74	L7_15	17.13 ± 1.09	14.08	20.22	6.37
L7_16	21.37 ± 1.55	17.44	25.51	7.25	L7_16	23.51 ± 1.45	19.52	27.66	6.15
L7_17	22.94 ± 1.74	18	28.01	7.58	L7_17	25.14 ± 1.56	21.07	29.73	6.22
L7_18	23.35 ± 1.76	18.77	27.88	7.53	L7_18	25.65 ± 1.57	21.52	30.09	6.13
L7_19	25.32 ± 1.81	20.23	30.25	7.16	L7_19	27.78 ± 1.66	23.45	32.75	5.97
L8_9	5.1 ± 0.4	3.93	6.22	7.92	L8_9	5.58 ± 0.38	4.46	6.65	6.86
L8_10	6.7 ± 0.47	5.34	7.95	7.02	L8_10	7.31 ± 0.42	6.13	8.54	5.76
L8_11	4.24 ± 0.38	3.20	5.31	8.93	L8_11	4.64 ± 0.36	3.55	5.69	7.82
L8_12	10.16 ± 0.89	7.7	12.44	8.74	L8_12	11.27 ± 0.8	9.05	13.6	7.11
L8_13	16.86 ± 1.24	13.12	20.35	7.39	L8_13	18.48 ± 1.14	15.42	21.77	6.19
L8_14	9.68 ± 0.96	7.01	12.31	9.93	L8_14	10.85 ± 0.88	8.31	13.27	8.07
L8_15	11.91 ± 0.98	9.04	14.63	8.25	L8_15	13.17 ± 0.91	10.49	15.89	6.95
L8_16	18.05 ± 1.35	14.43	21.69	7.45	L8_16	19.87 ± 1.27	16.22	23.52	6.40
L8_17	20.03 ± 1.55	15.57	24.43	7.72	L8_17	21.93 ± 1.39	18.37	26.07	6.35
L8_18	20.07 ± 1.57	15.85	24.19	7.84	L8_18	22.06 ± 1.41	18.28	26.12	6.38
L8_19	22.23 ± 1.62	17.62	26.49	7.27	L8_19	24.36 ± 1.45	20.37	28.78	5.97

Table S1(d).
Descriptive statistics for all ILDs (original data, after QC) separated by sex.

Males					Females				
ILD	**Mean ± sd**	**Min**	**Max**	**CV**	**ILD**	**Mean ± sd**	**Min**	**Max**	**CV**
L9_10	2.41 ± 0.32	1.48	3.31	13.08	L9_10	2.64 ± 0.3	1.78	3.55	11.51
L9_11	9.33 ± 0.7	7.46	11.35	7.48	L9_11	10.23 ± 0.66	8.34	12.11	6.48
L9_12	14.63 ± 1.16	11.5	18.14	7.94	L9_12	16.13 ± 1.07	12.97	19.22	6.65
L9_13	20.69 ± 1.47	16.44	25.13	7.13	L9_13	22.63 ± 1.4	18.33	26.77	6.18
L9_14	13.11 ± 1.13	9.85	15.82	8.63	L9_14	14.57 ± 1.08	11.32	17.44	7.41
L9_15	14.45 ± 1.13	11.13	17.81	7.84	L9_15	15.88 ± 1.07	12.67	18.81	6.75
L9_16	21.11 ± 1.54	16.59	25.44	7.30	L9_16	23.19 ± 1.45	19.28	27.18	6.24
L9_17	23.73 ± 1.75	18.78	28.61	7.39	L9_17	25.91 ± 1.58	21.36	30.52	6.08
L9_18	23.16 ± 1.74	18.44	28.14	7.52	L9_18	25.39 ± 1.61	20.87	29.92	6.33
L9_19	25.6 ± 1.82	20.54	30.38	7.09	L9_19	28.00 ± 1.67	23.1	33.01	5.98
L10_11	10.83 ± 0.78	8.5	13.06	7.17	L10_11	11.84 ± 0.7	9.75	13.68	5.94
L10_12	16.67 ± 1.27	13.16	19.82	7.62	L10_12	18.37 ± 1.14	15.12	21.8	6.22
L10_13	22.96 ± 1.59	18.31	27.57	6.93	L10_13	25.11 ± 1.45	21.03	29.36	5.79
L10_14	15.41 ± 1.27	11.65	18.58	8.24	L10_14	17.09 ± 1.16	13.72	20.28	6.80
L10_15	16.83 ± 1.25	13.14	20.02	7.41	L10_15	18.51 ± 1.16	15.26	21.95	6.25
L10_16	23.5 ± 1.66	18.86	28.29	7.07	L10_16	25.78 ± 1.56	21.29	30.4	6.04
L10_17	26.02 ± 1.89	20.53	31.3	7.26	L10_17	28.42 ± 1.65	23.89	33.19	5.82
L10_18	25.54 ± 1.89	20.32	30.99	7.41	L10_18	27.99 ± 1.7	23.48	32.98	6.06
L10_19	27.95 ± 1.95	22.35	33.19	6.98	L10_19	30.57 ± 1.75	25.7	35.76	5.72
L11_12	7.03 ± 0.72	5.15	9.18	10.32	L11_12	7.86 ± 0.68	5.80	9.8	8.68
L11_13	14.17 ± 1.12	10.92	17.41	7.92	L11_13	15.58 ± 1.05	12.44	18.73	6.77
L11_14	8.15 ± 0.87	5.94	10.38	10.66	L11_14	9.16 ± 0.79	6.90	11.34	8.64
L11_15	11.09 ± 0.92	8.65	13.49	8.29	L11_15	12.33 ± 0.88	9.72	15.01	7.15
L11_16	16.19 ± 1.26	12.73	19.77	7.77	L11_16	17.86 ± 1.19	14.33	21.35	6.68
L11_17	17.39 ± 1.39	13.39	21.1	8.02	L11_17	19.07 ± 1.27	15.67	22.73	6.64
L11_18	18.09 ± 1.47	13.78	22.21	8.12	L11_18	19.93 ± 1.31	16.45	23.86	6.57
L11_19	19.89 ± 1.49	15.96	24.25	7.48	L11_19	21.84 ± 1.37	18.11	25.84	6.27
L12_13	7.21 ± 0.58	5.57	8.91	8.02	L12_13	7.76 ± 0.51	6.63	9.33	6.55
L12_14	4.04 ± 0.38	3.04	5.05	9.29	L12_14	4.4 ± 0.33	3.51	5.33	7.53
L12_15	6.93 ± 0.57	5.26	8.59	8.25	L12_15	7.65 ± 0.53	6.05	9.23	6.99
L12_16	9.75 ± 0.72	7.53	11.75	7.39	L12_16	10.65 ± 0.69	8.69	12.72	6.52
L12_17	10.42 ± 0.81	7.91	12.86	7.81	L12_17	11.26 ± 0.75	9.47	13.51	6.67
L12_18	11.49 ± 0.9	8.9	14.15	7.79	L12_18	12.54 ± 0.8	10.37	14.89	6.39
L12_19	13.01 ± 0.87	10.77	15.52	6.71	L12_19	14.13 ± 0.82	11.87	16.46	5.84
L13_14	7.62 ± 0.54	6.1	9.19	7.03	L13_14	8.1 ± 0.56	6.74	9.73	6.97
L13_15	7.65 ± 0.55	6.12	9.06	7.24	L13_15	8.33 ± 0.53	6.79	9.84	6.39
L13_16	4.05 ± 0.36	3.01	5.15	8.88	L13_16	4.43 ± 0.34	3.51	5.35	7.75
L13_17	3.22 ± 0.53	1.89	4.68	16.32	L13_17	3.5 ± 0.52	2.14	4.92	14.85

Table S5(e).
Descriptive statistics for all ILDs (original data, after QC) divided by sex.

Males					Females				
ILD	**Mean ± sd**	**Min**	**Max**	**CV**	**ILD**	**Mean ± sd**	**Min**	**Max**	**CV**
L13_18	4.94 ± 0.53	3.46	6.42	10.64	L13_18	5.44 ± 0.49	4.17	6.78	9.05
L13_19	5.85 ± 0.58	4.33	7.43	9.97	L13_19	6.4 ± 0.57	4.81	7.80	8.90
L14_15	3.21 ± 0.29	2.36	4.03	9.12	L14_15	3.52 ± 0.29	2.72	4.4	8.38
L14_16	8.36 ± 0.58	6.6	10	6.92	L14_16	9.02 ± 0.58	7.43	10.65	6.46
L14_17	10.63 ± 0.72	8.67	12.83	6.78	L14_17	11.36 ± 0.66	9.73	13.14	5.83
L14_18	10.37 ± 0.74	8.33	12.39	7.14	L14_18	11.22 ± 0.7	9.3	13.35	6.21
L14_19	12.61 ± 0.82	10.21	15.01	6.48	L14_19	13.58 ± 0.72	11.54	15.61	5.33
L15_16	6.76 ± 0.51	5.28	8.13	7.48	L15_16	7.39 ± 0.52	5.97	8.86	7.09
L15_17	10.07 ± 0.75	7.96	12.31	7.48	L15_17	10.91 ± 0.66	9.11	12.77	6.05
L15_18	8.77 ± 0.71	6.86	10.94	8.11	L15_18	9.6 ± 0.66	7.69	11.4	6.89
L15_19	11.41 ± 0.8	9.19	13.63	7.01	L15_19	12.4 ± 0.71	10.49	14.48	5.76
L16_17	4.17 ± 0.38	3.02	5.2	9.11	L16_17	4.47 ± 0.31	3.63	5.3	7.01
L16_18	2.03 ± 0.39	0.90	3.17	19.05	L16_18	2.23 ± 0.37	1.2	3.32	16.82
L16_19	4.7 ± 0.42	3.41	5.76	8.91	L16_19	5.04 ± 0.4	3.96	6.2	7.90
L17_18	3.53 ± 0.26	2.81	4.28	7.49	L17_18	3.83 ± 0.23	3.18	4.5	6.10
L17_19	2.91 ± 0.23	2.21	3.6	7.87	L17_19	3.19 ± 0.25	2.59	3.92	7.77
L18_19	2.87 ± 0.21	2.24	3.49	7.47	L18_19	3.05 ± 0.2	2.48	3.63	6.70

Table S2(a). Effect of the sex on ILDs at the population (Pop) and family (Dam) level. Statistical significance (p-values) for the two-sided Student's t-test on 588 fish. Variables (ILDs) are ranked according to the combined degree of significance of all levels

Variable	Pop	Dam 1	Dam 2	Dam 3	Dam 4	Dam 5	Dam 6	Dam 7	Dam 8	Dam 9
log_L13_14	***	***	***	***	***	ns	***	**	***	***
L3_14	***	**	***	***	*	ns	***	ns	***	ns
L5_14	***	**	***	***	*	ns	***	ns	**	*
L8_14	***	*	***	***	*	ns	**	ns	***	**
L2_5	***	***	***	***	*	ns	*	ns	*	**
L4_14	***	*	***	***	*	ns	**	ns	**	*
L3_4	***	***	**	***	*	ns	ns	*	*	ns
L4_5	***	***	*	**	ns	ns	ns	ns	***	**
L7_14	***	*	***	**	ns	ns	**	ns	**	*
L10_14	***	*	***	**	*	ns	**	ns	**	ns
L1_3	***	***	*	***	**	ns	*	ns	ns	ns
L2_3	***	***	***	***	*	ns	ns	ns	ns	ns
L2_14	***	ns	***	**	*	ns	**	ns	**	ns
L11_14	***	**	**	*	ns	ns	*	*	**	*
L7_12	***	ns	***	***	ns	ns	*	ns	ns	**
L14_17	***	***	***	*	ns	ns	ns	*	*	ns
L1_5	***	***	ns	**	*	ns	**	ns	ns	ns
L3_12	***	ns	***	***	*	ns	*	ns	ns	ns
L5_16	***	*	ns	*	**	ns	**	ns	ns	**
L5_19	***	***	ns	*	*	ns	**	ns	ns	*
L6_7	***	**	***	*	ns	ns	**	ns	ns	ns
L6_8	***	***	ns	**	*	ns	ns	ns	ns	**
L8_12	***	ns	**	***	ns	ns	ns	ns	ns	***
L9_14	***	***	ns	*	ns	ns	**	ns	**	ns
L1_7	***	*	**	*	ns	ns	*	ns	ns	**
L1_14	***	ns	***	*	ns	ns	*	ns	**	ns
L2_4	***	**	***	**	ns	ns	ns	ns	ns	ns
L2_7	***	**	***	**	ns	ns	ns	ns	ns	ns
L5_12	***	ns	**	***	ns	ns	*	ns	ns	*
L14_19	***	ns	**	**	ns	ns	*	ns	**	ns
L3_8	***	**	**	**	ns	ns	ns	ns	ns	ns
L3_15	***	ns	*	***	*	ns	*	ns	ns	ns
L3_16	***	ns	*	*	**	ns	**	ns	ns	ns
L4_12	***	ns	*	***	ns	ns	*	ns	ns	*
L10_12	***	ns	***	**	ns	ns	*	ns	ns	ns
L3_19	***	*	ns	**	*	ns	*	ns	ns	ns
L4_7	***	*	**	**	ns	ns	ns	ns	ns	ns
L4_10	***	**	*	**	ns	ns	ns	ns	ns	ns
L5_15	***	*	ns	**	*	ns	*	ns	ns	ns
L6_14	***	ns	*	*	ns	ns	*	ns	**	ns
L7_16	***	ns	ns	ns	*	ns	**	ns	ns	**
L3_5	***	*	*	*	ns	ns	ns	*	ns	ns
L3_18	**	ns	ns	***	*	ns	*	ns	ns	ns

ns = not significant; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$;

Table S2(b). Effect of the sex on ILDs at the population (Pop) and family (Dam) level. Statistical significance (p-values) for the two-sided Student's t-test on 588 fish. Variables (ILDs) are ranked according to the combined degree of significance of all levels

Variable	Pop	Dam 1	Dam 2	Dam 3	Dam 4	Dam 5	Dam 6	Dam 7	Dam 8	Dam 9
L4_16	***	ns	ns	ns	*	ns	**	ns	ns	*
L7_15	***	ns	*	**	ns	ns	*	ns	ns	ns
L8_16	***	ns	ns	ns	*	ns	*	ns	ns	**
L11_12	***	ns	**	*	ns	ns	ns	ns	ns	*
L11_15	***	***	ns	*	ns	ns	ns	ns	ns	ns
L11_16	***	*	ns	ns	*	ns	ns	ns	ns	**
L16_17	***	**	**	ns	ns	ns	ns	ns	ns	ns
L18_19	***	***	ns	ns	ns	ns	ns	ns	*	ns
L1_8	**	ns	ns	*	ns	ns	ns	ns	ns	***
L1_12	**	ns	**	**	ns	ns	ns	ns	ns	ns
L2_8	**	**	ns	ns	ns	ns	ns	ns	ns	**
L2_10	***	ns	ns	ns	***	ns	ns	ns	ns	ns
L2_12	**	ns	*	**	ns	ns	*	ns	ns	ns
L4_9	**	ns	**	**	ns	ns	ns	ns	ns	ns
L4_15	***	ns	ns	*	*	ns	*	ns	ns	ns
L5_18	**	ns	ns	**	*	ns	*	ns	ns	ns
L6_11	**	**	ns	ns	ns	ns	ns	**	ns	ns
L7_19	***	ns	ns	ns	ns	ns	*	ns	ns	**
L8_15	***	ns	ns	**	*	ns	ns	ns	ns	ns
L11_19	**	*	ns	ns	ns	ns	ns	*	ns	**
L14_16	***	ns	**	ns	ns	ns	ns	ns	*	ns
L2_11	*	**	ns	ns	ns	ns	ns	**	ns	ns
L3_11	**	ns	*	*	ns	ns	ns	*	ns	ns
L7_18	**	ns	ns	**	ns	ns	*	ns	ns	ns
L8_19	*	ns	ns	ns	*	ns	ns	ns	ns	***
L9_12	***	*	ns	*	ns	ns	ns	ns	ns	ns
L1_11	*	*	ns	ns	ns	ns	ns	*	ns	*
L2_15	**	ns	ns	ns	*	ns	*	ns	ns	ns
L5_11	ns	ns	ns	*	ns	ns	*	**	ns	ns
L6_12	*	ns	*	*	ns	ns	ns	ns	ns	*
L9_10	ns	**	**	ns	ns	ns	ns	ns	ns	ns
L11_18	**	ns	ns	ns	ns	ns	ns	*	ns	*
L12_13	**	ns	**	ns	ns	ns	ns	ns	ns	ns
L12_15	**	**	ns	ns	ns	ns	ns	ns	ns	ns
log_L14_15	**	ns	ns	ns	ns	ns	ns	*	*	ns
L15_17	**	ns	**	ns	ns	ns	ns	ns	ns	ns
L1_4	**	ns	ns	ns	ns	ns	*	ns	ns	ns
L1_9	ns	*	**	ns	ns	ns	ns	ns	ns	ns
L2_9	*	ns	**	ns	ns	ns	ns	ns	ns	ns
L2_16	*	ns	ns	ns	*	ns	*	ns	ns	ns
L3_9	ns	ns	ns	***	ns	ns	ns	ns	ns	ns
L4_18	*	ns	ns	*	ns	ns	*	ns	ns	ns
L5_8	**	ns	*	ns	ns	ns	ns	ns	ns	ns

ns = not significant; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$;

Table S2(c). Effect of the sex on ILDs at the population (Pop) and family (Dam) level. Statistical significance (p-values) for the two-sided Student's t-test on 588 fish. Variables (ILDs) are ranked according to the combined degree of significance of all levels

Variable	Pop	Dam 1	Dam 2	Dam 3	Dam 4	Dam 5	Dam 6	Dam 7	Dam 8	Dam 9
L5_9	*	ns	ns	**	ns	ns	ns	ns	ns	ns
L7_10	**	*	ns	ns	ns	ns	ns	ns	ns	ns
L7_11	ns	ns	*	ns	ns	ns	ns	**	ns	ns
L8_18	*	ns	ns	**	ns	ns	ns	ns	ns	ns
L10_11	ns	ns	ns	ns	ns	ns	*	**	ns	ns
L10_16	**	ns	ns	ns	ns	ns	*	ns	ns	ns
L14_18	**	ns	*	ns	ns	ns	ns	ns	ns	ns
L17_18	*	*	*	ns	ns	ns	ns	ns	ns	ns
L3_10	ns	ns	ns	ns	**	ns	ns	ns	ns	ns
L3_17	ns	ns	ns	*	ns	ns	*	ns	ns	ns
L4_11	ns	ns	ns	ns	ns	ns	ns	**	ns	ns
L5_17	ns	ns	ns	*	ns	ns	*	ns	ns	ns
L6_13	*	*	ns	ns	ns	ns	ns	ns	ns	ns
L7_9	*	ns	*	ns	ns	ns	ns	ns	ns	ns
L7_13	*	ns	ns	ns	ns	ns	*	ns	ns	ns
L8_10	ns	ns	ns	ns	ns	ns	ns	ns	ns	**
L9_16	*	*	ns	ns	ns	ns	ns	ns	ns	ns
L9_19	ns	*	*	ns	ns	ns	ns	ns	ns	ns
L10_15	**	ns	ns	ns	ns	ns	ns	ns	ns	ns
L12_17	*	ns	*	ns	ns	ns	ns	ns	ns	ns
L13_15	*	ns	*	ns	ns	ns	ns	ns	ns	ns
L16_19	**	ns	ns	ns	ns	ns	ns	ns	ns	ns
L1_10	ns	ns	ns	ns	ns	ns	*	ns	ns	ns
L1_17	ns	*	ns	ns	ns	ns	ns	ns	ns	ns
L2_17	ns	ns	ns	ns	ns	ns	*	ns	ns	ns
L2_18	ns	ns	ns	ns	ns	ns	*	ns	ns	ns
L2_19	ns	ns	ns	ns	ns	ns	*	ns	ns	ns
L3_13	ns	ns	ns	ns	ns	ns	*	ns	ns	ns
L4_17	ns	ns	ns	ns	ns	ns	*	ns	ns	ns
L4_19	ns	ns	ns	ns	ns	ns	*	ns	ns	ns
L5_13	ns	ns	ns	ns	ns	ns	*	ns	ns	ns
L6_9	ns	ns	ns	ns	ns	ns	ns	ns	*	ns
L6_10	ns	ns	ns	ns	ns	ns	ns	ns	*	ns
L6_17	ns	*	ns	ns	ns	ns	ns	ns	ns	ns
L7_17	ns	ns	ns	ns	ns	ns	*	ns	ns	ns
L8_11	ns	ns	ns	ns	ns	ns	ns	*	ns	ns
L9_11	ns	ns	ns	ns	ns	ns	ns	*	ns	ns
L9_17	ns	ns	*	ns	ns	ns	ns	ns	ns	ns
L10_17	ns	ns	ns	ns	ns	ns	*	ns	ns	ns
L11_17	ns	ns	ns	ns	ns	ns	ns	ns	ns	*
L12_14	ns	ns	*	ns	ns	ns	ns	ns	ns	ns
L5_9	*	ns	ns	**	ns	ns	ns	ns	ns	ns
L7_10	**	*	ns	ns	ns	ns	ns	ns	ns	ns

ns = not significant; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$;

Table S2(d). Effect of the sex on ILDs at the population (Pop) and family (Dam) level. Statistical significance (p-values) for the two-sided Student's t-test on 588 fish. Variables (ILDs) are ranked according to the combined degree of significance of all levels

Variable	Pop	Dam 1	Dam 2	Dam 3	Dam 4	Dam 5	Dam 6	Dam 7	Dam 8	Dam 9
L15_19	ns	ns	*	ns	ns	ns	ns	ns	ns	ns
L1_2	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L1_6	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L1_13	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L1_15	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L1_16	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L1_18	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L2_6	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L2_13	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
log_L3_6	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L3_7	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L4_6	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L4_8	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L4_13	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L5_6	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L5_7	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L5_10	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L6_15	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L6_16	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L6_18	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L6_19	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L7_8	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L8_9	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L8_13	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L8_17	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L9_13	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L9_15	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L9_18	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L10_13	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L10_18	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L10_19	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L11_13	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L12_18	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L13_16	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L13_17	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L13_18	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L13_19	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L15_16	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L15_18	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L16_18	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
L17_19	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns

ns = not significant; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$;

GENERAL CONCLUSIONS

The present thesis has dealt with different aspects related to a common objective, the improvement of the livestock performances. Traditional methods applied to improve animal productions are derived from the quantitative genetics theory. The estimation of genetic parameters, such as the heritability and the genetic correlations, is the first step of the selection process applied on farmed populations by means of quantitative methods. To obtain genetic parameters, a reliable decomposition of the total phenotypic variance into its causal components is needed.

In the real practice, phenotypes are often available for individuals deriving from admixed populations characterized by complex structures of relationships. In these situations, the estimation of the variance components through the sample analysis of the covariance among individuals is not practical, if not impossible. The introduction of mixed linear models, the development of complex iterative algorithms for their solution and the achievement of the computational power necessary to get the convergence, has allowed in the last decades, to obtain reliable estimates also in the presence of complex population structures. The knowledge of genetic parameters has also had a great importance to determine the progress achieved for traits under selection or to predict the potential genetic gain achievable by comparing different breeding schemes in manner to quantify, also in economic terms, the results of different selection strategies.

In the last few years classical methods of selection have been implemented with the availability of new sources of information both at genetic and phenotypic level. The dramatic increase of information has been possible due to the huge reduction of the unitary cost per bit of information. The recent availability dense molecular marker maps, composed by many thousands of Single Nucleotide Polymorphisms has made possible the prediction of direct genomic breeding values by estimating the effect of the total number of segment in which the genome is divided into. Thanks to the reduction of the generation intervals and the improvement of the accuracy of estimations introduced by genomic selection, the genetic responses for important production traits have been nearly doubled. A side effects generated by the

genotyping technologies is the unbalancing between the number of variables, the markers, and the number of observations, the animals, which may have potential effects in the accuracy of estimates. Different methods to reduce this disproportion have been proposed, ranging from the reduction of the number of variables involved in estimations such the preselection of markers with a significant effect on the trait of interest or multivariate approaches for data reduction, as well as the increase of the number of observations by pooling animals of the same breed or having different origins (in this case with some problems related to the admixture) into common reference populations.

More recently, high throughput technologies have been implemented to collect phenotypes at high level of processivity, leading the animal breeding into to the phenotyping filed. New technologies for data recording are spanning from the field to the laboratory level with the aim to standardize, automate and make cost effective the measuring processes. New phenotyping methods are playing an important role also in animal selection by increasing the accuracy of breeding values and reducing the bottleneck created between the availability of genomic and phenomic data allowing at the same time a better comprehension of the genotype to phenotype pathways

In this thesis classics and innovative methods of animal selection, the last ones based on computer vision technologies, as well as the evaluation of the effect of data unbalancing generated by the availability of big genomic dataset has been addressed in different animals species, all characterized by a great economic importance and potential improvement by selection. An example has been the evaluation of the genetic parameters for innovative production traits on ewes of the Sarda breed. Milk coagulation traits are of particular interest for our territory considering that the sheep milk is totally transformed into cheese. A not negligible additive component has been highlighted in this study for the coagulation traits. However, considering their high genetic correlation, results suggested an individual use of MCPs. The negative genetic correlation of MCPs with ILCY, even though the latter was obtained by mechanical extraction on small milk samples, suggested a particular attention in using MCPs as cheese yield predictors in a hypothetical breeding program aimed to improve the individual cheese making ability.

In the second experimental contribution, the reduction of the unbalancing between the number of records and the number of variables was performed by applying the Principal Component Analysis. In this case, the accuracy of Direct Genomic Values obtained by estimating the effect of all substitutive predictors (the principal components) suggested both the initial panel densities and the way of markers repartition as main sources of potential inflation of the prediction accuracies. In the first case accuracies obtained from smaller panels outperformed accuracies obtained from panels with a higher marker density, in the second case however accuracies increased according to the degree of the data repartition. Both aspects are related to the balance between predictors and observations. Different strategies of data manipulations has demonstrated to be able to reduce this disproportions resulting in better estimations. Compared with traditional methods of estimations involving the use of the full set of SNP markers as predictors, PCA has demonstrated to be able to provide reliable estimates with the substantial difference that the multivariate approach implies the use of at least the half of the variables needed in the full model to achieved the same levels of accuracy.

The third and fourth experimental contributions were aimed at estimating body measurements on European sea bass by the integration of technologies based on the analysis of digital images. In the first case genetic parameters and genetic gains were estimated for the improvement of processing traits under different strategies of selection. Morphometric measurement in this case allowed the application of selection criteria normally not applicable directly on live candidates, because involving the selection for lethal traits. The development of indirect criteria of selection for processing traits will provide, through the use of improved regression coefficients, an increase in the genetic gain because selection will be directly applied on candidates, avoiding the needs of less reliable estimates performed on sibs.

In the second case, linear measurements derived from image analysis have been tested as sex indicators, supposing the presence of a sex-related morphometric variability existing among individuals in the early phases of their life, unverifiable by visual inspection but potentially detectable by method of image and data analysis. A previous work obtained in a similar experiment on European sea bass found significant differences among fingerlings in body mass at the same age. Results of

this study suggested the presence of morphometric differences particularly related to some body compartments between individuals of different sex suggesting the need to deepen with more attention the question of using morphometric differences as possible sex indicator for this species.