The Plant Cell, Vol. 16, 787–788, April 2004, www.plantcell.org © 2004 American Society of Plant Biologists

OPINION

Sequencing Maize: Just Sample the Salsa or Go for the Whole Enchilada?

With the complete sequence of the rice nuclear genome nearing completion, plant biologists have begun to wonder, discuss, and debate what genome is our next, best target? Although it is easy to list many worthwhile candidates among the angiosperms, the clear favorite for the next *complete* genome sequence should be maize. Not only does maize genetics offer a rich and varied history as the most esteemed model genetic system in the Kingdom Plantae, but more importantly, the maize genome offers an unparalleled opportunity to open new vistas on the complex organization, evolution, and dynamic behavior of what is arguably the most interesting and important genome whose complete sequence is within reach.

Why maize? And why should we desire a complete genome sequence? Sequencing the ''gene-space'' of maize by filtration methods that capture DNA sequences that exhibit undermethylation or slow reassociation kinetics is already providing much useful information that is stimulating plant biology immensely, and there is broad agreement that the resources committed to this effort have been an extremely worthwhile investment. Nonetheless, filtration sequencing does have two acknowledged limitations: (1) it will never provide the complete complement of genes encoded by the maize genome, and (2) it will leave us completely ignorant of some 90% of the genome and its unknown functions.

''Unusual'' genes—those not captured by these methods, for whatever reason, predictable or not—will remain undiscovered and unknown. The existence of genes with unusual properties, as well as unusual components of typical genes that control expression, is a known fact. A diverse class of unusual genes likely to be underrepresented in a filtered genome sequence comprises the so-called ''noncoding'' RNA genes, of which microRNA-encoding genes are just one example. Because noncoding RNA genes are typically small, they are poor targets for mutagenesis, which explains why so few of them have been identified to date through forward genetics. They also tend to be found in so-called ''intergenic'' regions, calling into question whether they will be reliably retrieved by filtration methods.

Most importantly, we cannot even begin to predict what might lie in the vast, putatively ''heterochromatic'' seas of repetitive sequences derived from transposable elements that comprise the bulk of the maize genome. Long distance (100 kb) effects on gene expression have been demonstrated in maize, but almost nothing is known about how complex arrays of multicopy sequences within which genes are interspersed contribute to the control of gene expression or to the generation of novel genetic diversity affecting this essential regulatory function. Repetitive sequences are distributed very nonrandomly in plant genomes, with different elements showing strong or total propensities for specific types of locations. The sequences and mechanisms that underlie this variation are essential to discovering their roles and functions in genome evolution, chromosome behavior, and the generation of novel genetic variation. Nothing less than a proper understanding in plants of the pivotal evolutionary processes that create and mold this variation is at stake here.

The only large, complex genome that has been sequenced to (near) completion is the human genome. What lessons does it offer that might be relevant to sequencing maize? An important one is that, like maize, the human genome abounds with segmental duplications, many of which went unrecognized in the original draft sequence assemblies. Human geneticists have begun to recognize that these duplications are not only numerous, but that they can also be specific to individual primate lineages and, most importantly, they have been fixed at extremely high rates in the human lineage, and so are obvious sources of the genetic variation that potentially contributed to the rapid morphological and functional evolution in this lineage.

Can the human genome, given that it is about the same size as the maize genome, tell us most of what we would want to know about large complex genomes? This seems unlikely because of vast differences in composition, organization, and evolution of the maize and human genomes, as well as in the distinct nature of genetic variation that exists within each species. More than 70% of the maize genome derives from long terminal repeat-retrotransposable elements, whereas the human genome possesses only three intact copies of such elements and there has been no transposition of these elements since divergence from chimpanzees 7 million years ago (Mya). The maize genome expanded rapidly in a huge burst of retrotranspositions, mostly within the past 3 million years, whereas the bulk of the human genome arose some 50 Mya, and then grew only gradually, mostly via segmental duplications. The maize genome possesses many duplicated genes (often in families of three or more) and these have not one, but two origins: a complete genome duplication due to an allotetraploidy event ~15 Mya and segmental duplications, both in tandem and via translocations. Perhaps most striking is the phenomenon of ''segmental aneuploidy'' in maize: haplotype variation of an extreme nature that is characterized by the complete absence of many genes in their expected locations. This observation is in sync with the likelihood that some lineages in the species *Zea mays* diverged more than 1 Mya, while long terminal repeat retrotranspositions have continued to occur on a massive scale. Thus, maize presents both a more complex gene duplication history and a more complex

OPINION

sequence environment than does the human genome.

There can be no argument over whether it will be sufficient to know just the bulk of the coding sequences and their immediately adjacent control signals. This is clearly not enough. We need to know the entire *neighborhood* of a gene in full detail; we need to know whether different genes exist in different *types* of neighborhoods that could influence not only their function but also their evolutionary potential; and we need to know all this not merely in the context of a single individual or genotype, but in the context of the *species*. The plant that evolved in a short 10,000 years from an obscure grass to the most important food crop in the Americas, that first offered the genetic ''oddities'' of transposons and paramutation that are now seen as central and fundamental properties of genomes, and that held the attention of world-class geneticists such as Barbara McClintock and Marcus Rhoades for entire lifetimes, is not likely to keep its secrets in plain view. The only way to fully understand (1) the nature of this astonishing complexity, (2) what aspects of it were responsible for the origin of the species *Zea mays*, and then its domestication and further improvement to produce the crop plant we now know as maize (*Z. mays* ssp *mays*), and (3) how the nature and properties of the genome might contribute to the future evolution and improvement of maize, will be to produce a high quality, reasonably *complete* genome sequence of an inbred maize line that can serve as a solid point of reference for comparison with other maize genotypes, as well as related species and genera.

Just as Linnaeus' historic achievement of classifying macroorganisms could never have anticipated the existence of an order of magnitude greater biotic diversity in the unseen world of microbes, the historic sequencing of the Arabidopsis genome served merely to intrigue us about the nature and function of the vastly greater complexity and diversity of more typical plant genomes such as that of maize. While the Arabidopsis and rice genomes unquestionably offer unimaginable riches that will require decades of exploration to fully appreciate, no amount of investigation of these can predict the nature and dynamical evolutionary behavior of the vastly different maize genome. Together, complete maize, rice, and Arabidopsis genome sequences will synergistically advance understanding of plant genomes an order of magnitude beyond that which could be gained from studies based on a single complete genome sequence. Considering that most flowering plant genomes are much more like maize than Arabidopsis or rice, a complete sequence of the maize genome will surely contribute disproportionately to better understanding of these genomes and their genetic processes than could Arabidopsis and rice, even together.

The simple fact that maize plays a huge role in feeding the world is perhaps justification enough for sequencing its genome to completion, but the implications of a complete maize sequence for the whole of plant biology provide the strongest justification. Can we really afford to bet that what we will miss by sequencing only the filtered DNA-space of maize will not be just as important and valuable, if not more so? Having already sampled an exciting variety of the spicy offerings available in the gene-space of maize, our appetites have merely been whetted—it's time to go for the whole enchilada!

Richard A. Jorgensen
Department of Plant Sciences
University of Arizona
Tucson, AZ 85721-0036
raj@ag.arizona.edu