Annu. Rev. Genomics Hum. Genet. 2004. 5:443–77
doi: 10.1146/annurev.genom.5.061903.180017

First published online as a Review in Advance on June 1, 2004

PLANT GENOMICS: The Third Wave

Justin O. Borevitz[1,2] and Joseph R. Ecker[1]
[1]*Genomic Analysis Laboratory, Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037; email: borevitz@salk.edu, ecker@salk.edu*
[2]*Department of Evolution and Ecology, University of Chicago, Chicago, Illinois 60637*

Key Words *Arabidopsis*, comparative genomics, functional genomics, microarray, knockout collection

■ **Abstract** Completing the primary genomic sequence of *Arabidopsis thaliana* was a major milestone, being the first plant genome and only the third high-quality finished eukaryotic genome sequence. Understanding how the genome sequence comprehensively encodes developmental programs and environmental responses is the next major challenge for all plant genome projects. This requires fully characterizing the genes, the regulatory sequences, and their functions. We discuss several functional genomics approaches to decode the linear sequence of the reference plant *Arabidopsis thaliana*, including full-length cDNA collections, microarrays, natural variation, knockout collections, and comparative sequence analysis. Genomics provides the essential tools to speed the work of the traditional molecular geneticist and is now a scientific discipline in its own right.

INTRODUCTION

The sequence is finished, we're done, and some now say, "Great, thank you very much, let's go back to doing interesting biology." But what do these strings of As, Ts, Cs, and Gs tell us? What are the words, sentences, paragraphs, chapters, and—most importantly—what are the messages in them? Others, such as Eric Lander, say, "I bought the book, but I can't read it." Clearly, much work remains to arrive at a comprehensive understanding of these completed genomes. So, the sequencing finish line abuts a new starting gate called functional genomics. If genome projects stopped after completion of the primary sequences, our goals to extract the meaning of these genome reference books, including the plant genome, would be only partially realized. To understand the complex instructions contained in raw sequence information of the plant genome, large-scale functional genomics projects are required. *Arabidopsis* is a supermodel genetic organism that is ideally suited and will serve as a reference for plant biology. Progress toward a complete understanding of gene regulatory networks shared among many plants is important for improving cultivated species and for understanding plant evolution.

The first wave of plant genomics was the era of single-gene sequencing, Restriction Fragment Length Polymorphism (RFLP)markers, low-density dot blot types of arrays (or northerns), and a one-gene, one-phenotype mentality. The second wave consisted of whole-genome sequencing, single nucleotide polymorphism (SNP) markers, and medium-density arrays with the continued goal of finding the genes that correspond to specific phenotypes. In the current third wave, we witness the comparative whole-genome sequencing from multiple related species that is merging with extremely high-density genotyping (such resequencing of individuals). Whole-genome arrays (WGAs) can monitor genome-wide transcription, alternative splicing, DNA binding, and epigenetic state. And slowly the third wave of genomics is giving rise to a new philosophy that the genome is dynamic and responds globally to developmental programs and environmental signals. We expect a network of genes to control complex phenotypes and look further into how these genes and genomes have evolved.

The annotation of the raw sequence shows that our current understanding of its function is continually improving. The high quality of the *Arabidopsis thaliana* annotation will strengthen broad comparisons involving proteome content, transcriptional patterns, and epigenetic state, with others plants and more distantly related model organisms. In this review, we discuss the initial findings of the *Arabidopsis* genome sequence and how, in the recent years following its completion, empirical annotation projects have dramatically improved the gene models. First, expressed sequence tags (ESTs), and full-length cDNAs are required to determine precise gene models. Collections of all open reading frames (ORFs) enable genome-wide biochemical studies. Developing high-density arrays to probe the entire genomic landscape enables the generation of a transcriptome atlas (135), new gene discovery (64, 157), identification of alternative splicing (27, 60), as well as DNA binding site analysis (21, 55, 112), DNA methylation profiles (84, 142), and natural polymorphism studies (14, 149). Knockout collections allow high-throughput reverse genetics studies and comprehensive forward genetic studies of the entire gene compendium (5, 126). Comparative genomics aims to identify function elements because they are more likely to be conserved through time while neutral mutations accumulate. Conserved regulatory regions, noncoding genes, and protein coding genes become evident when multiple genome sequences are aligned (29, 66, 131). Ultimately, to understand genome evolution, and processes leading to speciation, we must compare related genome sequences.

THE *ARABIDOPSIS* GENOME SEQUENCE

The sequence of the first plant genome was completed in December 2000, and it was the third complete genome of a higher eukaryote (140), after *Drosophila melanogaster* (1, 92) and *Caenorhabditis elegans* (30). Technically, the original *Drosophila* release was a high-quality draft and was finished after *Arabidopsis* (22). The following (unless indicated) summarizes aspects of a comprehensive article on

the *Arabidopsis* genome sequence (140). Details and references to original work can be found therein.

The *Arabidopsis* sequence represented 115 million base pairs (Mb) of euchromatin out of the estimated 125-Mb total. It was completed by the traditional bacterial artificial chromosome (BAC) by a BAC-based approach from a minimal tiling path of overlapping large insert clones. The sequence covered all ten chromosome arms, including parts of the centromeres. A major finding revealed by the five chromosome sequences was evidence for genome-wide duplication followed by gene loss and major translocations. Tandem duplications are also extensive. Overall, only one of three genes does not have a close family member. The *Arabidopsis* gene repertoire of 11,000–15,000 gene families is comparable in number to other sequenced organisms, highlighting the similarity of life's instructions that stem from our common single-celled ancestors. The total number of *Arabidopsis* genes was initially estimated at 25,490 and later revised to 30,700 (version 5 annotation). When the human genome was published (73, 144), what seemed remarkable was that *Arabidopsis*, the simplest plant genome, had a similar number of genes to humans! Or less remarkable was that the human genome had a similar number of genes to plants, depending on your perspective. The actual number of genes in any genome seems to be a contentious issue for some reason beyond bookkeeping. Certainly the number of genes does not reflect the complexity of the organism, as tempting as that initially seemed. Plants often contain many more genes than animals (mainly due to polyploidy or large-scale duplication). More likely, organism complexity is related to the levels of molecular interactions and regulatory circuitry using a similar genetic parts list. As new computation and empirical methods find more and more genes, and as the definition of a gene expands to include noncoding genes, we are not becoming more complex, even though our understanding is.

So what is unique about plants, as inferred from the first plant genome sequence? The *Arabidopsis* genome contains an amazing array of genes encoding enzymes involved in primary and secondary metabolism. These enzymes are the equipment of the plant chemical factories to build all required metabolic molecules and to generate an arsenal of specialized compounds. Because plants are sessile, they cannot move to avoid biotic attack or abiotic stress, or to find mating partners. Thus, they depend heavily on chemical signals. One example is the large gene family of cytochrome p450s, with more than 300 members involved in small molecule biosynthesis and detoxification (100). *Arabidopsis* also has a large number of transcription factors (~1500), many of which are in families unique to plants, such as the AP2/EREBP, RAV, NAC, ARF, and AUX/IAA families. Plants seem to lack many of the transcription factors families found in animals, such as nuclear steroid receptors. *Arabidopsis* has nearly 1000 serine/threonine kinases and more than 600 are receptor-like ser/thr kinases (91), but no receptor-like tyrosine kinases were found. Bacterial-like histidine kinases are also present. Some have been recruited as ethylene receptors or as red light photoreceptors, the phytochromes. Although showing sequence homology with histidine kinases, the plant phytochromes and several of the ethylene receptors have evolved ser/thr kinase activity (156, 159).

Plants have the extraordinary ability to photosynthesize, and hundreds of genes have been identified that are likely involved in light harvesting, chlorophyll biosynthesis, CO_2 fixation, or are a part of the two core energy-generating photosystems. Light also regulates development, in a process termed photomorphogenesis (94). The genome contains, in addition to the red light–absorbing phytochromes, blue light–absorbing cryptochrome, and phototropin photoreceptors, as well as hundreds of putative downstream light-signaling proteins (20). We know the function of only a few dozen light-signaling proteins.

Other unique gene functions of plants and yeast are creating electrochemical gradients using mainly proton-type ATPases, whereas *C. elegans* and *Drosophila* use mainly sodium-type ATPases. Thus, transport is usually coupled with protons rather than sodium ions. In addition, compared to animals, water channels (aquaporins) are highly overrepresented in the *Arabidopsis* genome. There are also cellular differences between plants and animals at the genome level. No genes encode intermediate filaments in *Arabidopsis*, whereas actin and α and β tubulin are present. The plant cell wall surely is unique, and homologues of animal cytoskeleton anchorage proteins (which link with the extracellular matrix) have not been seen in the *Arabidopsis* genome. Plants also seem to have a different repertoire of small G-protein-signaling molecules. Different mechanisms are used to protect plants and animals against their biotic environment. No major histoincompatibility complex genes or antibody-like genes were identified in plants; however, plant-specific nucleotide binding site leucine rich repeat (NBS-LRR)-type disease resistance (*R*) genes are abundant and can be grouped into subfamilies (89). The 149 *R* genes are present at several loci throughout the *Arabidopsis* genome and occur as singletons or in highly polymorphic clusters.

The sequence of the first plant genome reveled extensive horizontal transfer of genes from an ancestral cyanobacterium-like endosymbiont to the plant cell nucleus. New studies show that up to 18% of all plant genes originated from this engulfment, which resulted in the chloroplast (86). Mitochondrial genes have also been transferred to the nucleus. Strikingly, a large translocation of 620 kb brought several duplicated and rearranged copies of the mitochondrial genome near the centromere on Chromosome 2 (134). Compared to *Drosophila*, *C. elegans*, and yeast, there are about 150 protein families that are unique to *Arabidopsis*.

EXPRESSED SEQUENCES

Most of the analysis in this landmark genome sequence paper was based on computational gene predictions (140). Empirical knowledge of which regions in the genome truly encode genes was limited. In *Arabidopsis*, when more than 10,000 full-length cDNA sequences became available, 32% of the predicted gene models were incorrect (157). Gene prediction algorithms rely on comparison with known proteins, or long ORFs and consensus splice acceptor sites, or both. Comparison with known proteins is biased toward finding more of what is already known,

whereas gene discovery based on ORF models are biased for the parameters specified in the model. Therefore, small proteins or noncoding genes with nontraditional splice acceptor sites are often missed. A study of seven gene prediction programs showed bias in several aspects, including missed exons, wrong exons, and misprediction of exon boundaries, with error rates between 8% and 32% (116). Each program had more or less bias in different aspects, but multiple prediction methods can help (127). Empirical gene annotation removes this bias but relies on capturing an RNA that corresponds to the gene. Then a "correct" gene model can be identified and may lead to the discovery of new gene families. Similarly, ESTs often reveal transcription from "intergenic" regions and may not contain ORFs. These would not be identified as genes by gene prediction programs alone. The "correct" gene model, as determined by expression data, is only one of many that are possible from a given transcription unit; however, it exists, whereas predicted gene models may not. Alternative splicing, including alternative start and stop sites, are often found when enough sequence data are available. Usually, the number of alternative forms identified is related to the amount of expressed sequence that is available for a given organism rather than to the specifics of that organism (see Figure 3 for the number of ESTs from different organisms). Basically, the more you look, the more you find. For example, EST data are abundant compared to full-length cDNA sequence data, and alternative start and stop signals seem to be the common form of alternative splicing. When multiple copies of full-length cDNAs are available, internal alternative splicing is also found (163). Thus, an important complement to raw genomic sequence is a large collection of expressed sequences. EST collections and multiple clones of full-length cDNAs are very useful for empirical gene annotation.

Large collections of full-length cDNA have been sequenced in *Arabidopsis* (157), rice (67), *C. elegans* (110), *Drosophila* (130), mouse (19), and human (133). Version 4.0, and now version 5, of the *Arabidopsis* annotation are substantial improvements because they consider much of the full-length cDNA sequence data, many more ESTs, and homology to sequence from *Brassica oleracea* (154). Generally, the deeper the expressed sequence coverage is the more accurate the genome annotation will be. However, at some stage, depending on a particular cost-benefit analysis, continued sequencing is futile. Sequencing of expressed libraries eventually tops out and the number of new genes, or splice forms, found per additional read becomes too costly. In *Arabidopsis*, 7447 predicted genes are still hypothetical and have no evidence of expression from sequenced clones (157). Alternative methods (discussed below) are needed to confirm these and identify those not predicted.

Plant Gene Collection

A comprehensive catalogue of ORF clones is a valuable functional genomics tool and an important goal in several model organisms (19, 110, 130, 133, 157). For *Arabidopsis*, there is currently a collection of >11,000 full-length ORFs (http://signal.salk.edu), which are cloned into a recombination-based universal

plasmid vector (81, 157). These clones can be ordered from public stock centers and used for many purposes, including protein-protein interactions, protein expression, or in vivo experiments such as cell culture or in planta transformation. Vectors to overexpress fusion proteins containing epitope tags or fluorescent proteins are compatible with cDNAs cloned into recombination cassettes. Other uses of ORF collections include the systematic production of global two-hybrid interaction maps (43, 57, 143) or the construction of protein arrays that can test for protein interactions and small molecule binding (164). Another way to use full-length cDNA collections is activity screening. In a recent example, approximately 20,000 mouse cDNAs were tested for their ability to activate a reporter gene in high-throughput transfection assays (23).

WHOLE-GENOME ARRAYS ARE UNIVERSAL ANNOTATION TOOLS

One important use of a high-quality finished genome sequence is to construct whole-genome arrays (WGAs). WGAs are oligonucleotide arrays (or chips) that span or tile the entire genome sequence. The only limitations are the number of features (unique oligonucleotides) and the cost of the arrays. Current arrays contain a maximum of 1.3 million features and cost approximately $400. Robust statistical analysis methods are crucial to glean meaningful data from millions of data points. Fortunately, methods are being developed at rapid pace, such as http://www.bioconductor.org (56) and http://www.dchip.org (77). WGAs can capture the complete repertoire of a particular RNA sample and are used for gene discovery and characterization of known genes. Because WGAs are designed from full genome sequences they are not biased for previously known or predicted transcribed regions. Therefore, they have many uses beyond standard gene expressions. Soon it will be possible for a single array to contain features that cover nearly every base of the *Arabidopsis* genome. *Arabidopsis* is ideally suited for this type of universal array due to the high gene density (~4.5 kb/gene) and small genome size (125 Mb). As discussed below, results from several experiments measuring de novo RNA expression and DNA polymorphisms or binding can being integrated on a single platform. Thus, an *Arabidopsis* WGA is an ideal tool to provide broad genome annotation (Figure 1).

Gene Discovery and Gene Model Confirmation

In all organisms, new approaches are required to identify clones and confirm the remaining genes that are expressed at low levels and/or that are tissue specific. One approach aims to simply confirm expression of hypothetical genes via reverse transcription polymerase chain reaction (RT-PCR) from various cDNA populations and then to extend them via 5′ and 3′ rapid amplification of cDNA ends(155). This one-by-one approach is time-consuming and is biased for sequences initially detected

by the gene prediction algorithms. Long oligos can be designed to fit potential exons and can be probed for evidence of expression (129), but this is also limited to predicted exons. Arrays designed with PCR fragments spanning all unique regions of the chromosome have also been used (114). This study identified twofold more transcription on human chromosome 22 than expected based on ESTs or gene prediction algorithms. Another approach is to use high-density oligonucleotide tiling arrays that cover genomic DNA (entire chromosomes or the whole genome) with 25 base pair oligonucleotide features (25-mers) chosen in an unbiased way with regard to the potentially expressed sequences (64, 157). When cDNA populations are labeled and hybridized to these arrays it is possible to detect novel expression throughout the entire genomic region. Thus, this method can detect novel expression signatures and may be sensitive to genes expressed at low levels (claimed 1:100,000 to 1:200,000 sensitivity). Kapranov et al. (64) used tiling arrays designed against human chromosome 21 and 22. They predicted up to tenfold more expression than previously observed by cDNA sequencing or gene prediction. WGAs have also been used to determine expression patterns in the malaria parasite *Plasmodium falciparum* (75), although in this case the analysis was limited only to predicted genes. Here arrays that contain features for both DNA strands spaced at approximately 150-bp intervals were used to profile nine stages of the parasite in both the mosquito and human hosts. In *Arabidopsis*, Yamada et al. (157) profiled four tissues using WGA sets that cover both strands of the entire *Arabidopsis* genome at 25-bp resolution. This high-resolution hybridization information was used to correct computational gene models, determine novel 5′ and 3′ untranslated exons, identify many new intergenic and antisense transcripts, and identify new genes located in the centromere. Because WGAs can monitor genome-wide expression on both DNA strands, they are ideal tools to investigate the emerging regulatory mechanism of antisense transcription. Until now, antisense transcription from cDNA sequencing was seen only on a limited basis (69, 128, 139).

Gene models predicted from hybridization signals can be used to direct PCR primer synthesis. Transcription units can then be amplified by RT-PCR and the ORFs can be introduced into recombination-based cloning vectors. This way, one can confirm expression signatures on arrays, annotate their precise gene models, and capture ORF clones. New genes that escape gene prediction can also be captured in this way. The sensitivity of the arrays can be further increased by simply performing more replicates, through improved statistical models that account for innate hybridization differences among 25-mers, and by improved labeling and/or RNA subtraction techniques to capture rare messages. The use of whole-genome tiling arrays allows many tissues and treatments to be screened, increasing the chances to identify the rare transcripts. One WGA hybridization experiment may be comparable to deep EST sequencing (100,000 transcripts). Techniques to capture nonpolyadenylated RNA promise to reveal important atypical transcripts such as small RNAs or perhaps micro RNA precursors. However, mini-exons or mature micro or other small RNAs (<25 bp) will likely be missed using the array hybridization technique alone. Tiling arrays with features at higher densities (such

as every five bases) will provide better sensitivity and resolution to determine intron/exon boundaries more precisely.

Transcriptome Atlas

As more tissues and developmental stages are profiled a gene expression atlas can be created that describes the expression pattern of every gene in the genome. Such an atlas is a tremendously useful tool to the biologist interested in the expression profile of his/her favorite gene. An atlas can be queried for genes that fit a particular expression profile. Knowledge of the timing and expression pattern of genes allows potential networks to be created. The biological function of unknown genes in such a network can be inferred under the assumption that genes expressed similarly will be involved in a similar process. The larger the collection of samples, and the number of independent biological replicates for each sample in a gene expression atlas, the more useful and reliable it will be (162). Using a single comprehensive/universal array allows many groups to contribute data to expand the atlas. Commercially available high-density arrays are highly reproducible, a requirement for building to a unified transcriptome atlas.

One example is the publicly available gene expression atlas for mouse and human, which contains 91 samples run in duplicate (135). In *Arabidopsis*, an initial attempt to create a gene expression atlas utilized spotted PCR products on arrays containing 11,900 clones (152). Various groups contributed RNA samples that were hybridized to 534 arrays at the *Arabidopsis* Functional Genomics Consortium. Due to quality issues, only 397 arrays and 5698 spots could be analyzed. Raw data, various analysis scripts, and viewers are available (http://www.arabidopsis.org). Difficulties with background effects, spatial artifacts, normalization, and plate bias were found in these arrays and attempts were made to correct this (34). Another major problem was that biological replicates were not always included; sometimes only technical replicates were included. Without biological replicates the differences may not be reproducible and may merely reflect the normal variation inherent in biological systems. Overall, the utility of this initial *Arabidopsis* gene atlas may be limited; however, lessons learned can guide the construction of a future transcriptome atlas. A new *Arabidopsis* gene expression atlas has recently been created (AtGenExpress, ftp://arabidopsis.org/Microarrays/Datasets/AtGenExpress). This atlas was made on a high-density oligonucleotide array with 22,000 genes. It contains greater the 101 different developmental stages and/or genotypes, all collected in a single lab, with three biological replicates each (D. Weigel, M. Schmid, and J. Lohmann, unpublished data).

Alternative Splicing

WGAs should also be an effective tool for identifying alternative splicing. Mixtures of alternatively spliced messages could be detected on WGAs if exons were expressed at different levels within a single sample. There is sure to be detectable alternative splicing across tissues or developmental stages. In this case exons can

be tested directly for different expression levels once differences in overall gene expression are considered. The effect of a genetic mutation on alternative splicing can be observed when mutant and wild-type samples are compared. This is especially useful for analyzing potential splicing factor mutants (27). Furthermore, WGAs contain probes to previously unidentified exons because they scan entire genomic sequence. The RNA samples in which the new transcript forms are detected can then be used as templates for RT-PCR amplification and sequencing. Alternatively spliced clones can be captured in the same manner as new transcription units (see above). A recent study elegantly used exon junction arrays to detect alternative splicing in 74% of multiexonic genes in humans (60).

A few studies have looked at alternative splicing in *Arabidopsis* by analyzing ESTs and cDNA sequence (48, 165). New algorithms took advantage of recently released expressed sequence and could identify more than 1000 alternative splice forms as well as correct many of the gene models. Further improvements in analysis methods may help, but ultimately they are limited by the amount of expressed sequence data. We need new methods to identify alternative splicing that search a broader collection of samples. WGAs can both discover new and assay known alternative splicing from many diverse tissues, treatments, and developmental stages.

Chromatin Immunoprecipitation

Identifying the transcribed parts of the genome is another step toward understanding the genome's function.A greater level of understanding of the genome comes from knowledge of the sequences that function as DNA binding sites for various structural and regulatory proteins. Chromatin immunoprecipitation (ChIP) is a popular method to isolate DNA that is bound to a transcription factor or DNA binding protein of interest. Cross-linked DNA protein complexes are isolated from plant tissue and sonicated to shear the DNA to approximately 500-bp fragments. Using an antibody that is specific to a certain transcription factor, or to a fused epitope tag, the protein chromatin complex can be immunoprecipitated (IP). Then, cross-linking is reversed and generally PCR is used to detect whether a potential target sequence has been pulled down, providing in vivo evidence of a protein-DNA interaction that is either direct or indirect. When the ChIP product is labeled and hybridized to arrays (so-called ChIP chip), one can determine the genome-wide binding sites for a given transcription factor (sometimes also referred to as location analysis). It is important to consider the appropriate control for ChIP chip experiments, specifically one that controls the nonspecific binding of a particular antibody. One approach is to use a mutant that does not contain the epitope. Here nonspecific binding is pulled down in both the wild-type and mutant control using the same antibody. It is likely that even moderately specific antibodies can be used because only the difference in binding patterns along the genome between mutant and wild type are of interest. In this case a potential drawback is that secondary

effects caused by the mutation will also be identified. To get around the protein of interest is fused to an epitope-tag. Here the control is a line expressing only the epitope tag. Because approximately 500 bp of DNA that flank the DNA binding site is pulled down, it should be relatively easy to detect on WGAs that have probes every 25 bp or less. However, with this ~500-bp resolution it is then difficult to identify the specific binding sequence that interacts with contact residues of the DNA binding protein. This often requires in vitro methods such as affinity selection (97) or gel shift assays. Hybridizing labeled proteins to short oligonucleotide arrays, which can be made double stranded, could serve as a high-throughput method of affinity selection and could be complementary to in vivo ChIP chip. WGAs are potentially suited for this purpose because they may contain the sequence of the actual binding site.

Currently, arrays designed with genomic PCR products are used for ChIP chip in yeast (54, 112) and in humans (55, 87, 147, 148). In yeast, a comprehensive study was performed to identify genome-wide binding sites for nine cell cycle components. Between 30 and 290 sites were identified for each factor from arrays designed to yeast intergenic regions, many of which show cross-binding and can be ordered into a transcription factor binding site network (54). Other studies are performed with different arrays that may not cover all genomic regions, making comparisons across experiments difficult. Human chromosome 21 and 22 tiling arrays were used to identify binding sites for Sp1, cMyc, and p53. These sites resided within coding regions and 3′ and 5′ regions and co-localized with noncoding RNAs (21). So, WGAs are also the preferred tool for ChIP chip studies, providing consistency among arrays, comprehensive genome coverage, and multiple probes for each potential binding region.

An extension of ChIP chip, to determine RNA binding protein specificity, was recently reported (41a). Here epitope tags were added to five yeast RNA binding proteins. The immunoprecipitated RNA was labeled and hybridized to DNA arrays. Each RNA binding protein showed remarkable specificity to the subcellular location of the RNAs they bound. WGAs can also be used effectively for this purpose.

Methylome

Characterizing the chromosome-wide epigenetic state is also an important step in the functional annotation of the linear DNA sequence. Because *Arabidopsis* was the first methylated genome to be sequenced (140), it is an ideal model system for epigenetic studies (113). Cytosine methylation is often associated with repressed chromatin state including, but not limited to, heterochromatic telomeres and centromeres (58, 84). In addition, histone modification is also important in both global and local regulation of gene expression (32). In plants, methylation on histone lysine 9 and 3 can modify gene expression (113). The well-known histone modification, acetylation, is linked with DNA methylation as DNA methylases are found in histone deactylase complexes (137).

Can the states of DNA methylation and chromatin modification be surveyed on the genome-wide level? There are two popular methods for assaying DNA methylation. Southern blots analysis of DNA digested with methyl-sensitive or -insensitive enzymes can reveal different-sized fragments, indicating methylation. In a more direct approach, bisulfite treatment converts cytosine to uracil but methyl cytosine is protected. Subsequent PCR amplification, cloning, and sequencing can show which sites were methylated as cytosine residues remain in the sequence. Sequencing multiple clones can indicate the proportion of methylation at a particular base. Potentially, both methods could be used with WGAs to survey global methylation patterns. For example, WGAs could be hybridized with labeled DNA that has been digested with either a methyl-sensitive or -insensitive four-base cutter restriction enzyme. A significant increase in hybridization signal on array features from the methylation-sensitive treatment would indicate methylation. In this case, the number of sites surveyed is limited to the bases recognized by the enzyme. Labeling and strong hybridization of genomic DNA to oligonucleotide features after bisulfite treatment indicates heavy methylation if most cytosines in a 25-mer are protected from treatment. Other approaches for genome-wide chromatin analysis involve ChIP chip. Specific antibodies are available to modified histones with either acetylation or methylation on various lysine residues. ChIP chip studies using these antibodies should complement methylome analysis via differential enzyme digestion or bisulfite treatment to highlight the interplay between histone and DNA methylation (61).

In a recent study that profiled genome-wide methylation patterns, another approach was taken, using sucrose gradient sizing and labeling of small fragments of DNA that had been digested with a methyl-sensitive restriction enzyme. The products were then hybridized to an array that contained 384 PCR fragments across the genome (142). When compared to a methylation-deficient mutant *cmt3*, Tompa et al. determined that transposons were preferentially methylated by CMT3. Another study showed strand-specific methylation at the centromere but not at rDNA loci, both of which are heterochromatic (84). This analysis was done using bisulfite treatment and strand-specific PCR. The PCR reactions were sequenced in 20 cases or treated with methylation-sensitive restriction enzymes that had cytosine in the recognition site. The amount of product digested indicated the relative amount of methylation. Methyl cytosine antibodies were also used to precipitate methylated strands of denatured DNA. Here strand-specific PCR was used to determine which strand was methylated, and it provided consistent results relative to the strand-specific PCR after bisulfite treatment. Finally, to access genome-wide strand-specific methylation bias, nick transcription was used. The reaction products were hybridized to arrays made from 255 overlapping BACs on *Arabidopsis* chromosome 2 and 43 BACs on chromosome 4. The centromeres, but not the rDNA regions, again showed strand-specific bias in methylation. Methods such as these could be applied to whole-genome tiling arrays to comprehensively survey the methylome at high resolution in both wild-type and various DNA methylation mutants (18).

DNA Polymorphism

A further characterization of the raw genomic sequence aims to reveal the comprehensive pattern of DNA polymorphisms within a species. The natural variation in *Arabidopsis* is abundant in neutral DNA markers and also in those that cause phenotypic changes (6). Initial studies focused on single loci. Now several studies survey genome-wide polymorphisms. Schmid et al. sequenced more than 10,000 ESTs from libraries of six different *A. thaliana* accessions and more than 600 sequenced tagged sites from 12 accessions. They identified 8051 potential SNPs and 637 indel polymorphisms (124). Cereon Genomics performed a two-times shotgun coverage of the Landsberg *erecta* accession that found more than 50,000 potential polymorphisms (59). Magnus Nordborg has been funded to sequence 1500, ~500-bp fragments from 96 accessions. Currently, more than 900 fragments have been analyzed, and more than 15,000 polymorphisms have been identified (Magnus Nordborg, personal communication; http://walnut.usc.edu). Array hybridization can also be used to identify single-feature polymorphisms (SFPs). SFPs are identified when a particular 25-mer feature has a significantly different hybridization signal between at least two accessions (14, 149, 150). Recently, we demonstrated that 4% of features could easily be called SFPs between two accessions. We have made more than 19,000 SFPs available from 14 accessions using commercially available arrays (J.O. Borevitz & J.R. Ecker, unpublished data; http://naturalvariation.org/sfp). These SFPs fare well when compared with the available sequence data, revealing low false positive rates and moderate false negative rates (14; J.O. Borevitz & J.R. Ecker, unpublished data). Using WGAs to identify hybridization polymorphisms is analogous to a rough resequencing of the entire accession, allowing the identification of hundreds of thousands of SFPs in a single experiment.

If required, resequencing arrays can be used to determine the precise single nucleotide or deletion polymorphisms (24). In this case oligonucleotide arrays are designed as single base-pair tiles of both strands and include all three possible mismatches. Resequencing arrays interrogate 25 bp of sequence with 200 different oligos (25 bp × 4 bases × 2 strands). Currently they also require loci to be amplified specifically from long-range PCR products or from appropriate BAC or cosmid clones. A huge project used such arrays to resequence 22 Mb of human chromosome 21 from twenty unique chromosomes (104). Recently, these arrays were used to sequence the corresponding chimpanzee chromosome, revealing many insertion/deletion polymorphisms as well as SNPs (38). Perlegen has completely resequenced many human individuals across their entire genome with this method. Not all sequences (i.e., simple repeats and poor hybridizing sequences) can be read by hybridization, but as array density increases large regions of sequence can be read on single arrays (currently ~150 kb). This accuracy is costly compared to SFP genotyping: 200-fold in arrays and individual clone/product amplification. A single *Arabidopsis* WGA could identify very dense SFPs across the entire 120 Mb using whole-genome labeling methods.

Comparative Genome Hybridization

Comparative genome hybridization (CGH) has been used to detect deletions often in cancer cell lines. This is done by labeling DNA from different samples and hybridizing it to cDNA arrays (106) or to arrays of BAC clones that cover the entire genome (105). Using this approach, very large deletions and duplications have been identified. Oligonucleotide arrays have been used for the same purpose in microorganisms (65, 119) and recently in humans (13a). Naturally occurring gene-size insertion/deletion polymorphisms can easily be identified on high-density oligonucleotide arrays (14). Hundreds of deletions are found between any two *Arabidopsis* accessions that suggest candidate genes for causes of natural variation. Not surprisingly, many natural deletions are in genes that encode transposons (14). In a similar manner we determined the precise location of fast neutron-induced deletion mutations (J.O. Borevitz & J.R. Ecker, unpublished results). Thus, an additional use of whole-genome tiling arrays is high-resolution array CGH. Figure 2 shows how several experiments performed on a single WGA platform can be integrated to provide detailed functional annotation along the genome.

NATURAL VARIATION

New technologies are generating high-density polymorphism data that will be useful in identifying functional changes in candidate genes (9) and for haplotype analysis (11, 98, 108). Once polymorphism data is available across the genome it serves as background for genome-wide examination of the patterns of natural selection. Regions showing relatively high levels of polymorphism may be undergoing rapid evolution, whereas regions that lack variation may have undergone a recent selective sweep. When testing for selection the patterns of variation at candidate loci are compared against the genome-wide distribution, thus controlling for effects such as population expansion/contraction or recent migrations which should affect the entire genome in similar ways.

Several diversity estimates can be calculated genome wide. The Fst statistic measures between versus within population levels of diversity and has been calculated on 26,530 human SNPs that define the genome-wide distribution (2). When compared with coalescent simulations, 174 candidate genes showed significant evidence of being under selection, including the *CFTR* gene, which is associated with cystic fibrosis (115), and *PPARG*, which is associated with type 2 diabetes (7).

In addition to linkage disequilibrium studies in outbred populations, high-density polymorphism data can be used for mapping quantitative traits in pedigrees or in crosses between inbred lines. QTL studies with saturating molecular markers in mapping populations can more precisely define the interval of large-effect causative loci. However, abundant markers cannot overcome the need for recombination events to break up genetic intervals; this requires large sample sizes.

When complex traits have low heritability, replicate measures must be performed on many lines. One method to increase the number of lines without additional cost involves pooled genotyping or bulk segregant analysis (26, 90). We used bulk segregant mapping with high-density array genotyping to localize several Mendelian traits (14; J.O. Borevitz & J.R. Ecker, unpublished data) and moderate effect quantitative traits. For quantitative traits, selective genotyping is performed on pools of lines with extreme phenotypes, a method we call eXtreme Array Mapping (XAM) (152a). Because chip genotyping can identify tens to hundreds of thousands of SFPs it may also be useful in fine-mapping strategies where markers become limiting. Here recombinant genotypes are identified prior to pooling.

GENOME-WIDE KNOCKOUT COLLECTIONS

Several types of experiments can be performed when a saturating loss of function collection is available. For mapping studies, knockout lines can be ordered for all candidate genes in an interval, and multiple alleles are often available for each candidate gene. In addition, a knockout mutant can provide a crucial second allele when only a single EMS allele is available. Here observation of similar phenotypes in both alleles provides confirmation that the correct gene was identified. The null background is suitable for transgenic studies that investigate altered expression patterns or test altered proteins. Often redundancy can confound genetic studies by masking phenotypes. This problem can be dealt with by creating double, triple, or greater knockout mutations among multiple gene family members (5).

Functional genomics screens aim to identify all the genes contributing to a phenotype of interest. They can be performed quantitatively across the entire collection to ask how many genes affect a given process. This approach was taken in yeast using knockout collections (42, 132, 151). Forward genetic screens have also been performed in *C. elegans* (37, 63, 107) and *Drosophila* cell lines (28, 82) with RNAi knockdown approaches. Quantitative fitness data across all yeast genes globally revealed the extent of genetic redundancy due to gene duplication (47).

Several projects in *Arabidopsis* have created a near saturating collection of sequence-indexed knockout mutations. Alonso and colleagues have recovered flanking sequences for 145,417 insertion events that can be mapped to the genome (5). They reside in 21,858 known genes (including 500 bp of promoter), which have at least one insertion. Furthermore, more than 17,000 genes have two or more insertions. The T-DNA Express database (http://signal.salk.edu) provides a user-friendly tool to quickly identify insertion mutations for a gene of interest and has been widely used (more than 1.6 million hits since September 2001; H. Chen & J.R. Ecker, unpublished data). The many insertion mutations that do not currently map to known genes, i.e., those which lie within the intergenic "dark matter," may later prove useful to identify functions for new genes. These insertions may disrupt noncoding genes, such as miRNA precursors or important regulatory regions.

A comprehensive analysis of the pattern of these insertions across the genome revealed hot and cold spots, including clear under-representation in all five centromeres and bias toward promoter regions. All of the Salk T-DNA insertion mutations are freely available in public stock centers and do not require material transfer agreements. Analyzing segregation data for these and others lines could provide a rough estimate of the fitness consequences of the insertion if it differs widely from the expected 1:2:1 ratio. A collection of homozygous knockout lines in all of the predicted genes would allow for high-throughput functional screens, as has been performed in yeast, *C. elegans*, and *Drosophila*. We feel this is a high priority for the *Arabidopsis* community. Comprehensive phenotypic data from this deletion set will prove an essential part of functional genome characterization.

Several other studies have also created sequence-indexed collections of knockout mutations. Syngenta Inc. (126), a French program (INRA/Genoplante) (120), a German project (GABI-Kat) (79), the Japanese group in RIKEN (72a), the John Innes Center in the United Kingdom (141a), Cold Spring Harbor Laboratory (http://genetrap.cshl.org/), and the Wisconsin group (http://www.hort.wisc.edu/krysan/DS-Lox/) have all generated flanking insertion sequences. As of April 2004, there are more than 330,000 flanking sequences that can be mapped to the genome that hit >90% of the currently known genes. Mutant plants from some of these projects may require specific agreements; however, nearly all sequences can be searched on our web site (http://signal.salk.edu). Most of these collections have been made in the Columbia accession, where mutations in several genes can be combined without regard to background effect. That said, screening knockout alleles created in different backgrounds may reveal phenotypes modified by background effect. A complementary approach involves screening gain-of-function mutations, so called activation tagged lines (146) that can be easily created or ordered from stock centers.

COMPARATIVE GENOMICS

As genomics tools are developed in the reference plant *Arabidopsis thaliana*, sequence from other plants is becoming available. *Arabidopsis* will continue to be the genetic workhorse, where the initial function of most plant genes will be characterized. How can genomic sequence from other plants be used to improve the *Arabidopsis* annotation? Comparative genomics promises to identify the functional elements in a genome based on the assumption that these elements are preferentially conserved through evolutionary time (for more information, see the article in this issue on comparative genomics by Webb Miller: Comparative Genomics). According to the neutral model, sites will mutate at random. Changes that are even mildly negative should be eliminated by selection. Thus, over time, sequences not under selective pressure to remain the same will diverge due to drift. Protein coding regions tend to be conserved as most substitutions are deleterious; however, synonymous sites are often free to vary. Regulatory regions and noncoding genes

also tend to be conserved but are more difficult to identify. Although conservation within coding regions is well defined, regulatory regions do not follow known rules and thus are more difficult to model. The evolutionary distance between the species is important to consider as this determines the expected number of changes. For distantly related species it is likely that every base will have mutated at some time in the past. In this case, aligning intergenic DNA sequence is difficult and alignments are made in coding regions to the potential protein translations. Important functional domains in proteins can be identified by conservation, and tests exist to determine which sites are under selection (158). Positive Darwinian selection can also be identified from comparisons between closely related species or from variation between populations. Here an increase in amino acid substitutions is seen relative to synonymous changes.

Comparative genomics in plants is still in its infancy. Below we discuss methods to improve genome annotation using comparative techniques that are based on evolutionary models. We then discuss a broad evolutionary comparative study in mammals about the *CTFR* locus and recent genome-wide comparative analysis in yeast as an example for future studies in plants.

Ka/Ks Test to Identify Conserved Exons

DNA sequences that code for protein show different patterns of variation from noncoding regions. A simple evolutionary test, the Ka/Ks test, can be used to confirm exons by looking at the frequency of substitutions at sites that change amino acids (nonsynonymous) relative to the frequency of substitutions at silent sites that do not alter amino acids (synonymous). This test works because most amino acid changes result in a loss of protein function and are removed by purifying natural selection. Thus, a protein often has an excess of synonymous substitutions and the Ka/Ks ratio is around 0.065 (95). Here Ka here is the proportion of bases that result in amino acid changes out of the total possible sites that could result in amino acid change. Similarly, Ks is scaled by the total number of possible synonymous changes in the sequence. Thus, noncoding regions often show the neutral expectation, where Ka/Ks $=$ 1. Advanced Ka/Ks tests also consider GC content, differences in transition/transversion rates, and codon bias (46).

For coding region identification, two orthologous sequences should be at an ideal evolutionary distance such that there has been enough time for changes to occur. The larger the number of total base substitutions, the greater the statistical power of the Ka/Ks test. However, when divergence is too high, sequence alignment may be difficult and sites may have mutated more than once, violating the assumptions of standard Ka/Ks test. The Ka/Ks test was applied on a limited scale to mouse and human gene comparisons that had an average DNA identify of 86.5% (95). Later the test was used for whole-genome comparisons of 12,845 putative human mouse orthologs (145).

How well does the Ka/Ks test perform? To reject the neutral model and identify a true exon the Ka/Ks test should be significantly less than 1. In the original

comparison (95), 1244 exons were known. The Ka/Ks test identified 90.5% of them, revealing that the test has good sensitivity. As expected, the test performed best for longer exons. It also performed best when the sequences were between 10% and 15% diverged, thus illustrating the appropriate evolutionary range for calling exons when using Ka/Ks (95). How often is an exon called that is not an exon, i.e. what is the false positive rate? Simulations of random sequences at different exon lengths and divergence rates showed the false positive rate to be between 1%–5%, with longer exons and moderate divergence performing best. Multiple sequences provide more data and make this test extremely powerful even for short exons. Exons not under purifying selection such as pseudogenes will be missed by Ka/Ks tests. Loci under rapid evolution may have Ka/Ks ratios greater than 1 if amino acid substitutions are more prevalent than expected across the entire exon. Finally, the Ka/Ks test should be used in combination with other prediction methods to precisely call intron/exon boundaries.

Mammalian Comparative Study

We review two large-scale comparative studies in mammals and yeast to highlight the power of comparative genomics as examples that can soon be applied in plants. A major comparative analysis was recently reported covering 1.8 Mb of contiguous human sequence (across seven genes including the cystic fibrosis gene *CTFR*) from 13 vertebrates: human, chimp, baboon, cat, dog, cow, pig, rat, mouse, chicken, takifugu, tetraodon, and zebra fish (141). A substantial number of conserved noncoding regions (CNRs) were identified. Many of these could not have been found using pair-wise comparisons. This study illustrates the utility of examining a range of sequence distances for comparative analysis and emphasizes that the total branch length of the species tree is a good measure of the ability to find conserved sequences. As long as the sequences can be aligned, the longer the branch length the more power to identify conserved regions because more substitutions have occurred. In this study, human-fish genome alignments were mainly to coding regions, whereas human-chimp alignments were nearly identical and provided little information toward finding conserved regions. The human-chicken comparison was the most informative pair and allowed identification of 40% of the total conserved regions (identified from 13 vertebrate comparisons); however, sequence from multiple species is an important feature in obtaining the alignment. The genome sequences of human (73, 144), mouse (145), rat (42a), fugu (10), and low-pass sequence from dog (68) are now available. Currently, mammalian genomics is poised for broad scale comparative analysis that will improve annotation of all genomes. Evolutionary approaches offer a framework in which to consider multiple genomes and integrate multiple data sources.

Genome-Wide Yeast Comparative Analysis

In 2003 we saw the first genome-wide comparative studies where the entire genome sequence could be aligned from several species. Two papers reported the sequences

of either three additional *Saccharomyces* genomes at six- to seven-times shotgun coverage (66) or five additional genomes at two- to three-times coverage (29). Both papers improved the *S. cerevisiae* reference annotation with revised estimates of gene number and identification of many likely pseudogenes that were misannotated as genes. In addition to genes, phylogenetic footprinting was used to predict regulatory regions. These included known elements and many novel ones. Variation in gene expression can be explained by combinations of the identified regulatory regions. The Ka/Ks test was applied genome wide to the yeast data, yielding an average of 0.11. Rapidly evolving genes that showed an excess of amino acid changes had Ka/Ks values above 0.69 (across the entire gene) and included genes involved in stress response and sporulation (66). Thus, genome-wide Ka/Ks tests can also be used to screen for genes under positive selection because the genome-wide distribution is known.

In general, sequence conservation was high throughout the coding regions, an advantage of *Saccharomyces* as coding regions make up 70% of the genome. Indeed this conservation was used to align the mostly syntenic genomes. There are trade-offs when choosing the evolutionary distance between species for comparative studies. One would like to have as many substitutions as possible so that conservation is not the result of chance. The more genomes in the comparison the more diversity and the lower the probability of chance conservation, but this comes at a greater cost. In yeast where coding regions are extensive and genes are compact, alignments can be made but divergence is also high enough that variation at most sites has been sampled. This signal-to-noise ratio becomes a problem for large-genome organisms and for organisms with low gene density because alignments cannot be anchored by the coding regions (66). In this case, the only solution is to sequence many closely related individuals so that ample variation can be captured and alignments can still be made.

EXPRESSED SEQUENCE TAGS

Several large EST projects have been completed or are under way for a variety of plant species (http://www.ncbi.nih.gov/genomes/PLANTS/PlantList.html, http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html) (Figure 3). As of April 2004 there were nearly 21 million ESTs in GenBank, with mouse (4.1 M) and human (5.5 M) representing at least one order of magnitude more than the rest of the species. This representation bias may explain the preponderance of evidence for alternative splicing in mouse and human. There are more than 20,000 ESTs from 66 species, representing 28 from animal nodes and 29 from green plant nodes. Figure 3 shows a rough phylogenetic distribution [made with NCBI Taxonomy Browser and WebPhylip (80)] of these species and the current number of ESTs available for each (see below for three exceptions). The higher eudicots (rosids and asterids) and monocots are heavily represented (109) because

they are important crop species, whereas species are more evenly distributed along the animal lineages (Figure 3). The recent generation of ESTs in moss (96), pine (4), and cycads (17) has filled in a few of the more ancestral branches of the plant lineage. The National Plant Genomics Initiative has called for major funding to be allocated for EST sequencing from 50 important evolutionary nodes throughout the green plant lineage (93). In addition, 5–10 tractable species at important nodes should be developed, with fully functional genomics capabilities, including mapping populations, markers, arrays, and potential genome sequencing.

Orthologous Gene Comparisons

What can one do with this seemingly large and ever-growing collection of expressed sequences? There are several possibilities. For gene discovery, new sequences that may have important enzymatic or signaling properties will be identified. One example is in *Medicago truncatula*, where the nodule-specific cysteine-rich gene family with more than 300 members was discovered. These genes have not been identified in other plants species or any other legumes species, which suggests that this gene family may be specific to intermediate nodule legumes (88). As high-throughput methods become available, the longest nonredundant EST clones can be screened for specific protein activates (23, 74). Once considered junk sequences, noncoding transcripts, including functional and regulatory RNAs or antisense transcripts (69, 101), can also be discovered from species at different evolutionary distances. Sequence conservation suggests functional importance for these nontraditional RNAs. Studies that aim to understand the evolution of alternative splicing by determining the exon content of the ancestral sequence look very promising (70). Abundant sequence polymorphism can be identified within species when confirmed by sequences from multiple clones (71). When multiple libraries are made from different strains or from closely related species, SNPs can be directly assigned. SNPs identified from libraries made from outbred strains, such as human, can later be typed to determine haplotypes.

Comparative EST sequence analysis begins with the identification of related sequences within large databases. This must be done, often on translated nucleotides, with fast search algorithms such as an all-against-all basic local alignment search tool (BLAST). This is followed by clustering sequences with high scores coming from within and between species. It is important to distinguish between orthologs and recent paralogs or ancient paralogs (45). Orthologs are related by common descent between species, whereas paralogs are related through gene duplication. For example, α and β tubulin are ancient paralogs, whereas human α tubulin is orthologus to mouse α tubulin. Both α tubulin orthologs maintain the same function. Recent gene duplication results in both new paralogs being orthologous to a single sequence in another species. The new paralogs may be redundant

in function or may quickly diverge perhaps through altered expression patterns. Simple BLAST searches do not account for ancestry (paralogs versus orthologs) and phylogenetic methods that do are very computationally intensive. Initially, the clustersof orthologous groups (COGs) database (http://ncbi.nih.gov/COG/) (138) and several initial genome-wide comparative studies (25, 117, 140) relied on all-against-all BLAST searches. An important improvement came with the INPARANOID algorithm (111), which can account for recent paralogs; however, INPARANOID is limited to pair-wise genome comparisons. Alternative approaches for genome-wide identification of eukaryotic gene orthologs (EGOs) rely on triangular best matches (76). ESTs were first clustered to form contigs or singletons within each of 28 species, after which each species was compared to the others. EGOs are groups where reciprocal best BLAST matches were confined to three species. Li et al. (78) developed the Orthologous Markov Cluster algorithm (OrthoMCL), which identifies orthologous groups and accounts for recent paralogs. Additionally, OrthoMCL can be applied to more than two species. Recently, a genome-wide phylogenetic approach was developed (121). With this method, gene family clusters are made from the fast all-against-all BLAST searches when an identity threshold is applied. Parsimony trees are made from each cluster that are either unconstrained, or constrained such that all sequences from the same species form a single clade. Because recent paralogs fit well in the constrained tree, orthologus clusters are identified when the unconstrained tree is more parsimonious.

What can be done once these clusters of putative orthologs are identified? Sequence alignments can identify regions of the protein that have been conserved and will likely be functionally important. Analyzing nucleotide substitution rates can reveal protein domains under selection via the relative rates test (158). Globally, how are different protein families evolving? Are some under more selective pressure than others, and do these belong to certain pathways? Are certain pathways or gene ontologies evolving faster than others, and can we define lineages where gene families or entire biochemical/signaling pathways have rapidly expanded? Early genome sequence comparisons show expansion of transcription factors and enzymes families in plants (see above). These expansions could be traced to specific lineages as more data becomes available. Lastly, the large amount of sequence data can be concatenated to add power in phylogenetic studies that aim to better estimate branch lengths or discriminate early divergences (deep branches) (121). A complicating issue with EST sequence information is representation bias, i.e., choosing an ancient paralog for comparison when the correct ortholog has not yet been identified. If possible, it will be important to develop methods that are robust to this bias. Although representation bias is less of a problem for complete genome comparative analysis, large-scale gene loss (perhaps following wide duplication) can remove true orthologs. This effect becomes more pronounced as the evolutionary distance between the species being compared increases. Plants may be especially prone to large-scale genome duplication and deletion because polyploidization is common (44, 140, 160).

THE NEXT PLANT GENOME SEQUENCES

Rice

In 2002, two groups released the second plant genome sequence, rice. A four-times shotgun coverage of *Oryza sativa* ssp. *indica* (160) covered 361 Mb of the estimated 466 Mb. *Oryza sativa* ssp. *japonica* was sequenced to five-times shotgun coverage and resulted in 390 Mb of assembled sequence, including two bacterial pathogens (44). Earlier, Monsanto had generated 259 Mb of *japonica* sequence from 3391 BAC clones (12); however, this data was initially only available and searchable to registered users. In *japonica*, 37,777 genes were predicted, 77% of which had at least one internal paralog. This resulted in 15,000 distinct gene families, a comparable number to other sequenced genomes. The synteny between rice and other cereals was used to identify candidate genes for Quantitative Trait Loci (QTL) mapped in other cereals (44). Although rice has one of the smallest genomes of any grass, transposon sequences still make up about half of the genome. Transposons comprise about 10% of the *Arabidopsis* genome (140) whereas transposons comprise 80% or more of the genome in Maize (122). When *Arabidopsis* genes were compared with rice, homologous sequences were identified for nearly all the genes. An exception is that the TIR-NBS-LRR-type disease resistance gene family is missing in rice (44). However, when predicted rice genes are compared with *Arabidopsis*, almost half of the genes are unique to rice. This finding may have more to do with the substantial proportion of the rice genes showing high GC content (>0.65) (153, 160) than the actual absence of these genes in the *Arabidopsis* genome. The GC content is higher in rice at the 5′ ends of genes, an effect not seen in *Arabidopsis*. This GC gradient extends to the early introns as well and effects codon bias and amino acid usage in rice and maize but not *Arabidopsis* or tobacco (153). This has a compounding effect on gene prediction algorithms as well as Ka/Ks tests. In both rice and *Arabidopsis*, most highly repetitive sequences (transposons) were found in intergenic regions, whereas in humans they also accumulate in introns (160).

Comparing the *indica* shotgun sequence with the public *japonica* sequence revealed genome wide differences between the two subspecies. Approximately 16% of the genome cannot be aligned between these subspecies due to changes in repetitive sequences! This further illustrates the rapid evolution in genome size between the grasses, in part due to very active transposons. Within alignable regions between the two genomes, nucleotide variation was around 0.5%. When comparing different accessions within the same rice subspecies polymorphism rates were a similar 0.5% (160), revealing the large amount of variation also seen within putative subspecies.

Rice is now the second model plant. Transformation techniques are routine (51), large collections of mapping populations are available (50) as are sequenced-indexed collections of mutant lines (8) and wild accessions. Full-length cDNA collections have been made and sequenced (67) and genomics tools such as

high-density microarrays are being developed (3). As with *Arabidopsis*, whole-genome rice tiling arrays will be important for further annotating the rice genome. Rice and *Arabidopsis* are separated by ~200 million years, thus they provide a system to study the evolution of plant-signaling and developmental pathways. Comparing flowering time pathways between the facultative long-day plant *Arabidopsis* and the short-day plant rice has already revealed common molecular components (31). Natural variation between rice subspecies has been important in the discovery of the commonalities in the flowering time pathway (reviewed in 15, 125). It is important that the rice genome is completely finished to the quality of *Arabidopsis* (99), a goal expected to be reached by the end of 2004 (http://demeter.bio.bnl.gov/Shanghai_summary.html).

Future Plant Genomes

What other plants are in the pipeline for complete genome sequencing? The National Plant Genomics Initiative is giving top priority to the high-quality finishing of rice and deep draft coverage of maize, *Medicago truncatula* (the model legume), and tomato in their 2003–2008 outlook (93). In addition, the Initiative will develop 5–10 organisms at different evolutionary nodes as genetic models with full genomic tool kits. This will include extensive EST sequencing, microarrays, BAC libraries, physical and high-density genetic maps, mapping populations, wild accessions, and developing transformation techniques. Ultimately, as technology improves and sequencing costs fall, the genomes of these new model genetic organisms could be sequenced; thus, the Initiative is giving priority to organisms with relatively small genome sizes and minimal repetitive DNA content.

POPULUS The *Populus* genus contains 30 diploid species, including aspen, cottonwood, and poplar, and represents the model tree with a genome size of ~550 Mb (40 times smaller the pine). Recombinant lines and genetic maps are available. QTL and breeding studies are widely performed and *Populus* is easily propagated clonally (39). The Joint Genome Institute has essentially finished a deep draft genome sequence of *Populus trichocarpa* (cottonwood). Ten-times shotgun coverage, amounting to 5.8 Gb, is now available for download and BLAST searches (http://genome.jgi-psf.org/poplar0/poplar0.info.html). Approximately 80,000 ESTs are also available in *Populus* that will aid in gene annotation (Figure 3). Because *Populus*, along with *Arabidopsis*, is a member of the rosids, the comparative genome analysis should reveal recent changes but overall the genomes should be similar in makeup, much closer than *Arabidopsis* and rice.

MEDICAGO TRUNCATULA The model legume *Medicago truncatula* (estimated at 500 Mb), a close relative of alfalfa, is being sequenced in a multinational effort spearheaded by the University of Minnesota, Oklahoma University, and TIGR. The Nobel Foundation, the University of California, Davis, and centers in the UK, France and Hungary are also contributing BAC sequence data. Currently ~84 Mb

of sequence data exists, mainly from hundreds of completed BAC clones, limited shotgun genomic sequence, and BAC end sequences (http://www.genome.ou.edu/medicago.html). The *Medicago truncatula* genome project is using a BAC by BAC-based approach because of its highly repetitive nature and hopes to cover the 12 major euchromatin chromosome arms. More than 187,000 ESTs are currently available to aid in genome annotation and functional studies (62). The full genome sequence of the first legume will be the foundation for molecular genetics research, which aims to understand the process of legume/*Rhizobium* symbiosis resulting in the conversion of molecular nitrogen into usable organic forms.

MAIZE The maize genome is large (2500 Mb) and highly repetitive (80%). Strategies to sequence the complete gene space rely on filtering away heavily methylated DNA (99a) and normalization by copy number (148a, 161). With one million sequencing reads, an estimated two-times coverage of the gene space was obtained. A five-times coverage of the gene space from filtered or normalized libraries, together with low-pass BAC sequencing, should yield a useful working draft of Maize (13, 84a). Furthermore, sequence generated from several inbred lines will identify many polymorphisms. An alternative approach is to sequence off transposon ends because there is a bias for the Mutator transposon to insert into gene-rich areas. This approach relies on an engineered transposon called RescueMu, which contains a bacterial plasmid. So far, 70,000 RescueMu flanking sequences have been recovered, identifying many genes not present in EST collections (83).

TOMATO *Lycopersicon esculentum* (tomato) has been used as a model crop plant for decades. Genetic studies have focused on fruit ripening (40) and disease resistance (72, 85) as well as crop domestication (36). Large mutant collections are available (http://zamir.sgn.cornell.edu/mutants/), as are high-resolution genetic maps (41). Several mapping populations have been used for pioneering QTL studies (103). Introgression lines where segments from various related species have been bred into the cultivated tomato background are widely used (33). The *Lycopersicon esculentum* genome is ~900 Mb and is a priority for genome sequencing according to the National Plant Genomics Initiative. The Arizona Genome Institute is making a BAC tiling path from more than 88,000 fingerprinted BACs. More than 150,000 *Lycopersicon esculentum* ESTs have been generated as well as more than 10,000 ESTs from related *Lycopersicon* species. The tomato genome sequence will serve as a model for Solanaceae plants, which include commonly known potato, tobacco, pepper, eggplant, petunia, and nightshade (http://www.sgn.cornell.edu/).

BRASSICA The *Brassica* genus contains many diverse developmental forms, including cole crops (*Brassica olearacea*: cauliflower, cabbage, kohlrabi, broccoli, brussels sprouts, and kale) and oil seed rape (*Brassica napus*). As a crucifer, it is closely related to *Arabidopsis* and its sequencing may help with genome annotation, as discussed above (102). The multinational *Brassica* genome project aims to cover the 500 Mb of the *Brassica rapa ssp pekinensis* (Chinese cabbage)

genome by draft sequencing BAC clones and relying on synteny of *Arabidopsis*. *Brassica rapa* BAC libraries containing more than 100,000 clones will be end sequenced and 1000 BAC seeds will be chosen for initial draft sequencing (http://brassica.bbsrc.ac.uk/). In *Brassica napus* there are currently 37,000 ESTs. In addition, a 0.2-times shotgun sequencing project of *Brassica oleracea* was completed by TIGR and Cold Spring Harbor, which covers >70% of the *Arabidopsis* proteome (http://nucleus.cshl.org/genseq/comp_genomics/).

Species for Development as Model Systems for Ecology/Evolution Studies

MIMULUS *Mimulus* (monkey flower) has long been a model for studies of ecology and evolution aimed at determining the genetic mechanisms of speciation. Pollinator preference, floral morphology (16, 123), outcrossing rate (35), and altitude acclimation traits have been mapped as QTL. Several species have been studied, many of which are cross-compatible. Genetic maps and several mapping populations are available between different *Mimulus* species. The genome size, approximately 500 Mb, is similar to rice. Recently, NSF awarded $5 million to develop *Mimulus* as a model ecological genetic organism (http://www.biology.duke.edu/mimulus/). BAC libraries will be fingerprinted to create a physical map. High-density genetic markers will be developed that will work across species for comparative mapping studies. *Mimulus* is at the basal end of the Asterids, which includes the Solanaceae. It will be interesting to compare speciation processes in *Mimulus* and *Lycopersicon* as research advances in both these systems. For example, variation in mating systems is abundant in both genus and has been mapped in interspecific crosses.

AQUILEGIA *Aquilegia* (columbines) are in the Ranunculaceae family of the lower eudictos and contain about 70 species. They have a small genome size of ∼400 Mb and are studied as an ecological organism for adaptation to harsh serpentine soils (52) and for speciation involving pollinator preference and floral morphology (53, 53a). Recent adaptations in *Aquilegia* likely occurred independently, allowing studies of convergent evolution at the molecular level. For example are the same genes and similar mutations involved in the same adaptive process. Recombinant populations and low-density genetic maps are available. Floral traits influencing reproductive isolation between *Aquilegia formosa* and *Aquilegia pubescens* have been mapped (Scott Hodges, personal communication). *Aquilegia*, with a small genome and ancestral position in the lower eudicots, is uniquely suited for development of genomics tools, such as high-density genetic and physical maps, EST sequencing, microarrays, and development of genetic resources such as recombinant inbred lines (RILs). Currently no genetic models have been developed at this important early node in the flowering plant linage. *Aquilegia* is an attractive choice.

SELAGINELLA The lycophyte *Selaginella* is a seed-free plant that diverged 360 million years ago (mya) from the ancestor that gave rise to the angiosperms. It has a very small genome size equal to or less than *Arabidopsis* (49). The Arizona Genomics Institute recently developed a BAC library for *Selaginella* (132-Mb genome size) (http://www.genome.arizona.edu/BAC_special_projects/). Thus, small genome ancestral comparisons in plants will identify important functional regions in a manner analogous to *takifugu* (pufferfish) comparisons with the human genome (109).

Animal comparative genetics has sampled from a broad range of evolutionary lineages, whereas analogous plant studies are lagging behind. Because most of the focus has been on crop plants, we do not have comprehensive plant genome sequences or EST collections from species that span the deep branches of the green plant phylogeny. Future studies on the lower eudicot *Aquilegia* and the seed-free plant *Selaginella* would begin to remedy this situation. Their small genome size and their potential for development as model organisms make them ideal candidates.

CONCLUSION

The genome sequence of *Arabidopsis thaliana* has not been fully characterized, nor will such a large task be completed for some time. We have only touched the tip of the iceberg in our understanding of the information in the raw plant genome sequences. Deep comprehension of how the complex instructions of plant life are written in the linear code will be an enormous challenge for the future. Such an understanding will require the development of a third wave of new technologies to glean the next level of genomic complexity. Fortunately, new tools such as whole genome arrays (WGAs) can integrate genomic data on a common platform. Global genome data, including transcriptome atlases with alternatively spliced messages, DNA binding site profiles, and chromatin state surveys, will give a more holistic picture of the cells' activity. High-density polymorphism maps will reveal patterns of variation within a species. Our current philosophy has an appreciation for genomic complexity as we attempt to understand how plant cells function, how they interact, and how they send and receive systemic signals. This extends to the whole plant level and to interactions with other organisms and their environment.

We should not be satisfied with the current list of predicted genes and repeat this work by sequencing other crop genomes. We must consider how the sequencing of other species will result in a synergistic improvement our understanding of plant form and function (109). Thus, plant genomics proposals should include the development of genomics tools such at WGAs, large collections of mutant and wild strains, and the development of mapping populations in addition to genome sequencing. Such enabling tools and technologies are required for in-depth comparative analysis that many groups will accomplish by taking advantage of improved statistical and evolutionary analysis methods. These goals will not be achieved within single disciplines, but will require broad collaborations between molecular,

evolutionary, ecological, and computational biologists. The mammalian and *Drosophila* communities are well on their way; plant comparative genomics will need to regain pace.

ACKNOWLEDGMENTS

We thank Thomas Gal for help with Figure 3; Huaming Chen for assistance with current gene counts; and Jennifer Nemhauser, Sam Hazen, Isaac Mehl, and Todd Michael for comments on the manuscript. J.O.B. is supported by a fellowship from the Helen Hay Whitney Foundation. J.R.E. is supported by grants from the NSF, NIH, and DOE.

The *Annual Review of Genomics and Human Genetics* is online at http://genom.annualreviews.org

LITERATURE CITED

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–95
2. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12:1805–14
3. Akimoto-Tomiyama C, Sakata K, Yazaki J, Nakamura K, Fujii F, et al. 2003. Rice gene expression in response to N-acetylchitooligosaccharide elicitor: comprehensive analysis by DNA microarray with randomly selected ESTs. *Plant Mol. Biol.* 52:537–51
4. Allona I, Quinn M, Shoop E, Swope K, Cyr S St, et al. 1998. Analysis of xylem formation in pine by cDNA sequencing. *Proc. Natl. Acad. Sci. USA* 95:9693–98
5. Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, et al. 2003. Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* 301:653–57
6. Alonso-Blanco C, Koornneef M. 2000. Naturally occurring variation in *Arabidopsis*: an underexploited resource for plant genetics. *Trends Plant Sci.* 5:22–29
7. Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, et al. 2000. The common *PPARgamma* Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* 26:76–80
8. An S, Park S, Jeong DH, Lee DY, Kang HG, et al. 2003. Generation and analysis of end sequence database for T-DNA tagging lines in rice. *Plant Physiol.* 133:2040–47
9. Andersen JR, Lubberstedt T. 2003. Functional markers in plants. *Trends Plant Sci.* 8:554–60
10. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297:1301–10
11. Ardlie KG, Kruglyak L, Seielstad M. 2002. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 3:299–309
12. Barry GF. 2001. The use of the Monsanto draft rice genome sequence in research. *Plant Physiol.* 125:1164–65
13. Bennetzen JL, Chandler VL, Schnable P. 2001. National Science Foundation-sponsored workshop report. Maize genome sequencing project. *Plant Physiol.* 127:1572–78

13a. Bignell GR, Huang J, Greshock J, Watt S, Butler A, et al. 2004. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* 14:287–95
14. Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, et al. 2003. Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* 13:513–23
15. Borevitz JO, Nordborg M. 2003. The impact of genomics on the study of natural variation in *Arabidopsis*. *Plant Physiol.* 132:718–25
16. Bradshaw HD, Schemske DW. 2003. Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. *Nature* 426:176–78
17. Brenner ED, Stevenson DW, McCombie RW, Katari MS, Rudd SA, et al. 2003. Expressed sequence tag analysis in Cycas, the most primitive living seed plant. *Genome Biol.* 4:R78
18. Cao X, Jacobsen SE. 2002. Role of the *Arabidopsis* DRM methyltransferases in de novo DNA methylation and gene silencing. *Curr. Biol.* 12:1138–44
19. Carninci P, Waki K, Shiraki T, Konno H, Shibata K, et al. 2003. Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* 13:1273–89
20. Cashmore AR. 2003. Cryptochromes: enabling plants and animals to determine circadian time. *Cell* 114:537–43
21. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116:499–509
22. Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, et al. 2002. Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* 3:RESEARCH0079
23. Chanda SK, White S, Orth AP, Reisdorph R, Miraglia L, et al. 2003. Genome-scale functional profiling of the mammalian AP-1 signaling pathway. *Proc. Natl. Acad. Sci. USA* 100:12153–58
24. Chee M, Yang R, Hubbell E, Berno A, Huang XC, et al. 1996. Accessing genetic information with high-density DNA arrays. *Science* 274:610–14
25. Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, et al. 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 282:2022–28
26. Churchill GA, Giovannoni JJ, Tanksley SD. 1993. Pooled-sampling makes high-resolution mapping practical with DNA markers. *Proc. Natl. Acad. Sci. USA* 90:16–20
27. Clark TA, Sugnet CW, Ares M Jr. 2002. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* 296:907–10
28. Clemens JC, Worby CA, Simonson -Leff N, Muda M, Maehama T, et al. 2000. Use of double-stranded RNA interference in *Drosophila* cell lines to dissect signal transduction pathways. *Proc. Natl. Acad. Sci. USA* 97:6499–503
29. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, et al. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301:71–76
30. C_elegans_Sequencing_Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282:2012–18
31. Cremer F, Coupland G. 2003. Distinct photoperiodic responses are conferred by the same genetic pathway in *Arabidopsis* and in rice. *Trends Plant Sci.* 8:405–7
32. Elgin SC, Grewal SI. 2003. Heterochromatin: silence is golden. *Curr. Biol.* 13:R895–98
33. Eshed Y, Zamir D. 1995. An introgression line population of *Lycopersicon pennellii* in the cultivated tomato

enables the identification and fine mapping of yield-associated QTL. *Genetics* 141:1147–62

34. Finkelstein D, Ewing R, Gollub J, Sterky F, Cherry JM, Somerville S. 2002. Microarray data quality analysis: lessons from the AFGC project. *Arabidopsis* Functional Genomics Consortium. *Plant Mol. Biol.* 48:119–31
35. Fishman L, Kelly AJ, Willis JH. 2002. Minor quantitative trait loci underlie floral traits associated with mating system divergence in Mimulus. *Evol. Int. J. Org. Evol.* 56:2138–55
36. Frary A, Nesbitt TC, Grandillo S, Knaap E, Cong B, et al. 2000. fw2.2: a quantitative trait locus key to the evolution of tomato fruit size [see comments]. *Science* 289:85–88
37. Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J. 2000. Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* 408:325–30
38. Frazer KA, Chen X, Hinds DA, Pant PV, Patil N, Cox DR. 2003. Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.* 13:341–46
39. Frewen BE, Chen TH, Howe GT, Davis J, Rohde A, et al. 2000. Quantitative trait loci and candidate gene mapping of bud set and bud flush in populus. *Genetics* 154:837–45
40. Fridman E, Pleban T, Zamir D. 2000. A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. *Proc. Natl. Acad. Sci. USA* 97:4718–23
41. Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD. 2002. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* 14:1457–67

41a. Gerber AP, Herschlag D, Brown PO. 2004. Extensive association of functionally and cytotopically related mRNAs with puf family RNA-binding proteins in yeast. *Pub. Libr. Sci. Biol.* 2:E79

42. Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418:387–91

42a. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428: 493–521

43. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* 302:1727–36
44. Goff SA, Ricke D, Lan TH, Presting G, Wang R, et al. 2002. A draft sequence of the rice genome (*Oryza sativa L. ssp. japonica*). *Science* 296:92–100
45. Gogarten JP, Olendzenski L. 1999. Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.* 9:630–36
46. Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11:725–36
47. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* 421:63–66
48. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, et al. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucl. Acids Res.* 31:5654–66
49. Hanson L, Leitch IJ. 2002. DNA amounts for five pteridophyte species fill phylogenetic gaps in C-value data. *Bot. J. Linn. Soc.* 140:169–73
50. Harushima Y, Yano M, Shomura A, Sato M, Shimano T, et al. 1998. A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics* 148:479–94
51. Hiei Y, Ohta S, Komari T, Kumashiro T. 1994. Efficient transformation of rice

(*Oryza sativa L.*) mediated by Agrobacterium and sequence analysis of the boundaries of the T-DNA. *Plant J.* 6:271–82

52. Hodges SA, Arnold ML. 1994. Columbines: a geographically widespread species flock. *Proc. Natl. Acad. Sci. USA* 91:5129–32
53. Hodges SA, Arnold ML. 1994. Floral and ecological isolation between *Aquilegia formosa* and *Aquilegia pubescens*. *Proc. Natl. Acad. Sci. USA* 91:2493–96

53a. Hodges SA, Whittall JB, Fulton M, Yang JY. 2002. Genetics of floral traits influencing reproductive isolation between *Aquilegia Formosa* and *A. pubescens*. *Am. Nat.* 159:S51–60

54. Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, et al. 2002. Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. *Genes Dev.* 16:3017–33
55. Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM, Snyder M. 2002. GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc. Natl. Acad. Sci. USA* 99:2924–29
56. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucl. Acids Res.* 31:e15
57. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* 98:4569–74
58. Jacobsen SE, Meyerowitz EM. 1997. Hypermethylated SUPERMAN epigenetic alleles in *Arabidopsis*. *Science* 277:1100–3
59. Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL. 2002. *Arabidopsis* map-based cloning in the post-genome era. *Plant Physiol.* 129:440–50
60. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, et al. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302:2141–44
61. Johnson L, Cao X, Jacobsen S. 2002. Interplay between two epigenetic marks. DNA methylation and histone H3 lysine 9 methylation. *Curr. Biol.* 12:1360–67
62. Journet EP, van Tuinen D, Gouzy J, Crespeau H, Carreau V, et al. 2002. Exploring root symbiotic programs in the model legume *Medicago truncatula* using EST analysis. *Nucl. Acids Res.* 30:5579–92
63. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, et al. 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421:231–37
64. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, et al. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296:916–19
65. Kato-Maeda M, Rhee JT, Gingeras TR, Salamon H, Drenkow J, et al. 2001. Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res.* 11:547–54
66. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241–54
67. Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, et al. 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* 301:376–79
68. Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, et al. 2003. The dog genome: survey sequencing and comparative analysis. *Science* 301:1898–903
69. Kiyosawa H, Yamanaka I, Osato N, Kondo S, Hayashizaki Y. 2003. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* 13:1324–34
70. Kondrashov FA, Koonin EV. 2003. Evolution of alternative splicing: deletions, insertions and origin of functional

parts of proteins from intron sequences. *Trends Genet.* 19:115–19

71. Kota R, Rudd S, Facius A, Kolesov G, Thiel T, et al. 2003. Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). *Mol. Genet. Genomics* 270:24–33
72. Kruger J, Thomas CM, Golstein C, Dixon MS, Smoker M, et al. 2002. A tomato cysteine protease required for Cf-2-dependent disease resistance and suppression of autonecrosis. *Science* 296:744–47

72a. Kuromori T, Hirayama T, Kiyosue Y, Takabe H, Mizukado S, et al. 2004. A collection of 11 800 single-copy Ds transposon insertion lines in *Arabidopsis*. *Plant J.* 37:897–905

73. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
74. Langholz B, Richardson J, Rappaport E, Waisman J, Cockburn M, Mack T. 2000. Skin characteristics and risk of superficial spreading and nodular melanoma (United States). *Cancer Causes Control* 11:741–50
75. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, et al. 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 301:1503–8
76. Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, et al. 2002. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.* 12:493–502
77. Li C, Wong WH. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* 98:31–36
78. Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13:2178–89
79. Li Y, Rosso MG, Strizhov N, Viehoever P, Weisshaar B. 2003. GABI-Kat SimpleSearch: a flanking sequence tag (FST) database for the identification of T-DNA insertion mutants in *Arabidopsis thaliana*. *Bioinformatics* 19:1441–42
80. Lim A, Zhang L. 1999. WebPHYLIP: a web interface to PHYLIP. *Bioinformatics* 15:1068–69
81. Liu Q, Li MZ, Leibham D, Cortez D, Elledge SJ. 1998. The univector plasmid-fusion system, a method for rapid construction of recombinant DNA without restriction enzymes. *Curr. Biol.* 8:1300–9
82. Lum L, Yao S, Mozer B, Rovescalli A, Von Kessler D, et al. 2003. Identification of Hedgehog pathway components by RNAi in *Drosophila* cultured cells. *Science* 299:2039–45
83. Lunde CF, Morrow DJ, Roy LM, Walbot V. 2003. Progress in maize gene discovery: a project update. *Funct. Integr. Genomics* 3:25–32
84. Luo S, Preuss D. 2003. Strand-biased DNA methylation associated with centromeric regions in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* 100:11133–38

84a. Martienssen RA, Rabinowicz PD, O'Shaughnessy A, McCombie WR. 2004. Sequencing the maize genome. *Curr. Opin. Plant Biol.* 7:102–7

85. Martin GB, Brommonschenkel SH, Chunwongse J, Frary A, Ganal MW, et al. 1993. Map-based cloning of a protein kinase gene conferring disease resistance in tomato. *Science* 262:1432–36
86. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, et al. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. USA* 99:12246–51
87. Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, et al. 2003. Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc. Natl. Acad. Sci. USA* 100:12247–52

88. Mergaert P, Nikovics K, Kelemen Z, Maunoury N, Vaubert D, et al. 2003. A novel family in *Medicago truncatula* consisting of more than 300 nodule-specific genes coding for small, secreted polypeptides with conserved cysteine motifs. *Plant Physiol.* 132:161–73
89. Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. 2003. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* 15:809–34
90. Michelmore RW, Paran I, Kesseli RV. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. USA* 88:9828–32
91. Morris ER, Walker JC. 2003. Receptor-like protein kinases: the keys to response. *Curr. Opin. Plant Biol.* 6:339–42
92. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, et al. 2000. A whole-genome assembly of Drosophila. *Science* 287:2196–204
93. Natl. Plant Genomics Initiat. 2002. The National Plant Genomics Initiative: objectives for 2003–2008. *Plant Physiol.* 130:1741–44
94. Neff MM, Fankhauser C, Chory J. 2000. Light: an indicator of time and place. *Genes Dev.* 14:257–71
95. Nekrutenko A, Makova KD, Li WH. 2002. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.* 12:198–202
96. Nishiyama T, Fujita T, Shin IT, Seki M, Nishide H, et al. 2003. Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: implication for land plant evolution. *Proc. Natl. Acad. Sci. USA* 100:8007–12
97. Niu X, Guiltinan MJ. 1994. DNA binding specificity of the wheat bZIP protein EmBP-1. *Nucl. Acids Res.* 22:4969–78
98. Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, et al. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* 30:190–93
99. Palmer LE, McCombie WR. 2002. On the importance of being finished. *Genome Biol.* 3:COMMENT2010
99a. Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, et al. 2003. Maize genome sequencing by methylation filtration. *Science* 302: 2115–17
100. Paquette SM, Bak S, Feyereisen R. 2000. Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. *DNA Cell Biol.* 19:307–17
101. Patankar S, Munasinghe A, Shoaibi A, Cummings LM, Wirth DF. 2001. Serial analysis of gene expression in *Plasmodium falciparum* reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite. *Mol. Biol. Cell* 12:3114–25
102. Paterson AH, Lan TH, Amasino R, Osborn TC, Quiros C. 2001. *Brassica* genomics: a complement to, and early beneficiary of, the *Arabidopsis* sequence. *Genome Biol.* 2:REVIEWS1011
103. Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD. 1988. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* 335:721–26
104. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294:1719–23
105. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* 20:207–11

106. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, et al. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* 23:41–46
107. Pothof J, van Haaften G, Thijssen K, Kamath RS, Fraser AG, et al. 2003. Identification of genes that protect the *C. elegans* genome against mutations by genome-wide RNAi. *Genes Dev.* 17:443–48
108. Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69:1–14
109. Pryer KM, Schneider H, Zimmer EA, Ann Banks J. 2002. Deciding among green plants for whole genome studies. *Trends Plant Sci.* 7:550–54
110. Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M, et al. 2003. *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* 34:35–41
111. Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314:1041–52
112. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* 290:2306–9
113. Richards EJ. 2002. Chromatin methylation: who's on first? *Curr. Biol.* 12:R694–95
114. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, et al. 2003. The transcriptional activity of human Chromosome 22. *Genes Dev.* 17:529–40
115. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, et al. 1989. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 245:1066–73
116. Rogic S, Mackworth AK, Ouellette FB. 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* 11:817–32
117. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, et al. 2000. Comparative genomics of the eukaryotes. *Science* 287:2204–15
118. Deleted in proof
119. Salamon H, Kato-Maeda M, Small PM, Drenkow J, Gingeras TR. 2000. Detection of deleted genomic DNA using a semiautomated computational analysis of GeneChip data. *Genome Res.* 10:2044–54
120. Samson F, Brunaud V, Balzergue S, Dubreucq B, Lepiniec L, et al. 2002. FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of *Arabidopsis thaliana* T-DNA transformants. *Nucl. Acids Res.* 30:94–97
121. Sanderson MJ, Driskell AC, Ree RH, Eulenstein O, Langley S. 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. *Mol. Biol. Evol.* 20:1036–42
122. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274:765–68
123. Schemske DW, Bradshaw HD Jr. 1999. Pollinator preference and the evolution of floral traits in monkeyflowers (*Mimulus*). *Proc. Natl. Acad. Sci. USA* 96:11910–15
124. Schmid KJ, Sorensen TR, Stracke R, Torjek O, Altmann T, et al. 2003. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res.* 13:1250–57
125. Schultz TF, Kay SA. 2003. Circadian clocks in daily and seasonal control of development. *Science* 301:326–28
126. Sessions A, Burke E, Presting G, Aux G, McElver J, et al. 2002. A high-throughput *Arabidopsis* reverse genetics system. *Plant Cell* 14:2985–94
127. Shah SP, McVicker GP, Mackworth AK, Rogic S, Ouellette BF. 2003. GeneComber: combining outputs of

gene prediction programs for improved results. *Bioinformatics* 19:1296–97
128. Shendure J, Church GM. 2002. Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol.* 3:RESEARCH 0044
129. Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, et al. 2001. Experimental annotation of the human genome using microarray technology. *Nature* 409:922–27
130. Stapleton M, Liao G, Brokstein P, Hong L, Carninci P, et al. 2002. The *Drosophila* gene collection: identification of putative full-length cDNAs for 70% of *D. melanogaster* genes. *Genome Res.* 12:1294–300
131. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, et al. 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *Pub. Libr. Sci. Biol.* 1:E45
132. Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, et al. 2002. Systematic screen for human disease genes in yeast. *Nat. Genet.* 31:400–4
133. Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, et al. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci. USA* 99:16899–903
134. Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, et al. 2001. Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: implication of potential sequencing errors caused by large-unit repeats. *Proc. Natl. Acad. Sci. USA* 98:5099–103
135. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. USA* 99:4465–70
136. Deleted in proof
137. Tamaru H, Selker EU. 2001. A histone H3 methyltransferase controls DNA methylation in *Neurospora crassa*. *Nature* 414:277–83
138. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, et al. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucl. Acids Res.* 29:22–28
139. Terryn N, Rouze P. 2000. The sense of naturally transcribed antisense RNAs in plants. *Trends Plant Sci.* 5:394–96
140. The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
141. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424:788–93
141a. Tissier AF, Marillonnet S, Klimyuk V, Patel K, Torres MA, et al. 1999. Multiple independent defective suppressor-mutator transposon insertions in *Arabidopsis*: a tool for functional genomics. *Plant Cell* 11:1841–52
142. Tompa R, McCallum CM, Delrow J, Henikoff JG, van Steensel B, Henikoff S. 2002. Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3. *Curr. Biol.* 12:65–68
143. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403:623–27
144. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
145. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–62

146. Weigel D, Ahn JH, Blazquez MA, Borevitz JO, Christensen SK, et al. 2000. Activation tagging in *Arabidopsis*. *Plant Physiol.* 122:1003–13
147. Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ. 2002. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.* 16:235–44
148. Wells J, Graveel CR, Bartley SM, Madore SJ, Farnham PJ. 2002. The identification of E2F1-specific target genes. *Proc. Natl. Acad. Sci. USA* 99:3890–95
148a. Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, et al. 2003. Enrichment of gene-coding sequences in maize by genome filtration. *Science* 302:2118–20
149. Winzeler EA, Castillo-Davis CI, Oshiro G, Liang D, Richards DR, et al. 2003. Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* 163:79–89
150. Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, et al. 1998. Direct allelic variation scanning of the yeast genome. *Science* 281:1194–97
151. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, et al. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285:901–6
152. Wisman E, Ohlrogge J. 2000. *Arabidopsis* microarray service facilities. *Plant Physiol.* 124:1468–71
152a. Wolyn DJ, Borevitz JO, Loudet O, Schwartz C, Maloof J, et al. 2004. Light response QTL identified with composite interval and eXtreme array mapping in *Arabidopsis* thaliana. *Genetics* 2004 167: In press
153. Wong GK, Wang J, Tao L, Tan J, Zhang J, et al. 2002. Compositional gradients in *Gramineae* genes. *Genome Res.* 12:851–56
154. Wortman JR, Haas BJ, Hannick LI, Smith RK Jr, Maiti R, et al. 2003. Annotation of the *Arabidopsis* genome. *Plant Physiol.* 132:461–68
155. Xiao YL, Malik M, Whitelaw CA, Town CD. 2002. Cloning and sequencing of cDNAs for hypothetical genes from chromosome 2 of *Arabidopsis*. *Plant Physiol.* 130:2118–28
156. Xie C, Zhang JS, Zhou HL, Li J, Zhang ZG, et al. 2003. Serine/threonine kinase activity in the putative histidine kinase-like ethylene receptor NTHK1 from tobacco. *Plant J.* 33:385–93
157. Yamada K, Lim J, Dale JM, Chen H, Shinn P, et al. 2003. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 302:842–46
158. Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19: 908–17
159. Yeh K-C, Lagarias JC. 1998. Eukaryotic phytochromes: Light-regulated serine/threonine protein kinases with histidine kinase ancestry. *Proc. Natl. Acad. Sci. USA* 95:13976–81
160. Yu J, Hu S, Wang J, Wong GK, Li S, et al. 2002. A draft sequence of the rice genome (*Oryza sativa L. ssp. indica*). *Science* 296:79–92
161. Yuan Y, SanMiguel PJ, Bennetzen JL. 2003. High-cot sequence analysis of the maize genome. *Plant J.* 34:249–55
162. Zak DE, Gonye GE, Schwaber JS, Doyle FJ III. 2003. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network. *Genome Res.* 13:2396–405
163. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, et al. 2003. Impact of alternative initiation, splicing,

and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.* 13:1290–300

164. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, et al. 2001. Global analysis of protein activities using proteome chips. *Science* 293:2101–5
165. Zhu W, Schlueter SD, Brendel V. 2003. Refined annotation of the *Arabidopsis* genome by complete expressed sequence tag mapping. *Plant Physiol.* 132:469–84

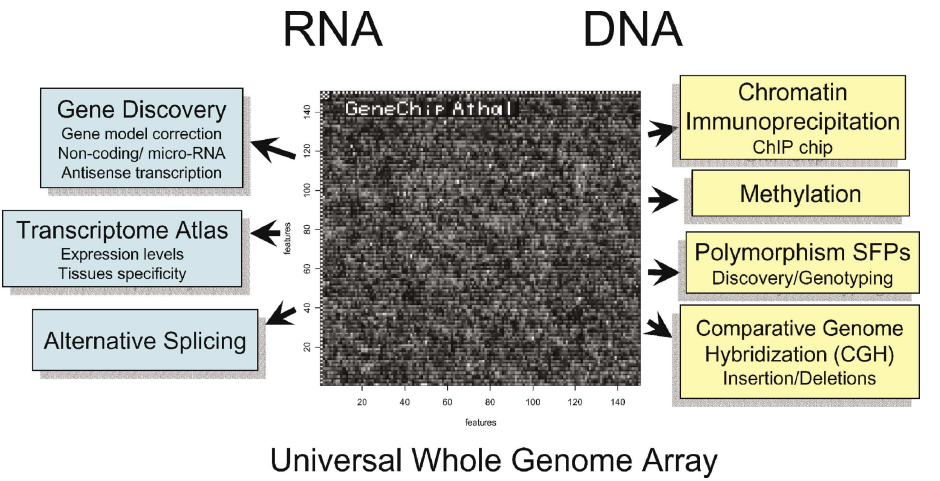

Figure 1 High-density oligonucleotide arrays can be used to characterize a fully sequenced genome and provide a common platform for comparisons among different biological processes.

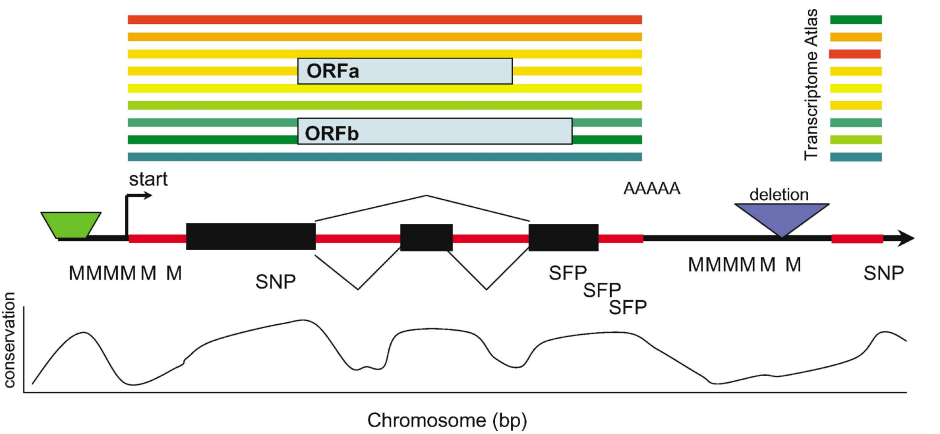

Figure 2 An example showing detailed genome annotation including exons (*black boxes*), sites of transcription (*red*), transcript levels in different tissues (*rainbow colors*), alternative splicing, locations of DNA binding proteins (*green box*), DNA methylation (M), DNA polymorphism (SNP, SFP), and a comparative measure of divergence. An integrated view of the genome annotation can be obtained with various experiments using WGAs as a single technology platform.

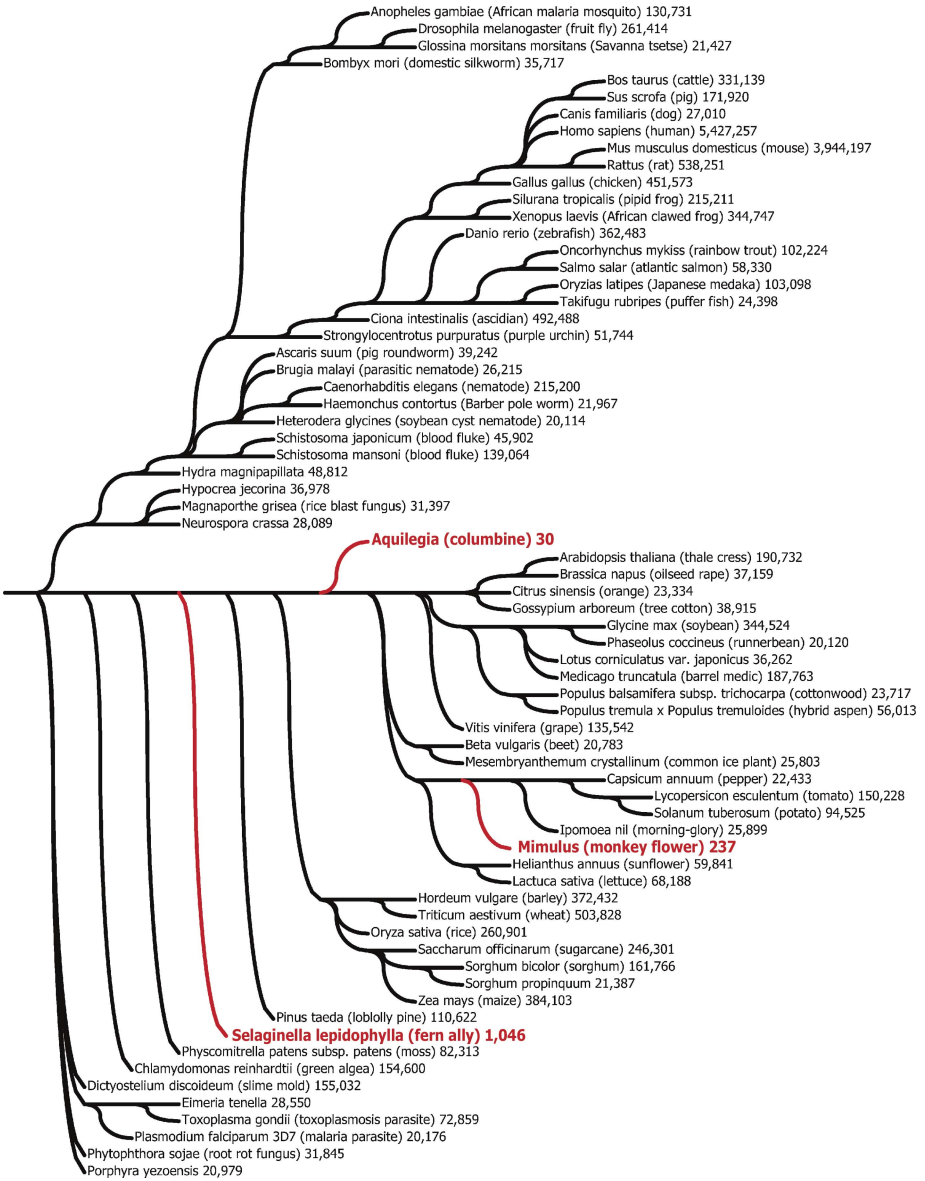

Figure 3 Phylogeny of species with more than 20,000 ESTs in GenBank as of November 14, 2003. The evolutionary sampling along the animal lineage is broader than along the plant lineage. Mimulus, Aquilegia, and Selaginella are at important nodes that should be sampled extensively and perhaps developed into model genetic organisms. The small genome size of Mimulus (~500 Mb), Aquilegia (~400 Mb), and Selaginella (120 Mb) (109) makes these key organisms potential candidates for full genome sequencing. Tree made using http://biocore.unl.edu/WEBPHYLIP/ (80) and the Taxonomy Browser at NCBI.