

# Association mapping

## Speeding up discovery in plant genetics and breeding

by Madeline Fisher

Not long after joining the USDA-ARS in 1998, maize geneticist Ed Buckler set out to bridge a scientific divide

that helped shape his career ever since. On the one hand, he saw breeders tapping natural genetic diversity through linkage analysis, a technique that ties phenotypic traits to specific regions in the genome, allowing those regions to be targeted in breeding efforts. On the other hand, he knew molecular geneticists were busy cloning individual genes, including some from corn. But the link was missing. The genes isolated by geneticists had no connection to the natural variation breeders were working with. And linkage mapping, while useful for identifying broad genomic regions underlying a trait, couldn't pinpoint the actual genes involved. Or perhaps one such gene had been identified in maize by 1998, muses Buckler, who is based at Cornell University. "So, one gene out of 50,000," he says.

Today, many more genes have been tied to important agronomic traits in corn, thanks in large part to Buckler and a group of U.S. collaborators,

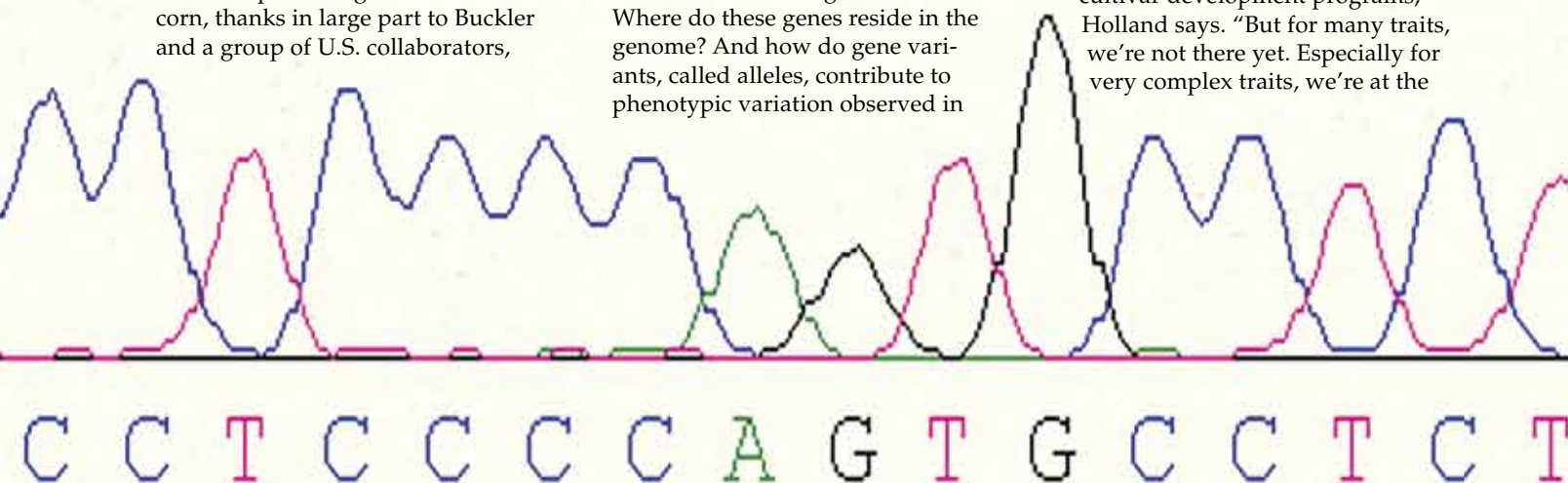
known as the Maize Diversity Project. Aided by advances in DNA sequencing technology and computational methods, the group has introduced powerful new techniques for finding genes that influence specific phenotypes, including association mapping—a complement to linkage analysis widely employed in human genetics—and nested association mapping—which unites the two former approaches. A related project also produced the first "haplotype" map in corn: a comprehensive catalog of genetic diversity in the maize germplasm pool.

What these resources and techniques now offer is a faster way to tackle fundamental questions in plant genetics and breeding: For example, do just a few genes with large effects influence leaf architecture, yield, and other "complex" traits, or are tens to hundreds of genes involved? Where do these genes reside in the genome? And how do gene variants, called alleles, contribute to phenotypic variation observed in

the field? Corn hasn't been the only crop to benefit either. Barley, alfalfa, rice, and several other crop plants have been the focus of association mapping analyses now. And an Australian team, led by David Jordan of the Queensland Alliance for Agriculture and Food Innovation, recently reported a nested association mapping study of sorghum in *Crop Science*.

A remaining question is how plant breeders will benefit. Even when they know little about the genetics of breeding populations, breeders can still select for many traits quite efficiently by evaluating phenotypes, says CSSA member Jim Holland, a close collaborator of Buckler's with USDA-ARS at North Carolina State University. So, will tracking genes speed the process, especially when many genes are involved?

"Yes, we do think this will be useful for breeding, and we can point to some traits where gene-based selection is already fully integrated into cultivar development programs," Holland says. "But for many traits, we're not there yet. Especially for very complex traits, we're at the



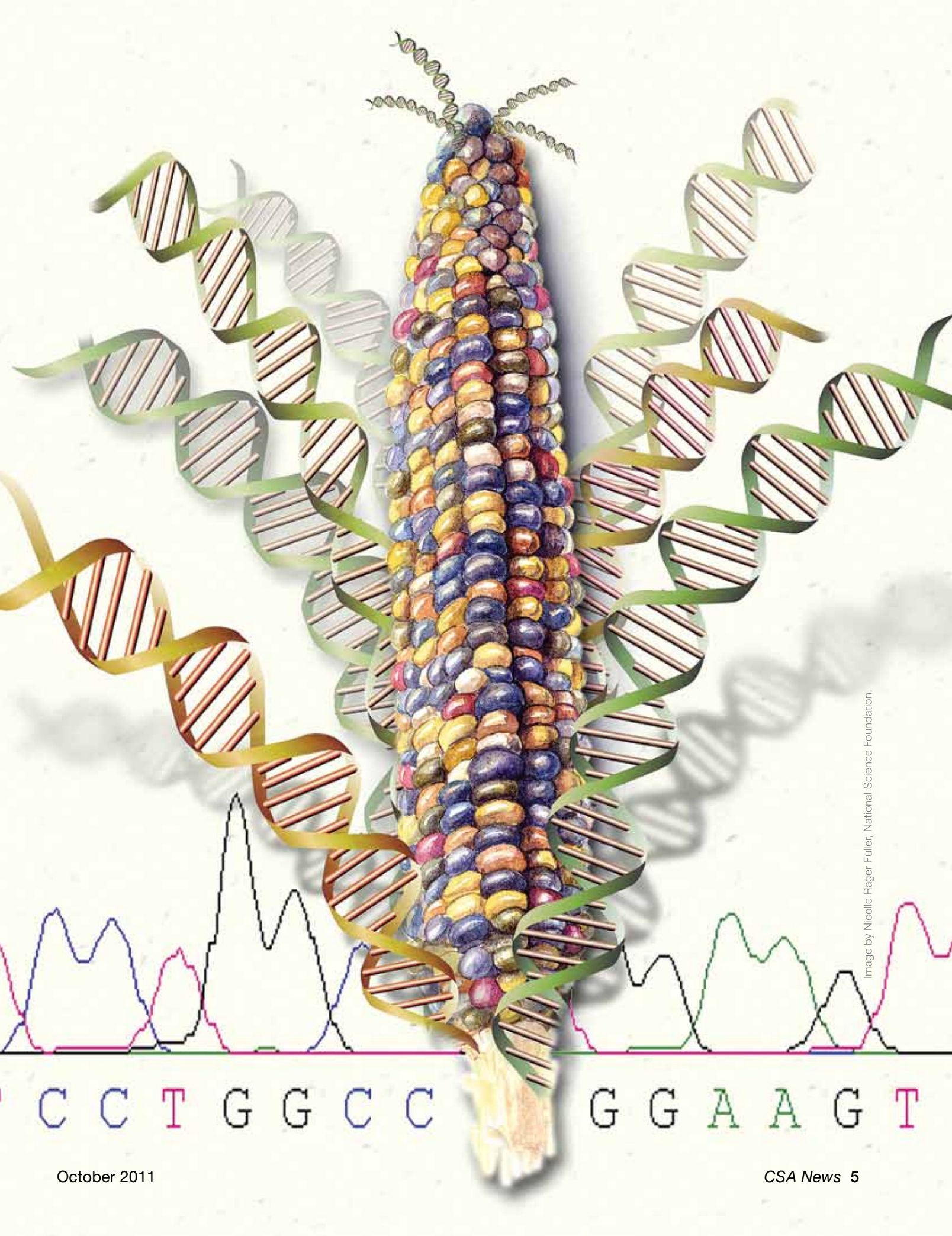


Image by Nicolle Rager Fuller, National Science Foundation.



## Basics of Genetic Mapping

Linkage analysis, or quantitative trait loci (QTL) analysis, was developed decades ago to link two types of information: measurements of traits, or phenotypes, and data on genetic identities, or genotypes. The goal is to identify stretches of DNA (known as QTLs) that underlie complex traits—those that vary continuously between individuals, are influenced by many genes, and interact with the environment. Linkage analysis begins with a cross between two parents with different phenotypes; for example, a short plant and a tall one.

Measuring the heights of the offspring easily yields a continuum of phenotypes from tall to short, but how are the genotypes assessed? Here geneticists make use of DNA sequence variations or “markers” that occur across the plant genome, such as single nucleotide polymorphisms (SNPs) or simple sequence repeats (SSRs). The key is that these sequence variations in a biparental population typically exist in two possible states, or alleles—one from the female parent and one from the male. For example, the female version (allele) of a SNP

might be “A” in the DNA sequence “...AAGGCTATT...” while the male allele is “T” in “...ATGGCTATT...”.

Alleles at, say, 100 SNP locations are then scored in all of the offspring from the cross, after which geneticists establish any “marker-trait associations.” To do so, they use statistics to identify instances where one allele of a marker, say the SNP “A” above, consistently correlates with tall plants, while the alternate form shows up in short ones.

When such a relationship is uncovered either for a single SNP or more than one, breeders then hypothesize that the marker (or markers) is linked to the trait plant height. And because they know where the SNP resides in the genome, they can map the trait associated with it to a specific genomic location, or QTL. Once all of this has been accomplished, breeders can then use the marker associated with the QTL to screen offspring quickly for a desired trait, rather than having to rely on subtle differences in phenotypes that can be difficult to discern.

stage now of basic studies that are setting up the framework for gene-based selection in the future.”

### Higher-Resolution Maps

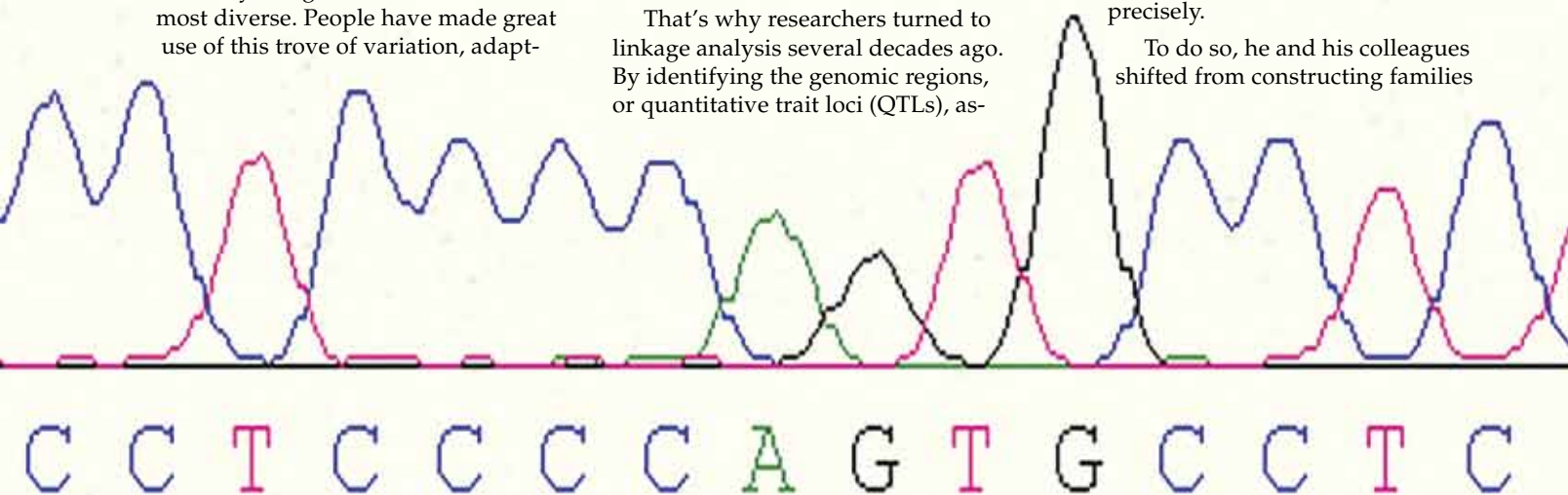
The traits of maize are manifestations of an astounding genetic versatility. On average, two maize lines are as genetically different as humans are from chimpanzees, and the crop plant is widely recognized as the world’s most diverse. People have made great use of this trove of variation, adapt-

ing maize to environments the world over. But the plant’s deep gene pool can also be overwhelming, especially when it comes to complex traits controlled by many genes. Scores of genes contribute to yield, for instance, each of which exerts just a small influence and interacts strongly with the environment. In cases like this, bringing all of the needed genes together in one plant is extremely challenging using conventional breeding methods.

That’s why researchers turned to linkage analysis several decades ago. By identifying the genomic regions, or quantitative trait loci (QTLs), as-

sociated with a trait, breeders could screen plants for genetic markers linked to QTLs rather than relying solely on phenotypic characteristics (see sidebar above). What has limited the method is the size of these regions: In general, QTLs span 10 to 20 million DNA base pairs—or up to a quarter of a chromosome in length—and contain hundreds of genes, Buckler says. When he joined USDA, he hoped to pin down QTLs much more precisely.

To do so, he and his colleagues shifted from constructing families



## Pinpointing Genes in a QTL

Association mapping builds on this technique by reducing the length of the trait-associated QTL regions that breeders target. That is, while linkage mapping typically identifies QTLs that include hundreds of genes, association mapping can pinpoint a QTL to a stretch of just a few genes, or even a single gene, by exploiting historical recombination and wider genetic diversity. The nested association mapping strategy developed by the Maize Diversity Project integrates the advantages of both linkage analysis and association mapping.

A powerful genetic resource for both linkage analysis and association mapping is a haplotype map, which scientists create by identifying SNPs (or other markers) that are inherited together in blocks known as haplotypes. The unique map, or pattern, of haplotypes in each individual can then be compared to the patterns of others, offering a picture of nearly all the genetic diversity present in a sample (i.e., how much the individuals differ genetically from each other), as well as a catalog of the genetic recombination events that have occurred in the sample's history.

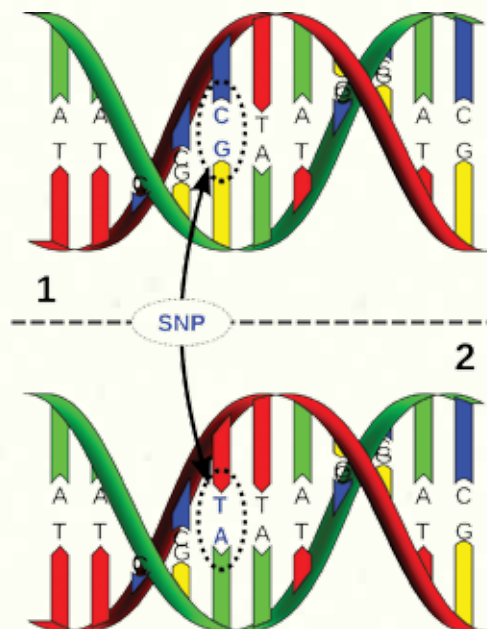


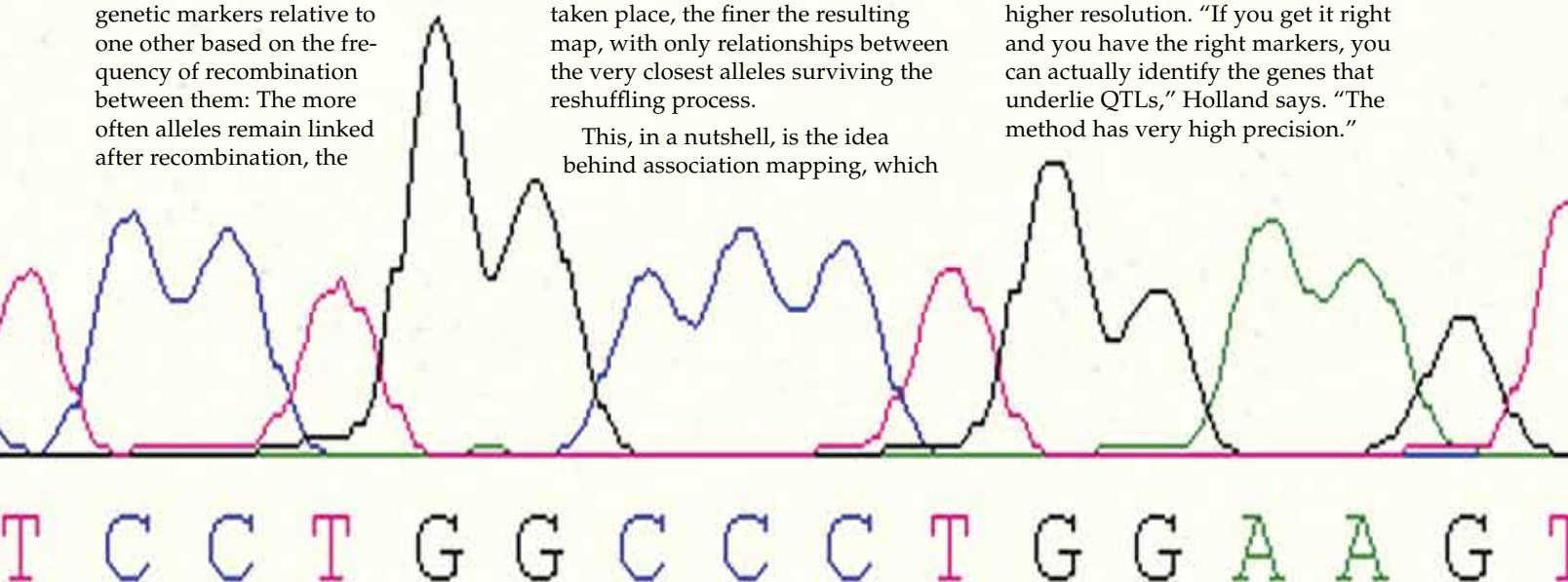
Illustration of a single nucleotide polymorphism where DNA molecule 1 differs from DNA molecule 2 at a single base-pair location. Credit: David Hall/ Wikipedia.

by crossing two parent plants—the first step in linkage analysis—to assembling populations of diverse individuals. The reason has to do with genetic recombination: the pairing of DNA strands (chromosomes) and swapping of corresponding DNA segments that occurs each generation. Researchers map genes or genetic markers relative to one other based on the frequency of recombination between them: The more often alleles remain linked after recombination, the

closer they sit on a chromosome; the more frequently they separate, the further apart they are. If recombination hasn't been given much chance to occur, however—such as in a two-parent cross—recombination frequencies overall will be low, limiting the resolution of the map. Conversely, the more recombination that has taken place, the finer the resulting map, with only relationships between the very closest alleles surviving the reshuffling process.

This, in a nutshell, is the idea behind association mapping, which

Buckler's group first reported in *Nature Genetics* in 2001 and reviewed in *The Plant Genome* in 2007. By analyzing diverse populations rather than the progeny of a single cross, the technique not only exploits a wider range of natural variation, but also a vast number of historical recombination events. The outcome is a map of much higher resolution. "If you get it right and you have the right markers, you can actually identify the genes that underlie QTLs," Holland says. "The method has very high precision."





Plant geneticist Edward S. Buckler uses high-throughput robotics to efficiently sample the DNA variation of thousands of genes in maize. *Photo by Peggy Greb (USDA-ARS).*

But the scientists soon realized it also had drawbacks, particularly a tendency to find false positives. Sometimes populations showing a desired trait also carry a specific gene variant not because the variant actually controls the trait, but due to genetic relatedness. This generally isn't a problem in linkage analysis because researchers know the genetic structure of the family they created. But in association mapping, where relationships between diverse populations aren't necessarily well understood, marker-trait associations arising from kinship

and evolutionary history can easily be mistaken for causal ones.

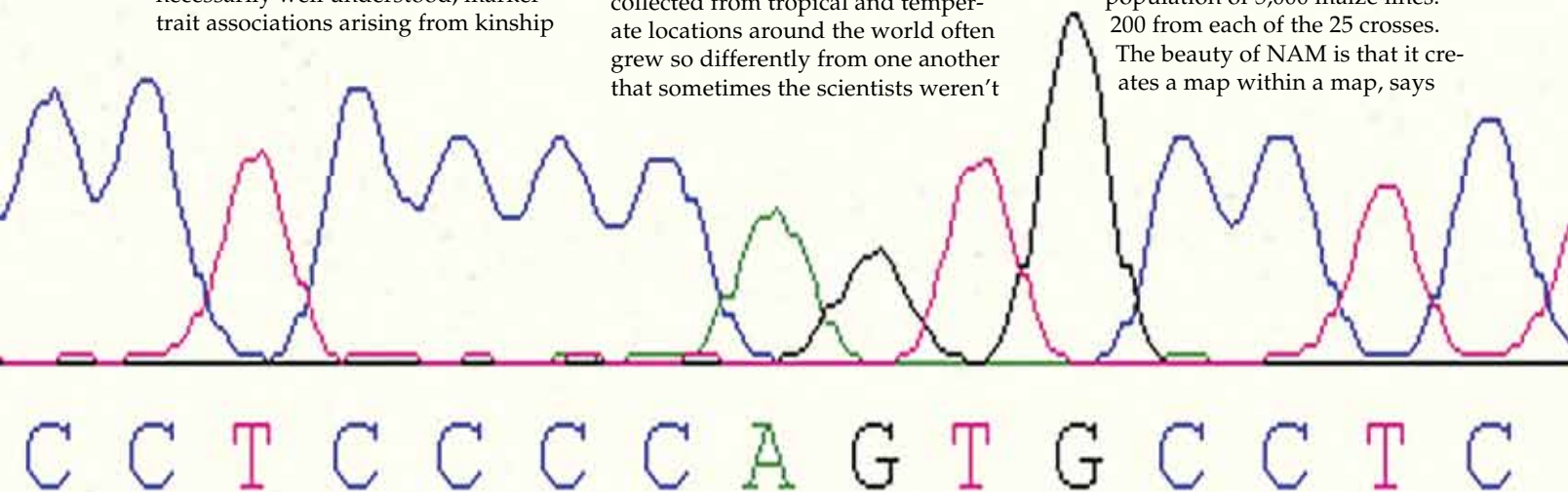
In 2006, Buckler and his former postdoc, ASA and CSSA member Jianming Yu, now at Kansas State University, published a "unified mixed model" method in *Nature Genetics* that dropped this false association rate dramatically. But as he walked among the corn plants in his research plots near Ithaca, NY, Buckler soon noticed another weakness of association mapping. Cultivars and land races collected from tropical and temperate locations around the world often grew so differently from one another that sometimes the scientists weren't

sure they were even scoring the same trait. Comparing yield, for example, in plants that flower three weeks apart is tough. "Sometimes we have too much diversity," Buckler laughs.

## Nested Association Mapping

After contemplating these agronomic challenges for awhile and talking with each other, the scientists then had an idea. Working with diverse populations in association mapping let researchers exploit more diversity and produce detailed maps, while the controlled crosses of linkage analysis eliminated false positives and produced homogenized populations whose phenotypes were easier to compare. So, they reasoned, why not blend the two approaches? Before long, Buckler, Holland, and their other main collaborator, USDA-ARS geneticist Mike McMullen, devised a plan to combine the techniques by crossing each of 25 maize lines gathered from around the world and the elite commercial cultivar B73.

The ambitious project would require a massive data-crunching and field effort, however, and Holland initially thought it would take more resources than a group of publicly funded researchers could pull together. Nevertheless, in a 2009 *Science* paper, the group introduced the technique, nested association mapping (NAM), along with their NAM population of 5,000 maize lines: 200 from each of the 25 crosses. The beauty of NAM is that it creates a map within a map, says





“Using association mapping, you can also take advantage of the ancient recombination to map at high-resolution ... sometimes down to the actual gene.”

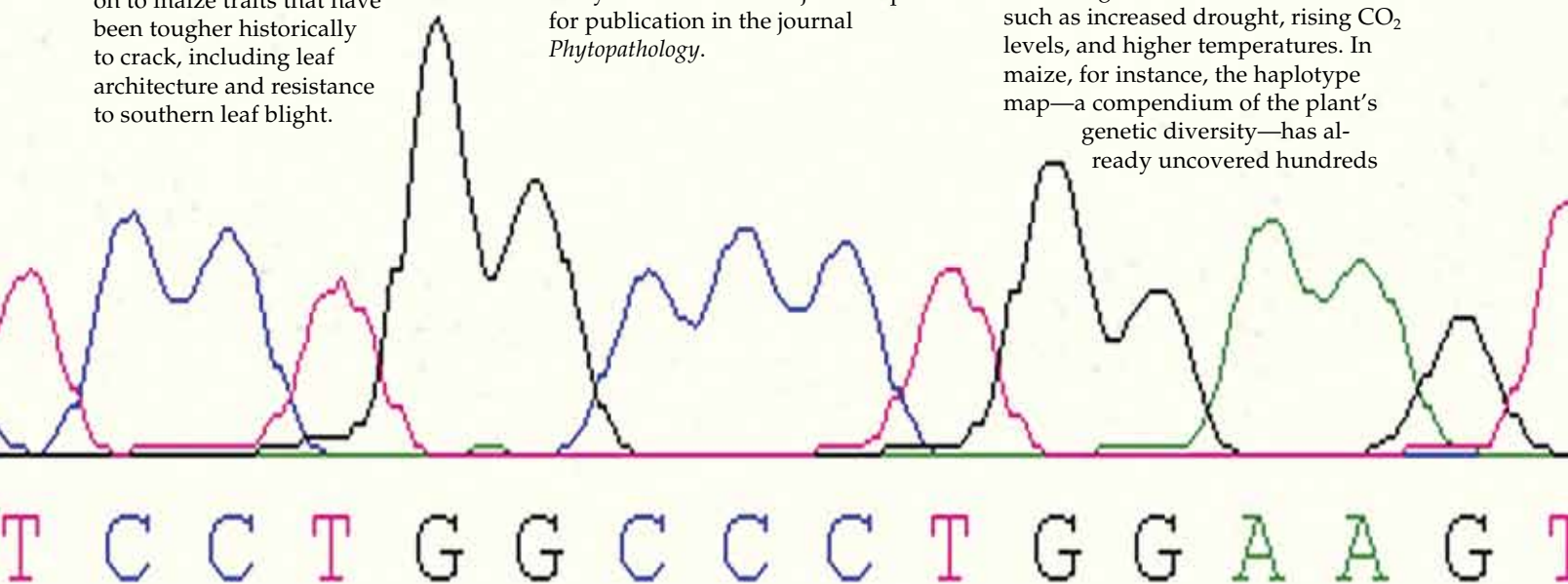
ASA, CSSA, and SSSA member Mike Gore, a former graduate student of Buckler's, now with the USDA-ARS in Maricopa, AZ. Because researchers know the parents of each population (B73 and one of the 25 lines), they can easily spot the recent recombination events in each population's history, letting them identify which chromosomal regions are controlling complex traits with high accuracy. “Then using association mapping, you can also take advantage of the ancient recombination to map at high-resolution within those larger regions you identified,” Gore says, “sometimes down to the actual gene.”

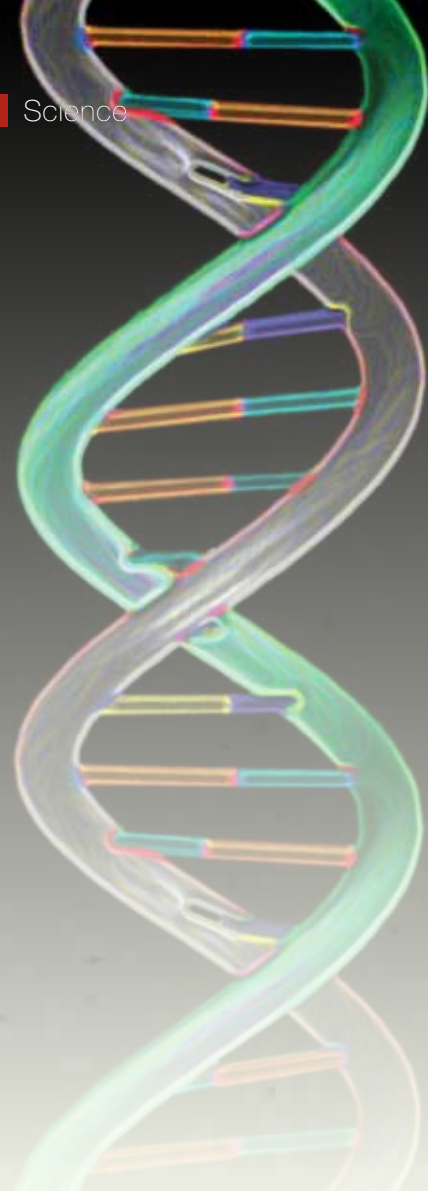
In a companion study, also published in *Science* in 2009, the team tested NAM's ability to find QTLs and genes by applying it to flowering time—a trait whose genetics are already fairly well understood and phenotype is easy to measure, Buckler says. They've since moved on to maize traits that have been tougher historically to crack, including leaf architecture and resistance to southern leaf blight.

That's one of the great benefits of association mapping and NAM, adds Yu, an ASA and CSSA member. Assembling a diverse sample of lines, genotyping them, and analyzing their genetic backgrounds takes time, of course. But once this framework is in place, “you can use the same population for a lot of different purposes,” he says, so long as it contains genetic variation for the trait in question. A few years ago, for example, he helped his USDA-ARS colleague, ASA and CSSA member Guihua Bai, use the technique to identify genes underlying resistance to wheat head blight in roughly 350 winter wheat lines that Bai had collected from the USDA winter wheat performance trials. As they were carrying out that study, however, the lines also became unexpectedly infected with soil-borne mosaic virus. So, Bai and Yu scored the plants for resistance to this disease, too, and performed a second association mapping analysis. The work was just accepted for publication in the journal *Phytopathology*.

### Flexibility Could Help Breeders Adapt Crops to Climate Change

It's because of this flexibility, say the scientists, that the methods could also help breeders adapt crop plants to shifting environmental conditions, such as increased drought, rising CO<sub>2</sub> levels, and higher temperatures. In maize, for instance, the haplotype map—a compendium of the plant's genetic diversity—has already uncovered hundreds





of genomic regions that vary between temperate and tropical varieties, says Gore, who led the haplotype map project as Buckler's student. Some of these regions undoubtedly include genes that promote growth under hot, arid conditions. And if those genes can be identified, breeders can then

"I want to push association mapping in plants so it's at the same level as in human and animal genetics. But our share of funding can't compare with human genetics. All we can do right now are little pieces."

use marker-assisted selection to target them, while leaving behind less desirable variation.

Still, it won't be easy. More than 30 QTLs contribute to southern leaf blight resistance alone, Holland says, and the same is likely true for many other complex traits, at least in corn. "So how you use this in breeding is somewhat more complicated," he says. "It's not as simple as just having one gene to worry about."

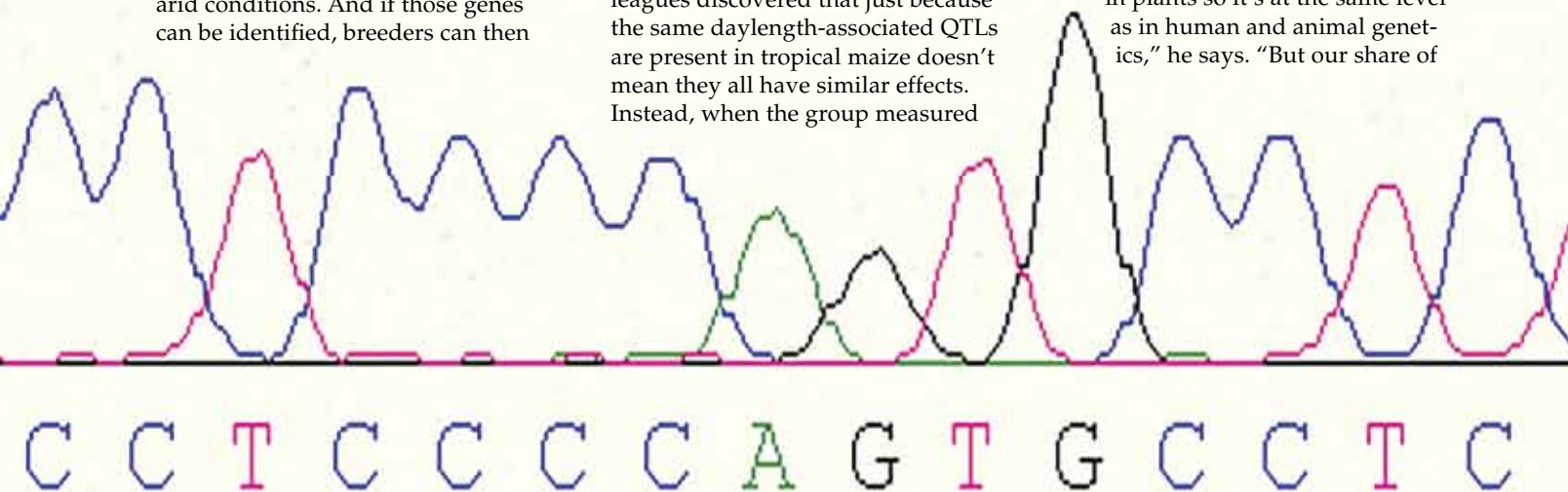
There can be other complications, as well, as Holland learned during a study of four QTLs that underlie the daylength response in maize. Tropical maize from nearer to the equator is extremely sensitive to daylength, and it responds to longer summer days in the U.S. Corn Belt by flowering extremely late. Breeders would like to know the QTLs involved in order to select against them when working with tropical germplasm. But in work published in the May-June 2011 issue of *Crop Science*, Holland and his colleagues discovered that just because the same daylength-associated QTLs are present in tropical maize doesn't mean they all have similar effects. Instead, when the group measured

the phenotypic effects of these QTLs in maize lines bred in Mexico and Thailand, they found that each one had different impacts on flowering time depending on the variety it came from. In one case, in fact, a tropical QTL actually caused plants to flower *earlier*, rather than later, as predicted.

The unexpected findings argue again for the need to home in on genes. "This project was geared at trying to make us better at incorporating tropical germplasm," Holland says. "But it's really just showing us that we don't know enough yet."

### Plenty of Work Ahead

Indeed, there seems to be plenty left to do. For his part, Yu would like to see many more large-scale studies: ones in which thousands to millions of genetic markers are scored in hundreds to thousands of individuals. Research dollars are tight, however. "I want to push association mapping in plants so it's at the same level as in human and animal genetics," he says. "But our share of





funding can't compare with human genetics. All we can do right now are little pieces."

At the same time, as the price of DNA sequencing and other genomics technology continues to drop, phenotyping rather than genotyping is becoming the cost-limiting factor, Gore says, especially as plant scientists push to analyze ever-larger populations. That's why as he works in Arizona to improve the heat- and drought-tolerance of Southwest-grown cotton, he's pursuing two tacks. One is to develop a NAM population for cotton, and possibly a haplotype map, as well. The other is to create a high-throughput phenotyping method in collaboration with Pedro Andrade-Sanchez at the University of Arizona. Their invention consists of infrared thermometers mounted on a tractor, which read the canopy temperature in cotton—a reflection of its heat tolerance—as the tractor moves through the field. Since the tractor and plots are also on GPS coordinates, the setup immediately yields a continuum of heat tolerance data for hundreds of cotton cultivars, which Gore can then relate to his genetic maps with statistics.

And where does Buckler see the field going? Not surprisingly, he'd like to see even more integration. Most traits that have been examined so far don't vary much with environmental conditions, he says; the next step will be to examine those that do, like yield.



In Arizona, Michael Gore, Pedro Andrade-Sanchez, and John Huen are working on tractor-based proximal remote sensing. This high-throughput phenotyping system is being used to collect plant height, canopy temperature, and canopy reflectance data from cotton plants. Phenotyping is still very expensive so the group hopes that strategies such as this will lower the costs. *Photo courtesy of Michael Gore.*

What this will involve is identifying all the variants of all the genes that control a trait; running trials at, say, 100 locations around the globe; and then asking how variance in the genes interacts with different conditions to control phenotypes. "There won't be one optimal genetic allele," he predicts. "There will be genetic variants that will really need to be optimized for every environmental zone." The Maize Diversity Group has already started collaborating with researchers in China, Mexico, and Africa on the project, he adds, and it should

be producing results for breeders and geneticists within a decade.

If it sounds too big, like it can't be done, recall that Holland felt the same way when the team first came up with the idea for NAM. "That's why Ed is great because he'll say, 'We can do it!'" Holland laughs. "And then we do."

*M. Fisher, lead writer for CSA News magazine*

