Next in the sequence: the rice genome project aims to be complete by 2004.

MARK EDWARDS/STILL PICTURES

Now for the hard ones

Arabidopsis was an obvious choice for the first plant genome project, but it will never feed the world. David Adam reports on efforts to harvest the genomes of rice and other important crop plants.

No plant scientist would deny that the complete genome sequence of the thale cress *Arabidopsis thaliana* is a landmark. The information it contains should give a boost to just about every area of plant science. But many agricultural researchers would not have made *Arabidopsis* their number-one choice. As one cereal biologist asks mischievously: "When are we going to sequence a real plant?"

Arabidopsis was tackled first because it is a popular model organism. Just as its rapid life cycle and small size made it ideal for laboratory studies, its relatively small genome recommended it for sequencing.

Insights from the *Arabidopsis* sequence are likely to boost crop science, but the genomics of crop plants is a much bigger challenge. Many have genomes larger than our own (see table, opposite), thanks to a tendency for plants to carry duplicate copies of large sections of DNA. Often all their chromosomes are duplicated several times over, a phenomenon known as polyploidy.

Nevertheless, the rice genome should be available within four years. And although researchers are divided over the need to sequence more crops once *Arabidopsis* and rice are in the bag, others may eventually follow. "It will take a lot of money and improved technology, but I believe that the full genome of maize, at least, will be sequenced eventually," says Ed Coe, a maize researcher at the University of Missouri in Columbia.

Gigantic genomes

The genetic baggage of many plants is thought to be an evolutionary insurance against an unpredictable environment. As little as a quarter of a plant's genes may be essential for growth. But having multiple copies of genes increases the chance that a plant will possess a particular variant that will make the difference between, for example, surviving a drought or perishing.

This genetic extravagance makes plants difficult genomic targets, and not just because they have so much DNA to trawl through. The many stretches of identical 'junk' DNA throughout the genome can send sequencers round in circles as they try to stitch sequenced fragments together.

At about 400 million base pairs long, the rice genome is some four times larger than that of *Arabidopsis*, but still relatively small for a cereal crop. Its manageable size is one reason why the International Rice Genome Sequencing Project (IRGSP) was launched two years ago. More importantly, rice is a staple food for half of the world's population. And although global rice production has doubled over the past 30 years — thanks largely to the introduction of new varieties — a better knowledge of its genetics will be needed to continue this progress.

The bulk of the rice genes discovered so far were identified using fragments of complementary DNA, produced from the messenger RNA strands made when a gene is expressed — a good way to spot important genes. Some 40,000 of these expressed sequence tags (ESTs) have been published for rice so far, and many more are held in the proprietary libraries of agribiotech companies.

But studying genomes using ESTs alone is like doing astronomy with the naked eye. This is why Japan's Ministry of Agriculture, Forestry and Fisheries (MAFF) took the lead in setting up the IRGSP to decode the plant's 12 chromosomes. Originally intending to finish the job within a decade, this date has since been brought forward.

"We are very optimistic that we can finish by the end of 2004," says Takuji Sasaki, leader of MAFF's rice genome research programme. The project is making good progress, he adds, but success still depends largely on the scale of future funding, particularly in the United States.

The commercial sector also has rice in its

sights. In April 1999, Celera Genomics of Rockville, Maryland, caused a stir with its boast that it could sequence the entire rice genome in just six weeks. That project never got off the ground. "We never started it because the other organizations that we approached to cooperate all had their own programmes," says Bill Tucker, manager of business development and licensing at Celera's agricultural genomics division.

The multinationals Novartis and DuPont have built up their own rice DNA sequence databases, and the agribusiness giant Monsanto caught many plant researchers napping when it announced that it had completed a draft rice sequence earlier this year. The work was done under contract by a team led by Leroy Hood, then at the University of Washington in Seattle. Even more surprising was the company's pledge to turn all of its hard-won data over to the public IRGSP.

Six months later, however, there are doubts over the Monsanto sequence's accuracy and extent. "The quality is not so good in some places and we have to be very careful when using it," says Sasaki. This reflects the company's commercial interests, he suspects. Some gene-rich regions are thoroughly done, for example, but other areas are covered only briefly. "The draft should only be considered a scaffold to accelerate the sequencing of the rice genome," Hood says.

The draft should save the IRGSP — originally expected to cost US$200 million — time and money. Monsanto also stands to benefit, as the public project will finish the company's work by improving accuracy and plugging gaps — the most time-consuming part of any sequencing project.

The finished segments will then be released to the open databases, but the deal between the IRGSP and Monsanto allows the company to view the data as they are being accumulated. "Monsanto will learn a lot more about the rice genome by putting the information in the public's hands," says Benjamin Burr, an IRGSP member at the Brookhaven National Laboratory in New York state.

After the rice genome is finished, sequencers must decide whether to go after other cereal crops with larger genomes. Some experts believe it would be a waste of effort. Although different cereals have significantly different amounts of DNA, they share a common set of genes.

Wheat, rye, barley, maize, sorghum and millet all seem to have a similar genetic layout to rice — the genes are in much the same order but the larger genomes have more junk DNA between genes. This should mean that much of the information from the rice genome can be applied to other plants. "It would be stupid to sequence maize and wheat," argues Burr.

Species		Genome size (base pairs)
Brassicas		
Thale cress	*Arabidopsis thaliana*	1.0×10^8
Oilseed rape/ canola	*Brassica napus*	1.2×10^9
Cereals		
Rice	*Oryza sativa*	4.2×10^8
Barley	*Hordeum vulgare*	4.8×10^9
Wheat	*Triticum aestivum*	1.6×10^{10}
Maize/corn	*Zea mays*	2.5×10^9
Legumes		
Garden pea	*Pisum sativum*	4.1×10^9
Soya bean	*Glycine max*	1.1×10^9
Solanaceae		
Potato	*Solanum tuberosum*	1.8×10^9
Tomato	*Lycopersicon esculentum*	1.0×10^9
Human	*Homo sapiens*	3.2×10^9

Against the grain

But others disagree. "Rice is a useful framework for studying other cereals, but it isn't perfect," says Rob Martienssen of the Cold Spring Harbor Laboratory on Long Island. He thinks that efforts to sequence maize will be under way within ten years — and by 2003, if current trends continue, advances in technology will make it cheaper to sequence maize than it is currently to do rice.

Whether full-blown sequencing takes off or not, work on cereal ESTs will continue. Monsanto, Novartis and AstraZeneca are all investigating ESTs in maize; and Pioneer Hi-Bred, a seed company owned by DuPont and based in Des Moines, Iowa, claims to have identified about 80% of the genes expressed in maize. Publicly funded researchers are also making progress. A public database of maize ESTs coordinated by Virginia Walbot at Stanford University in California, for instance, now has more than 70,000 entries.

While cereal scientists debate the necessity of sequencing crops other than rice, those working on brassicas such as oilseed rape (canola) could soon be reaping benefits from the *Arabidopsis* sequence, as the model organism is a member of the same family. The insights they gain will flow in large part from a public functional genomics project called *Arabidopsis* 2010. Its ambitious aim is to work out the function of each of the plant's 25,000 genes within a decade, and enter them into a database with information about the proteins they code for.

As the functional information piles up, researchers working on other crops will discover how much the results from *Arabidopsis* apply to their species. Russel Kohel of the US Department of Agriculture and Texas A&M University in College Station, for instance, is relying on comparisons with *Arabidopsis* for his work on cotton. "We're not even considering sequencing," he says.

HOLT STUDIOS

Ear splitting: researchers are divided over whether the maize genome should be sequenced.

In the long run, many plant scientists would like to sequence at least one member of two other important plant groups: the legumes, which include peas, beans and other pulses; and the Solanaceae, which include potatoes and tomatoes. But for now, ESTs are the only game in town.

Randy Shoemaker at Iowa State University in Ames is striving to improve soya bean crops — funded in part by a cooperative of local farmers — and has banked over 130,000 publicly available ESTs. He aims to double this within 18 months, but does not expect the plant to be fully sequenced in the foreseeable future.

Meanwhile, plant genomics is spawning a series of public–corporate collaborations. DuPont is investing up to £15 million (US$22 million) in work on wheat at the John Innes Centre in Norwich, and the centre has agreed a £50 million, ten-year collaboration with AstraZeneca. In France, several research institutes and companies joined forces in 1998 to launch a $200 million initiative called Génoplante. And in 1999, Novartis signed a $50 million agreement with the Department of Plant and Microbial Biology at the University of California, Berkeley.

Like Monsanto, these companies are realizing that raw sequence information may be of little use without the context that academic researchers can provide. "There has been a realization that sharing and collaboration offer a way forward," says Burr. ■

David Adam is a member of *Nature*'s science writing team.

Web links:

Arabidopsis Genome Initiative
http://www.arabidopsis.org/agi.html
International Rice Genome Sequencing Project
http://rgp.dna.affrc.go.jp/Seqcollab.html
Monsanto Rice Genome Project
http://www.monsanto.com/monsanto/biotechnology/background_information/00apr03_rice.html
Arabidopsis 2010 Project
http://nasc.nott.ac.uk/garnet/2010.html
Maize EST database
http://www.zmdb.iastate.edu