Breeding beyond genomics

From the Statistical point of view, I tend to think of Quantitative Genetics and related fields as domains where two main 'pillars' cohabitate: **Inference** and **Prediction**. Even if the same tool, e.g., penalized linear models, can be used for both tasks and inference and prediction may reinforce each other, they are distinct concepts. It is interesting to observe how these two pillars have reacted to big data, i.e., the *large p small n paradigm*. While studies where the main target is inference have tried (unsuccessfully) to protect against false positives, prediction practitioners have embraced the new era with joy. Why is that so? Very simple: Prediction is falsifiable via cross-validation whereas inference validation is not that straightforward, and an increase in variables easily leads to confounding. Most relevant distributional properties in inference validation depend on knowing the actual, 'true' model. Both inference and prediction are, however, encountering serious problems.

First, consider 'Inference'. For many years, inference in breeding involved a few parameters and two or very few carefully chosen models, say including or not maternal effects. Today, literature is flooded with reports of genome wide association (GWAS) signals and studies on selective footprints. In a standard GWAS, i.e., when markers are individually estimated without penalization, a main issue is controlling false positive rates. Identifying selective sweeps is also tricky, numerous statistics coexist, each pinpointing to different genome regions. Further, significance is not well defined in this task. Do not get me wrong, I am responsible for some GWAS and a few selective sweep studies. Large scale GWAS in unrelated individuals from populations with a large effective size can be very useful. Understanding patterns of DNA variability is based in solid theory. In most livestock studies, though, one should take results with caution as few signals have been replicated in independent studies.

Prediction in turn has been blessed with multidimensionality. As long as penalization and crossvalidation are properly employed, having more variables is more desirable than having only a few. Success in prediction as number of predictors increased is astounding in the livestock and plant breeding fields, and genetic progress has accelerated since the application of genomic selection. This has been possible, I insist, because prediction is falsifiable and, therefore, pragmatism dominates. Of note, a model may predict well even without including the causative mutations and so a better performing model may not be the one that is closest to 'biological causality'. The Achilles heel of prediction is **interpretability**. Most prediction machines are 'black boxes', although degree of 'opacity' varies. GBLUP allows at least recovering marginal marker effects, whereas convolutional neural networks do not. In all, interpretability is a non-negligible issue regarding communication of breeding methods to industry and society. Further, numerous prediction methods are available, yet they tend to perform similarly. Have we reached a 'methodological' plateau?

Quantitative Genetics skills are in high demand worldwide, yet Breeding is a mature field where scientific advances seem incremental. As in many disciplines, animal breeders' population is rather inbred, and scientific progress will likely increase by looking for

inspiration outside our own field of science. Where are the main challenges of livestock breeding, then?

I am optimistic. I do see many exciting prospects in several areas and let me just mention a few. Phenomics, the automatic measurement of numerous phenotypes, is by far the main and most attractive challenge, in my opinion. Highly unstructured, massive and heterogeneous datasets can now be cheaply produced by sensors. New opportunities exist both for developing algorithms that transform raw data into meaningful phenotypes and for implementing breeding programs based on high dimensional data. Among phenotypes, analyzing individual and group behavior via, say, video recording is an exciting problem. Impact of breeding on behavior is a topic of utmost interest in terms of research, industry, and society.

Breeding programs are accelerated evolutionary experiments and provide unique biological knowledge. This is a second domain where I foresee relevant discoveries, once longitudinal phenomic and genomic datasets are available. Animal genomes are highly resilient but also responsive; the same selective pressure is likely to result in (slightly) different allele frequency changes. Besides, response to selection has almost never been exhausted. This intriguing observation highlights the relevance of new mutations and that distinct physiological mechanisms may be activated in concerted action but at different stages.

Finally, domestication of terrestrial species has been a rare phenomenon in human history. Only a handful of species have been domesticated, likely because of behavioral and reproductive constraints. This scenario is completely different in aquaculture, where dozens of species recently have started to be grown in captivity, and many more are in the process. The aquaculture industry is in general more advanced technologically than terrestrial farming and poses new practical and methodological challenges. But domestication can be extended even more broadly, e.g., insects can be used for animal and human feeding. There are numerous uncharted territories for the curious breeder.

I finish by thanking numerous discussions with Miguel Toro, Daniel Gianola, Gustavo de los Campos and Andrés Legarra throughout the years.

Miguel Pérez-Enciso
ICREA, Passeig de Lluís Companys 23, 08010 Barcelona, Spain
and
Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, 08193 Bellaterra, Barcelona, Spain