In particular, in systems such as these, where the normal balance between multiple mechanisms is an important factor, the ability to assess the roles of individual components in a physiological setting will offer important insights into how cell migrations are controlled.

Acknowledgements

Special thanks to L. Cooley, D. Montell and members of our laboratory for comments on the manuscript. Work in our laboratory is supported by grants from the NIH, the ACS, and the Lucille P. Markey Charitable Trust.

References

1 Keynes, R. and Cook, G.M.W. (1995) *Cell* 83, 161–169
2 Massagué, J. and Pandiella, A. (1993) *Annu. Rev. Biochem.* 62, 515–541
3 Hynes, R.O. (1996) *Dev. Biol.* 180, 402–412
4 Sulston, J.E. and Horvitz, H.R. (1977) *Dev. Biol.* 56, 110–156
5 Hunter, C.P. and Wood, W.B. (1990) *Cell* 63, 1193–1204
6 Kenyon, C. (1986) *Cell* 46, 477–487
7 Thomas, J.H., Stern, M.J. and Horvitz, H.R. (1990) *Cell* 62, 1041–1052
8 Trent, C., Tsung, N. and Horvitz, H.R. (1983) *Genetics* 104, 619–647
9 Stern, M.J. and Horvitz, H.R. (1991) *Development* 113, 797–803
10 Burdine, R.D., Chen, E.B., Kwok, S.F. and Stern, M.J. (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94, 2433–2437
11 DeVore, D.L., Horvitz, H.R. and Stern, M.J. (1995) *Cell* 83, 611–620
12 Terranova, V.P. *et al.* (1985) *J. Cell Biol.* 101, 2330–2334
13 Sutherland, D., Samakovlis, C. and Krasnow, M.A. (1996) *Cell* 87, 1091–1101
14 Burdine, R.D., Branda, C.S. and Stern, M.J. (1998) *Development* 125, 1083–1093
15 Kornfeld, K. (1997) *Trends Genet.* 13, 55–61
16 Clark, S.G., Stern, M.J. and Horvitz, H.R. (1992) *Nature* 356, 340–344
17 van der Geer, P. and Hunter, T. (1994) *Annu. Rev. Cell Biol.* 10, 251–337
18 Sundaram, M., Yochem, J. and Han, M. (1996) *Development* 122, 2823–2833
19 Chen, E.B., Branda, C.S. and Stern, M.J. (1997) *Dev. Biol.* 182, 88–100
20 Lee, T., Feig, L. and Montell, D.J. (1996) *Development* 122, 409–418
21 Hedgecock, E.M., Culotti, J.G., Hall, D.H. and Stern, B.D. (1987) *Development* 100, 365–382
22 Culotti, J.G. (1994) *Curr. Biol.* 5, 587–595
23 Steven, R. *et al.* (1998) *Cell* 92, 785–795
24 Zheng, Y., Cerione, R. and Bender, A. (1994) *J. Biol. Chem.* 269, 2369–2372
25 Nobes, C.D. and Hall, A. (1995) *Cell* 81, 53–62
26 Ridley, A.J. (1995) *Curr. Opin. Genet. Dev.* 5, 24–30
27 Zipkin, I.D., Kindt, R.M. and Kenyon, C.J. (1997) *Cell* 90, 883–894
28 Kimble, J. (1981) *Dev. Biol.* 87, 286–300
29 Sulston, J.E., Schierenberg, E., White, J.G. and Thomson, J.N. (1983) *Dev. Biol.* 100, 64–119

E.B. Chen *is at the Center for Molecular Medicine, Emory University, 1462 Clifton Rd, Decatur, GA 30322, USA.*
M.J. Stern *is at the Department of Genetics, Yale University, PO Box 208005, New Haven, CT 06520-8005, USA.*

Of genes and genomes and the origin of maize

SHAWN WHITE (white090@tc.umn.edu)

JOHN DOEBLEY (doebley001@tc.umn.edu)

The crop plant maize (corn) is remarkably dissimilar to its recent wild ancestor, teosinte, making it an extremely interesting model for the study of evolution. Investigations into the evolution of maize are currently being performed at the molecular and morphological levels. Three independent lines of research are poised to shed light on the molecular basis of this spectacular transformation: (1) determining the structure and origin of the maize genome; (2) understanding the role of transposable elements in maize evolution; and (3) elucidating the genetic basis for morphological differences between maize and its wild ancestor teosinte.

Maize exhibits an extreme morphological divergence from its apparent wild progenitor, teosinte (*Zea mays* ssp. *parviglumis*), despite the fact that the domestication of maize from teosinte occurred only about 7000 years ago, a mere eye-blink on the evolutionary timescale. To add to this paradox, maize and teosinte are completely interfertile, and show no greater divergence in their chromosomes, gene structures, or nucleotide sequences than one is apt to observe between two varieties of maize[1,2]. Discrepancies of this nature are not an uncommon observation, and perhaps the best known example involves our own species. Humans and chimpanzees show a phenomenal departure in morphological and behavioral traits, although the human and chimp genomes are remarkably similar[3].

How can morphological evolution race forward while the genome lags behind? Over two decades ago, King and Wilson[3] proposed that a small number of mutations affecting gene regulation could account for major differences between humans and chimps. Even earlier, Britten and Davidson[4] had postulated that alterations in the patterns of gene regulation were more important in evolution than changes in protein function, and that these alterations were brought about by the expansion and dispersion of repetitive DNA sequences. Over 20 years after these seminal papers, the relationship between molecular events and morphological evolution is still poorly understood. In this paper we review how the convergence of research on three fronts – the evolution of the maize genome, the evolution of its genes, and the evolution of its morphology – is beginning to help uncover this relationship.

Copyright © 1998 Elsevier Science Ltd. All rights reserved. 0168-9525/98/$19.00

PII: S0168-9525(98)01524-8

Evolution and structure of the maize genome

It has been known for some time that the maize genome is composed largely of highly repeated sequences and that the genes themselves exist as small islands in this sea of repetitive DNA (Refs 5, 6). However, much about the nature and structure of the repetitive sequences has remained obscure until recently. In a series of papers, Jeff Bennetzen's group (Purdue Univ.) and Sue Wessler's group (Univ. of Georgia) have shown that the sea of repeats surrounding maize genes is largely composed of transposable elements. In one set of experiments, reverse-Southern analysis (in which labeled genomic DNA is used as a probe in hybridizations with cloned DNA fragments immobilized on a membrane) of a large DNA segment surrounding the *Adh1-F* locus on chromosome 1 of maize revealed that the DNA flanking this gene is composed of uninterrupted stretches of repetitive sequences that compose approximately 82% of the 280 kb region[7]. Characterization of the individual repeats indicates that they are often reiterated more than once in the 280 kb region, are present elsewhere in the maize genome, and represent several different repeat elements. A similar analysis of the flanking regions of ten other maize clones revealed that they are also composed of clusters of diverse repetitive elements distributed throughout the maize genome in copy numbers ranging from 600 to 54 000 per haploid genome[8].

The sequencing of some of the repeats in the *Adh1-F* region has revealed that they consist of the long terminal repeat (LTR) retrotransposon class of transposable element[9]. Twenty retrotransposons (falling into ten families) were found to comprise 150 kb (62%) of the 240 kb analysed. Most of these elements appear to be intact, although some individual elements are interrupted by insertions of other retrotransposons or solo LTRs, and some incomplete, defective elements were identified. The copy numbers of the elements range from 10 to 30 000 and, together, are estimated to make up about a quarter of the maize genome. Because the repeat DNA found in the *Adh1* region is probably indicative of that surrounding other maize genes, it appears that the general structure of the gene-containing regions of the maize genome is one of single-copy genes separated from one another by (~30 kb) tracts of repetitive DNA composed primarily of LTR retrotransposons (Fig. 1).

Another approach to studying the flanking regions of plant genes has led to the discovery of a variety of transposable elements near maize genes. Known transposable element sequences were used as probes in 'electronic Southerns' (nucleic- and amino-acid-level searches of GenBank and EMBL databases), revealing Tourist (Refs 10, 11) and Stowaway (Ref. 12) inverted-repeat elements, LTR retrotransposons[13], LINEs (S. White, unpublished), and SINEs (T. Bureau, pers. commun.) in the flanking sequences and introns of maize genes. However, in contrast to the LTR retrotransposons in flanking regions of the *Adh1* gene, most of the elements found via database searches are members of the small (80–343 bp) Tourist and Stowaway element families (also known as 'MITEs' – miniature inverted repeat transposons).

To date, over 110 of approximately 376 genomic maize gene sequences have been found to harbor at least one transposable element (Ref. 13; S. White, unpublished; and Thomas Bureau, pers. commun.). Many elements are found less than 1 kb from the start site of transcription. The position of transposable elements so close to genes suggests that they might be involved somehow as *cis* regulator elements and that their insertion or deletion can drive the evolution of gene expression. Some cases of maize genes whose putative regulatory regions have been supplied by transposable elements have been described[13]. Transposable elements make a particularly attractive agent of evolution because their movement can be triggered by environmental or genomic stresses, such as UV light[14,15], microbial or viral infection[16,17] and interspecies hybridization[18].

That transposable elements have been involved in the expansion of the maize genome is indicated by comparisons of the nuclear genomes of maize and sorghum, its relative in the *Andropogoneae* tribe of grasses. The maize genome is approximately three times larger than that of sorghum varieties with the same chromosome number (n = 10)[19]. Restriction-fragment-length polymorphism (RFLP) mapping of these two grass genomes has shown that the content and order of their genes are very similar[20,21]. Analysis of intergenic regions surrounding the *Adh1* gene in maize and the orthologous locus in sorghum show that the intergenic regions of sorghum are smaller than the element-rich intergenic regions of maize[22], suggesting that transposable element insertions are responsible for the increase in the length of maize intergenic regions.

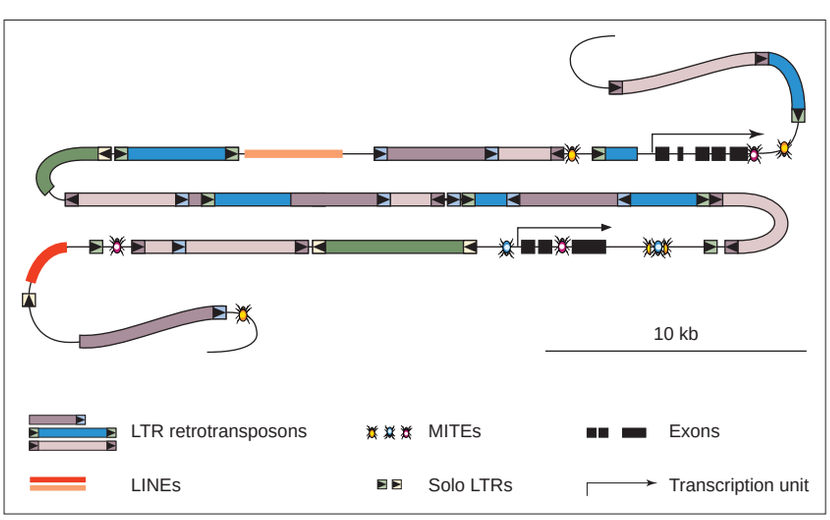

FIGURE 1. The structure of the genic regions of the maize genome as composed of expressed genes in a sea of LTR retrotransposons with smaller miniature inverted repeat elements (MITEs) located within and adjacent to functional genes. Different retrotransposon and MITE families are represented by different colors.

Did the maize genome originate from an ancient polyploid?

Beyond transposable element activity, polyploidy is another mechanism likely to have been involved in the shaping of the maize genome. Maize has long been thought to be an ancient tetraploid whose genome has reverted over time to functional diploidy and thus lacks two clear sets of duplicated chromosomes. Part of the evidence for this interpretation is that while maize has ten gametic chromosomes, several members of the *Andropogoneae* tribe to which maize belongs have only five. In addition, maize has two unlinked copies of many genes, each of which is found on duplicated chromosomal segments[23].

Recently, Gaut and Doebley[24] revisited this issue, applying a molecular evolutionary analysis to duplicated sequences in the maize genome. They attempted to distinguish among three possible modes of polyploidy, each of which would predict a distinct pattern of divergence between duplicated sequences. Genomic allotetraploids, such as wheat or cotton, arise via the fusion of the genomes of two distinct species. Because the two genomes are distinct, their chromosomes do not pair and evolve separately within the cells of the tetraploid. The divergence times for the duplicated genes of a genomic allotetraploid would be equivalent to the divergence time of its two diploid ancestors. Autotetraploids, such as potato (*Solanum tuberosum*), arise from the doubling of the chromosome number within a species. Because the two chromosome sets are not distinct and can pair at meiosis, the plants will exhibit tetrasomic inheritance and the duplicated loci will fail to diverge. If maize arose by this means, it must have become diploidized over time, switching from tetrasomic to disomic inheritance. The divergence times of the duplicated sequences would represent the time of the switch from tetrasomic to disomic inheritance. Finally, segmental allotetraploids arise from the hybridization of species whose genomes are only partially distinct, so that they exhibit a mixture of disomic and tetrasomic inheritance. Accordingly, the divergence times of their duplicated genes will form two groups: one representing the time of divergence of the diploid ancestors (for genes whose initial inheritance pattern was disomic) and the other the time of the switch to disomic inheritance (for genes whose initial inheritance pattern was tetrasomic).

Gaut and Doebley compared the divergence times for 14 pairs of duplicate maize loci, revealing two statistically different groups. This result is consistent with a segmental allotetraploid origin but not with either genomic allotetraploidy or autotetraploidy. By applying a molecular clock to the sequence data, they approximated when the two ancestral diploids diverged and when the polyploidy event occurred. The divergence of the parental diploids was estimated to have occurred 20.5 million years ago, and the time of polyploid formation to have occurred at least 11.4 million years ago (Fig. 2). The former date is earlier than the estimated time of divergence between maize and sorghum, indicating that the sorghum genome is more closely related to one of the maize subgenomes than the two maize subgenomes are to each other.

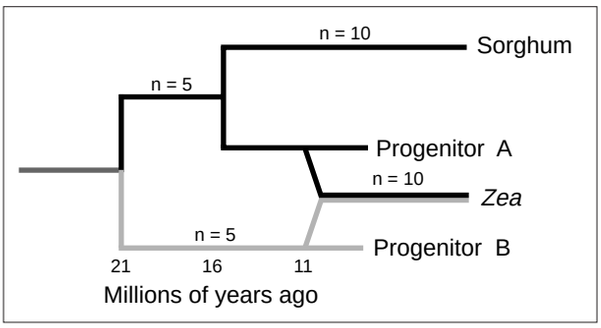

FIGURE 2. Phylogenetic reconstruction of the origin of the maize genome from two putative ancestral diploid species. As shown, DNA sequence data suggest that one of the two ancestral diploids of maize was more closely related to modern sorghum than it was to the other ancestral diploid.

From molecules to morphology

Given the dynamic nature of the maize genome with its vigorous transposable element activity and the presence of two subgenomes (up to 72% of the genome may be duplicate[20,25]), it is perhaps not surprising that maize shows a wide departure in morphology from its wild ancestor. Did transposon activity and duplicated genes provide the grist for the evolutionary mill that generated the ear of maize from its teosinte forerunner? Research over the past decade has brought us closer to an answer.

To begin to investigate the molecular basis of the morphological evolution of maize, Doebley *et al.* used quantitative trait locus (QTL) mapping to determine the number, location and relative level of effects of genes controlling the morphological differences between maize and its wild ancestor, teosinte. Two different maize-teosinte F_2 populations were analysed in these mapping experiments, both involving crosses of teosinte and primitive varieties of maize[26,27]. The results from these two populations were generally congruent, revealing that traits distinguishing maize and teosinte are controlled primarily by five chromosomal segments. Each of the segments identified has an effect on several traits, suggesting that they could carry either a single gene with pleiotropic effects, or several linked genes.

For two of the chromosomal regions, it has been determined that a single locus is responsible for major effects on the traits studied[28,29]. One locus, found on chromosome 4, controls differences in structures associated with the kernels of maize and teosinte, and has been named *teosinte glume architecture1* (*tga1*). The second locus, on chromosome 1, is responsible for differences in plant architecture, and corresponds to a previously described maize mutant named *teosinte branched1* (*tb1*).

Teosinte glume architecture

In teosinte, a shiny, stone-like fruitcase surrounds the kernel and protects it from herbivory by humans as well as other animals. Therefore, a crucial step in the evolution of maize as a crop plant was the reduction and softening of the fruitcase, allowing access to the kernel for use as food. The teosinte fruitcase is composed of two structures, an invaginated rachis internode (cupule) in which the kernel sits, and a hardened

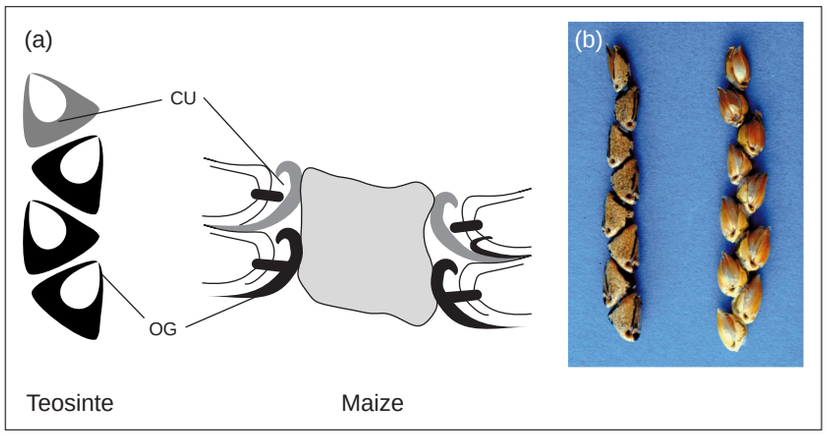

FIGURE 3. Maize and teosinte ears showing the effects of *tga1*. (a) Schematic drawings of longitudinal cross sections through teosinte and maize ears, showing the outer glume (OG) and cupule (CU). The teosinte ear has deeply invaginated cupules and hardened outer glumes that are curved upward, parallel to the axis of the ear, while the maize ear has flattened cupules and softer, perpendicularly oriented glumes. (b) Comparison of ears from wild-type teosinte (left) and teosinte carrying two copies of the maize *tga1* allele (right). Adapted from Ref. 30.

(indurated) glume that curves up and over the kernel to cover it and close the opening of the cupule (Fig. 3a). Maize kernels, on the other hand, are not encased in a fruitcase, because maize cupules are collapsed and its glumes are thinner, shorter and less curved.

A QTL identified on chromosome 4, corresponding to *tga1*, principally governs the difference between maize and teosinte glume hardness, size and curvature. The relative effects of the maize and teosinte alleles at *tga1* have been compared, both in maize and in teosinte genetic backgrounds, using nearly isogenic lines created by backcross breeding[28]. Recently, Dorweiler and Doebley[30] have described the developmental bases of these differences from which several inferences can be drawn. First, *tga1* might contribute to the shiny hard surface of the teosinte glume and fruitcase by controlling silica deposition in the epidermal cells (Fig. 4). With the maize allele of *tga1*, silica deposition is restricted to a subset of the epidermal cells while the teosinte allele conditions silica deposition in all epidermal cells. Second, the induration of teosinte glumes might be due, in part, to the effects of *tga1* on the number of glume mesoderm cells that become lignified (impregnated with lignin, a complex macromolecule that hardens cells). There are more of these lignified cells in the isogenic line carrying the teosinte allele than in the one with the maize allele. Third, in lines with the maize allele of *tga1*, the glumes and rachis internodes appear to grow at a slower rate, accounting for the decreased size of these structures and their inability to surround the kernel fully. Based on the multiple unrelated processes in which *tga1* seems to be involved, Dorweiler and Doebley suggest that it is a regulatory gene. While this remains to be proven, it is clear from a simple visual comparison of teosinte and teosinte carrying the maize allele of *tga1* that the effects of *tga1* are sufficient to represent a significant step in maize domestication (Fig. 3b).

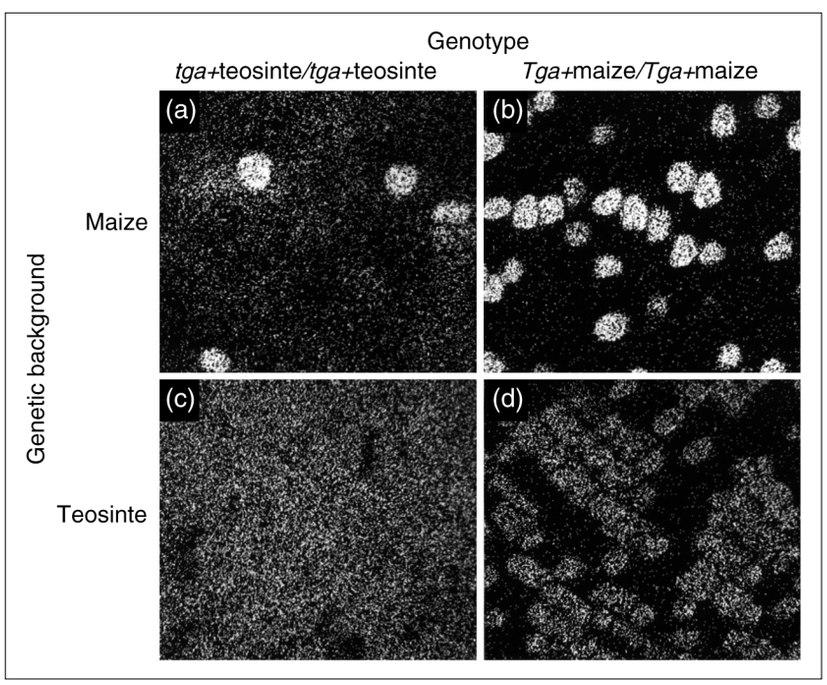

FIGURE 4. Silica distribution in the epidermis of maize and teosinte glumes. Concentrations of silica are detected by X-ray microanalysis and correspond to white dots on the micrograph. (a) Maize homozygous for the teosinte allele (*tga+*teosinte), (b) wild-type maize, (c) wild-type teosinte, (d) teosinte homozygous for the maize allele (*Tga+*maize).

Teosinte branched

Teosinte and maize plants also differ in plant architecture as expressed in the number and length of their primary lateral branches (branches that grow off the main stem) and the sex of the inflorescences that terminate these branches (Fig. 5). Teosinte has many long lateral branches that are terminated in male inflorescences (tassels), while the branches of maize are short and tipped with female inflorescences (ears). Teosinte ears are borne on secondary branches (branches that grow off the primary lateral branches), which are rarely found in modern maize. It appears that during maize evolution there was selection for increased repression of the growth of branches (increased apical dominance). The QTL associated with a major fraction of this difference between maize and teosinte has been identified as the gene *tb1* (Ref. 29), a gene initially discovered as a mutant of maize in 1959 (Ref.

31). This original *tb1* maize mutant (hereafter referred to as the 'maize *tb1* mutant') resembles teosinte in overall plant architecture but, unlike teosinte, usually bears only sterile tassel-like inflorescences rather than ears on its secondary branches.

Recently, the *tb1* gene has been cloned via *Mutator* transposon tagging[32]. Northern blot analysis using RNA from various maize organs shows that the gene is expressed in the ear primordia, the branch subtending the ears, and in the husks that surround the ear. A comparison of RNA levels in ear primordia between isogenic lines carrying either a maize or a teosinte allele revealed approximately twice the level of message accumulation is associated with the maize allele. So, what role might *tb1* have played in the evolution of maize plant architecture? Doebley *et al.*[32] note that the branches and their leaves (husks) in which *tb1* is expressed are reduced in size in wild-type maize relative to teosinte and maize *tb1* mutants. From this observation, they hypothesize that *tb1* functions as a repressor of the growth of these organs. During maize evolution, selection for a higher level of expression would cause increased repression leading to shorter branches and husks instead of fully developed branches and leaves. The twofold increase in message level associated with the maize allele relative to the teosinte allele in the same genetic background supports this interpretation.

FIGURE 5. (a) A teosinte (*Zea mays* ssp. *mexicana*) plant and (b) one of its primary lateral branches with terminal tassel (T). Silks (stigmas, S) are shown emerging from teosinte ears hidden within the leaf sheaths. (c) A wild-type maize plant. (d) A wild-type maize ear shoot. (e) A *tb1* mutant maize plant and (f) its primary lateral branches with terminal male inflorescences but no ears.

Even if this interpretation is correct, much remains to be learned. For example, maize *tb1* mutant plants do not form normal ears on their secondary branches as does teosinte. This suggests that *tb1* also has an additional function in ear development. This and other mysteries surrounding *tb1* should be answered in future studies.

Conclusion

What do the recent studies discussed above tell us about the evolution of maize and its genome? First, while the maize genome might be quite dynamic, much of the ebb and flow of genome size resulting from the amplification and reduction in the copy number of LTR retrotransposons in the intergenic regions might have little consequence to the plant. The same might not be true for MITEs whose common intrusion into the regulatory regions (perhaps resulting from target-site preference for AT-rich regions) suggests an important role in the evolution of gene expression. Second, the duplicate nature of the maize genome might have played a key role in the evolution of this most unusual crop. Two copies of each gene might enable one copy to carry on with the ancestral function while the other copy acquires a new role. In fact, several of the duplicated pairs studied by Gaut and Doebley are known to have distinct specializations. Third, accepting the inferences that both *tga1* and *tb1* are regulatory genes and that a change in regulation of *tb1* underlies some of the morphological differences between maize and teosinte, maize evolution would appear to fit the model of King and Wilson, and Britten and Davidson, that changes in gene regulation comprise the mainstay of evolutionary change.

Much more, of course, remains to be done to tie up this story. The question remains as to what specific molecular events changed the expression or function of teosinte genes to give rise to maize phenotypes and whether transposable elements have been involved. Although a multitude of various, stable transposable element insertions are found very close to maize genes, these elements might still be neutral with respect to their effect on gene regulation. Hence they might simply bracket, but not intrude into, gene regulatory regions. This question will be answered by comparative analyses of multiple alleles of individual genes to determine

whether these insertions have any functional consequences. In this regard, it will be particularly interesting to define the mutations responsible for changes in the regulation or function of genes, such as *tga1* and *tb1*, and so learn what molecular magic caught the eye of ancient teosinte farmers some 7000 years ago.

References

1 Kato, T.A. (1976) *Research Bulletin Massachusetts Agricultural Experiment Station* 635
2 Doebley, J. (1990) *Econ. Bot.* 44 (Suppl. 3) 6–27
3 King, M-C. and Wilson, A.C. (1975) *Science* 188, 107–116
4 Britten, R.J. and Davidson, E.H. (1971) *Quart. Rev. Biol.* 46, 111–133
5 Hake, S. and Walbot, V. (1980) *Chromosoma* 79, 251–270
6 Freeling, M. (1984) *Annu. Rev. Plant Physiol.* 35, 277–298
7 Springer, P.S., Edwards, K.J. and Bennetzen, J.L. (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 863–867
8 Bennetzen, J.L. *et al.* (1994) *Genome* 37, 565–576
9 SanMiguel, P. *et al.* (1996) *Science* 274, 765–768
10 Bureau, T.E. and Wessler, S.R. (1992) *Plant Cell* 4, 1283–1294
11 Bureau, T.E. and Wessler, S.R. (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 1411–1415
12 Bureau, T.E. and Wessler, S.R. (1994) *Plant Cell* 6, 907–916
13 White, S.E., Habera, L.F. and Wessler S.R. (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 11792–11796
14 Neuffer, M.G. (1966) *Genetics* 53, 541–549
15 Bradshaw, V.A. and McEntee, K. (1989) *Mol. Gen. Genet.* 218, 465–474
16 Pouteau, S., Grandbastien, M.A. and Boccara, M. (1994) *Plant J.* 5, 535–542
17 Mottinger, J.P., Dellaporta, S.L. and Keller, P.B. (1984) *Genetics* 106, 751–767
18 Bregliano, J.C. and Kidwell, M.G. (1983) in *Mobile Genetic Elements* (Shapiro, J.A., ed.), pp. 363–410, Academic Press
19 Laurie, D.A. and Bennett, M.D. (1985) *Heredity* 55, 307–313
20 Whitkus, R., Doebley, J. and Lee, M. (1992) *Genetics* 132, 1119–1132
21 Moore, G., Devos, K.M., Wang, Z. and Gale, M.D. (1995) *Curr. Biol.* 5, 737–739
22 Avramova, Z. *et al.* (1996) *Plant J.* 10, 1163–1168
23 Helentjaris, T., Weber, D. and Wright, S. (1988) *Genetics* 118, 353–363
24 Gaut, B.S. and Doebley, J.F. (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94, 6809–6814
25 Ahn, S. and Tanksley, S.D. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 7980–7984
26 Doebley, J., Stec, A., Wendel, J. and Edwards, M. (1990) *Proc. Natl. Acad. Sci. U. S. A.* 87, 9888–9892
27 Doebley, J. and Stec, A. (1993) *Genetics* 134, 559–570
28 Dorweiler, J., Stec, A., Kermicle, J. and Doebley, J. (1993) *Science* 262, 233–235
29 Doebley, J., Stec, A. and Gustus, C. (1995) *Genetics* 141, 333–346
30 Dorweiler, J.E. and Doebley, J.F. (1997) *Am. J. Bot.* 84, 1313–1322
31 Burnham, C. (1959) *Maize Genet. Coop. Newslett.* 33, 74
32 Doebley, J., Stec, A. and Hubbard, L. (1997) *Nature* 386, 485–488

S. White *and* ***J. Doebley*** *are at the Department of Plant Biology, University of Minnesota, St Paul, MN 55108, USA.*

Genetwork

Using metabolic pathway databases for functional annotation

Newly obtained nucleotide and protein sequences are searched routinely against databases, and the World Wide Web has made such queries simple to perform[1]. Improved searching and scoring methods detect more subtle similarities than ever before, often allowing a researcher to make reasonable guesses about the possible role(s) of new gene sequences. Unfortunately, functional annotation of gene sequences can be fraught with difficulties, the most pernicious of which can be erroneous descriptions of database entries[2]. Therefore, the results of any database search need careful examination, and it is essential to understand the functions of the matched proteins. Metabolic pathway databases can help in providing this understanding and also offer the context for further explorations of a functional assignment. Here, we describe what you might do when you find database matches that suggest your new protein has some similarity to, say, ketol-acid reductoisomerase and you have little idea what these words even mean.

The SWISS-PROT database[3], maintained by Amos Bairoch, is the most complete general resource for information about individual proteins. SWISS-PROT annotations have descriptions of the function of a protein, its domain structure, post-translational modifications, variants, reactions catalyzed by this protein, active site residues, similarities with other sequences and more. The database entries are linked to the ENZYME database[4], which contains short descriptions of each enzyme and the reaction it catalyzes. ENZYME is the primary reference point for the Enzyme Classification (EC) numbers and, unlike SWISS-PROT, includes enzymes that have not yet been sequenced.

To put an enzyme name into a biochemical perspective, it is valuable to consider the metabolic pathways to which it contributes. Perhaps the most familiar way to do this is using the popular poster of biochemical pathways distributed by the Boehringer Mannheim[5], which is now available on the WWW[6]. This online map can be searched for both the enzyme and the metabolite names, and it links to the ENZYME database. If you still prefer the paper version, you can request it by sending an e-mail message to biochemts_us@bmc.boehringer-mannheim.com. The Kyoto Encyclopedia of Genes and Genomes (KEGG)[7] was developed especially for the Web and offers the additional ability to focus on the metabolic reactions in specific organisms. This frequently updated site presents a comprehensive set of metabolic pathway charts, both general and specific for each of the completely sequenced genomes, as well as for *Caenorhabditis elegans*, *Drosophila* and human. Before getting links to the pathways for a specific organism, it is necessary to step down through the text hierarchy. However, on the charts, the enzymes that

Copyright © 1998 Elsevier Science Ltd. All rights reserved. 0168-9525/98/$19.00
PII: S0168-9525(98)01523-6