Gene discovery in potato

W. De Jong[1], J. Davidson, G. Bryan, P. Birch, R. Waugh & D. Marshall

All characteristics of potato, including tuber shape, size, number, colour, skin finish, texture, flavour, cooking characteristics, utilisation of nutrients, susceptibility to low-temperature sweetening and resistance to pathogens, are controlled to a greater or lesser extent by its genes. Like other higher eukaryotes, potato probably encodes 50,000 to 100,000 genes, the vast majority of which have never been characterised. Obviously, knowing what the genes look like (in sequence terms), where each is encoded in the genome, and what each gene does, would greatly enhance our ability to improve potato, both through conventional crossing and selection schemes, and via targeted genetic modification.

Large international teams have recently determined the complete nucleotide sequence (and thus gene content) of two model organisms, yeast (15×10^6 base pairs) and the soil nematode *C. elegans* (70×10^6 bp). The first complete genome sequence of a higher plant, *Arabidopsis thaliana* (120×10^6 bp), has just been completed, and the sequence of the human genome (3×10^9 bp) is already available as a 'first draft'. This incredible volume of sequence information, and parallel developments in miniaturization and information management technology, are radically altering the questions that researchers can ask. For example, it is now possible, using microarray technology, to simultaneously measure the expression level of every yeast gene on a single microscope slide. To obtain a quick overview of how yeast regulates its metabolism, investigators have grown yeast under a wide variety of conditions, measured the expression level of every gene under each condition, and, by determining which genes are coordinately regulated, provided a remarkably detailed understanding of yeast metabolic control. Using the complete sequence as a guide, international teams are also in the process of systematically knocking out each gene, and then evaluating each mutant yeast line for altered behavior.

The current cost of sequencing, although much lower than just a few years ago, is still too high to justify sequencing the entire potato genome (1×10^9 bp), but it has nevertheless reached a point where large scale efforts targeted at specific portions of the genome can be considered realistically. In most higher eukaryotes, only a small portion of the genome actually codes for genes (for example, assuming that the average gene is 1000-bp in length, only 5-10% of the potato genome would consist of genes). What the rest of the genome does is not well understood – some of it performs necessary structural functions (e.g. sequences comprising the centromeres, telomeres etc.), while other sequences probably help to regulate the expression of the genes themselves, and some sequences appear to be 'junk' DNA (e.g. transposons). The regions corresponding to genes are unique in that these sequences are transcribed into messenger RNA, exported to the cytoplasm, and then translated into proteins, the functional actors of the cell. By isolating messenger RNA, reverse transcribing it into complementary DNA (cDNA), and then sequencing cDNA clones, it is possible to focus sequencing efforts on the genes themselves. The use of this approach on a very large scale was pioneered in 1993 by The Institute for Genomics Research (TIGR), based in the USA, who called the resulting sequences 'expressed sequence tags', or ESTs. Prior to their report, the scientific community generally assumed that sequencing was

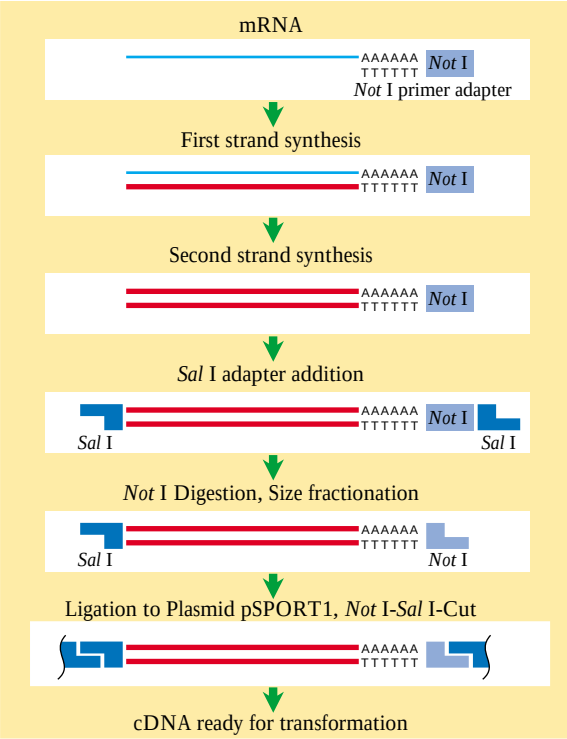

Figure 1 Summary of the cDNA cloning procedure.

[1] Cornell University, Ithaca, New York

Figure 2 Biomek 2000

best conducted by completely sequencing one gene at a time. TIGR demonstrated, however, that with automation and an emphasis on high throughput, more sequence information could be obtained per unit time/unit cost by sequencing fragments of randomly selected genes, even when the inevitable redundancy is taken into account. Since their seminal report, the EST approach to gene discovery has been applied to many organisms, and is now generally recognized as a logical first step in genome characterisation.

To date, despite its importance as the world's fourth most important food crop, very little effort has been directed at sequencing the potato genome. To help fill this gap, we recently initiated an EST program in potato. So far, we have constructed three cDNA libraries, one from shoots, and two from tubers, and have sequenced approximately 1000 clones from each library. An outline of the cloning procedure used is shown in Fig. 1. All of the sequencing has been from the 5´ end of randomly selected cDNA clones, to maximize the chance that at least some protein coding sequence would be detected in each clone. Sequencing this relatively large number of clones has required the development of some automation, which has so far been focused on template preparation, but will eventually be applied to sequencing reaction set-up and downstream data processing. In particular, with a Biomek 2000 robot (Fig. 2), we now routinely and inexpensively isolate template DNA from 384 bacterial cultures at a time, using an alkaline lysis isolation procedure first described by Bruce Roe of Oklahoma State University (http://www.genome.ou.edu/protocol_book/protocol_index.html). All of the sequences have been determined by examining reaction products on an Applied Biosystems 377 sequencer. With our current protocols, one technician can easily generate over 1000 sequences per month. The volume of sequence data has necessitated the development of a wide range of bioinformatics support at SCRI, to automate removal of vector sequences and homology searches to determine if any genes of known function are related to each EST, and to organize all of the information and database search results in an easily accessible, user friendly database – SPUDBase.

Since its inception, the potato EST program has served primarily as a tool of 'gene discovery', a source of sequences that can be browsed for genes of value to other projects at SCRI. For example, several ESTs have been identified that are clearly related to known disease resistance genes. Since we are currently trying to characterise genetically several sources of resistance to the white potato cyst nematode, each of these ESTs has been mapped, to determine whether any co-localize with resistance loci. In another project looking for potato genes activated during the defense response against the late blight pathogen, *Phytophthora infestans*, colleagues isolated a cDNA fragment corresponding to an induced gene. This cDNA was not full-length, nevertheless comparing it against our sequence database revealed that a full-length cDNA clone had already been characterised in the EST program.

Over the past few years, colleagues at SCRI have developed over 150 simple sequence repeat (SSR) markers from potato. Because of their highly-polymorphic, multi-allelic nature, SSR markers are very well suited for genetic studies in highly heterozygous, autotetraploid potato (see Ann. Rep. 1996/7, 96-98). About 30 of the 3000 EST sequences obtained to date have been found to contain SSRs. Most of the gene-derived SSRs are trinucleotide repeats, which is not surprising since only trinucleotide motifs can contract or expand without altering the reading frame of the gene. An example of an EST containing a microsatellite is shown in Fig. 3. In collaboration with colleagues at the International Potato Center in Peru, we are currently investigating how frequently these coding sequence SSRs reveal polymorphism, and for

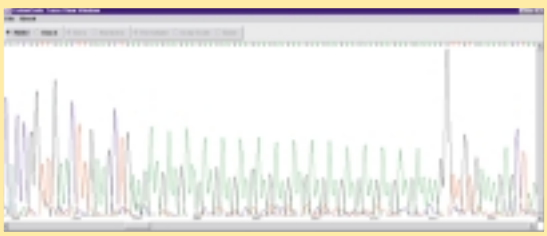

Figure 3 Sequence electropherogram of EST containing Simple Sequence Repeat with17 repeats.

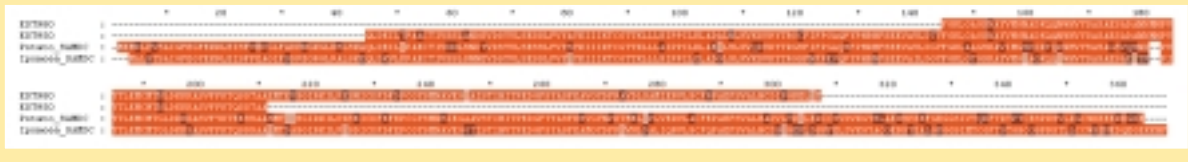

Figure 4 Alignment of SAMDC sequences.

those which do, determining genetic map locations. At least in theory, SSR markers based on coding regions should be useful in a wider range of germplasm than other SSRs, because primer annealing sites will evolve more slowly in coding than non-coding sequences.

One of the more interesting results of the exploratory sequencing so far is that several ESTs closely related to well known potato genes are even more closely related to genes from other species, suggesting a hitherto unrecognized ancient polyploidization or gene multiplication event in potato. For example, the potato gene encoding S-adenosyl methionine decarboxylase (SAMDC) has been intensively studied at SCRI[1]. Two overlapping ESTs (950 and 980) bear obvious sequence similarity to potato SAMDC, but appear to be more closely related to SAMDC from Japanese morning-glory *(Ipomoea)* and other plant species. Figure 4 shows an alignment of the two SAMDC-like potato ESTs against SAMDC sequences from the databases. This suggests the possibility of the existence of a second potato SAMDC gene - whose function still needs to be determined at the biochemical level. This was not suspected before it was sequenced, although in hindsight, it may provide an explanation for the residual SAMDC activity observed in transgenic potato plants expressing the original SAMDC in an antisense orientation, even though SAMDC transcripts could not be detected. Other interesting potato ESTs may represent a second hexokinase gene and a possible second sucrose phosphate synthase gene.

Sequencing several thousand potato genes is only a beginning, but still a very useful foundation. In the future, we hope to build on this work in at least three ways. The first is to use non-redundant ESTs to develop a very dense gene map of potato. In a typical mapping experiment, genes controlling a trait are localized to a chromosomal region that may contain anywhere from tens to thousands of genes. Knowing what genes are present in that region will greatly accelerate functional correlations between traits and genes. Secondly, high density microarrays of potato ESTs will be constructed to allow the expression of thousands of genes to be monitored simultaneously in a wide variety of experimental conditions, e.g. during tuber transition from dormancy to sprouting, or during response to pathogen attack. Thirdly, as increased sequence information becomes available and bioinformatics expertise develops, it will become possible to conduct large scale comparisons of all potato ESTs with the completely sequenced *Arabidopsis* genome. A significant issue in working with any crop plant is how to use functional information obtained in the intensively studied model plant *Arabidopsis*. Over the next five to ten years, systematic mutant screening will allow functions to be assigned to literally thousands of genes in *Arabidopsis*. When a gene in *Arabidopsis* is known to have only one clear relative in potato, inferences can be made with some confidence, but it becomes more complicated when the gene is part of a gene family in either or both species. It will also be possible to compare the extent of conservation of genome structure between potato and *Arabidopsis*. We have already seen occurrences of groups of several potato ESTs showing 'hits' to the same large segment ('contig') of *Arabidopsis* genome sequence. If we can demonstrate that such sets of ESTs, or a subgroup thereof, are closely linked in potato, suggesting some degree of genome collinearity, this will further increase the applicability of sequence information from the model to crop plant systems. Thus, large scale potato EST sequencing will help to provide a much needed context for judging gene relationships between the two species, and thus ensure that better use can be made of the very rapid progress in *Arabidopsis* and other model systems.

References

1 Pedros, A.R., MacLeod, M.R., Ross, H.A., McRae, D., Tiburcio, A.F., Davies, H.V. & Taylor, M.A. (1999). *Planta* 209, 153-160.