Opinion

Goodbye to 'one by one' genetics

Athanasios Theologis

Address: Plant Gene Expression Center, Buchanan Street, Albany, CA 94710, USA. E-mail: theo@nature.berkeley.edu

Published: 6 April 2001

Genome ***Biology*** 2001, **2(4)**:comment2004.1–2004.9

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/4/comment/2004

Abstract

The completion of the *Arabidopsis thaliana* (mustard weed) genome sequence constitutes a major breakthrough in plant biology. It will revolutionize how we answer questions about the biology and evolution of plants as well as how we confront and resolve world-wide agricultural problems.

Historical perspective

One of the major problems that humans face as we enter the twenty-first century is how to feed an overpopulated planet. Currently, the world population is approximately six billion people, and it is estimated that with the current birth rate it will double by 2050 [1]. The consequence of such growth is that humans will be forced to carry out agriculture using less, and lower quality, farming land. In order for us to be successful under these adverse conditions, a highly sophisticated knowledge of plant biology will be required that will allow the development of agronomically important species suitable for producing more food per unit farming area. The introduction of *Arabidopsis* as a model plant species fifteen years ago has revolutionized plant biology and provides opportunities for achieving this goal. *Arabidopsis* was adopted as a model organism by plant geneticists because of its small diploid genome, low repetitive DNA content, and rapid reproductive cycle [2]. *Arabidopsis* appears to contain homologs of most of the genes found in agronomically important crop plants, including rice, maize, soybean, and tomato [3].

It is axiomatic that technology advances biology and *vice versa*. The sequencing of the mustard weed genome [4], together with the genomes of other eukaryotes - yeast [5,6], worm [7], fly [8] and human [9,10], reaffirms the validity of this axiom. It has been only 48 years since the structure of the DNA was elucidated [11] and almost 30 years since the first DNA molecule was cloned and propagated in *Escherichia coli* [12]. Subsequently, recombinant DNA technology was developed that allowed biologists to clone and study genes 'one by one', laying the foundations for the biotechnology industry [13,14]. After 20 years, however, the 'one by one' approach was not revealing information fast enough to allow understanding of the complexity of biological systems. It was also very expensive. It costs $10 to $20 per base to sequence your favorite gene in a timely manner. More importantly, genomic sequencing of intensely studied eukaryotic organisms with a small fraction of redundant genes, such as *Saccharomyces cerevisiae,* has indicated that no more than 30% of an organism's genes can be identified by classical genetic analysis [6]. It was the advent ten years ago of genomics, a new scientific discipline established by the emergence of three independent fields - biology, engineering and computer science - that eliminated the shortcomings of the 'one by one' approach and opened up new and exciting possibilities for advancing biology.

The *Arabidopsis* Genome Initiative (AGI)

When established in 1989, the Human Genome Project included all the best-known model organisms (*E. coli*, yeast, worm, fly, human and mouse) but omitted the mustard weed, for reasons that are poorly understood [15]. The *Arabidopsis* sequencing project was initiated in 1994 by the European Community (now the European Union, EU) under the leadership of Michael Bevan at the John Innes Institute in the UK. Funds were secured in 1994 that led to the establishment of a consortium of European laboratories (ESSA), led by Bevan, to sequence chromosome 4 [16]. Shortly thereafter, Satoshi Tabata of the Kazusa DNA Research Institute in Japan was funded for sequencing chromosome 5 [17]. The flame that

initiated the American sequencing effort was kindled one humid night - on July 8 1993 - at the Cold Spring Harbor Laboratory (CSHL), where the instructors of the *Arabidopsis* Molecular Genetics course (Joe Ecker, Joanne Chory, and myself) were entertaining Elliot Meyerowitz after his evening lecture at Blackford Hall. While we were drinking beer together with Rob Martienssen and Venkatesan Sundaresan, two members of the CSHL Plant Biology group, and Gary Drews (another instructor) and a few students from the course, I remember asking Elliot when the sequencing of the *Arabidopsis* genome would be initiated. His answer was, "Well, we need a few sequencing machines to be given to each of you and the project can start." That evening's conversation was somehow transmitted to Jim Watson by Rob and Venkatesan, and the rest is history.

Jim Watson's deep interest in genome sequencing led to a Banbury conference in the spring of 1994 at CSHL that convinced the National Science Foundation (NSF) and some prominent skeptics of the urgency and importance of sequencing the *Arabidopsis* genome. The most obvious objections at that time, from some of the leaders in the *Arabidopsis* community, were whether funds would be removed from other aspects of plant biological research and whether it was intellectually stimulating to sequence the *Arabidopsis* genome. Following the Banbury conference, a workshop on sequencing the *Arabidopsis* genome was held at NSF in the summer of 1994, where the guidelines were established for finding resources for the project. The resources were allocated in 1995 and the sequencing effort was initiated in the fall of 1996 by three US groups (see Figure 1). An historic meeting was held in the summer of 1996 at NSF, where the six sequencing groups from Europe, Japan and the US, representatives of the *Arabidopsis* community, NSF, the Department of Energy (DOE) and the US Department of Agriculture (USDA) met to coordinate an international effort to sequence the 125 Mb *Arabidopsis* genome. That meeting established the *Arabidopsis* Genome Initiative (AGI), an international effort to sequence the first plant genome by 2004 (Figure 1).

The sequencing strategy

The AGI discussed various strategies for sequencing the *Arabidopsis* genome, including the whole-genome shotgun-sequencing approach [18], which was proposed by Ron Davis, one of the principal investigators (PIs) of the SPP consortium (see Figure 1). The approach was rejected at that time as risky, and more secure strategies were adopted, such as BAC-end sequencing (see below) [19], and approaches based on physical maps [16,20]. In retrospect, if the shotgun approach had been adopted in 1996, the discovery of all the genes in the *Arabidopsis* genome would have preceded the corresponding studies of the worm, fly and human genomes. It was decided that chromosomes 4 and 5 would be sequenced using a map-based strategy, built on existing physical maps made using yeast artificial chromosome (YAC), cosmid and plasmid P1 clones [16,20]. Chromosomes 1, 2, and 3 would be sequenced using the BAC-end strategy (Figure 2a). In 1996 two BAC libraries became available to AGI [21,22]. In a collaborative effort among three AGI groups – Genoscope [23], a member of the EU-2 consortium, The Institute for Genomic Research (TIGR) [24] and SPP [25] (see Figure 1) - the end-sequences for almost all the BACs (approximately 18,300 clones, representing 14-fold coverage) in the two libraries were determined and made publicly available.

The BAC-end sequencing strategy (Figure 2a) is based on extension from a few fully sequenced BAC clones ('seed' or nucleation points) using a minimum set of overlapping BAC clones selected from a set of end-sequenced BACs. All sequencing groups used a variation of the same strategy for sequencing individual BACs, P1 or cosmid clones, as shown in Figure 2b. Shotgun, plasmid, or M13 libraries were constructed and an appropriate number of clones were sequenced to 7-10-fold coverage (that is, each base is sequenced an average of 7-10 times). Two major software programs were used for assembly and editing by most of the AGI members (except TIGR, which used the in-house 'TIGR assembler' [26]): Phred/Phrap [27,28] for sequence assembly and Gap4 [29] or Consed9 [30] for viewing and editing. The AGI members used almost all the same annotation programs for gene prediction and annotation of the genome: programs such as Genefinder, Grail, Genscan, Xpound, tRNAscan-SE, BlastN, BlastX, Gene Mark HMM, Glimmer A, NetGene 2, Splice Predictor, Pedant and Repeat Masker (see [4]). The entire genome was also reannotated upon completion, by two AGI members, TIGR [24] and MIPS [31] (a member of the EU consortia), to ensure uniformity of the final product [4].

The adopted strategies allowed different groups around the globe to sequence different regions of the various chromosomes at the same time. Existing incomplete physical and genetic maps of each of the five chromosomes were used in selecting the seed BACs necessary to initiate the sequencing process. The incomplete physical chromosome maps were supplemented by fingerprint and hybridization data using 24,000 BACs, to yield a complete map of the genome with 99% coverage and resulting in an acceleration of the sequencing process [32,33].

A few additional crucial decisions were also made during the 1996 NSF meeting, regarding data release policy, acceptable error in the final sequencing product, and acceptable degree of completion. It was agreed that the sequence produced by the AGI should be immediately deposited in GenBank [34] even before it was finished and annotated. Accordingly, phase I genomic sequence - comprising raw sequence containing gaps and of unknown orientation - was available to the plant biology community at

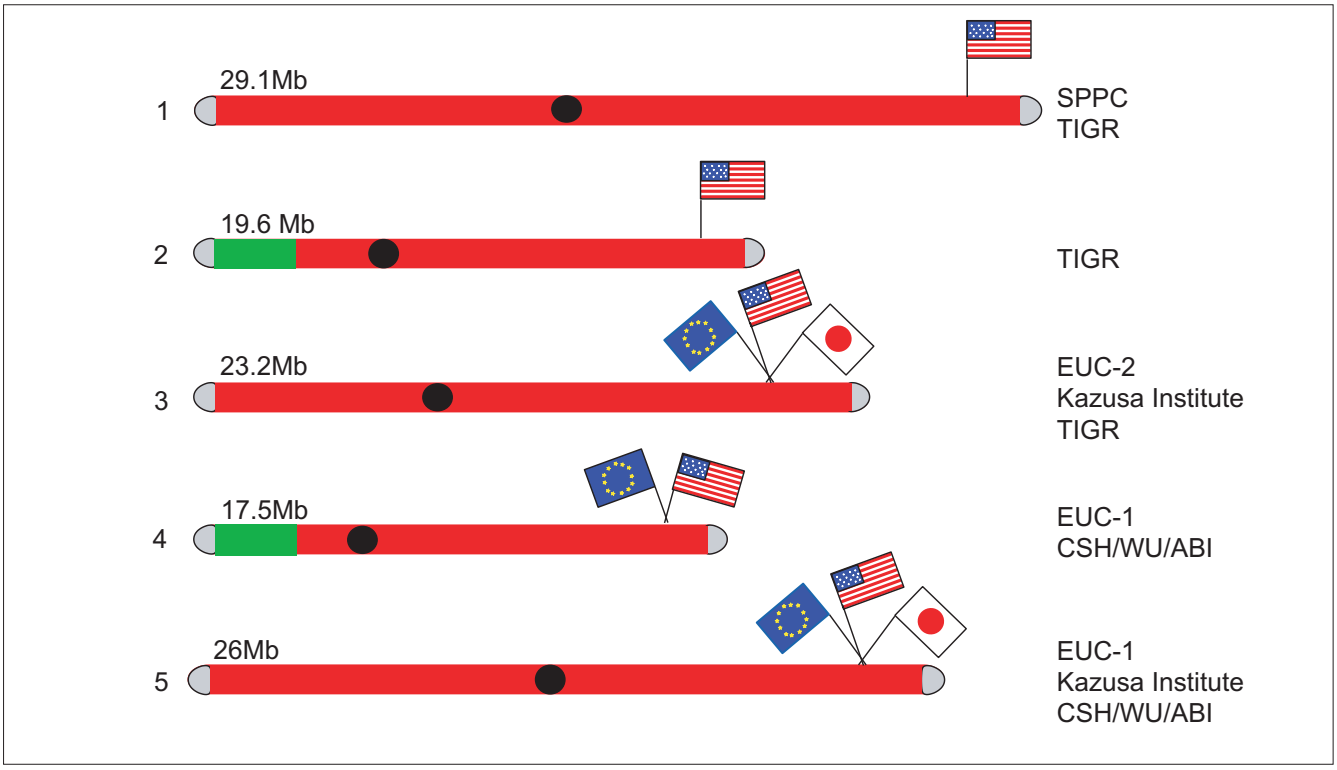

Figure 1
The *Arabidopsis* Genome Initiative (AGI), an international project for sequencing the genome of the flowering plant *Arabidopsis thaliana*. For each chromosome, black represents centromeres; grey telomeres; green rDNA repeats; and red contiguous DNA sequence. Sequencing groups are: TIGR, The Institute for Genomic Research (PIs: Craig Venter and Claire Fraser) [24]; Kazusa DNA research Institute (PI: Satoshi Tabata); SPPC, The SPP Consortium, Stanford, University of Pennsylvania, PGEC-USDA and University of California Berkeley (PIs: Ronald Davis, Joseph Ecker, and Athanasios Theologis) [25]; EUC-1 and EUC-2, European Union Consortia for Chromosomes 3, 4 and 5 (PIs: Michael Bevan and Francis Quetier); and the CSH/WU/ABI consortium of Cold Spring Harbor Laboratory, Washington University and Applied Biosystems, Inc. (PIs: Richard McCombie, Richard Wilson and Ellson Chen). PI, principle investigator.

the HTGS section of GenBank [34] a few days after the shotgun sequencing was completed. This immediate-release policy allowed plant molecular geneticists to clone their favorite gene by 'walking' much faster than before. There are numerous examples for which the availability of unfinished genome sequence allowed the cloning of genes, such as *axr3* [35] and *shy2* [36] (two examples that I know about because of my own research interests). It was also agreed that the acceptable sequencing error rate would be no more than 1 error per 10 kb, and that the finished sequence should be at least 97% double-strand sequenced, with the remaining 3% to be pseudodouble-stranded (defined as the sequence of a single clone obtained with two different chemistries or the sequence of two clones with the same chemistry). Regarding the extent of completion, the AGI agreed that the genome sequence would be considered complete when each chromosome was represented by only two contigs - chromosomal arms - separated by the centromeric region. In addition, each arm should end at the telomere repeat. The rDNA clusters of chromosomes 2 and 4 were also excluded from the finished product [37].

The DNA molecules

The AGI completed the *Arabidopsis* genome sequencing project four years ahead of schedule. Chromosomes 2 and 4 were published in December 1999 [38,39] and the remaining chromosomes 1, 3 and 5 [40-42], along with a uniform annotation and analysis of the entire genome [4], were published in December 2000. The accelerated pace was primarily due to the acceleration of funding by the NSF two years into the project, thanks to the vision of the NSF administrators [43], Mary Clutter, Machi Dilworth and the late DeLill Nasser. The rapid progress of the project during the first two years, and the excess of sequencing capacity of the participating groups, warranted such an action.

The final product represents the most completely sequenced eukaryotic genome thus far [4]. The ten chromosomal arms are without gaps, except for three at the bottom arm of chromosome 1 where there are highly repetitive DNA sequences. These gaps will be closed this year. The combined length of the ten arms, which extend from either the telomeres or the ribosomal DNA repeats to the 180 base-pair

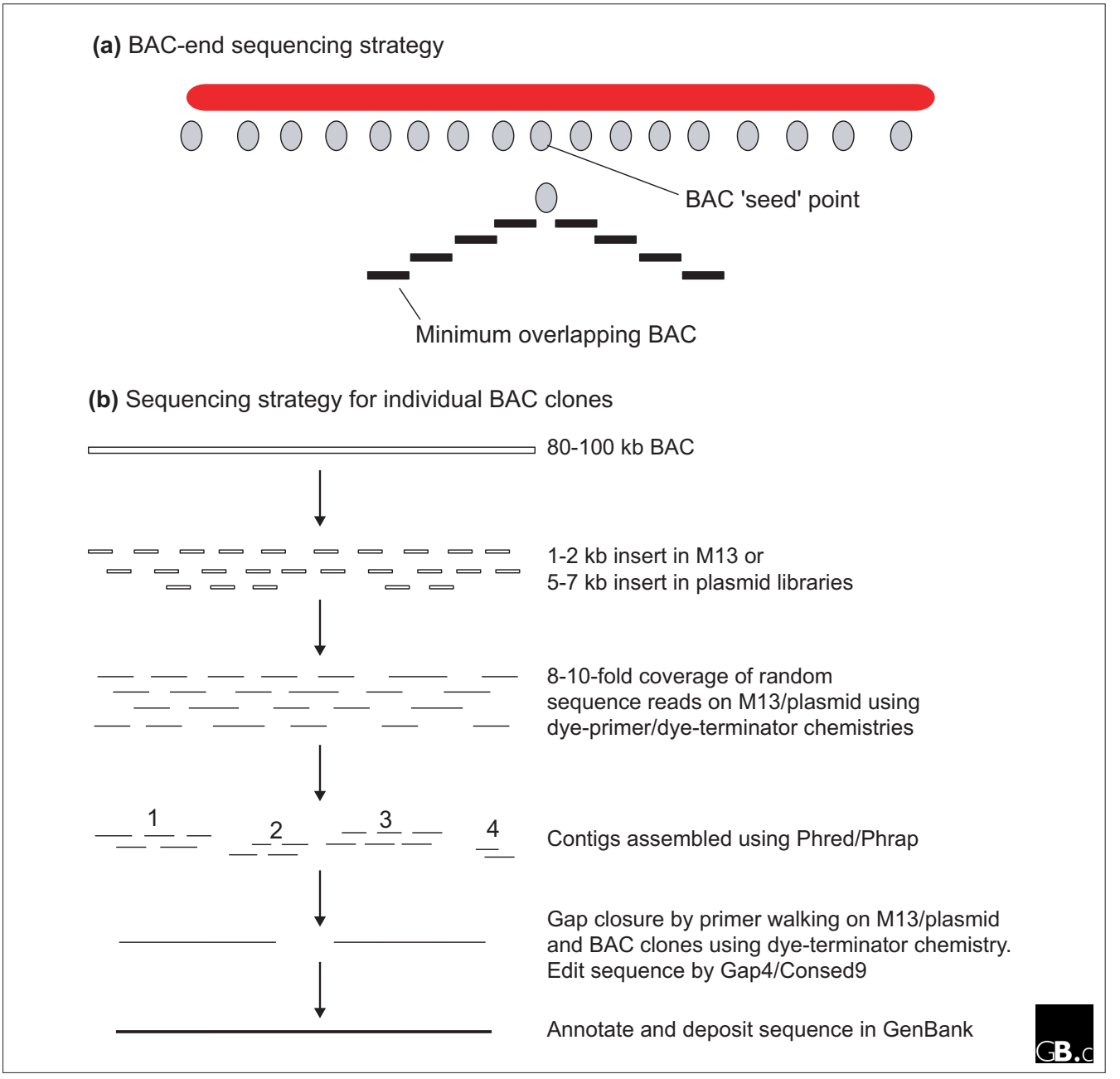

Figure 2
Sequencing strategies for **(a)** chromosomes and **(b)** individual BACs.

centromeric repeat is approximately 115 Mb (Table 1). The unsequenced centromeric and rDNA repeat regions are estimated to represent approximately 10 Mb, resulting in a genome size of about 125 Mb. The overall GC content is low (35%) and uniform across the five chromosomes, which made the DNA relatively easy to sequence using the available chemistry. The sequencing error rate is no more than 1 error per 20 kb, better than that agreed upon by the AGI in 1996. The collinearity of the assembled chromosomes was verified using BLAST alignment analysis of the BAC-end sequences against the assembled chromosomal contigs. In addition, the location of a large number of sequenced chromosomal markers on the assembled chromosomes is collinear with their location on the recombinant inbred linkage map [44]. The relationship between physical and genetic distance is around 200-250 kb per cM and is uniform across the five chromosomes [4]. The sequencing cost is estimated at 60¢ per base ($70 million for 115 Mb), corresponding to approximately 10¢ per citizen of

Europe, Japan and the USA (a combined population of about 600 million people).

Gene prediction programs and database searches were used to annotate and determine some of the features of the five chromosomes. The genome contains 25,500 putative genes at a density of 1 gene per 4 kb (Table 1). *Arabidopsis* contains as many genes as humans and twice as many as the fruitfly (Table 2). The average gene is 2 kb long and contains five exons. Thus, almost 50% of the chromosomal DNA is covered with genes (Table 1a). Approximately 20% of the genes are intronless, and some of them have been annotated as hypothetical. The genome includes genes of very different sizes. For example, chromosome 1 contains a gene (F5F8.4) with 68 exons that encodes a 500 kDa polypeptide that is a ubiquitous putative membrane protein found in yeast, worm, fly and human. The same chromosome contains a very small gene that encodes the L41 protein of the 60S ribosomal complex, with only 25 amino acids. At least 15,300 genes (60%) are expressed, as indicated by the presence of corresponding cDNAs or expressed sequence tags (ESTs) in various databases (Table 1). Genes that are most highly represented by ESTs include proteins of the photosynthetic and protein synthesis machineries and few metabolic enzymes, such as catalase. The distribution of ESTs is more or less the same across the length of the chromosomes but decreases dramatically in the regions flanking the centromeres [4].

The chromosomes encode approximately 600 tRNAs, with chromosome 1 bearing the majority of them (236 tRNAs [4,40]). The tRNA gene content of chromosome 1 is very similar to that of the entire fly genome [8]. The tRNA genes are evenly distributed along the chromosomes, except for two regions in chromosome 1 where tRNA gene clustering is observed [40]; the two clusters are located at 9.989 Mb and 19.282 Mb, respectively. The first contains 26 genes encoding $tRNA^{Pro}$ and the other 27 tandem repeats of the tri-repeat $tRNA^{Tyr}$- $tRNA^{Tyr}$- $tRNA^{Ser}$. The tRNA genes of eukaryotes in general occur as multigene families with a diverse arrangement. In some cases they are spread throughout the genome, whereas others are clustered at single chromosomal sites [45]. It has been suggested that tRNA gene clustering may reflect their tissue-specific co-regulation [45].

There are large-scale inter- and intra-chromosomal duplications in *Arabidopsis* chromosomes [4]. The duplicated regions constitute 68 Mb, or 60% of the genome. The number of homologous genes in the duplicated regions varies considerably, ranging from 20 to 50% [4]. This may be due to either tandem duplications or gene loss after segmental duplication [4]. In addition to the detection of the large-scale intra- and inter-chromosomal duplications, analysis reveals a plethora of diverse repetitive elements comprising 10% of the sequence. Retroelements include members of the LINE-like Ta11 family of elements, and long-terminal repeat (LTR) elements of both the Ty3-gypsy and the Ty1-copia families. Representative members of various transposable element families are scattered throughout the chromosome; for example, the Ac/Ds-(*Hat*/*mariner*), En-Spm-mutator and Tc1 type elements all occur. In addition, a number of simple and low-complexity repeats are found throughout the chromosomes. There is an inverse relationship between gene and retroelement density in the borders of the centromeric region, a hallmark of such chromosomal regions [46].

Table 1

Some features of the *Arabidopsis* chromosomes and proteome

	Number	%
(a) The five DNA molecules		
Total length (Mb)	~115	
Number of genes	~25,500	
Average gene density (kb per gene)	~4.5	
Average gene length (kb)	~2	
Average peptide length (amino acid)	430	
Number of genes with ESTs	15,300	60
Number of ESTs	105,733	
(b) The proteome		
Total proteins	~25,500	100
Proteins with similarity to GenBank entries	17,850	70
Unknown proteins	5865	23
Hypothetical proteins	7140	28
Proteins with putative functions	12,495	49
(c) Classification of proteins with putative functions		
Cellular metabolism		22.5
Transcription		16.9
Defense		11.5
Signaling		10.4
Growth		11.7
Protein fate		9.9
Intracellular transport		8.3
Transport		4.8
Protein synthesis		4.1

~ denotes approximate numbers.

Table 2

Gene content among sequenced eukaryotes

Organism	Approximate number of genes	Reference
Yeast	6,000	[6,54]
Worm	19,000	[7]
Fruitfly	14,000	[8]
Mustard weed	25,500	[4]
Human	25,000-40,000	[9,10]

The proteome

Computational analysis of the chromosomal sequences reveals that they encode 25,500 putative proteins (Table 1b). Approximately one third (28%) of the proteins are 'hypothetical', meaning that they are predicted by various gene-prediction programs but do not have corresponding ESTs or other evidence of expression. A quarter (23%) have unknown function, but they are known to be transcriptionally active because an EST corresponds to each of them. Thus, only half the predicted proteins have an identifiable putative function. The same analysis also reveals that 70% of the annotated proteins have some similarity to other hypothetical, unknown or putative-function proteins from plants and other eukaryotes, such as yeast, worm, fly and human, or include protein-family signature or motif sequences. Table 1b shows a functional classification of the proteins based on the amino acid motifs. The analysis was generated by PEDANT [47] and shows that almost 30% of this class of proteins participate in cellular metabolism and another 50% are involved in transcription, plant defense, signaling, and growth and development (Table 1b).

Comparison of the predicted proteins of the five chromosomes with other proteins from available complete genome sequences reveals that the absolute number of *Arabidopsis* gene families and singletons (proteins with no paralogs) is in the same range as in worm and fly, indicating that a proteome of 11,000 types is sufficient for multicellular life [4]. This analysis also reveals that proteins involved in RNA splicing, tRNA biosynthesis, translation and metabolism are highly conserved throughout the eukaryotic kingdom. Although numerous families of proteins are common among all eukaryotes, *Arabidopsis* has 150 unique protein families encoding transcription factors, structural proteins, enzymes and proteins of unknown function. The chromosomes carry 1,528 gene families with 4,140 gene members arranged in tandem. The number of adjacent members per family varies from 2 to 23 (Figure 3). Thus, 17% of all *Arabidopsis* genes are arranged in tandem arrays, and these are distributed throughout the chromosomes. For example chromosome 1 contains 300 gene families with tandem gene arrangement of their members; 175 families are unique, meaning that their members encode a set of defined protein isoforms. The remaining 123 are superfamilies consisting of more than one multigene family, varying in number of members from 2 to 5 [40].

The future

The completion of the first plant genome sequence is a milestone for biology, in general, and for plant sciences in particular. Although the sequence provides a wealth of information, considerably more experimental work is required in order for it to contribute to the advancement of plant science. Most of the genes and their predicted functions should be interpreted with caution. Mapping the transcriptional units of the *Arabidopsis* chromosomes is currently underway [48]. It will verify the annotation, experimentally. Oligonucleotide and cDNA chip technology [49,50] will allow mapping of the transcriptional units, leading to the isolation of full-length cDNA clones for all the proteins encoded by the *Arabidopsis* chromosomes - a set that has been dubbed the 'ORFeome'. Construction of the ORFeome will eliminate the need for cDNA library construction, and each transcribed gene will be represented at

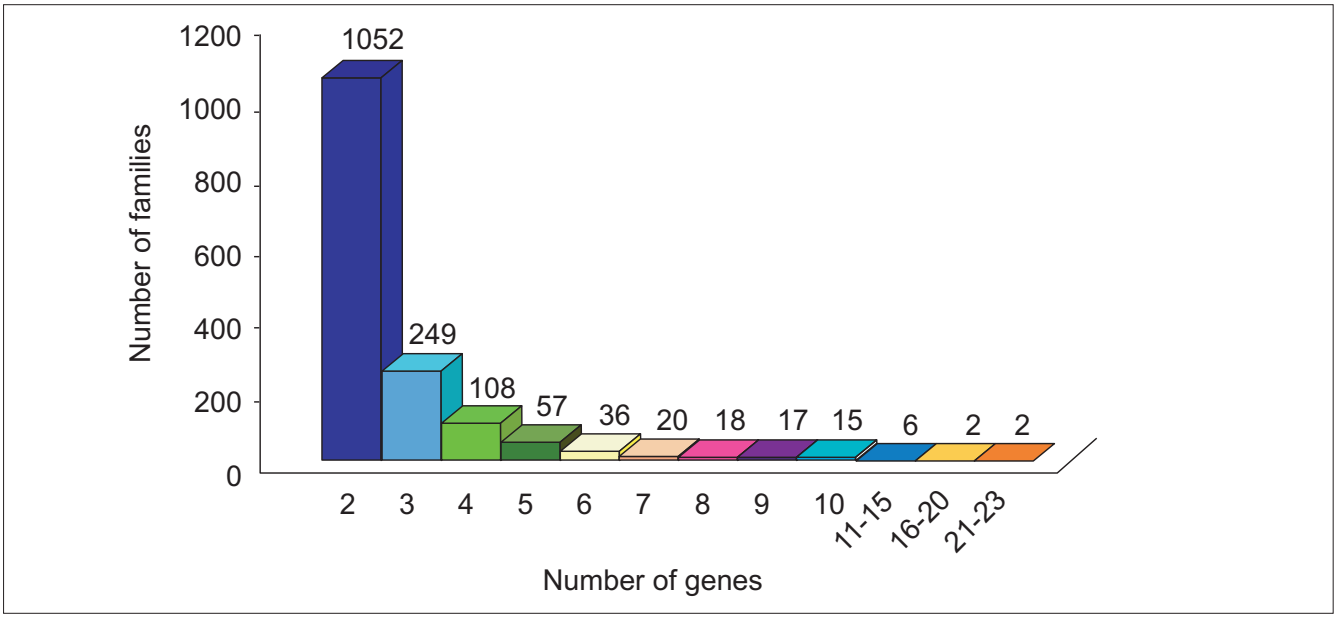

Figure 3
Frequency distribution of genes in multigene families with tandem gene arrangements. Figure adapted from [4].

equimolar concentration in the ORFeome. In addition, it will lead to the development of an *Arabidopsis* protein chip [51]. Concomitantly, the DNA sequence and chip technology will eliminate the need for Southern and northern hybridizations in *Arabidopsis*. Furthermore, the isolation of T-DNA insertional mutants for all the genes in the genome [52] will offer additional resources for elucidating the function of the genes by reverse genetics.

While the value of the *Arabidopsis* genome sequence will be greatly enhanced by the resources described above, its full potential will be realized only when the technique of gene transplacement by homologous recombination is developed in plants [53]. Only then will plant biology flourish and tremendous advances in agriculture be achieved. All the resources generated in the post-sequencing era will allow the elucidation of the biological and biochemical function of the *Arabidopsis* proteome. More importantly, we will be able to do more productive and meaningful experimentation leading to a deep and genuine understanding of how plant cells function [54].

A crucial task for the future will be the trying to understand the biological significance of the numerous multigene families. Why do so many gene products encode isoforms of the same polypeptide? This fundamental question applies for gene families with tandem gene arrangement as well as for families with dispersed gene arrangements, such as the *ACS* genes encoding ACC synthases (1-aminocyclopropane-1-carboxylate synthases). This family has two members (ACS2 and ACS10) on chromosome 1 and eight other members on the other four chromosomes. The question therefore arises as to why *Arabidopsis* has ten different *ACS* isoforms. It has been postulated that multiple *ACS* isoforms reflect tissue-specific expression of each, to satisfy the biochemical properties of the cells/tissues in which each is expressed. For example, if a group of cells or tissues has low concentrations of S-adenosyl methionine (Ado-Met), then these cells would express a high affinity (low K_M) *ACS* isoform. Accordingly, the distinct biological function of each isoform is defined by its biochemical properties, which in turn determine its tissue-specific expression pattern. Such a concept can accommodate gene families encoding enzymes as well as structural proteins [55].

The most frequent explanation for the presence of large numbers of multigene families in *Arabidopsis* is that it reflects functional redundancy: if something goes wrong with one gene product there is another to take over the lost function. But no two isoforms have completely overlapping functions [56], and most evolutionary biologists doubt the existence of functional redundancy. John Maynard Smith [56] has used theoretical considerations to deduce highly contrived situations in which it could occur, but many prominent geneticists consider his arguments strong evidence that it does not occur in nature. And there is experimental evidence to support such a conclusion: individual knockout of any one of the seven different oxysterol-binding protein genes in yeast yields a different expression profile, even though all seven genes have to be knocked out in order to reveal a lethal phenotype [57].

New technological breakthroughs in genomics will be required for elucidating the complex and repetitive structure of the centromeres. This will lead to the construction of artificial chromosomes. Furthermore, the tertiary structure of intact chromosomes has to be elucidated in order to understand how the 'packaging' of chromosomal DNA occurs within the nucleus [58]. Eventually, we will be able to chemically synthesize new chromosomes consisting of desirable sets of genes, package them *in vitro* and construct new plant species with superior agronomical properties.

The sequencing of the *Arabidopsis* genome signals the dawn of a golden era. Numerous genomes will be sequenced as sequencing technologies improve and the cost per base-pair decreases. We sequenced many individual genes during the 'one by one' era and many genomes will be sequenced in the era of genomics. It is only a matter of time. Only sequencing will reveal how plants evolved and will validate the various phylogenetic trees constructed using limited molecular information [59]. I believe that knowing the evolution of plants is just as important to science as knowing the evolutionary history of the human species.

The AGI proved to be a successful venture and was an appropriate one to achieve such a landmark. Thousands of young, bright, individuals have dedicated their intellectual and technical energies over the last five years to complete this project, using state-of-the-art molecular, engineering and computational technologies to achieve their goal. The AGI effort led to the development of high-throughput robotic instruments that were used for sequencing chromosome 1 [60]. Instruments currently made available to the genomics community by Gene Machines, Inc. [61], such as an M13 template preparation robot, the Mantis plaque and colony picker, the RevPrep 96-well plasmid-preparation robot and the PolyPlex 96-well oligonucleotide synthesizer**,** were developed and tested for the *Arabidopsis* sequencing project by the Stanford Genome and Technology Center, a member of the SPP consortium [25]. The entire effort was a communal undertaking, making a refreshing change from the competitive nature of modern science. The successful outcome of the AGI fulfills the expectation of Francis Crick [62] "You do not win battles by debating exactly what is meant by the word battle. You need to have good troops, good weapons, a good strategy, and then hit the enemy hard. The same applies to solving a difficult scientific problem." I am proud to have been a part of this battle whose victory holds such potential rewards for future generations.

References

1. **UN Long-Range World Population Projections** [http://www.undp.org/popin/wdtrends/longrange/lrfig1.htm]
2. Meyerowitz, EM: **Structure and organization of the *Arabidopsis thaliana* nuclear genome.** In *Arabidopsis.* Edited by Meyerowitz EM, Somerville C. Cold Spring Harbour: Cold Spring Harbor Press; 1994, 21-36.
3. Martienssen RA: **Weeding out the genes: the *Arabidopsis* genome project.** *Func Integr Genomics* 2000, **1:**2-11.
4. *Arabidopsis* Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**796-815.
5. Dujon B: **The yeast genome project: what did we learn?** *Trends Genet* 1996, **12:**263-270.
6. Goffeau A, Barrell GB, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston, M, *et al.*: **Life with 6000 genes.** *Science* 1996, **274:**546-567.
7. *C. elegans* Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, **282:**2012-2018.
8. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.*: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287:**2185-2195.
9. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, Fitzhugh W, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
10. Venter JC, Adams MD, Myers EW, Li P, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
11. Watson JD, Crick FHC: **Molecular structure of nucleic acids - a structure for deoxyribose nucleic acid.** *Nature* 1953, **171:**737-738.
12. Thomas M, Cameron JR, Davis RW: **Viable molecular hybrids of bacteriophage lambda and eukaryotic DNA.** *Proc Natl Acad Sci USA* 1974, **71:**4579-4583.
13. Maniatis T, Fritsch EF, Sambrook J: *Molecular Cloning - A Laboratory Manual.* Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 1982.
14. Sambrook J, Fritsch EF, Maniatis T: *Molecular Cloning: A Laboratory Manual, 2nd edition.* Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 1999.
15. Roberts, L: **Controversial from the start.** *Science* 2001, **291:**1182-1188.
16. Bevan M, Bancroft I, Bent E, Love K, Goodman H, Dean C, Bergkamp R, Dirkse W, Van Stavereb M, Stiekma W, *et al.*: **Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*.** *Nature* 1998, **391:**485-488.
17. Sato S, Kotani H, Nakamura Y, Kaneko T, Asamizu E, Fukami M, Miyajima N, Tabata S: **Structural analysis of *Arabidopsis thaliana* chromosome 5. I. Sequence features of the 1.6 Mb regions covered by twenty physically assigned P1 clones.** *DNA Res* 1997, **4:**215-230.
18. Venter JC, Adams MD, Sutto, GG, Kerlavage AR, Smith HO, Hunkapiller M: **Shotgun sequencing of the human genome.** *Science* 1998, **280:**1540-1542.
19. Venter JC, Smith HO, Hood L: **A new strategy for sequencing.** *Nature* 1996, **381:**364-366.
20. Kotani H, Sato S, Fukami M, Hosouchi T, Nakazaki N, Okumura S, Wada T, Liu Y-G, Shibata D, Tabata S: **A fine physical map of *Arabidopsis thaliana* chromosome 5: construction of a sequence-ready contig map.** *DNA Res* 1997, **4:**371-378.
21. Choi S, Creelman RA, Mullet JE, Wing R: **Construction and characterization of a bacterial artificial chromosome library of *Arabidopsis thaliana*.** *Plant Mol Biol Rep* 1995, **13:**124-128.
22. Mozo T, Fischer S, Shizuya H, Altmann T: **Construction and characterization of the IGF *Arabidopsis* BAC library.** *Mol Gen Genet* 1998, **258:**562-570.
23. **Genoscope** [http://www.genoscope.cns.fr].
24. **The Institute for Genomic Research (TIGR)** [http://www.tigr.org].
25. **SPP *Arabidopsis* Genome Sequencing Page** [http://sequence-www.stanford.edu/ara/SPP.html].
26. **TIGR Assembler** [http://www.tigr.org/softlab/assembler].
27. Ewing B, Green P: **Base-calling of automated sequencer traces using *Phred* II. Error probabilities.** *Genome Res* 1998, **8:**186-194.
28. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using *Phred* I. Accuracy assessment.** *Genome Res* 1998, **8:**175-185.
29. **UK Human Genome Mapping Project Resource Centre** [http://www.hgmp.mrc.ac.uk].
30. Gordon D, Abajian C, Green, P: **Consed: a graphical tool for sequence finishing.** *Genome Res* 1998, **8:**195-202.
31. **Munich Information Center for Protein Sequences (MIPS)** [http://mips.gsf.de]
32. Marra M, Kucaba T, Sekhon M, Hillier L, Martienssen R, Chinwalla A, Crockett J, Fedele J, Grover H, Gund C, *et al.*: **A map for sequence analysis of the *Arabidopsis thaliana* genome.** *Nature Genet* 1999, **22:**265-270.
33. Mozo T, Dewar K, Dunn P, Ecker JR, Fischer S, Kloskal S, Lehrach H, Marra M, Martienssen R, Meier-Ewert S, *et al.*: **A complete BAC-based physical map of the *Arabidopsis thaliana* genome.** *Nature Genet* 1999, **22:**271-275.
34. **GenBank** [http://www.ncbi.nlm.nih.gov/Genbank/]
35. Rouse D, Mackay P, Stirnberg P, Estelle M, Leyser O: **Changes in auxin response from mutations in an *AUX/IAA* gene.** *Science* 1998, **279:**1371-1373.
36. Tian Q, Reed JW: **Control of auxin-regulated root development by the *Arabidopsis thaliana SHY2/IAA3* gene.** *Development* 1999, **126:**711-721.
37. Goodman H, Ecker JR, Dean C: **The genome of *Arabidopsis thaliana*.** *Proc Natl Acad Sci USA* 1995, **92:**10831-10835.
38. Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, *et al.*: **Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*.** *Nature* 1999, **402:**761-768.
39. Mayer K, Schueller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Duesterhoeft A, Stiekema W, Entian KD, Terryn N, *et al.*: **Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*.** *Nature* 1999, **402:**769-777.
40. Theologis A, Ecker JR, Palm CJ, Federspiel NA, Kaul S, White O, Alonso J, Altafi H, Araujo R, Bowman CL, *et al.*: **Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**816-820.
41. Salanoubat M, Lemcke K, Rieger M, Ansorge W, Unseld M, Fartmann B, Valle G, Blocker H, Perez-Alonso M, Obermaier B, *et al.*: **Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**820-822.
42. Tabata S, Kaneko T, Nakamura Y, Kotani H, Kato T, Asamizu E, Miyajima N, Sasamoto S, Kimura T, Hosouchi T, *et al.*: **Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**823-826.
43. **National Science Foundation (NSF)** [http://www.nsf.gov]
44. Lister C, Dean, C: **Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*.** *Plant J* 1993, **4:**745-750.
45. Beier D, Stange N, Gross HJ, Beier H: **Nuclear tRNA^{Tyr} genes are highly amplified at a single chromosomal site in the genome of *Arabidopsis thaliana*.** *Mol Gen Genet* 1991, **225:**72-80.
46. Copenhaver GP, Nickel K, Kuromori T, Benito M-I, Kaul S, Lin X, Bevan M, Murphy G, Harris B, Parnell LD, *et al.*: **Genetic definition and sequence analysis of *Arabidopsis* centromeres.** *Science* 1999, **286:**2468-2474
47. **PEDANT** [http://pedant.mips.biochem.mpg.de]
48. **Salk Institute Genomic Analysis Laboratory (SIGNAL)** [http://signal.salk.edu]
49. Lockhart DJ, Winzeler EA: **Genomics, gene expression and DNA arrays.** *Nature* 2000, **405:**827-836.
50. Young RA: **Biomedical discovery with DNA arrays.** *Cell* 2000, **102:**9-15.
51. Kodadek T: **Protein microarrays: prospects and problems.** *Chem Biol* 2001, **8:**105-115
52. Bouché N, Bouchez D: ***Arabidopsis* gene knockout: phenotypes wanted.** *Curr Opin Plant Biol* 2001, **4:**111-117.
53. Scherer S, Davis RW: **Replacement of chromosome segments with altered DNA sequences constructed *in vitro*.** *Proc Natl Acad Sci USA* 1979, **76:**4951-4955.
54. Johnston M: **The yeast genome: on the road to the golden age.** *Curr Opin Genet Dev* 2000, **10:**617-623.
55. Rottmann WE, Peter GF, Oeller PW, Keller JA, Shen NF, Nagy BP, Taylor LP, Campbell AD, Theologis A: **1-aminocyclopropane-1-carboxylate synthase in tomato is encoded by a multigene**

family whose transcription is induced during fruit and floral senescence. *J Mol Biol* 1991, **222:**937-961.
56. Nowak MA, Boerlijst MC, Cooke J, Smith JM: **Evolution of genetic redundancy.** *Nature* 1997, **388:**167-171.
57. Beh CT, Cool L, Phillips J, Rine J: **Overlapping functions of the yeast oxysterol-binding protein homologs.** *Genetics* 2001, **157:**1117-1140.
58. Bamford DH, Gilbert RJC, Grimes JM, Sturat DI: **Macromolecular assemblies: greater than their parts.** *Curr Opin Struct Biol* 2001, **11:**107-113.
59. Pryer KM, Schneider H, Smith AR, Cranfill R, Wolf PG, Hunt JS, Sipes SD: **Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants.** *Nature* 2001, **409:**618-622.
60. Marziali A, Willis TD, Federspiel NA, Davis RW: **An automated sample preparation system for large-scale DNA sequencing.** *Genome Res* 1999, **9:**457-462.
61. **Gene Machines Inc.** [http://genemachines.com]
62. Crick, F: *The Astonishing Hypothesis, The Scientific Search for the Soul.* New York: Scribers: 1994.