CROP BREEDING AND APPLIED BIOTECHNOLOGY

ARTICLE

Biometry in plant breeding

Cosme Damião Cruz[1*], Pedro Crescêncio Souza Carneiro[1] and Leonardo Lopes Bhering[1]

Crop Breeding and Applied Biotechnology
21(S): e380621S5, 2021
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332021v21Sa18

Abstract: *In this manuscript we discuss the purpose and scope of biometry and its interactions, complementation, and even overlap with several other areas of genetics. We emphasize that biometry is an area of genetics that enables researchers to analyze, process, and interpret biological phenomena from data, usually obtained from experimental tests, in an improved way to guide strategies and decision making for optimization of resources. We also highlight the importance of the biometry professional in the context of breeding and the need for continual training, due to new demands for and challenges from inclusion of different types of information for processing and analysis paradigms to better interpret these paradigms.*

Keywords: *Biometry, genetics, breeding, data analysis, quantitative genetics.*

INTRODUCTION

Challenges have arisen from an increase in the demand for food because of population growth and difficulties in increasing production through the costs and adversities of climate change. These challenges are quite evident and have required the use of extensive technical, human, and financial resources in the attempt to overcome them. It is necessary to produce greater amounts of food with quality, at low cost, and under increasingly diverse environmental conditions.

Breeding has a long history and was for many years practiced by ancestors who had the ability and art of identifying individuals of superior performance, and they passed traits on to new crops through the descendants of these individuals. The success of this strategy certainly depended on the ability of those dedicated and competent farmers who often adopted subjective and not very accurate criteria in their choices, based on what was pleasing to the eye and had good flavor. As time went on, breeding became more intensive and included the need for always overcoming barriers and achieving new levels of production (Allard 1971). This purpose required an intense effort of educating professionals able to exercise the activity of breeders and act in public and private institutions exclusively dedicated to the activity of breeding. Choosing good genotypes was no longer an amateur activity but became a professional one; it was no longer exclusively an art, but came to be a science formed within the principles of genetics and experimentation. Observation was subjected to the critique of accuracy and suitability for making decisions.

From Mendel's laws, we discovered how the factors responsible for inheritance of the simplest traits were transmitted; however, many questions still remained regarding the mechanism of inheritance of more complex characteristics, such

***Corresponding author:**
E-mail: cdcruz@ufv.br
ORCID: 0000-0003-3513-3391

Received: 26 April 2021
Accepted: 10 May 2021
Published: 06 July 2021

[1] Universidade Federal de Viçosa (UFV), Departamento de Biologia Geral, Campus UFV, 36.570-900, Viçosa, MG, Brazil

as those involved in production of meat, milk, grain, or fruit, among other products. Therefore, a specific area in genetics arose to deal with this matter, which was quantitative genetics, that is, an area that deals with inheritance and variation of quantitative traits, which are strongly affected by the environment (Hallauer and Miranda Filho 1988, Vencovsky 1987, Vencovsky and Barriga 1992, Falconer and Mackay 1996). This area was the precursor of biometry.

From the postulates of quantitative genetics, we began to understand how genetic and environmental factors interact in control of these complex traits, and we came to understand the enormous difficulty faced by breeders in the process of choosing and discarding genetic material. These choices are based on prediction of real genetic values of individuals in competition, whether through the action of the environment, the genotype × environment interaction, or even some genetic effects that hinder the selection process, such as dominance and epistasis effects. These factors hinder establishing a good relationship between good performance of an individual (plant or animal) and the good quality of their gametes, which are the biological links passed from one generation to another.

Quantitative genetics has made an enormous contribution, generating valuable information placed at the service of breeders to obtain superior plants and animals in a shorter time and with fewer requirements of physical and financial resources. All this information was obtained from costly experimental trials, conducted in a careful manner, which always adopted and observed the scientific criteria of experimentation (Ramalho et al. 1993, Souza Júnior 2001, Oliveira et al. 2005). The large amount of data and the need for more appropriate analyses of this information came to require a new area of science, which meant the need for a new knowledge concentration area and a new contingent of qualified professionals. This new area is called "biometry", and the professionals acting in this area are called "biostatisticians".

Within this perspective, many groups of researchers perceived that they could provide a more robust format to the area of biometry, effectively contributing to dissemination of knowledge and training of professionals in the area, leading to scientific advances from the proposal and refinement of quantitative methods for data analysis and interpretation of genetic parameters.

We now understand biometry as the area of genetics that allows analysis, processing, and interpretation of biological phenomena from data, generally obtained from experimental trials in a more refined manner, to direct strategies and decision making for optimization of resources.

In the process of data analysis, the biometric area presents and refines models (in the case of statistical procedures) and computational architecture (in the case of computational intelligence approaches) for better data processing. By the action of data processing, biometry deepens studies on effective computational algorithms that make the use of quantitative methods viable and generate information useful for interpretation of results. Because of its interpretive activity, biometry requires knowledge of all biological areas, especially quantitative genetics, allowing understanding of biotic and abiotic factors that affect the phenomenon studied and directing the choice of better strategies and decisions aiming at optimal use of physical, financial, and human resources.

BIOMETRY AND QUANTITATIVE GENETICS – SCOPE AND COMPLEMENTARITY

One of the main objectives of breeding for crop and livestock production is an increase in yield, along with improved nutritional quality of individuals (animals or plants) primarily directed to consumption. Such objectives can be achieved through improving environmental conditions or through improving the genetic potential of individuals or populations (Bernardo 2002, Borém et al. 2017).

By changing the environmental and technological conditions, it is possible to increase yield and advance other benefits, because such changes allow the individual, plant or animal, to take better advantage of more favorable environmental conditions. Various branches of knowledge contribute to improvements in the environment. In addition, improving the genetic value of plants and animals contributes to achieving the desired aims. Most species have wide genetic diversity, and it is possible to select and recombine more adapted, better quality, and more efficient genetic types.

In many situations, breeding is the only means of attaining increases in yield and quality and, in relation to techniques of an environmental nature, it has the advantage of bringing about hereditary changes, that is, it is able to transmit the phenotypes of traits of interest to descendants. Thus, in many cases, the agronomic or livestock advantages can be perpetuated. In contrast, environmental improvement is normally quite costly and sometimes is not widely adopted

because of technical, personnel, and financial problems, and it may only provide good results if the genetic material available is improved in a way to be able to express its full potential within the adequate environmental conditions. In general, the researcher should be concerned with the combination of environmental and genetic improvement.

Professionals need to have a holistic view regarding breeding and environmental balance. Competent and qualified actions depend on aggregating scientifically based knowledge with practical experience and a global vision that will lead to satisfactory results. An understanding of the trait under selection and the instruments of generating information that will guide decision making is fundamental. In this context, training professionals in genetic areas, especially quantitative genetics and biometry, is indispensable.

Quantitative genetics is the part of genetics that studies quantitative traits, emphasizing their inheritance and the components that determine their variation. Quantitative traits are generally controlled by various genes and are highly affected by the environment, thus exhibiting continuous (and sometimes discontinuous) variations. Qualitative traits, however, have simpler inheritance (conditioned by one or few genes) and are little or not affected by the environment (Falconer and Mackay 1996). Biometry is complementary, because it pervades areas such as modeling, experimentation, and computational processing to generate information that, illuminated by the theories of quantitative genetics, allows interpretation of phenomena and decision making.

The study of inheritance and of variation in qualitative traits is based on analysis of generations, separating individuals in classes and evaluating their proportions in the results of determined crosses. Nevertheless, information on the individual is not of great value in the study of inheritance of quantitative traits, due to the random effect of the environment. If the effect of the environment can both increase and diminish the phenotypic manifestation of a trait, the mean value of a set of individuals will be a more reliable measurement, because the effects of the environment tend to cancel each other out. Thus, quantitative traits are studied at the population level. In addition to the mean, another measure used to define a population is the variance. Therefore, in studying quantitative traits, we evaluate which fractions of the mean and of the variance are inheritable.

Measurements of central position and of dispersion are routinely dealt with in statistical analyses; however, understanding regarding the mean and genotypic variance is a matter of great importance in the context of both quantitative genetics and biometry. In quantitative genetics, we seek to understand and interpret the meaning of values from this information for the breeder, whereas in biometry, we seek to establish the models that allow better estimation of these values based on experimentation and observance of the laws of genetics.

We know, in a very simple way, that the mean represents the sum of all observations divided by their total number. In some areas, this is sufficient. In quantitative genetics, the concern is not necessarily knowing how to estimate the mean, but understanding its value when obtained in whatever population or in a population in Hardy-Weinberg equilibrium. Breeders have clear expectations of the consequences regarding a mean derived from self-fertilization of a population or regarding a cross between populations. Thus, important concepts emerge regarding inbreeding depression and regarding heterosis that transcend the process of obtaining such estimates.

Understanding of the genotypic variation of a population calls attention to the existence of two basic models of genetics. In one model, the genotypic value of an individual is predicted by the quantity (2, 1, or 0 alleles) of a favorable allele found in the individual. In another model, this genotypic value refers to the consequence of the union of maternal and paternal gametes, with different genetic information, that combine in fertilization. The decomposition of this genotypic variation in additive variation and variation due to dominance in a monogenic model, and an epistatic variation component when considering more than one locus, is fundamental for directing breeding programs based on sexual recombination of the selected individuals and predicting the heterotic potential of hybrid combinations between superior and divergent parents. From the perspective of biometry, the question of estimation is also fundamental, because correct estimation and adequate interpretation are attributes of this area. Nevertheless, biometry also aggregates statistical information, such as concepts of fixed and random effects, and information regarding statistical and genetic designs and breeding strategies, especially those referring to the definition of testing and recombination units. From this, a more appropriate value of genotypic variation can be obtained, which, elucidated by quantitative genetics, will be interpreted and made available for breeders to use in breeding programs. Thus, as an illustration, genotypic variance in the context of quantitative genetics can be expressed through the following equation:

$$\sigma^2_G = a^2[(D+R) - (D-R)^2] + d^2H(1 - H) - 2adH\,(D - R) \quad (1)$$

for any population that has the genotypes AA, Aa, and aa with *D*, *H*, and *R* frequency and genotypic values $\mu + a$, $\mu + d$, and $\mu - a$, respectively, and

$$\sigma^2_G = 2pq\alpha^2 + (2pqd)^2 \quad (2)$$

for a population in Hardy-Weinberg equilibrium with the frequency of the **A** e **a** alleles given by p and q, respectively. In this expression, α is the mean effect of gene replacement, given by $\propto = a + d(q - p)$.

While the expression σ^2_G has genetic interpretations, its estimator requires knowledge from the researcher that goes beyond genetics, and therefore involves themes permeated by biometry. Thus, considering a genetic design of analysis of generations that includes the P_1 and P_2 parent generations, which are homozygotes and contrasting, and the derived F_1 and F_2 populations, there is the estimator of σ^2_G obtained through the information of variation measured in the populations evaluated using the following equation:

$$\hat{\sigma}^2_G = V(F2) - \frac{V(P1)+V(P2)+2V(F1)}{4} \quad (3)$$

However, if the population is structured in families and evaluated in a statistical design, such as randomized or completely randomized blocks, the estimator of σ^2_G would depend on information of an analysis of variance referring to the treatment mean squares (*MST*), residual mean squares (*MSR*), and mean squares of the number of replications of each family in the design, that is

$$\hat{\sigma}^2_G = \frac{MST - MSR}{r} \quad (4)$$

Thus, from the data of experimentations and from the basic principles of biometry, along with knowledge of the breeding objectives, biometry allows us to obtain information regarding the genetic variability of the population using equations 3 and 4, which will provide valuable interpretations according to the expressions, just as described in equations 1 and 2.

Another type of information that is essential in breeding is gain from selection (*GS*), the result of careful and meticulous activity of the breeder in identifying the best genotypes and recommending them for recombination, generating the improved population (Paterniani 1996). Quantitative genetics deals with this matter by affirming that the *GS* is the consequence of the displacement of the genotypic mean, since individuals that carry favorable alleles through the action of the breeder have a greater chance of leaving descendants than others do. With the practice of selection, the mean of a population for a determined trait of interest will be given by

$$\mu_s = \mu_0 + GS$$

where μ_s is the mean value of the improved population; μ_o is the original mean of the population; and GS is the gain obtained from selection. Considering that both the original population and the improved population are in Hardy-Weinberg equilibrium, if the frequency of the favorable allele in the improved population is $p' = p + \Delta_p$, the mean values of the populations in question will be given by

$$\mu_0 = \mu + (2p - 1)a + 2p(1 - p)d$$

and

$$\mu_s = \mu + (2p' - 1)a + 2p'(1 - p')d$$

Thus, so that the value of μ_s is superior to μ_o, resulting from the use of a selection technique, the values of gain from selection (GS) must be expressive. These gains depend on the genetic structure of the population and on the genetic effects of the alleles that control the trait under selection. In general, there are difficulties in knowing the values of gene frequency, as well as the genetic effects in a determined population. However, estimation of additive variance, based on appropriate genetic and statistical designs, will enable the prediction of said gain.

Assuming that for selection of quantitative traits, the value of Δ_p (given by p' – p) is expected to be small and that, consequently, its quadratic value (Δ^2_p) can be considered null, the equation is

$$GS = 2\Delta_p [a + d(1 - 2p)] = 2\Delta_p \alpha \quad (5)$$

Thus, it is understood by quantitative genetics, based on equation 5, that the increase in the mean of a population, through selection, is directly proportional to the variation in the gene frequency imposed by the practice of selection and to the value of exchange between the alleles given by the mean effect of gene replacement ($\propto$). Expression 5 in itself is interpretable, but it may also give rise to other relevant information, such as the fact that the GS will essentially depend on the magnitude and the significance of the additive variance component, which, for its part, is a quadratic function of $\propto$.

A complementation of the interpretation of the definition of the GS can be obtained by the biometric area, considering aspects of the trait, of the experimentation, and of the selection techniques, among others. Thus, we can understand the gain from selection by means of the following expression:

$$GS = H^2 DS \quad (6)$$

where H^2 is the heritability of the trait and DS is the differential of the selection carried out, expressed by the difference between the mean values of the individuals selected and of the original population under selection. A more extensive manner of interpretation is the expansion of expression 6, which results in the following equation:

$$GS = ip\, \sigma_G H \quad (7)$$

By equation 7, we find that the GS is the consequence of some factors that go beyond genetic factors. Thus, we can visualize that it is dependent on the intensity of the selection carried out; as it grows, it can lead to greater increases in gains, but it is known that it can lead to loss of variability in future generations through reduction in the effective size of the population under selection. It will depend on the breeding arrangement adopted, expressed by the parental control (p), where different units of selection and of recombination may be involved. It will depend on selective accuracy (expressed by H, which is the square root of heritability), indicating that well-conducted experimental trials and adoption of replications and appropriate designs may favor the GS. Finally, there is the need for genetic variability (σ_G, which expresses the genotypic standard deviation) to have gains from selection. The conclusion is that expressions 5, 6, and 7 are interpretations of the same phenomenon (GS) from different angles. Nevertheless, the main point here is to affirm that there is no interest in establishing that they are subjects from different areas, but rather that we can complement, deepen, and diversify knowledge to enrich our information and better support our decision making when areas complement each other.

Heterosis is defined as the superiority of the descendant in relation to the mean of its parents. If this definition is treated on the basis of quantitative genetics, we can consider two populations, P1 and P2, in Hardy-Weinberg equilibrium, with genotypic means given by the following equations:

$$\mu_1 = a(p - q) + 2pqd$$

$$\mu_2 = a(p' - q') + 2p'q'd$$

The genotypic mean of the F_1 population will be given by

$$\mu_{F1} = ap(p - \Delta) + d[2pq + \Delta(p - q)] - a[q(q + \Delta)]$$

Thus, heterosis (h) is given by

$$h = \mu_{F1} - \mu_p = \mu_{F1} - \frac{\mu_1 + \mu_2}{2} = d\Delta^2 \quad (8)$$

From expression 8, it can be concluded that the heterosis manifested in intervarietal hybrids is directly in accordance with the genotypic value of the heterozygote (d), expressed by the mean degree of dominance and of the difference in gene frequency, that is, of genetic diversity between the populations crossed. It is also understandable that heterosis will be at a maximum when one allele is fixed in one population and the other in the other population, as well as that if the populations do not differ in gene frequency, there will be no heterosis. Many studies of evaluation of genetic diversity among parents have been substantiated based on this expression, seeking those of good performance and that exhibit diversity, recommending their crosses with the expectation that the hybrid will manifest heterosis and that the segregating populations will show variability and transgressive individuals. Quantitative genetics stimulates researchers

to seek better understanding regarding heterosis, presenting and discussing some hypotheses of dominance and/or overdominance that would explain the appearance of heterosis. Furthermore, biometry takes the route of modeling and, as in the studies of diallel analysis proposed by Gardner and Eberhart, presents us with concepts and estimators of mean, varietal, and specific heterosis that are very useful in a program that aims to recommend hybrid combinations for commercial growing.

Finally, we emphasize one of the most important items of information in breeding, which is heritability. Quantitative genetics invites us to an understanding of this phenomenon from different equally important angles. We can understand heritability as being the proportion of phenotypic variation (σ^2_F) that has a genetic nature and indicates the degree of difficulty in obtaining gains from selection in accordance with the existence of genuinely genetic variability in the population of interest.

$$H^2 = \frac{\sigma^2_G}{\sigma^2_F} \qquad (9)$$

Yet, heritability is also a measure of the accuracy of the selection process, indicating the degree of accord between the phenotypic value manifested by the individual and its true genetic value. Thus, heritability can be quantified by the square of the correlation between observable phenotypic values and true genotypic values, that is

$$H^2 = r^2_{FG} \qquad (10)$$

where $F = G + M$, in which the phenotypic value (F) is given by the genotypic value (G) under the effect of the medium (M). In expression 10, it is assumed that $\text{cov}(F, G) = \sigma^2_G$, since replication and experimental randomness and random action of the environment ensure that the association between genotype and environment is null.

A third equally important concept refers to the fact that heritability measures how much the variations of the parents are reflected in the variations passed on by descendants. In this context, a linear relation is established between the genetic values predicted and manifested in the progeny (F) and the phenotypic values manifested in its parents (P), that is

$$F = \beta_0 + \beta_{PF}P + \varepsilon$$

Thus,

$$H^2 = \beta_{PF} \qquad (11)$$

Finally, a widely used concept is that which relates the genotypic performance of the individuals of an improved population (Y) resulting from the recombination of superior individuals, identified in comparative tests between individuals of an original population under selection, manifesting phenotypic values (X). In this case, the following predictive model is used:

$$(Y - \overline{Y}) = \hat{\beta}(X - \overline{X})$$

where $(Y - \overline{Y})$ is the gain from selection in the selected progeny and $(X - \overline{X})$ is the differential of selection practiced in the population under breeding. Also,

$$H^2 = \hat{\beta} = \beta_{Ut,Um} \qquad (12)$$

where $\beta_{Ut,Um}$ is the regression coefficient established by the relation between the phenotypic values of the test unit (Ut) and the genotypic values of the improved unit (Um).

We see that the definitions presented in expressions 9 to 12 are genetically based and are closely related to the questions of breeding. However, obtaining the value of this heritability requires a great deal of other information, such as type of family, genetic and statistical design, replications, breeding strategies, and others. Thus, supposing the evaluation of a set of g families in randomized block experiments with b blocks, in which n plants were evaluated per plot, the biometric approaches that take into consideration all the genetic aspects, as well as the information of the breeding strategy adopted and experimental particularities, would allow estimation of different heritability coefficients from the same experiment. These coefficients are not only interpreted under the standards of quantitative genetics, but also meet the different requirements of the breeder in decision making, such as how to adopt selection between and within a family, or stratified selection, combined selection, or simply family selection, among others.

BIOMETRY – KNOWLEDGE GUIDING THE SELECTION STRATEGY

Breeding programs require intensive experimentation so that accurate genetic values can be obtained and used as selection criteria. This has been a major challenge for breeders, that is, recognizing which evaluated items will have the best *per se* performance or that will provide the best descendants in segregating generations. In the early days of breeding, the individual phenotypic information was taken as the indirect value of the genotype; however, quantitative genetics has shown that most of this value, and especially most of the variation among these values, could have an environmental cause, and the adoption of these values as a criterion could lead to selection that is not very accurate, especially for polygenic traits strongly affected by the environment (Paterniani 1966, Falconer and Mackay 1996).

The adoption of the basic principles of experimentation, such as the need to conduct trials with replication, randomness, and, if necessary, local control, led to a revolution by providing more accurate mean values, information regarding experimental accuracy, and estimates of genetic and environmental parameters useful in guiding breeding strategies. This may have been the moment that the foundations of biometry emerged in breeding, in which there was complementation between genetic and statistical principles for a common purpose.

Proposals for improvement in the information regarding individuals and populations have been presented through diverse biometric approaches. For decades, questions regarding indirect selection, simultaneous selection, and family selection were raised, studied, and substantiated. Thus, theories of correlated responses or selection indices have indicated that aggregating information on auxiliary traits that are generally easier to measure and with greater heritability can provide criteria that lead to more balanced gains in a set of traits of interest to breeders. Another approach is one that seeks to increase selective accuracy through criteria that involve information on the individual, the focus of selection, and of its kin, generating family indices. In this case, the experimentation and the information regarding the genealogy of those evaluated become fundamental. Some genetic designs were especially established so that, in a single experiment, information could be obtained on the individuals and their kin, such as diallel crosses and designs I and II of Comstock and Robinson, in which the presence of both full sibs and half sibs are found. These designs provide information that assists breeders in selection of parents seeking to obtain base (segregating) populations of greater potential, as well as an understanding of the genetic effects involved in genetic control of the traits under study.

Along with the intra-allelic (dominance) and inter-allelic (epistasis) interactions, the genotype × environment interaction (G×E) is among the three main genetic difficulties faced by the breeder. The biometric area provided modeling that allowed a basis for better understanding of this G×E interaction, as well as alternatives to mitigate its harmful effects in the selection of genotypes aiming at recommendation. In this phase, the analyses of environmental stratification and studies on adaptability and stability are prominent. For these studies, various methodologies have been proposed, which allow recommendation more suitable for a wide region or for specific conditions. However, from the perspective of genotype selection aiming at recombination in recurrent selection programs, in which the effect of the G×E interaction can be an advantage, biometry also responded with appropriate modeling, presenting selection indices that capitalize on the G×E interaction (Cruz et al. 2011a, Cruz et al. 2012, Cruz et al. 2014).

The experimental statistics and biometry areas have contributed a large number of models, some of which achieve complexity by their parametrization with the sole purpose of ensuring the good choice of genetic material and of establishing a predicted value of performance of the individual or of its descendant in the best manner possible. Processing these data has become viable by the availability of increasingly powerful computational resources and the establishment of computational routines that incorporate the features of the biological phenomenon under study in the statistical models.

More recently, breeding has aggregated additional information of great value for establishment of selection criteria. Massive datasets of molecular markers, of spectroscopies, such as near infrared (NIR) imaging, and of images have allowed much more parametrized models to be fitted, with substantial gains in measurements of quality of fit. However, the use of models with datasets of large dimensions requires knowledge of more effective approaches and, for that purpose, some basic principles of prediction through statistical paradigms, computational intelligence, machine learning, and even non-Boolean logic are being incorporated in the biometric area (Cruz and Nascimento 2018).

THE BIOMETRIC AREA AS A RESPONSE TO ADVANCES IN THE AGE OF MOLECULAR GENETICS

Genetics is constantly evolving, overcoming new challenges and responding to the desires of the population regarding the most varied themes, emphasizing some themes related to heredity and to genetic control of important traits. This is also true in the biometric area. With the advent of modern molecular biological techniques, various methods have arisen for detection of genetic polymorphism directly at the DNA level (Cruz et al. 2011b). As of that time in the 1980s, this gave rise to a new cycle of knowledge generation, of fitting models, and development of algorithms appropriate for data processing, with a view toward the need to aggregate these data to the values of a phenotypic nature that are traditionally measured in plants and animals under selection.

It should be noted that this new information required new studies and scientific effort and resulted in advancement both in breeding activities, with information from genetic mapping and detection of genes controlling quantitative traits in plants and animals, and in analysis of population dynamics, with studies on diversity, regular mating systems, degree of kinship, and others.

Groups that act in the biometric area responded to world demand and formulated new models and adapted or refined procedures to aggregate the data and information coming from molecular genetics studies, which resulted in valuable information, including the prediction of genomic values of unrealized individuals and hybrids. They also helped clarify important phenomena, such as epistasis and hybrid vigor.

Among the areas of biometric study that involve the use of DNA molecular markers, the contributions in the strategy of genome-wide selection (GWS) are noteworthy. It allows aggregation of DNA information in selection of superior genotypes and provides greater genetic gains with greater effectiveness, lower cost, and time savings (Resende 2008, Silva et al. 2015, Borém et al. 2017, Sant'Anna 2019a, Sant'Anna 2019b).

Genome-wide selection (GWS) has been adopted for the development of more accurate methods of prediction of phenotypes of plants and animals, and to make inferences regarding their regulators (Meuwissen et al. 2001, Resende 2008). Although GWS is efficient, some challenges are routinely faced by professionals in the biometric area, which include genetic questions inherent to the use of molecular markers, statistical questions regarding the use of different data analysis paradigms, and especially computational questions arising from the requirement of analysis of relatively large datasets.

Thus, we perceive the importance of having professionals with a wide vision and ability to associate themes from areas such as the exact sciences and biological sciences to overcome great challenges, such as understanding genetic polymorphism, gene linkage, gametic phase disequilibrium, and genetic control of traits, so that biological models can be well established. Other challenges are in regard to presuppositions that must be assumed *a priori* to perform suitable stochastic analyses and obtain good interpretation of results. In addition, researchers must be very careful in relation to the dimensions of the model adopted (or parametrization), the presence of harmful effects of multicollinearity among the markers (explanatory variables), and the complexity of the quantitative traits under study (response variables), for which, without the assistance of appropriate computational resources, proposing or presenting solutions becomes unfeasible. Fortunately, various biostatisticians have sought to solve genetic issues, overcome computational limitations, and make analyses more viable and faster. They have endeavored to obtain accurate information regarding response variables, adopt more robust methods, and make use of strategies that can accommodate the information available in high dimensional molecular marker matrices.

Once more, we can highlight that the biometric area, applied to the questions of GWS, allows one to exercise the full knowledge of models presented by quantitative genetics, the uses of which were simplified over the years; and some effects known to be important, such as dominance and especially epistasis, were neglected. Thus, the possibility of greater parametrization of GWS models contributing to advance information is also noteworthy since, most of the time, additive or additive-dominance prediction models were used (Cruz et al. 2012, Cruz et al. 2014). In view of that, the possible effects due to the intra- and interallelic interactions were not detected to be suppressing genetic information and, in the statistical context, inflating the residue associated with the prediction model adopted. In this context, biostatisticians presented new statistical models and sought new modeling alternatives, which have been used in breeding programs (Gianola et al. 2011, Silva et al. 2014, Carneiro et al. 2017, Sant'Anna et al. 2019, Sant'Anna et al.

2021), and unlike the conventional stochastic modeling used up to then, they are based on the principles of machine learning and computational intelligence (Silva 2014).

By incorporating the use of methodologies and the application of new paradigms to breeding, such as computational intelligence and machine learning, biometry has provided new alternatives of analyses to assist in cultivar selection. Artificial intelligence methods are rapidly becoming essential for data analysis, especially as a support for decision-making processes (Carneiro et al. 2017).

THE BIOMETRIC AREA ADAPTED TO NEW PARADIGMS OF ANALYSIS AND DATA INTERPRETATION

Biometry applied to breeding is based on genetic principles and the purpose of meeting demands for interpretation of biological phenomena and providing information that can guide strategies and optimize resources. Thus, processing, data analysis, and interpretation of results is the activity inherent to the work of biostatisticians. The accumulation of information and advance of new technologies, especially in the area of computation, has made the biometric area attentive to new analysis techniques with diverse objectives, especially for analyses of prediction, classification, and recognition of patterns (Resende 2002, Resende et al. 2014).

Biometry, like data science, has earned prominence worldwide. Both are interdisciplinary areas directed to the study and analysis of data aiming at detection of patterns and/or obtaining information to assist in decision making. Biometry differs by the type of data it uses, by the phenomena emphasized, and, essentially, by being practiced in total observance of the principles of genetics and the purposes of breeding. In 2019, data scientists appeared in first place in the ranking made by the American recruiting site Glassdoor, which lists the best jobs in the United States. Biostatisticians are part of this select group of researchers.

Many statistical procedures summarize or allow interpretation of phenomena through measuring core trends, generally represented by the mean (arithmetic, weighted arithmetic, geometric, and harmonic), median, and mode. Their objective is to represent an entire set by a single value, and many hypotheses are associated with the existence of equality of these means in different sources of variation. Another measurement often requested in statistical studies refers to dispersion, represented by information regarding amplitude, variance, and standard deviation. Based on these core measurements, we construct estimates of error of a study and analyze the dispersion of its data.

Statistical models attempt to explain a response variable, with the purpose of fitting or classification through a set of variables or independent effects beyond an experimental error. Presuppositions regarding these errors are necessary and indispensable for establishment of estimators that lead to values that will serve as a basis for guiding strategies and adoption of procedures for optimization of resources. Biometric procedures based on principles and statistical models are widely applicable in breeding. Nevertheless, other currents of thought have been adopted, leading to solutions from different perspectives and giving researchers the opportunity to adopt the solution of greatest interest.

Differentiated solutions for problems in the biometric area can be achieved within the area of artificial intelligence (AI), or computational intelligence, which deals with automation of intelligent behavior and is divided into different paradigms, notably the symbolic, connectionist, and evolutionary paradigms. The symbolic paradigm is based on symbolic transformations (numbers, letters, words, and symbols) to establish a logical route until discovering a determined solution. The most successful form of symbolic AI is specialist systems, which use a network of production rules. The connectionist paradigm includes the procedures of neural networks and fuzzy logic inspired by the operation of the human brain. The term was introduced by Donald Hebb in the 1940s. The essence of these procedures is the learning algorithm that allows modification of the weights of connections and extraction of linear and non-linear information from the problem under study; it is therefore of great interest to breeding. Finally, there is the evolutionary paradigm, which is composed of a series of algorithms inspired by natural evolution, called genetic algorithms.

The neural network is an area of computational intelligence that meets the need for generating solutions related to numerous problems, including those of classification and prediction, which are routine activities in breeding. In contrast with the statistical approach, information is not summarized, but each example, or piece of information, is relevant in a learning process in which each input (which corresponds to the independent variables in statistical vocabulary) has

weights, called adjustable synaptic weights. Among the types of neural networks important in solving classification problems, the multilayer perceptron and radial basis function neural networks are prominent.

In neural networks with the purpose of fitting or classification, the variations in a response variable can be explained by a set of inputs whose linear and non-linear actions are captured through abstract variables, with information generated by neurons in hidden or intermediate layers with a variable number of both layers and neurons per layer. Such potentialities allow analysis of complex traits in a more accurate manner and better understanding of important phenomena such as dominance and epistasis.

Curently, biostatisticians make use of the resource of machine-learning approaches, which allow machines to develop models and make predictions without the need for reprogramming. As the machines are exposed to new data, they learn more and adapt in an independent manner. Machine-learning procedures can be classified as supervised or non-supervised. In supervised learning, the algorithms learn which model is most suitable for predicting the variable of interest, based on dependent variables (or inputs), through a set of examples that are submitted to the system. It is said that this type of learning involves the participation of an "external agent" and is generally used in situations in which the historic data foresee possible future occurrences. The decision tree procedure and its refinements and neural networks are prominent in this type of learning. In unsupervised learning, the procedure generates response patterns from measurements in variables of interest based on some structure and measure of similarity. The main procedures are reduction of dimensionality (principal component analysis, multidimensional scaling), cluster analysis (K-means, hierarchical methods), and Kohonen self-organizing maps.

CONCLUSION

Biometry is an area of genetics in continuous evolution, which has contributed to generating information and finding solutions to diverse questions in genetics and breeding. The professional in this area must know the features of the basic factors of heredity, population dynamics, mating systems, and strategies for conducting populations, among others. In addition, (s)he must be attentive to all the tools of analysis and of data processing arising from the rapid and continuous evolution of this area. It is an area that essentially depends on the critical perspective and attitudes of good breeders that are data-generating agents and the great beneficiaries of the information generated, but their attentive eye is on the methods of data analysis and on the algorithms that allow data processing supported by the biometric area.

A view of the current scenario leads to the conclusion that as many associated areas advance, an increasing volume of data will become available, which will require careful analysis and interpretation. These data now also result from the accumulation of information deposited in historical databases, and from aggregating and prospecting information of diverse natures, involving data on genetic materials, climate, soil, etc. Also prominent in this scenario are globalization of information and capturing of new data, especially acquisition and interpretation of image and spectral data.

There is an urgent need for further advances, now supported by biometric genetics, in the context of data processing and in computational intelligence and machine-learning approaches. This is associated with emerging areas in breeding, especially phenomics, and further developments in the question of handling big data.

REFERENCES

Allard RW (1971) **Princípios do melhoramento genético das plantas**. São Paulo, 485p.

Bernardo R (2002) **Breeding for quantitative traits in plants**. Woodbury, Stemma, 368p.

Borém A, Miranda GV and Fritsche-Neto R (2017) **Melhoramento de plantas**. 7th end, Editora UFV, Viçosa, 543p.

Carneiro VQ, Silva GN, Cruz CD, Carneiro PCS, Nascimento M and Carneiro JES (2017) Artificial neural networks as auxiliary tools for the improvement of bean plant architecture. **Genetics and Molecular Research 16**: gmr16029500.

Cruz CD and Nascimento M (2018) **Inteligência computacional aplicada ao melhoramento genético**. Editora UFV, Viçosa, 414p.

Cruz CD, Carneiro PCS and Regazzi AJ (2014) **Modelos biométricos aplicados ao melhoramento genético**. Editora UFV, Viçosa, 668p.

Cruz CD, Ferreira FM and Pessoni LA (2011b) **Biometria aplicada ao estudo da diversidade genética**. Suprema, Visconde do Rio Branco, 620p.

Cruz CD, Regazzi AJ and Carneiro PCS (2011a) **Modelos biométricos aplicados ao melhoramento genético**. Editora UFV, Viçosa, 514p.

Falconer DS and Mackay TFC (1996) **Introduction to quantitative genetics**. Longman Group Limited, Edinburgh, 464p.

Gianola D, Okut H, Weigel KA and Rosa GJ (2011) Predicting complex

quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC Genetics 12**: 87.

Hallauer AR and Miranda Filho JB (1988) **Quantitative genetics in maize breeding**. Iowa State University Press, Ames, 468p.

Meuwissen THE, Hayes BJ and Goddard ME (2001) Prediction of total genetic value using genome wide dense marker maps. **Genetics 157**: 1819-1829.

Oliveira AC, Furtado DF and Ramalho MAP (2005) **Experimentação em genética e melhoramento de plantas**. UFLA, Lavras, 300p.

Paterniani E (1966) Genética e melhoramento do milho. In Krug CA (ed) **Cultura e adubação do milho**. Instituto Brasileiro de Potassa, São Paulo, p. 109-148.

Ramalho MAP, Santos JB and Zimmermann MJO (1993) **Genética quantitativa em plantas autógamas: aplicações ao melhoramento genético do feijoeiro**. UFG, Goiânia, 271p.

Resende MDV (2002) **Genética biométrica e estatística no melhoramento de plantas perenes**. Embrapa, Brasília, 975p.

Resende MDV (2008) **Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**. Embrapa Florestas, Colombo, 330p.

Resende MDV, Silva FF and Azevedo CF (2014) **Estatística matemática, biométrica e computacional**. Editora UFV, Viçosa, 881p.

Sant'anna IC and Cruz CD (2019a) **Redes neurais artificiais na predição de valores genéticos: aplicação de inteligência computacional e seleção genômica no melhoramento de plantas**. Novas Edições Acadêmicas, Mauritius, 128p.

Sant'anna IC, Nascimento M, Silva GN, Cruz CD, Azevedo CF, Gloria LS and Silva FF (2019b) Genome-enabled prediction of genetic values for using radial basis function neural networks. **Functional Plant Breeding Journal 1**: a1.

Sant'anna IC, Silva GN, Nascimento M and Cruz CD (2021) Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks. **Acta Scientiarum. Agronomy 43**: e46307.

Silva GN and Cruz CD (2015) **Redes neurais artificiais: Novo paradigma para a predição de valores genéticos**. Schaltungsdienst Lnag o.H.G, Berlin, 92p.

Silva GN, Tomaz RS, Sant'anna IC, Nascimento, M, Bhering LL and Cruz CD (2014) Neural networks for predicting breeding values and genetic gains. **Scientia Agricola 71**: 494-498.

Souza Júnior CL (2001) Melhoramento de espécies alógamas. In Nass LL, Valois ACC, Melo IS, Valadares Inglis MC (org) **Recursos genéticos e melhoramento de plantas**. Fundação MT, Rondonópolis, p. 159-199.

Vencovsky R (1987) Herança quantitativa. In Paterniani E and Viégas GP (ed) **Melhoramento e produção do milho**. Fundação Cargill, Campinas, p. 137-214.

Vencovsky R and Barriga P (1992) **Genética biométrica no fitomelhoramento**. Revista Brasileira de Genética, Ribeirão Preto, 486p.