Annual Review of Genetics

On the Road to Breeding 4.0: Unraveling the Good, the Bad, and the Boring of Crop Quantitative Genomics

Jason G. Wallace,¹ Eli Rodgers-Melnick,² and Edward S. Buckler³,⁴

¹Department of Crop and Soil Sciences, The University of Georgia, Athens, Georgia 30602, USA; email: jason.wallace@uga.edu
²Corteva Agriscience, DowDuPont, Johnston, Iowa 50131, USA
³United States Department of Agriculture, Agricultural Research Service, Ithaca, New York 14853, USA
⁴Institute for Genomic Diversity, Cornell University, Ithaca, New York 14853, USA

Keywords
quantitative genetics, breeding, agriculture, adaptation, heterosis, deleterious alleles

Abstract
Understanding the quantitative genetics of crops has been and will continue to be central to maintaining and improving global food security. We outline four stages that plant breeding either has already achieved or will probably soon achieve. Top-of-the-line breeding programs are currently in Breeding 3.0, where inexpensive, genome-wide data coupled with powerful algorithms allow us to start breeding on predicted instead of measured phenotypes. We focus on three major questions that must be answered to move from current Breeding 3.0 practices to Breeding 4.0: (a) How do we adapt crops to better fit agricultural environments? (b) What is the nature of the diversity upon which breeding can act? (c) How do we deal with deleterious variants? Answering these questions and then translating them to actual gains for farmers will be a significant part of achieving global food security in the twenty-first century.
INTRODUCTION

The outlook for future food security is well known: By 2050 there will be approximately 9 billion people on the planet, and 9–11 billion by 2100 (50). The linear growth in food production seen historically will not be enough to satisfy global demand by 2050 (142), especially with increasing demand for high-quality protein by a growing global middle class (171). This argument is as old as Malthus, yet to date agricultural production has managed to keep ahead of population growth thanks to mechanization, fertilization, plant breeding, and other agronomic advances. However, it is dangerous to assume that just because production has outpaced population growth in the past that it will inevitably do so in the future. Most of the easily arable land has already been brought into cultivation, and owing to land degradation and urbanization, its area is actually shrinking (198). Most major sources of water—both surface and groundwater—are being overdrawn, and water shortages are likely to get worse in the next few decades (54). In addition, extreme weather events, such as droughts, floods, and damaging storms, are predicted to increase as climate change moves through the twenty-first century (17, 25, 27). In short, it is getting harder and harder to grow crops, yet more and more people rely on us doing so.

THE FOUR STAGES OF PLANT BREEDING

Only 1–3% of modern industrialized society is directly involved in food production (187), a sharp decline from rates in historical societies. Much of this shift is the result of management improvements (e.g., plows, planters, fertilizers), but much is also due to improvements in crop genetics through breeding (37). Although some important agricultural traits have discrete genetics—such as Mendel’s famous pea phenotypes or major disease resistance loci—most are highly quantitative, such as plant architecture, maturity, nutritional quality, and yield. Although quantitative genetics is roughly a century old (43, 47, 188), the principles it encompasses have been applied throughout the history of plant breeding.

Based on the techniques involved, we have split plant breeding into four major stages: three that have already been achieved and one that we will likely reach in the near future (Figure 1). Each of these stages builds upon earlier ones by integrating established techniques with new technologies to increase breeding efficiency.

Breeding 1.0 began 10–12 thousand years ago, as people from around the world explored and cultivated nearly 7,000 species of food plants (81). Professional breeders were rare or absent, but phenotype-based selection by local, independent farmers eventually resulted in the dramatic phenotypic changes seen in modern crops (36, 120).

Breeding 2.0 began in the late nineteenth and early twentieth centuries when inbreeding depression was recognized (32), Mendelian genetics was (re)discovered (26, 34, 173), and quantitative genetics theory was established (43, 188). Many advances in plant breeding during this time were in the science of breeding itself, including replicated field trials, controlled crossings, statistical analyses, formal experimental designs, hybrid breeding, pedigree-based estimates of breeding values, and precise measurement of yield at scale (e.g., with multirow combines).

Roughly 30 years ago we entered Breeding 3.0 when molecular markers and genomic data began to complement phenotypic data. This stage started with marker-assisted backcrossing and pedigree confirmation, then moved to dissecting complex traits with linkage mapping (91). The introduction of high-throughput genotyping then expanded the quantitative genetics tool kit to dissect variation in natural populations (genome-wide association) (147) and to select on genome-estimated breeding values (genomic selection) (119).
Figure 1
Four stages of plant breeding. Breeding efforts can be divided into four existing or near-future stages based on available methodology. Breeding 1.0 is mostly incidental selection by farmers. Breeding 2.0 involves using statistics and experimental design to improve selection efforts. Breeding 3.0 includes genetic and genomic data and is the current state of the art. Breeding 4.0 will probably arrive soon (at least for some crops) and is marked by the genome-wide ability to combine any known alleles into desirable combinations.

We are now on the cusp of Breeding 4.0, a new level of breeding where functional genetic variants can be rationally combined both faster and better than ever before. This level of breeding is catalyzed by major technological advances in genetics and information systems. For example, genome resequencing studies can now cost less than a replicated yield trial, and genome editing is
expected to enable parallel, precise modifications to many sites (possibly hundreds) per generation. High-throughput phenotyping can measure numerous traits with unprecedented spatiotemporal resolution, and machine learning approaches permit the processing and interpretation of agronomic data at a level far beyond what humans can assimilate.

Although one can conceive an even more advanced level (Breeding 5.0) involving de novo design of genes, pathways, and traits, its realization is so far in the future that we will not spend time on it here.

OVERVIEW

In this review, we focus on several major questions facing quantitative genetics as it transitions from Breeding 3.0 to Breeding 4.0. Crop quantitative genetics has always straddled the boundary between basic and applied science, so it is no surprise that these questions touch on basic mechanisms of evolution, domestication, and development even as they impact plant improvement practices and agronomic performance.

We have chosen three key questions to consider: (a) How do we adapt crops to better fit agricultural environments? (b) What is the nature of the diversity upon which breeding can act? (c) How do we deal with deleterious variants? We still have only partial answers to each of these questions, and fleshing them out is a prerequisite to fully harnessing the tools of Breeding 4.0. Although many of our examples come from the grasses owing to the authors’ familiarity with them, these issues apply across all crops.

HOW DO WE ADAPT CROPS TO BETTER FIT AGRICULTURAL ENVIRONMENTS?

At the risk of stating the obvious, crops did not evolve in agricultural settings. Modern agriculture puts plants in a very different environment from their ancestors, with many selective pressures operating in different ways (Figure 2). As a result, new alleles are pushed to high frequencies, and the resulting selective sweeps have been found in maize (72), wheat (18), rice (59, 190), soybean (100, 202), tomato (84, 102), and sunflower (20), among others. Importantly, no crop has yet reached the top of its adaptive peak, especially since changing agronomic practices keep moving it. Finding ways to move crops closer to their peak is the ultimate goal of modern plant breeding. We focus here on three aspects of agriculture where increased understanding is likely to benefit breeding in the near term: new mega-environments, interplant competition, and soil interactions.

New Mega-Environments

A mega-environment is a group of environments sharing similar climatic conditions (day length, rainfall, temperature, etc.). Most crops originated in very localized regions but were spread by humans throughout the world (82). Adaptations to new mega-environments are thus among the largest changes crops have gone through since domestication.

The fundamental problem with finding the genes involved in these adaptations, however, is that population structure is usually highly confounded with the same adaptations one wants to map (160). Although statistical methods can reduce false positives due to background population structure (137), the correlation between adaptation and structure usually means they also reduce true positives. Such confounding can be mitigated by decoupling population structure and adaptation, either through experimental design or by selecting populations where they are naturally decoupled.
In agriculture, the gold standard experimental designs to study adaptation usually follow either a Multiparental Advanced Generation Intercross (MAGIC) (87) or Nested Association Mapping (NAM) (117) design. Although the details differ, both methods separate adaptation from population structure by using a small number of parents to create a large number of recombinant progeny. The resulting population has a balanced structure and retains the statistical power of other controlled mating designs, allowing adaptive alleles to be identified with high confidence. Such analyses have been used to identify genes controlling flowering time and photoperiod sensitivity (14, 73, 107, 114, 155), drought tolerance (96), salinity tolerance (150), cold tolerance (152), and development rate (124).

An alternate approach is to take advantage of populations where structure is naturally decoupled from selection. This situation usually occurs in organisms with large population sizes that live across environmental gradients with few barriers to gene flow. Environmental genome-wide association studies using such populations have proven extremely powerful and include plant height and inflorescence architecture in sorghum (94, 122); photoperiod-, altitude-, and drought-related adaptation in maize (4); climatic adaptation in conifers (195); a host of environmental conditions in Arabidopsis (55, 93); and even salinity tolerance in Atlantic herring (112). Although not every organism is suitable for this sort of analysis, taking advantage of those that are can identify genes and alleles that are adaptive for different environmental conditions. Breeders can then target these genes for improvement in related organisms.

**Interplant Competition**

The wild ancestors of crops rarely, if ever, existed in genetically homogenous stands, and even traditional landraces are usually heterogeneous mixes of genotypes. Modern crops, however, often...
Group selection:
a state where selection acts on the fitness of the entire group instead of individuals

grow in genetically uniform stands where the combined performance of a field is more important than any individual plant. This imposes a form of group selection on modern cultivars, so that plant architecture is optimized for whole-field production through steeper leaf and root angles, shorter plants, better tolerance of high planting densities, and similar traits (35, 37).

A competitive plant is one that is good at taking resources for itself at the expense of its neighbors. Although several authors have noted that less-competitive plants yield better in modern agricultural settings (e.g., 35, 127, 151), the genetic loci controlling competitiveness have only rarely been mapped. This may be because the basic traits that underlie competition (root angle, leaf angle, metabolite production) are usually interesting in their own right, so they are often mapped without regard for competition (11, 16, 197). The few studies directly mapping competition confirm that the beneficial loci in low-competition settings are different from ones in high-competition settings (90, 166, 196). Especially at early breeding stages, both breeding and research environments tend to reward plants that strongly compete with their genetically different neighbors, even though these plants do not consistently deliver the best yields when planted in uniform stands (41). Integrating selection for lower competition into most breeding pipelines requires either alternative field schemes or (more likely) in silico modeling (203) to determine optimal trait combinations to include in breeders’ selection indices.

Soil Interactions

Soil is not just a substrate for crop growth; it is arguably one of the most complex ecosystems on Earth (30). Repeated tilling leaves the soil in a state of perpetual disturbance (both physically and biologically) with highly altered microbial profiles that are further shifted by fertilizers, irrigation, crop rotation, and other agronomic practices (33, 57, 58, 113, 128, 162, 184). Such disturbance creates a situation where plants’ evolved responses are maladapted to the ecology of modern agricultural soil. Nascent research seeks to understand the genetics of microbial associations, such as in recruiting beneficial rhizobia (2, 29, 80, 164) and mycorrhizae (2, 95), establishing microbial communities (3, 19, 174), excluding pathogens (5, 118), and potentially influencing food quality (12). Microbial communities can alter host phenotypes (132, 179), although very little is known about the quantitative genetics involved. Microbial profiling and metagenomic analyses of environmental (170) and crop-associated (40, 49, 135) microbial communities could open a window onto how to harness these communities for future breeding efforts. It should be noted, however, that research in several plants indicates that the environment is probably the biggest driver of plant microbial communities (77, 106, 135), especially in the rhizosphere. This means it may be easier to adjust management practices or apply specific microbes directly to seeds or plants than to breed for improved associations.

WHAT IS THE NATURE OF THE DIVERSITY UPON WHICH BREEDING CAN ACT?

Although breeding can treat genetics like a black box (and did so in Breeding 1.0 and much of 2.0), understanding the nature of plant genetic variation can significantly improve breeding efforts. One of the best tools for this is the plant’s own genome because it contains a history of environmental adaptation. Comparing different genomes within a species thus gives clues to which loci are important; this sort of analysis is especially powerful in crops with hundreds to thousands of genomes available (e.g., 1, 15).

A major conclusion of whole-genome comparisons is that plant variation is significantly driven by variation in gene content and copy number instead of just differences in protein sequences.
This hypothesis was first put forward nearly 20 years ago (154) and suggested that the presence and absence of genes might underlie important phenomena, such as hybrid vigor (46). Presence–absence variation explains significant amounts of phenotypic variation (22, 104, 180), although many presence–absence genes have low RNA expression (68) and even fewer of them are translated into proteins (182). It thus seems that, even though many genes can have variations in copy number, not all such variations are important. Most of the important presence–absence variation in plants appears to stem from either polyploidy or tandem duplications, as opposed to other mechanisms like transposon duplication.

Polyploidy

There have been at least two major polyploid events in all angiosperms, and many more in most lineages (79, 186). Although this implies that polyploidy is the norm, most polyploid events are actually evolutionary dead ends (115), and it is only the lucky few that survive in the long term.

Even though plants readily undergo polyploidization, the fitness consequences of polyploidy are still not completely understood (108). One known consequence is that many of the duplicated genes are rapidly lost via mutation, deletion, or other mechanisms. This loss is not random; rather, one syntenic segment (subgenome) from a given region tends to remain relatively intact, whereas the other is preferentially mutated, deleted, or otherwise degraded (156) (Figure 3). Both maize and *Brassica rapa*, for example, show clear evidence of differential genome retention (21, 145). Bread wheat (*Triticum aestivum*) is an example of this process in action, as a substantial portion of its allelic variation stems from dosage-altering loss-of-function mutations stemming from its polyploidy approximately 10,000 years ago (88). Subgenome dominance is apparently established very rapidly with both gene expression differences and epigenetic marks of dominance appearing in the very first generation (39).

Tandem Duplication

Tandem duplication of genes occurs because of replication errors or unequal crossing-over during meiosis (143). Since tandemly duplicated genes are often still under the control of their native regulatory elements, these duplications provide an easy mechanism for modification of gene dosage. For example, variation in carotenoid degradation in maize is due to a dioxygenase gene that first transposed approximately 2 Mb away and then was tandemly duplicated up to 23 times, providing significant quantitative variation in carotenoid degradation (168). Tandem duplication has also been implicated in aluminum tolerance in maize (109), salt tolerance in wheat (200), and dwarfing in both wheat (99) and sorghum (125).

The Size of Mutational Space

Given that both polyploidy and tandem duplication can lead to large increases in gene content (including significant variation within a species), how big is the actual mutational space in a crop? Or in other words, how many different mutations can shift a crop toward the same goal? Although population genetics highlight the opportunities for convergent evolution (140), older studies suggest the space for mutations can be remarkably small. For example, mutations in numerous genes can produce the sweet corn phenotype, but three of four independent evolutions were all in one cleft of a single enzyme (172). Meanwhile, rice aromas have developed via 10 independent mutations in the same gene (86), and the nonshattering phenotypes of domestic sorghum, rice, and maize are all due to independent mutations in the same orthologous genes (103). These were all recent
Duplicated genomes

Subgenome dominance. Immediately after a polyploidy event, the genome contains two complete sets of genes (A and B, center). Over time, many genes are reduced to a single copy owing to mutations and deletions (red Xs). Although one would expect this process to be random (left), most genomes show evidence of genes that were preferentially retained or lost in blocks (right). The dominant subgenome consists of the blocks that are preferentially retained (blue boxes). Note that which subgenome a block belongs to depends only on how well it is retained, not on where it came from. In other words, a species with both A and B genomes due to polyploidy can have a dominant subgenome consisting of a mix of A and B segments, instead of A being completely dominant to B or vice versa.

Quantitative trait loci (QTLs): locations in the genome that have been identified as influencing a quantitative trait

selections caused by humans, but natural variation follows the same pattern. For example, sorghum and maize diverged about 12 million years ago, during which time maize underwent a whole-genome duplication from which it currently retains 20–30% of the duplicated genes (156). Despite this distance, large mapping projects in both species show consistent alignment of the quantitative trait loci (QTLs), where one locus in sorghum is matched by two syntenic QTLs in maize (107). All of this indicates that, in practice, there are only a limited number of routes to alter a given phenotype.

What about raw genome size? Angiosperm genomes can vary over three orders of magnitude (8), from the carnivorous herb Genlisea tuberosa (0.061 GB) (44) to the canopy plant Paris japonica (149 GB) (136). At first glance, these differences in genome size can appear to have strong impacts on QTL, especially ones that act through gene regulation. For example, in Arabidopsis (0.125–0.150 GB genome), almost all QTLs are within 5 kb of the affected gene (185). By contrast, two of the best-characterized QTLs in maize (~2.6 GB genome)—teosinte branched 1 and vegetative to generative transition 1—are in enhancer elements approximately 60 kb away from the genes they affect (24, 153). However, even though the distance between important DNA elements can vary dramatically across plant genomes, the actual genomic space that is important seems to be much smaller and more constant. For example, nearly 90% of maize phenotypic variation could be assigned to the 3% of the genome that is either protein-coding or noncoding open chromatin (149). This gives a much smaller sequence space to search for variation. Similarly, only 5–7%
of the rice genome is in open chromatin regions that are likely to impact function (199). If this pattern holds across other plants, it implies that the functional mutational space may be quite small indeed.

**HOW DO WE DEAL WITH DELETERIOUS VARIANTS?**

Human geneticists often focus on deleterious mutations because of their role in diseases, but most plant breeders do not think about them much. Any obvious deleterious mutation is eliminated early, with little thought given to molecular mechanisms or how many milder mutations go unseen. Although *Arabidopsis* is estimated to accumulate 1 mutation per generation (131), maize appears to accumulate nearly 90 of them (23). Assuming 5–10% of the genome has a functional role (149), this means five to nine mutations per generation could potentially affect a phenotype, and most of them would probably be detrimental. Identifying, controlling, and repairing these mutations is likely to be a major aspect of research as we move toward Breeding 4.0.

**Deleterious Alleles and Domestication**

When researchers compare domesticated species with their wild relatives, they find that domestication is often associated with an increase in the number of apparently harmful alleles. This pattern can be seen in both plants and animals, including rice (105, 126), sunflower (144), tomato (84), dogs (110), and horses (159). On the basis of population genetic theory, there are a few processes that may explain this increase in detrimental alleles. First, domestication alters selection pressures so that traits favored in the wild become neutral or disfavored under domestication (74). Second, purifying selection is frequently reduced after domestication bottlenecks (83). Third, inbreeding of domesticated varieties further reduces their effective population size and effective recombination rates (105). Although the relative contribution of each force varies by species, all of them probably act to some extent on domesticated crops.

Emerging evidence suggests that this cost of domestication can be tempered by modern improvement practices. For example, modern inbred lines of maize carry fewer nonsense mutations than their wild relatives (22) and have less genetic load than traditional outcrossing landraces (193). Young alleles generally appear to be under more stringent purifying selection in maize than in its ancestor, teosinte (7). In cassava, meanwhile, domestication loci have fewer deleterious mutations in cultivated varieties compared to progenitor populations even as drift has increased plants’ overall genetic load (141).

**The Problem with Chromosomes**

One problem with deleterious variants is that the biology of chromosomes makes them hard to breed out. Organizing genes into large linear chromosomes helps cells properly segregate complete genomes during division, but this structural constraint also reduces the efficiency of selection by linking alleles that may have different (and often opposite) fitness values. This means that any given individual has a mix of good and bad alleles linked together within a haplotype, and recombination is often not efficient enough to ever create a single ideal combination. Instead, multiple suboptimal haplotypes selectively interfere with one another so that none reach fixation. This phenomenon is called Hill-Robertson interference (65) and results in negative linkage disequilibrium (repulsion phase) between beneficial variants in low-recombination regions (42).

Hill-Robertson interference has important (but often overlooked) consequences for plant breeding. First, interference reduces the efficacy of selection on any individual site (65). This
Heterosis: the tendency of a hybrid offspring to be more fit than either of its inbred parents.

allows slightly deleterious variants to accumulate and reduces the probability of fixation for any given beneficial allele, a process that has been shown in rice (105), maize (148), and sunflower (144). Second, having beneficial alleles spread across multiple nonrecombining haplotypes reduces genetic variance by eliminating extreme phenotypes, which in turn decreases the raw material for selection to act upon. Third, assuming most deleterious alleles are at least partly recessive, the low-recombination regions where interference is most severe should benefit more from heterozygosity than high-recombination regions. This benefit arises because low-recombination regions have a greater chance to complement deleterious alleles. Artificial inbreeding should thus favor individuals that retain heterozygosity in low-recombination regions, as has been verified in oat (71) and maize (117). Interference and its consequences are easiest to see around centromeres because they are so large, but the same processes also occur in localized regions across the genome.

One can assume that at least some QTLs of agronomic importance are located in similar low-recombination regions, making their improvement extremely difficult. One work-around is to select for haplotypes that complement each other in the hybrid state, which is what results in distinct heterotic groups in maize and other hybrid breeding programs (169). Looking toward Breeding 4.0, technologies that allow breeders to precisely manipulate recombination sites or alter specific alleles through genome editing could bypass the Hill-Robertson effect entirely.

Implications for Heterosis

Heterosis has been known for 150 years (31), but its molecular underpinnings are still debated (157). Because of the complementation mentioned earlier, regions with strong Hill-Robertson interference are expected to behave as a single overdominant locus (98) even though the individual alleles are strictly dominant. This effect is called pseudo-overdominance.

Experimental work confirms that at least some apparently overdominant loci are actually pseudo-overdominant. Several generations of random mating in a biparental maize family eliminate the appearance of overdominance (48, 121), and grain QTLs consistently localize to the centromere and pericentromere (92, 158), which are both predicted to show strong pseudo-overdominance. Confirming this prediction requires fine-mapping traits, something that by definition is difficult to do in low-recombination regions but has been managed occasionally. For example, a single overdominant QTL near the centromere of maize chromosome 5 could be separated into two dominant QTLs in repulsion phase (53). Meanwhile, a sorghum height QTL separated into two genes with opposite effects approximately 3 Mb apart (98); both alleles were dominant and resulted in apparent overdominance in the hybrid.

Although we have focused on low-recombination regions, these regions usually have relatively few genes in them (6, 38, 75). Therefore, even if the most obvious genetic load is in low-recombination regions, the most overall load is likely in high-recombination regions. Selection is more efficient in these regions, but other factors associated with high recombination can oppose movement up the fitness landscape. Most notably, GCbiased gene conversion likely maintains some deleterious mutations at frequencies much higher than expected under mutation-selection-drift equilibrium (52, 149).

The importance of low-recombination regions for heterosis may also depend on the species. In bread wheat, for example, heterosis appears to be more driven by epistatic effects than by dominance effects (78). Significant heterosis has also been seen from gene-dosage effects in rice (70), maize (194), and Arabidopsis (45), indicating that the relative amounts of different genes are just as important as dominant and recessive relationships. Heterosis also changes in different environments (101), which makes it more difficult to pin down consistent mechanisms.
DELIVERING QUANTITATIVE GENETICS TO THE FIELD

How important is it to understand the adaptive and deleterious variants across the genome? This sort of knowledge has already proven useful in a few specific pathways, especially for traits that are difficult and/or expensive to score and that are controlled by a small number of genes. Such traits include many disease resistance loci, submergence tolerance in rice (189), and carotenoid accumulation in maize (56, 192). For Breeding 3.0 and 4.0 to achieve their full potential, we must identify how to translate basic knowledge into real-world results.

Fisher-Orr and the Diminishing Returns of Breeding

The basic unit of modern crop breeding is the QTL. The more beneficial QTL you can breed into a variety, the better that variety becomes and the higher up the adaptive peak it climbs. First-generation QTL mapping focused on the big domestication loci (e.g., 13, 85, 103, 139, 165, 167, 183) and disease resistance genes (e.g., 69, 111, 163, 191). These efforts were very successful, but as time passed, the pattern emerged that newer QTLs generally had smaller effect sizes, especially in outcrossing species. Smaller QTLs require more effort to map and provide less benefit, putting crop breeding in a state of diminishing returns.

Several phenomena contribute to this pattern. First, scientists were rational: They cloned biggest QTL first. But another major factor was that most QTL effects appear to follow the Fisher-Orr geometric model (129, 130). In brief, this model says (a) large-effect mutations are more likely to be harmful than helpful and (b) selection is less efficient for complex traits than for simple ones. In breeding contexts, this means that alleles with large beneficial effects rapidly become fixed, after which only successively smaller and smaller allele effects actually move plants closer to the adaptive peak (Figure 4). In addition, most beneficial alleles for complex traits (e.g., yield) have small effects in the first place.

Large mapping panels have confirmed that most alleles have small effects (14, 116, 133, 180). This implies that, especially in major crops, many of the large-effect QTLs have probably already been identified and fixed in elite germplasm, leaving breeders to work with progressively smaller effects that are harder to identify. This appears to be especially true in outcrossing species like maize (14, 134) and even cattle (reviewed in 64). Self-pollinating species appear to have more large-effect genes, such as the dwarfing genes of the Green Revolution (60), flowering time in barley (114), and drought tolerance in rice (178). It may be that in the constant reshuffling of haplotypes, each generation in outcrossing species selects for small-effect alleles that play nice with other genes. By contrast, self-pollinating crops pass down complete haplotypes to their offspring, which may more readily allow large-effect alleles to evolve.

Genomic Prediction

One method of dealing with increasingly small QTLs is to stop trying to map them and instead work with the entire genome. This is the approach of genome-wide prediction and genomic selection (Figure 5), a pair of incredibly successful paradigms that are affecting almost all sectors of breeding (63). Genome-wide prediction is the process of using genetic data to predict an individual’s phenotype; genomic selection is simply using those predictions to make breeding decisions. This sort of selection scheme was first demonstrated by Meuwissen et al. (119), and its goal is to improve breeding by both reducing the cost of phenotyping and reducing the cycle time for early generation selections.

Genomic selection has proven extremely beneficial to dairy cattle, and signs indicate it will be similarly valuable for crop breeding (28, 63, 176). However, there are still many unknowns, such...
Figure 4
The Fisher-Orr model states that the closer an organism is to an adaptive peak, the more likely a large mutation is to be harmful. The blue individual is far from the peak, so many different mutations result in a net benefit. By contrast, the green individual is close enough to the peak that even moderately sized changes can overshoot the ideal and leave its fitness lower than before.

as the optimal allocation of resources or choice of experimental designs. Krchov & Bernardo (89), for example, found that some breeding designs were always better under genomic selection, some were always better under phenotypic selection, and many could shift between the two based on budget, resource allocation, trait heritability, and other factors. Heffner et al. (61) estimate that genomic selection outperforms marker-assisted selection in terms of gain per unit time as long as the trait has a heritability of 0.2 or greater. Meanwhile, the choices of statistical models for genomic selection are many and growing (51); although with real-world data, there seem to be little practical differences among them (62). Genomic selection for hybrid performance is also becoming more powerful (76, 97), which is likely to make genomic selection of hybrids possible even for historically inbred crops like wheat (201).

Genomic selection has great potential for improving crops, but in most cases, that potential is still unrealized. This is most true for species outside of the major row crops and outside of the developed world, largely due to simple logistics. Genomic prediction requires high-throughput genotyping systems and bioinformatics expertise that many programs do not have, and breeding decisions for annual crops must occur in two to three brief windows for genomic selection to be worthwhile. The cost of genotyping can also be a major determinant of whether genomic selection is worthwhile (89, 146), though the still-falling price of sequencing implies that may not be true for long. Prediction models are also limited to highly related germplasm; models using different breeding programs rarely have value, and accuracies can rapidly drop even beyond half-sib family structures (10). All of these barriers need to be overcome for genomic selection to see truly global deployment.
Figure 5
Genomic selection. (a) Traditional selection involves cycles of making crosses to get new genetic combinations, evaluating the new varieties by phenotyping, and using those evaluations to select the next generation of parents. (The point at which a variety is spun off into production is not shown.) (b) Genomic selection adds several layers to this scheme and splits the process into training and selection cycles. When training, the breeder must go around the entire outside loop: making crosses, getting genotype and phenotype information from them, building a mathematical model, and finally using that model to select future parents. After that, the breeder can skip the training portion and go directly from new material to genotype to choose parents based on the model, with no need to phenotype. Although the model must be retrained every few generations, breeders generally save significant time and money by skipping the (usually laborious and expensive) phenotyping step two or three times.

Democratizing Breeding 3.0
Many crop breeding programs are still at the Breeding 1.0 or 2.0 stages, especially in the developing world, yet widely deploying Breeding 3.0 is critical to providing benefits for the global population. This democratization process is already underway, as inexpensive genotyping catapults formerly neglected crops into twenty-first century genetics. Millets (9, 175, 177, 181), cassava (138), cacao (123), strawberry (67, 161), sweet potato (66), and many others are benefiting from this technology, but it must reach even further to have truly global benefits. DNA sequencing is likely to drop to $1/Gb in the next several years, which implies that virtually any breeding line can be skim sequenced for little cost. High-throughput sequencing facilities, meanwhile, will probably reduce the price of genotyping seeds to only $1 or $2 per sample. However, while sequencing and
genotyping are dropping to negligible prices, field work and the logistics of tracking and processing thousands of samples are not. New ideas and informatics systems are needed to enable small programs to deal with these logistics. Most breeders do not have the time or skills to process the data themselves, and ideally breeders should never even need to look at a nucleotide sequence. Projects like the Genomic Open-source Breeding Informatics Initiative are starting this effort, but a larger international community of both public and private entities is needed to make Breeding 3.0 a global reality.

MOVING TO BREEDING 4.0

Beyond just democratizing Breeding 3.0, what do we need to move to Breeding 4.0 and capitalize on the ability to manipulate individual genes and even single base pairs? The last decade has seen tremendous strides for identifying functional variants through genome-wide associations, evolutionary constraint, comparative genomics, and a range of molecular biology assays. For breeding, however, each of these approaches needs to be carefully measured in terms of its usefulness. Unlike human genetics, crops do not have one focus species but rather hundreds of species, each with many different breeding targets around the globe. A major question for the next decade is how to leverage knowledge across these species and breeding programs to make the most use of limited resources. For example, how can maize breeding learn from rice trials, what lessons can cassava take from apple or cotton breeding, and how much of what is learned in Arabidopsis can we apply to bananas, blueberries, oats, or loblolly pine?

A detailed understanding of every gene’s function is probably not needed for Breeding 4.0. We do, however, need to be able to estimate the effect of each functional variant in a range of target environments. On a practical level, how do we get there? Although a species could have 50 million common variable sites in the genome, most of them probably do not matter very much. To reach Breeding 4.0, we need to reduce this sea of variation to a few tens of thousands of high-probability functional sites.

What types of data are the most useful for this filtering? We present some recommendations below:

1. Genetic mapping of complex target traits is by far the most expensive but also the most important approach to identifying functional variants. Complex traits such as yield rarely resolve to single genes, and trying to resolve any given QTL to a single nucleotide is rarely worth the cost. Instead, the central goal of mapping should be to resolve a few key QTLs to the gene level to highlight pathways and processes not previously considered. However, the greatest value of genome-wide association studies and genetic mapping may well be the data sets of genotypic and phenotypic data they produce. These can be used to benchmark every other approach, especially in the context of multiple environments, and can be integrated into ongoing breeding efforts.

2. Genomic annotation is the process of distilling our molecular understanding of the genome down to machine processable data. It includes annotations with protein domains, gene ontologies, methylation patterns, chromatin states, transcription factor binding sites, expression levels, and many others. Although all of these measures may be biologically interesting, there are very strong correlations among them, and many can be predicted from each other. The community needs to rigorously define the cost-effectiveness of each annotation to identify where to best put limited resources.

3. The mapping of intermediate phenotypes—for example, RNA transcripts or metabolite abundance—gives much higher resolution than the mapping of terminal phenotypes such
as yield and frequently identifies the exact gene involved or even specific causal variants. Related approaches such as chromatin profiling and proteomics could dramatically improve our ability to identify functional variants. The key challenge is to make these technologies cheap enough that crops can be profiled across a range of genotypes and environments.

4. Evolution is the ultimate yield trial, as it integrates the success of various genotypes over millions of years. The greatest limitation of current approaches is that sequencing distant evolutionary species only provides information on the most conserved elements. Saturation of closely related species is needed to fully understand regulatory conservation. Although crops are grown in environments where they did not evolve, the careful choice of landraces or wild analogs (as opposed to elite lines) may still be able to address these questions.

Each of these approaches can enrich for functional variants but not identify them for certain. Intelligent integration of these approaches, however, could provide whole sets of high-confidence variants in a cost-effective manner. This integration requires characterizing global germplasm resources both genetically and phenotypically, developinginformatics tools to share this information, and using appropriate methods (e.g., machine learning or similar) to integrate these diverse data sets.

A final key to Breeding 4.0 is large-scale genomic editing (tens to hundreds of sites per generation). Direct genome editing will then almost certainly replace crosses as the most efficient way to tailor genetic variation into optimal combinations. Better still, it can do so without linkage drag destabilizing the decades of breeding that went before. Of course, such edited crops would probably need to overcome consumer resistance to engineered foods, but that is an entirely different aspect to global food security.

Breeding has always been a numbers game, so we do not need to identify every variant with 100% accuracy. Even if we are only right 10% of the time, it may be enough to push crop breeding faster and more cost-effectively than we could otherwise. The hunt for these variants is already underway in many labs around the world. Integrating all this work into breeding pipelines will be key to providing an adequate, nutritious, and sustainable food supply for the entire globe throughout the twenty-first century.

SUMMARY POINTS

1. Breeding can be divided into four stages based on available technology; we are currently in Breeding 3.0.
2. To reach Breeding 4.0, we need to identify specific alleles that are responsible for desirable variation in crops.
3. Crops are still only partly adapted to agriculture and can be improved by breeding them to better fit modern growing environments.
4. Variation that is relevant to agriculture is not randomly spread through the genome, so finding the relevant portions can improve the focus of breeding efforts.
5. Much of Breeding 4.0 probably revolves around identifying and purging deleterious variants.
6. More effort is needed to democratize current and future breeding technologies so that they benefit all of global agriculture.
DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We wish to thank Sara Miller for help preparing the manuscript, and the United States Department of Agriculture-Agricultural Research Service, the University of Georgia, and DowDuPont for funding support that made the preparation of this article possible.

LITERATURE CITED

1. 3,000 Rice Genomes Project. 2014. The 3,000 rice genomes project. Gigascience 3:7


