Genomics and Synteny

Susan R. McCouch*

Plant Breeding Department, 418 Bradfield Hall, Cornell University, Ithaca, New York 14853–1901

Building on early observations about natural plant variation in time and space, early geneticists, including Darwin, Mendel, and Vavilov, posed fundamental questions about the origin, structure, and evolution of genetic diversity. They postulated that an underlying reservoir of innate and heritable genetic possibilities delineated the options for the growth, development, and reproduction of organisms at both the individual and population levels. The field of plant genetics today continues to address many of the same questions while integrating developments in molecular and computational biology. Over the last 15 to 20 years, new, highly automated tools have created unprecedented opportunities for generating and analyzing large biological data sets, and the systematic processing of nucleic acid and protein sequence information from many different organisms has fundamentally changed the way that biologists approach the study of living things.

GENOMICS

The term genome (derived from the words genes and chromosomes) was first used by Winkler (30) to signify the complete set of chromosomes and their genes. The term genomics was first used in 1986 to describe the enterprise that aimed to map and sequence the human genome (18). The field of genomics takes advantage of the common biological language represented by DNA and RNA and uses high throughput sequencing strategies, microchip arrays, digital technology, and computationally intensive analysis to understand the structure, function, and evolution of diverse organisms. Applications of genomic technologies in all areas of biology have lowered the barriers that once separated the plant, animal, and microbial research communities.

LINKAGE MAPPING

The use of restriction fragment length polymorphisms (RFLPs) as genetic markers made it possible to map, for the first time, an almost unlimited number of randomly distributed polymorphic loci in a single population and provided the foundation for efficient, whole-genome studies at the molecular level. The application of RFLP technology for genetic mapping was pioneered in humans by Botstein et al. (8). The distributed nature of restriction enzyme sites and the neutral nature of restriction enzyme polymorphism turned out to be equally applicable for genetic map construction in plants. This was first demonstrated by Bernatzky and Tanksley (4) and Helentjaris et al. (12), whose work established the foundation for molecular mapping in a wide variety of plant species over the next decade.

QUANTITATIVE TRAIT LOCUS (QTL) ANALYSIS

The advances in molecular linkage mapping unleashed a series of powerful new methodologies for studying genotype-phenotype relationships. Given the emphasis on diversity (rather than on a single species) in the plant kingdom, low- or medium-resolution molecular maps were constructed for numerous plant species over the last 15 years. One of the rationales for constructing these maps was to facilitate genetic analysis and characterization of genes underlying both simply and quantitatively inherited traits. The plant science community was quick to take advantage of the fecundity, versatility of mating habit (both inbreeding and outcrossing are possible in most plant species), and ability to manipulate the reproductive cycle of plant populations to pioneer QTL analysis using molecular markers (for review, see 26). The first use of a complete RFLP map for this purpose was reported by Paterson et al. (22) in tomato (*Lycopersicon* sp.), and this study launched an epoch of accelerated mapping and QTL analysis in plants.

GENOME DUPLICATION

At the same time RFLP maps were being used to identify the position of genes and QTLs along the chromosomes of higher plants, these same maps were defining internally duplicated chromosome segments within plant genomes. The location and extent of internal duplications were first documented in the maize (*Zea mays*) genome (13), supporting earlier hypotheses describing modern, diploid maize as an ancient polyploid. More recently, studies of the Brassicaceae family, of which Arabidopsis is a member, have emphasized the themes of genome duplication and rearrangement (for review, see 6, 17). Polyploid ancestry in all biological kingdoms is probably more universal than once was believed, as postulated by Ohno (20) and demonstrated in yeast (*Sac-*

* E-mail srm4@cornell.edu; fax 607–255–6683.

charomyces cerevisiae; 31). As increasing resolution is achieved in genetic, physical, and sequence-based mapping studies, detection of genome-wide and fine-scaled duplication is providing new insights regarding the timing of duplication events and the types of selection pressures affecting specific genes and/or genomic regions of plant, animal, and microbial genomes over the course of evolution.

COMPARATIVE MAPPING AND HOMEOLOGY-BASED GENE ISOLATION

The term synteny (from the Greek; syn = together, taenia = ribbon) is used in genetics to indicate the presence of two or more loci on the same chromosome. The original relevance of the term dates to a pregenomics era when locating genes to chromosomes was accomplished without the advantage of whole-genome mapping technologies. Today, the concept of synteny has been expanded to address questions of homeology (residual homology of originally completely homologous chromosomes). The earliest whole-genome comparative maps for plants were developed among species in the Solanaceae family. Bonierbale et al. (7) demonstrated that cDNA markers along the 12 chromosomes of tomato and potato (*Solanum tuberosum*) were largely collinear, differing only by five detectable paracentric inversions, whereas pepper (*Capsicum* sp.) appears to have a greater propensity for rearrangement. Studies in the Gramineae family are perhaps the most developed. Common sets of low-copy cDNA "anchor probes" were used to develop comparative maps among seven or more different grass family members (27; for review by Devos and Gale, see 9) and regions of conserved gene order with corresponding positional conservation of phenotypes, attributed to mutants and QTLs, have been well documented (for references, see 23). However, numerous exceptions where linked markers do not map to the predicted locations within collinear regions have also been reported (3, 21, 29) and raise interesting questions about the predictive nature of comparative maps at both low and high levels of resolution. Kilian et al. (14) were the first to attempt to clone a gene in one plant species based on detailed positional and sequence information (i.e. on microsynteny, as it is now called) in a homeologous region of another genus. Although the corresponding chromosome segments of barley (*Hordeum vulgare*) and rice (*Oryza sativa*) were clearly homeologous in the region of the barley stem rust resistance gene *Rpg1,* the target gene could not be identified in the predicted location in rice. This study illustrated some of the difficulties associated with the application of comparative maps to gene isolation efforts, and is also consistent with evidence suggesting that plant resistance genes may evolve more rapidly than other kinds of genes (19). Studies of gene collinearity across ever more divergent genomes have moved from hybridization-based Southern analysis of cDNA clones (within families) to more highly automated forms of sequence-based analysis based on comparisons with model genomes. In a recent study by Ku et al. (16), microcollinearity between tomato and Arabidopsis was evaluated en route to identifying a candidate gene for the *ovate* mutant controlling fruit shape in tomato. A 105-kb region of the tomato genome was sequenced and compared with the almost-completed genomic sequence of Arabidopsis. Rather than a pattern of one-to-one collinearity in a defined homeologous region, they described distributed networks of synteny that reflect the various genome duplication events that have marked the evolution of Arabidopsis since divergence from its last common ancestor with tomato. The extent of monocot-dicot homeology has also been addressed, but conclusive evidence on this question awaits analysis of the emerging genomic sequences of rice and Arabidopsis.

SEQUENCES AND SEQUENCE DATABASES

Today, the search for a gene of interest often starts with sequence information, including expressed sequence tags (ESTs), genomic sequence, and protein sequence. A major step in making sequence information publicly available for large-scale analysis was the formation of GenBank and its counterpart, the European Molecular Biology Laboratory data library, in 1982 (5). As more and more data is deposited, these and other databases have become increasingly useful because there is an ever greater chance of finding sequence similarity (and inferring potential homology) to a newly sequenced gene. Key to making sequence information useful to biologists has been the development of rigorous and efficient algorithms for searching large databases and distinguishing biologically significant relationships from chance similarities (2, 24, 25). The value of cDNA sequencing was first summarized by Adams et al. (1) where they argued that EST sequencing would be the most efficient and cost-effective way of tagging most of the genes in an organism long before complete genomic sequence was available. This approach had the additional appeal of focusing only on the most conserved portions of diverse genomes and on sequences that were likely to be of functional significance. Full genomic sequencing of even the most compact plant genome(s) represents a major commitment of time and resources, but in contrast with EST sequencing, is designed to reveal all the genes in an organism. In addition, genomic sequencing provides information about global genome structure and organization, regulatory regions, transposable elements, and noncoding sequences. Thus, the availability of fully assembled and annotated genomic sequence for divergent model plant species, such as Arabidopsis and rice, in addition to many thousands of ESTs, bacterial

artificial chromosome ends, and genomic sequences in multiple other species, helps drive genomics research in the plant biology community today.

FUNCTIONAL GENOMICS

Genetics is the study of variation and inevitably involves a quest to understand the relationship between genotype and phenotype. For many years, genetic studies were initiated based on observations of either naturally occurring or consciously derived phenotypic mutations in an organism. Insertional mutagenesis, which may involve activation of native transposons or the introduction of foreign transposons or T-DNA elements into a genome via plant transformation, provides a way of knocking out, reducing, enhancing, or completely altering the expression patterns of specific genes (for review, see 15, 28). The systematic application of insertional mutagenesis techniques in a genomics context aims to interrupt or alter the expression of every gene in an organism as the basis for studying the relationship between phenotype and genotype. The efficiency of this approach ultimately depends on the ability to recognize altered phenotypes and rapidly associate phenotypic changes with specific sequence alternations (i.e. to identify functional nucleotide polymorphisms). With the availability of well-saturated genetic and comparative maps, abundant EST and genomic sequence information, and an increasingly sophisticated set of computational tools, these populations are increasingly useful as sources of novel genetic variation and as a tool for studying structure-function relationships across diverse organisms. Today, efforts to understand the functional significance of a gene or suite of genes in virtually any biological context involve both keen observation of phenotype and intensive computational comparisons of the structural properties of DNA or protein. Linear approaches to sequence alignment are increasingly complemented by assessing the more highly conserved, three-dimensional structure of a gene's protein product. In a recent paper by Frary et al. (10), a major QTL determining fruit size in tomato was positionally cloned, and in an effort to understand the gene's function, the predicted structure of the protein product was compared with others in a protein structure database. The tomato protein revealed a three-dimensional structure with remarkable similarity to a human oncogene RAS (rat sarcoma, named because it was first discovered in rat) protein, known to deregulate cell division in humans. This work presents the tantalizing hypothesis that a common, ancient gene family, whose identity predates the divergence of plants and animals, may be responsible for controlling fruit size in tomatoes and tumor development in humans. In a similar vein, work in plant disease resistance strongly suggests that the defense system in plants has evolutionary origins in common with the immune system in mammals (for review, see 11). The ability to recognize these similarities and to comprehend the logic that supports the arguments lies at the root of biologists' intrigue with studies of genomics and synteny today.

ACKNOWLEDGMENTS

I gratefully acknowledge help from my graduate students, Michael Thomson and Jeremy Edwards, during the preparation of this manuscript.

LITERATURE CITED

1. **Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie RWR, Venter JC** (1991) Science **252:** 1651–1656
2. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ** (1990) J Mol Biol **215:** 403–410
3. **Bennetzen JL, SanMiguel P, Chen M, Tikhonov A, Francki M, Avramova Z** (1998) Proc Natl Acad Sci USA **95:** 1975–1978
4. **Bernatzky R, Tanksley SD** (1985) Genetics **112:** 887–898
5. **Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FI, Rindone WP, Swindell CD, Tung C-S** (1986) Nucleic Acids Res **1988:** 1861–1864
6. **Blanc G, Barakat A, Guyot R, Cooke R, Delseney M** (2000) Plant Cell **12:** 1093–1101
7. **Bonierbale MW, Plaisted RL, Tanksley SD** (1998) Genetics **120:** 1095–1103
8. **Botstein D, White RL, Skolnick M, Davis RW** (1980) Am J Hum Genet **32:** 314–331
9. **Devos KM, Gale MD** (1997) Plant Mol Biol **35:** 3–15
10. **Frary A, Nesbitt TC, Frary A, Grandillo S, vander Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB, Tanksley SD** (2000) Science **289:** 85–88
11. **Hammond-Kosack KE, Jones JDG** (1997) Annu Rev Plant Physiol Plant Mol Biol **48:** 575–607
12. **Helentjaris T, Slocum M, Wright S, Schaefer A, Nienhuis J** (1986) Theor Appl Genet **72:** 761–769
13. **Helentjaris T, Weber D, Wright S** (1988) Genetics **118:** 353–364
14. **Kilian A, Chen J, Han F, Steffenson B, Kleinhofs A** (1997) Plant Mol Biol **35:** 187–195
15. **Krysan PJ, Young JC, Sussman MR** (1999) Plant Cell **11:** 2283–2290
16. **Ku H-M, Vision T, Liu J, Tanksley SD** (2000) Proc Nat Acad Sci USA **97:** 9121–9126
17. **Lagercrantz U** (1998) Genetics **150:** 1217–1228
18. **McKusick VA** (1997) Genomics **45:** 244–249
19. **Michelmore RW, Meyers BC** (1998) Genome Res **8:** 1113–1130
20. **Ohno S** (1970) *In* Evolution by Gene Duplication. Springer-Verlag, New York, pp 137–146
21. **O'Neill CM, Bancroft I** (2000) Plant J **23:** 233–243
22. **Paterson AH, Lander ES, Hewitt JD, Peterson S, Linxcoln SE, Tanksley SD** (1988) Nature **335:** 721–726

23. **Paterson AH, Lin Y-R, Li Z, Schertz KF, Doebley JF, Pinson SRM, Liu S-C, Stansel JW, Irvine J** (1995) Science **269:** 1714–1718
24. **Pearson WR, Lipman DJ** (1988) Proc Nat Acad Sci USA **85:** 2444–2448
25. **Smith TF, Waterman MS** (1981) Adv Appl Math **2:** 482–489
26. **Tanksley SD** (1993) *In* Annual Review of Genetics, Vol 27. Annual Reviews Inc., Palo Alto, CA, pp 205–233
27. **Van Deynze AE, Sorrells ME, Park WD, Ayres NM, Fu H, Cartinhour SW, Paul E, McCouch SR** (1998) Theor Appl Genet **97:** 356–369
28. **Walbot V** (1992) Annu Rev Plant Physiol Plant Mol Biol **43:** 49–82
29. **Wilson WA, Harrington SE, Woodman WL, Lee M, Sorrells ME, McCouch SR** (1999) Genetics **153:** 453–473
30. **Winkler H** (1920) *In* Verbreitung und Ursache der Parthenogenesis im Pflanzen-und Tierreiche. Gustav Fischer, Jena, Germany, pp 1–3
31. **Wolfe KH, Shields DC** (1997) Nature **387:** 708–713