

General Information:

The potato genome

Every organism has a genome, a chemical 'instruction book' or 'blueprint' that describes how all the genes should be put together. This is written down as a DNA sequence, a long sentence made up of the chemical letters A, C, T and G. This sequence contains many tens of thousands of genes which can be thought of as 'words' in the sentence. Each gene controls different aspects of how the organism grows and develops. Slight changes in these instructions give rise to different varieties - each individual has a slightly different version of the DNA sequence for the species.

Understanding the complete genome sequence, the exact spelling of the DNA letters, for potato will help scientists develop a better understanding of how potato grows and develops, leading to improved crops worldwide.

Each copy of the potato genome consists of 12 chromosomes and has a length of approximately 840 million base pairs, making it a medium-sized plant genome.

A high quality, well-annotated genome sequence of potato will provide a valuable foundation which can be combined with existing knowledge of potato genetics and the continuing advances in analysing which genes are switched on or off and which chemicals are produced when. Observing how these changes are affected by changes in the genome will allow scientists to identify different variants of genes which are responsible for important quantitative traits in potato.

The Potato Genome Sequencing Consortium (PGSC) seeks to provide such a resource to the potato research and breeding community in the near future, allowing the full potential of biotechnology-based improvement of this important crop plant to be realized.

The Potato Genome Sequencing Consortium aims to read the complete genome sequence for potato by the year 2010.

About potato

Potato is a member of the *Solanaceae*, a plant family that includes several other economically important species, such as tomato, eggplant (aubergine), petunia, tobacco and pepper. Potato is an important global food source. After wheat and rice, potato is the third most important food crop, with a world-wide production of

309 million tons in 2007. By 2020 it is estimated that more than two billion people worldwide will depend on potato for food, feed, or income. Improving potato varieties so that they can better cope with environmental challenges such as drought, and pests or diseases are key objectives of global potato breeding programs.

The potato has one of the broadest genetic diversities of any cultivated plant. Wild species of potato are very widely distributed in the Americas, from the South Western USA to Southern Chile and Argentina and from sea level to the highlands of the Andes Mountains. Many wild species can interbreed directly with the common potato and possess a wide range of valuable traits such as resistance to pests and diseases or tolerance to frost and drought, making them a useful resource for breeding new varieties.

Worldwide, an economic loss on the potato crop of about €3 billion per year is estimated from diseases such as late blight. These diseases are still largely controlled by frequent application of fungicides. *It is expected that one of the first benefits of knowing the potato genome sequence will be a major breakthrough in our ability to characterize and select genes involved in disease resistance.*

However, potato, like man, has two, slightly different, copies of the genome (it is polyploid). It gets one copy of its genome sequence from the mother plant, and a separate and slightly different one from the father. This makes studying potato genetics complicated and many important traits are poorly understood. Yet, an understanding of its genetic composition is a basic requirement for developing more efficient breeding methods. The potato genome sequence will provide a major boost to gaining a better understanding of how potato traits are linked to genes, underpinning future breeding efforts. Currently, non-genome led potato breeding takes about 10-12 years to develop a new variety. It is expected that being able to use the genome information will dramatically shorten the time taken to breed new varieties as well as reducing the cost.

Potato Genome Sequencing Consortium

The international Potato Genome Sequencing Consortium (PGSC) is a collaboration between 16 research groups in 13 countries; Argentina, Brazil, China, Chile, India, Ireland, The Netherlands, New Zealand, Peru, Poland, Russia, the United Kingdom and the United States. The PGSC has its basis in long-standing research on the molecular genetics of potato within the partner

organizations, and includes partners with world-leading expertise in genome sequencing and computational analysis.

Each partner raises the funding needed to contribute to the project independently, mostly through grants from government research agencies and industry bodies.

Technical details

The PGSC is sequencing two genotypes:

RH89-039-16 (RH), a diploid, heterozygous potato

DM1-3 516R44 (DM), a doubled monoploid.

RH89-039-16

The PGSC originally started out with sequencing RH genotype. This part of the project builds on a diploid potato genomic bacterial artificial chromosome (BAC) clone library of 78,000 clones, which has been fingerprinted and aligned into ~7000 physical map contigs. In addition, the BAC-ends have been sequenced and are publicly available. Approximately 30,000 BACs are anchored to the Ultra High Density genetic map of potato, composed of 10,000 unique AFLP™ markers.

From this integrated genetic-physical map, between 50 to 150 seed BACs have currently been identified for every chromosome. Fluorescent in situ hybridization experiments on selected BAC clones confirm these anchor points. The seed clones provide the starting point for a BAC-by-BAC sequencing strategy. This strategy is being complemented by whole genome shotgun sequencing approaches using both 454 GS FLX and Illumina GA2 instruments. Assembly and annotation of the sequence data will be performed using publicly available and tailor-made tools. The availability of the annotated data will help to characterize germplasm collections based on allelic variance and to assist potato breeders to more fully exploit the genetic potential of potato.

DM1-3 516R44

Sequencing of DM was started because the overall progress in RH was slow. The heterozygosity of RH has limited the progress of physical mapping and will

complicate the assembly of the genome sequence. Whole genome shotgun sequencing of DM1-3 516R44 (CIP801092), a doubled monoploid potato clone, is expected to eliminate the complexity in assembly.

Sequencing technologies

The PGSC have used three different technologies to obtain the genome sequence.

Sanger sequencing - the traditional 'one read at a time' technology that was used to sequence the human genome

Solexa sequencing - A 'Next Generation Sequencing' platform where millions of pieces of the genome can be read at the same time, but it is not known where in the genome they come from.

454 sequencing - Another 'Next Generation Sequencing' platform where hundreds of thousands of sequences can be read at the same time.

Sequencing in layman's terms:

What is a genome sequence?

A genome is the complete set of DNA letters that describe how an organism is made up. DNA is a long string of four different chemicals: Adenine, Cytosine, Guanine and Thymine. This four letter alphabet spells out the instructions, or genes, that control how an organism such as a plant or a person works. These instructions are many thousands of letters (known as bases) long, and are embedded in a library that is hundreds of millions of bases long. It is estimated that potato contains 840 million bases, about one sixth the size of the human genome which contains 5 billion bases. Sequencing the genome is finding out the exact order of all the bases so we can spell out the entire genome. To make things more complicated, potato, like human, gets one copy of the genome from one parent, and one copy from the other parent. These are slightly different, and different combinations of these differences are responsible for the differences we see between potato varieties, just as people differ from one another and from their parents.

How was the genome sequenced?

Getting the exact sequence is very challenging. Even with the best technologies available we can only read a few hundred letters at a time so we have to find the sequence of small pieces then put them together, like reading the complete works of Shakespeare a few words at a time. The initial strategy that was used was to break the genome sequence up randomly into smaller chunks of about 100,000 bases. These can be separated and grown up in bacteria as bacterial artificial chromosomes (BACs). To use our Shakespeare analogy, this is like taking one page at a time. We can use careful fingerprinting to match them to each other and work out which of these pieces go with others, much like taking all the jigsaw edge pieces or all the sky pieces and putting them into separate heaps when trying to put a jigsaw together. Or looking for key words, such as the name of characters, to work out which Shakespeare play is on the page we are reading. At this point we still do not have the sequence, we just know which parts go together and, in some but not all cases, the approximate order they line up in the genome.

Each of these smaller pieces can then be sequenced one at a time. They are still too big to read in one go so they are broken up into many small pieces which can be read individually. Each of these small sequence reads can then be compared to each of the others and where the 'words' in one are identical to the 'words' in another, joined together to form longer sentences, and then paragraphs known as 'contigs'. This technique is known as 'shotgun sequencing'. With enough small sequence reads, the full sequence of each of these BACs can be determined then placed together to form the final genome sequence. This is how the human genome was sequenced.

There are drawbacks with this approach. Each tiny fragment of sequence must be read one at a time, a time consuming and costly approach. Also you do not want to sequence lots of BACs that cover the same portion of the genome as this would be rereading the same parts over and over again - a waste of time and money.

A few years ago a major new set of sequencing technologies appeared. Instead of reading the sequence of one piece of DNA at a time, they could read the sequence of millions of pieces of DNA at the same time. Instead of having to prepare tens of thousands of individual BACs, the whole genome could be read at once - an approach known as 'whole genome shotgun'. This is not without difficulty. Whilst we can read many individual short sequences, and in the case of potato over 1 billion short sequences have been read, putting the jigsaw back

together is very difficult. It becomes even more difficult when there are two slightly different (about one letter in every 60-100) versions in the mixture so a special research strain of potato was identified that has only one version of the genome. This is the variety that has now been sequenced.

Putting together this huge jigsaw puzzle required a large computer and novel computer programs. Potato is one of the largest plant genomes to be sequenced in this manner. In the first stages sequence overlaps were used to create short contigs. These can be joined together using other information, such as knowing which short reads are at opposite ends of the same fragment of DNA (we can only read the sequence from one end of a piece of DNA at a time) and using lots of information from the large collection of BACs that have been fingerprinted and part-sequenced. These then form larger scaffolds which have regions where we now know the sequence joined by gaps where we don't know the sequence but we do know how big the gaps are.

Once the parts where we know the sequence are of a usable size, typically where they can contain the full instructions for an individual gene, then they are useful to the rest of the research community and are released as a draft sequence. Much like a jigsaw where we have most of the major features put together and can see what the overall picture is, but still with many tricky pieces remaining to be placed to fill in the detail of the gaps. This is the stage the potato sequence is at in September 2009 where the majority of the genome is in several hundred large pieces and we estimate that 95% of genes have their complete sequence in one sequence piece.

The next stage is to improve the draft assembly. We are actively joining contigs and scaffolds together, and working out the order in which they sit in the genome. There will be some parts (hopefully very few) where we have made minor mistakes in piecing things together. The extensive analysis we are now beginning will help us to recognize and correct these so when the genome sequence is properly published around the end of 2009, the research community will be able to have full confidence in it. We are making the draft available early as there is already much information that can be used by other researchers.

What do you do when you have the sequence?

Getting the potato sequence is only the beginning. We now have to identify where all the genes, the sets of instructions for how the plant grows, are located. We need to work out what they do, both using sophisticated computer programs to predict their function and by individual scientists testing these genes in the lab. Then we need to see how these change across varieties by taking regions of interest and looking to see what those short sequences are like in other varieties. Important areas for research are identifying genes that affect the nutritional quality of potato, its resistance to pests, especially Potato Cyst Nematode in the UK and Phytophthora Infestans, the potato blight. Of interest to agriculture are genes that allow potatoes to be drought tolerant, and many, many more properties. Having the genome is just the start of being able to read the complete genetic instruction book for the potato.

What benefits will knowing the genome sequence bring?

All the properties described above could be developed without knowing the genome sequence, but it would take a long time and be very costly in terms of trial and error. If we know the genetic fingerprint for the properties we desire then these can be selected very early on in breeding programs, saving many years from the current 10-12 year development cycle and reducing the resources needed to develop these new strains. This genetic selection approach is very promising and technology to exploit the genome sequence immediately is already being prepared in both the UK and the Netherlands.

Understanding the genetic blueprint for potato also raises the option of genetically modifying the crop to engineer specific properties. The commercial prospects for future development of genetically modified potato remain unclear at present.