PLANT GENOME MAPPING: STRATEGIES AND APPLICATIONS

Andrew H. Paterson
Plant Genome Mapping Laboratory, University of Georgia , Athens GA 30602, USA

Keywords: comparative genetic mapping, DNA marker, genetic linkage mapping, genome sequence, physical mapping, polymerase chain reaction, Quantitative trait locus**.**

Contents

Summary

The manipulation of plant genomes based on sexual crosses, taking advantage of the Mendelian principles of segregation and recombination, was central to crop domestication by aboriginal peoples as well as to the ongoing improvement of crops by scientific plant breeding, and is substantially responsible for the ability of humanity to sustain nearly 7 billion people (and growing). For much of its history, plant improvement was practiced based solely on phenotype, its effectiveness reduced by the vagaries of non-genetic factors such as heterogeneous testing environments (see also – *Conventional Plant Breeding for Higher Yields and Pest Resistance*). Demonstration of the chromosomal basis of heredity set the stage for the development of diagnostic tools to track the movement of genes and genomic regions, and DNA-based markers made it possible to routinely apply genetic diagnostics in crop improvement and other applications (see also – *Plant Breeding and Molecular Farming*). Several types of DNA markers are widely used, with different strengths and weaknesses and also different requirements for technological infrastructure and *a priori* knowledge of the subject organism. A range of breeding strategies can be employed to develop different types of genetic populations, suitable to address different questions in inbreeding or outcrossing taxa. The limited resolution of genetic approaches to mapping of plant genomes are increasingly complemented by physical approaches, based on physical breakage of DNA molecules to varying degrees and their computational 'reassembly' to deduce the organization of the native DNA. The fundamental similarity among angiosperm (flowering) plants resulting from descent from a common ancestor that lived perhaps 140-170 million years ago is reflected in the gene repertoire and arrangement of modern plants. It is becoming feasible to deduce the probable genome organization of previously un-studied plants based on comparison of well-studied botanical models, and methods are well-established for developing DNA markers that work well across divergent taxa. Technological progress sets the stage for choosing organisms for study based more heavily on their intrinsic importance, reducing the need that they be facile for genomics.

1. What is a genome, and why map it?

The word "genome" describes the total repertoire of DNA in a particular organelle. Animals have one genome in the nucleus, and a second, very different genome in the mitochondrion. Plants have yet a third genome, in the chloroplast. While each of these are important, the nuclear genome is by far the largest of the three, and imparts the vast majority of characteristics to an organism.

Genome mapping is a widely-applicable approach to scanning the genetic information of an organism for genes that are responsible for a specific trait. Higher plants are thought to have 25,000 or more genes, the vast majority of which remain of unknown function (although rapid progress is being made toward their characterization in selected botanical models). A particular strength of genome mapping, is that it facilitates isolation of genes based simply on measurement of their effect(s) on phenotype -- requiring no *a priori* knowledge of the biochemical function performed by a gene.

A "genome map" can be thought of much as a roadmap, reflecting the relative proximity of different landmarks to one another. Genome mapping is made possible by the fact that the nuclear genome in higher organisms is organized and transmitted as linear units, called chromosomes. Just as mileposts guide the motorist along a highway, "DNA markers" provide reference points that define specific places along each chromosome.

Two broad categories of genome mapping approaches offer different levels of resolution at which genomes can be studied.

a. Genetic mapping is based on recombination, literally the naturally-occurring 'breaking and rejoining' of chromosomes to determine the relative proximity of DNA landmarks to one another based on the frequency at which they co-occur on the same chromosome segment. Recombination is a biological phenomenon and can occur at very different rates in different genomic regions, populations or environments. Nonetheless, it directly measures the transmission of genes (or groups of genes) from parent to progeny, a central component of crop improvement.
b. Physical mapping involves determination of the proximity of DNA landmarks by direct measurement of the physical quantity of DNA that lies between them. This is often similar to genetic mapping, involving *artificial* breakage of DNA molecules using radiation, DNA-modifying enzymes, or shear forces, and determination of the frequency at which DNA landmarks co-occur on the same chromosome segment. Alternatively, cytomolecular mapping is a form of physical mapping in which DNA landmarks are directly visualized on chromosomes using microscopy, and the distance between them measured directly. Physical mapping is less subject to the vagaries of different genomic regions, populations or environments than genetic mapping. Further, it generally offers finer resolution than can be reasonably achieved by genetic mapping, facilitating cloning of specific genes and sequencing of entire genomes.

A third approach, comparative mapping, takes advantage of map information (either genetic or physical, the latter including DNA sequence) from one species to deduce the

probable gene arrangement in another species. Most frequently, this approach uses detailed information such as, for example, the completely-sequenced genome of a botanical model to deduce the probable gene arrangement in the genome of an 'orphan crop' for which DNA-level information is lacking.

2. Genetic mapping

Genetic mapping uses the Mendelian principles of segregation and recombination to determine the relative proximity of DNA markers along the chromosomes of an organism. Most major crops and botanical models now enjoy detailed "reference" genetic maps of specific populations that were prioritized based on economic considerations and/or experimental facility, as well as less-detailed maps of many additional populations made for specific research or crop improvement goals (for example, the identification of a disease resistance locus). Ideally, the "reference maps" are a primary source of DNA markers for the less-detailed maps, providing a conduit for meta-analyses that permit researchers to take advantage of the cumulative results from studying many different populations, in different environments, for a range of traits.

2.1 Genetic markers

A primary consideration in genetic mapping is the degree of differentiation between the parents of a population to be studied. For a genetic marker to be 'informative' in a particular population, it must be 'polymorphic', revealing different forms (alleles) in the respective parents. Genetic mapping is naturally closely-allied with crop breeding (see also – *Plant Breeding and Molecular Farming*), and indeed, many populations made for crop breeding have been genetically mapped. However, many crop breeding populations are derived from crossing the "best with the best" parents, and the parents are often closely related. Therefore, reference maps tend to be made from populations made by crossing parents that are maximally divergent at the DNA level, providing for cost-efficient discovery and characterization of large numbers of DNA markers. This is done with the tacit, and usually correct, assumption that different subsets of markers from the reference cross will be informative in different breeding populations.

2.1.1 Visible, or morphological markers

Genetic mapping was practiced for nearly four decades prior to the demonstration that DNA was the hereditary molecule, taking advantage of the phenomenon of "genetic linkage" explained by Morgan using "visible markers." A visible marker is simply a mutation in a particular gene, which imparts a discrete, easily-identified phenotype to an organism. While most visible markers were randomly transmitted from parent to progeny, the principle of "genetic mapping" was based on the observation that specific alleles at an occasional pair of markers were usually co-transmitted (for example, progeny tending to have the characteristics of either one parent or the other, rather than random assortment). For such gene pairs, the frequency of the rare 'recombinants' (progeny that have the characteristic of one parent at one locus and the other parent at the other locus) reflects the proximity between the markers along the chromosome. Students of Drosophila, bean, and a variety of other organisms had demonstrated the basic principles now associated with genetic mapping, and in fact had constructed

partial genetic maps, well before it was established that DNA was the molecule they were studying.

Several limitations of "visible markers" obstructed the progress that could be made by early practitioners of genetic mapping. First, visible markers often had deleterious effects on the organism -- after all, they represented mutations in a gene that had an obvious function. (For example, consider the competitiveness of wingless fruitflies, in nature). While laboratory strains of organisms were carefully maintained in non-competitive environments, a particular strain could typically only survive under the "genetic load" of a few visible markers. Because hundreds of such markers were needed to provide mileposts for all regions of all chromosomes, extensive genetic mapping experiments simply were not feasible. Moreover, visible markers were rare -- in the day of Morgan (circa 1910), identification of visible markers was a serendipitous event. Genetic linkage analysis required that visible markers be assembled into "multiply marked stocks" by tedious breeding procedures that might take longer than the experiment. Today, such "visible markers" can be created and identified in an efficient manner– while they remain obsolete as genetic markers, they do contribute to quickly and efficiently isolating genes associated with a particular phenotype.

2.1.2 Protein markers

With the advent of electrophoresis (the separation of molecules based on differential migration in an electrical field), macromolecules encoded in the hereditary information of an organism supplemented, and eventually largely replaced visible markers. Most such markers were specific proteins, for which enzymatic assays were known that yielded colorimetric substrates. 'Isozyme' analysis (see also – *Enzyme Production*) led to numerous fundamental insights into genome organization and population biology as well as practical uses as diagnostic tools for specific traits, and remains in practice today for applications that only require a modest number of genetic markers per organism. Many of the underlying proteins function in primary metabolism and the same 'staining' procedures work across a wide range of taxa, so these methods are among the quickest and cheapest marker systems to develop, being readily applied to new taxa or new questions with a minimum of upfront investment. However, colorimetric stains are only available for a modest number of proteins, limiting isozyme studies to systems for which sufficient information could be obtained from only a small sampling of loci in a genome.

2.2 DNA markers

The demonstration that DNA was the hereditary molecule, the unraveling of the genetic code, and discovery that fortuitous properties of specific 'molecular shears' called 'restriction enzymes', permitted the reproducible isolation and 'cloning' (replication of the DNA in a bacterium) of specific segments of DNA, laid the groundwork for modern plant genomics. By basing genetic analysis directly on DNA, rather than on naturally occurring "visible markers", no longer did scientists have to endure long and tedious breeding experiments to assemble visible markers into "multiply marked stocks" -- since members of a taxon shared a common DNA language. Instead, one might take advantage of naturally-occurring "spelling errors" (mutations) in the language -- most of

which had no effect on the viability or fitness of the organism, thus could accumulate freely in a lineage without impairing viability. Moreover, because one had millions, sometimes billions of "letters" in which "spelling errors" (mutations) might occur, the number of potential "DNA markers" was far greater than visible or protein markers.

2.3 Key characteristics of DNA markers

In setting the stage for a comprehensive comparison of the spectrum of DNA marker types, I must first define some key characteristics that differentiate among them.

2.3.1 Separation method

Most DNA markers have traditionally required some form of electrophoresis, the separation of different macromolecules by differential migration in an electrical field. The medium through which the molecules migrate largely determines the resolution that can be obtained, Agarose electrophoresis (see also – *Physical Methods applied in* Biotechnology) has been, and remains widely practiced, routinely permitting separation of DNA fragments ranging in size from about 0.2 – 20 kb, with specialized methods available for separations of up to 2000 kb or more. One can routinely differentiate between bands that differ in size by about 5-10%, with specialized (and costly) grades of agarose used at high concentrations permitting DNA bands differing in size by 2% to be resolved in the range of 200-1000bp. Acrylamide electrophoresis, involving more costly reagents and more complex gel preparation, offers single-nucleotide resolution with some commercial instruments partially automating data collection. Capillary electrophoresis, now routine for high-throughput sequencing, also offers single-nucleotide resolution and automates both the gel preparation and data collection processes, but requires time on a costly instrument with a costly service contract. Nonetheless, instrument depreciation and service contract can be amortized over much data – in principle, the instrument can perform ~1500-2000 separations per day, each of which may contain 4 loci.

There are significant advantages of DNA marker methods that do not require electrophoresis, being based instead on the ability to detect mismatches in hybridization of one DNA molecule to another. Relatively short (ca 20 nt or less) DNA fragments are particularly sensitive to mismatches, which can alter the 'melting' (denaturation) temperature of double-stranded DNA by 5 C. In principle, this permits one to distinguish 'match' versus 'mismatch' for thousands of samples at a time. First implemented in 1983, such methods have recently been adapted to 'microarray' platforms that permit many thousands of loci to be screened for polymorphism on a single microscope slide.

2.3.2 Biallelism

In most diploid or allopolyploid plants (the latter behaving essentially as diploids at meiosis), an individual may have one of three possible genotypes at a locus. Specifically, it can be heterozygous with one allele from each of its parents, or it can be homozygous for one parental allele or the other (assuming that inbreeding can be done, as is true for most diploid and allopolyploid plants) (see also – *Genetic Engineering of*

Plants). Complete information, i.e. the presence in the parents of different alleles that can each be uniquely distinguished in the progeny, offers substantial advantages in genetic mapping. In several DNA marker systems, one parental allele is null, represented by the absence of a band. In this situation, one can know that if a segregating individual has no band, it is homozygous for the allele of the parent with the null polymorphism. However, if the band is present, it could be either heterozygous or homozygous (most assays lacking the sensitivity to quantitatively distinguish one copy from two). In genetic populations in which al three possible genotypes are still present (see further discussion below), this substantially reduces the genetic map information that can be obtained from a population.

2.3.3 Distribution

The organization of different types of DNA elements in plant genomes studied to date is highly structured, with the pericentromeric regions of chromosomes tending to be rich in repetitive DNA and the distal regions tending to be gene-rich. Accordingly, DNA markers that are based on genes and gene-like sequences are expected to be enriched in distal regions of the chromosomes. While repetitive DNA is generally rich in the pericentromeric space, some classes such as most DNA transposons tend to be distributed more like genes. It is particularly important to have adequate marker density in the gene-rice regions, as these also tend to experience more recombination and therefore need more markers to adequately diagnose genotype. Ideally, one might employ two or more complementary 'classes' of DNA markers to obtain good genome coverage in a reference map.

-
-
-

Bibliography

Botstein, D., R. White, M. Skolnick, and R. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314-331. [A seminal description of methods for molecular mapping of a genome]

Bowers, J.E., C. Abbey, S. Anderson, C. Chang, X. Draye, A.H. Hoppe, R. Jessup, C. Lemke, J. Lennington, Z. Li, Y.R. Lin, S.C. Liu, L. Luo, B.S. Marler, R. Ming, S.E. Mitchell, S. Kresovich, K.F. Schertz, and A.H. Paterson. 2003. A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics 165:367-386. [One of the most detailed reference maps in plants]

Bowers, J.E., M.A. Arias, R. Asher, J.A. Avise, R.T. Ball, G.A. Brewer, R.W. Buss, A.H. Chen, T.M. Edwards, J.C. Estill, H.E. Exum, V.H. Goff, K.L. Herrick, C.L.J. Steele, S. Karunakaran, G.K. Lafayette, C. Lemke, B.S. Marler, S.L. Masters, J.M. McMillan, L.K. Nelson, G.A. Newsome, C.C. Nwakanma, R.N. Odeh, C.A. Phelps, E.A. Rarick, C.J. Rogers, S.P. Ryan, K.A. Slaughter, C.A. Soderlund, H.B. Tang, R.A. Wing, and A.H. Paterson. 2005. Comparative physical mapping links conservation of

microsynteny to chromosome structure and recombination in grasses. Proceedings of the National Academy of Sciences of the United States of America 102:13206-13211. {One of the most detailed physical maps in plants]

Feltus, F.A., H.P. Singh, H.C. Lohithaswa, S.R. Schulze, T. Silva, and A.H. Paterson. 2006. Conserved intron scanning primers: Targeted sampling of orthologous DNA sequence diversity in orphan crops. Plant Physiology 140:1183-1191. [A promising method for development of pan-taxon DNA markers from genome sequences of botanical models]

Gao, W.X., Z.J. Chen, J.Z. Yu, R.J. Kohel, J.E. Womack, and D.M. Stelly. 2006. Wide-cross whole-genome radiation hybrid mapping of the cotton (Gossypium barbadense L.) genome. Molecular Genetics and Genomics 275:105-113. [One of the few radiation hybrid mapping efforts in plants]

Hulbert, S.H., T.E. Richter, J.D. Axtell, and J.L. Bennetzen. 1990. Genetic-Mapping and Characterization of Sorghum and Related Crops By Means of Maize Dna Probes. Proceedings of the National Academy of Sciences of the United States of America 87:4251-4255. [An early example of comparative mapping in plants]

Kynast, R.G., R.J. Okagaki, M.W. Galatowitsch, S.R. Granath, M.S. Jacobs, A.O. Stec, H.W. Rines, and R.L. Phillips. 2004. Dissecting the maize genome by using chromosome addition and radiation hybrid lines. Proceedings of the National Academy of Sciences of the United States of America 101:9921-9926. [Another radiation hybrid mapping efforts in plants]

Lee, M., N. Sharopova, W.D. Beavis, D. Grant, M. Katt, D. Blair, and A. Hallauer. 2002. Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. Plant Molecular Biology 48:453-461. [An application of intermating procedures to increase recombination]

Liu, S.C., S.P. Kowalski, T.H. Lan, I.A. Feldmann, and A.H. Paterson. 1996a. Genome-wide high-resolution mapping by recurrent intermating using Arabidopsis thaliana as a model. Genetics 142:247-258. [Basic description of the intermating method to increase recombination]

Liu, S.C., S.P. Kowalski, T.H. Lan, K.A. Feldmann, and A.H. Paterson. 1996b. Genome-wide high-resolution mapping by recurrent intermating using Arabidopsis thaliana as a model (vol 142, pg 247, 1996). Genetics 143:1861-1861. [Basic description of the intermating method to increase recombination]

Marra, M., T. Kucaba, N. Dietrich, E. Green, B. Brownstein, R. Wilson, K. McDonald, L. Hillier, J. McPherson, and R. Waterston. 1997. High-throughput fingerprint analysis of large-insert clones. Genome Research 7:1072-1084. [Basic description of BAC fingerprinting methods]

Paterson, A.H., Y.R. Lin, Z.K. Li, K.F. Schertz, J.F. Doebley, S.R.M. Pinson, S.C. Liu, J.W. Stansel, and J.E. Irvine. 1995. Convergent Domestication of Cereal Crops by Independent Mutations at Corresponding Genetic-Loci. Science 269:1714-1718. [An early description of the comparative mapping of phenotypes/QTLs]

Rong, J.-K., F.A. Feltus, V.N. Waghmare, G.J. Pierce, P.W. Chee, X. Draye, Y. Saranga, R.J. Wright, T.A. Wilkins, O.L. May, C.W. Smith, J.R. Gannaway, J.F. Wendel, and A.H. Paterson. 2007. Meta-analysis of polyploid cotton QTLs shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. Genetics 176:2577-2588. [An early description of QTL meta-analysis across populations and environments]

Soderlund, C., S. Humphrey, A. Dunham, and L. French. 2000. Contigs built with fingerprints, markers and FPC V4.7. Genome Research 10:1772-1787. [Computational tools for physical mapping]

Thornsberry, J.M., M.M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E.S. Buckler. 2001. Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28:286-9. [An early description of association genetics methods and their application in plants]

Woo, S.-S., J. Jiang, B. Gill, A. Paterson, and R. Wing. 1994. Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. Nucleic Acids Res 22:4922-4931. [The first BAC library for a plant]

Wu, K., W. Burnquist, M. Sorrells, T. Tew, P. Moore, and S. Tanksley. 1992. The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theor Appl Genet 83:294-300. [Genetic mapping methods for heterozygous plants]

Zamir, D. 2001. Improving plant breeding with exotic genetic libraries. Nature Reviews Genetics 2:983-989. [New approaches to introgression of DNA from wild relatives of plants]

Biographical Sketch

Professor Andrew H. Paterson is jointly appointed in three Departments, Crop and Soil Science; Botany; and Genetics, at the University of Georgia. He directs the Plant Genome Mapping Laboratory, a multi-college unit that includes about 30 scientists, staff, and students. He received his B.S. (1982, Summa Cum Laude) in agriculture from the University of Delaware, and his MS (1986) in plant breeding and PhD (1988) in plant genetics from Cornell University. From 1989-1991 he was employed by the E. I. DuPont Company in agricultural biotechnology, also serving as an adjunct faculty member in the Department of Plant and Soil Science at the University of Delaware. In 1991, he joined the faculty of Texas A&M University, where he was appointed to the Christine Richardson Endowed Professorship in 1996. He moved to the University of Georgia in 1999, where he was appointed a Distinguished Research Professor in 2002.

Prof. Paterson conducts research in the area of plant genetics, using genomic tools and approaches to study crop improvement, molecular biology, and plant biodiversity. His lab has played especially important roles in the study of cotton, sorghum, sugarcane, peanut, Bermuda grass, rice, Arabidopsis, broccoli, cabbage, cauliflower and papaya, and contributed to knowledge of other crops. He has authored or co-authored 198 refereed publications, 46 book chapters, 6 patent applications, 2 books (with another pending), and given 146 invited presentations. As of August 2006, his work had been cited 6,727 times (according to Web of Science). He has led 'virtual centers' of national scope for sorghum and cotton, and participated in additional virtual centers for cotton, Brassica, maize, and rice. In addition to a long list of primary research grants, he has secured two competitive training grants (one as PI, one as co-PI), one regional economic development grant, and has helped 31 scholars to secure $1.14 million to study in his lab. He was the PI on a non-cash award from the US Dept of Energy Joint Genome Institute 'Community Sequencing Program' to sequence the 736 Mb sorghum genome, and a another non-cash award to conduct pilot-scale (500 Mb) sequencing of *Gossypium raimondii*, toward the formulation of an efficient strategy to sequence its entire genome. He has trained 89 undergraduates, 49 postdoctoral scientists, and conferred (or co-conferred) 38 advanced (M.S. or Ph.D.) degrees with 5 more in progress.

Prof. Paterson is a member of the American Association for the Advancement of Science, Genetics Society of America, Crop Science Society of America, and several other professional organizations. He is presently on the editorial boards of three professional journals (Genetics, Theoretical and Applied Genetics, and Tropical Plant Biology), and is frequently an *ad hoc* reviewer and/or panelist for many additional journals and granting agencies. He has served in several administrative roles at Univ GA, numerous elected roles in the International Cotton Genome Initiative including being the present co-chair of structural genomics workgroup, presently chairs the (worldwide) Sorghum Genomics Executive Committee and has served in scientific advisory roles to several national and international research initiatives. His contributions to science have been recognized with numerous awards, most recently including a Guggenheim Foundation Fellowship in Plant Sciences (2007), and the Cotton Genetics Research Award conferred by the National Cotton Council (2008).