PROJECT SUMMARY

*"**Coffee and Genomics**"*

The Coffea arabica genome sequencing project

Project coordination by Prof. Giorgio Graziosi, DNA Analytica Srl (a Trieste University spin off), in collaboration with Padua University, Udine University and the Udine Applied Genomics Institute.

Research commissioned by: Lavazza and illycaffé.

Sequencing of the *Coffea arabica* genome

DNA sequencing is the process of determining the order of the various nucleotides (and therefore of the four nitrogenous bases which distinguish between them, adenine, cytosine, guanine and thymine) that make up nucleic acid, which contains all genetic information. Determining the genes present in the sequence, and the instructions for their expression in time and space, is a fundamentally important aspect of research into the whys and hows of living organisms.

The sequencing of the coffee genome is therefore of great interest not only for science, but also for agriculture and food production, and has economic and other knock-on effects for both producer and consumer countries. Understanding the coffee genome can, for example, help identify the genes which contribute to improving resistance to diseases and infections, fruit ripening synchrony, plant size and adaptation to difficult conditions. Caffeine content and the thickness and solubility of cell walls, which affect the extraction of coffee as a drink from ground beans, are other aspects that are controlled either directly or indirectly by genes. Obtaining good genetic information about the plant is therefore indispensable for producing a good quality product and, therefore, good coffee.

This highly complex research project intends, for the first time, to sequence and reconstruct the genome of a tetraploid or 4n (quartets of similar

chromosomes) organism deriving from the union of two progenitors, most probably *Coffea canephora* and *Coffea eugenioides.* This polyploidy complicates the reconstruction of the genome by turning it into something of a jigsaw puzzle.
The aim of the project is to decode the DNA of *Coffea arabica* and store the results in special databases.
The *Arabica* genome is made up of about 2.6 billion bases with a physical length of about 140 cm. Because of this sheer size, work has been split into various stages, with different strategies developed for each of them.
The five main stages into which the research is divided are: 1) Preparation of a BAC genomic library; 2) Physical DNA sequencing from BAC clones; 3) Physical sequencing of entire genomic DNA; 4) Computational reconstruction of sequences; 5) Annotation and identification of encoding genes.

1) Preparation of a BAC genomic library

The size and complexity of *Coffea arabica* DNA make direct sequencing of the entire genome very complicated, most importantly because of its polyploidy. As a result of this, some sequence pairs are very similar but not identical, each deriving from one of the two ancestral genomes. For this reason, an innovative hybrid strategy has been used that combines traditional direct sequencing of the entire genome with BAC clone sequencing. The genome therefore had to be fragmented into lots of smaller segments, each of which was then inserted into a vector made up of bacterial artificial chromosomes (BAC). This strategy was used to obtain a collection of bacterial colonies (clones), containing at least one copy of all the DNA sequences present in the genome, referred to as a genomic library. The initial stage ends with the creation of a BAC genomic library containing about 175,000 clones, each of which contains a fragment of *Arabica* DNA made up of about 100,000 nucleotides.

2) Physical sequencing of the BAC DNA

Each clone was placed in a microbiological culture so that the bacterial cells

multiplied, automatically multiplying the copies of the fragment of *Arabica* DNA, in order to produce a sufficient quantity of DNA for successful sequencing. Work then began on sequencing proper, which for reasons of cost and genomic library redundancy regarded only 36,864 of the 175,000 clones obtained, forming 96 pools containing 384 clones each.

3) Physical sequencing of the entire genomic DNA

The strategy used involved the production of collections of fragments of genomic DNA of various sizes, each of which was sequenced to both ends, obtaining about 100 gigabytes of total sequence.

4) Computational reconstruction of the sequences

The end result of the sequencing process is a broad collection of small random sequences that have to be assembled to obtain a consensus sequence that covers the entire genome. In order to do this, the bacterial vector DNA is removed from the raw sequence data, leaving only the sequences of the fragments of *Coffea arabica* DNA. The sequence of every individual fragment of *Arabica* was then compared with the sequence of all the other fragments to establish if the tail of one fragment matched the head of any other fragment and the two could be joined in at least a partial reconstruction of the genome. Over 50% of Arabica DNA has been reconstructed using this process and the work now underway to combine the two strategies described in points 2 and 3 should get closer to a full reconstruction.

5) Annotation and identification of the encoding genes

This stage of the project mainly involves the application of various bioinformatic systems to analyse the genomic sequencing data, in particular for gene prediction and annotation purposes. Prediction involves the identification of the genes, or in other words those parts of the protein encoding DNA that are actually responsible for the characteristics of the plant and the beans. Annotation refers to the functional description of the

proteins coded by each individual gene identified in the prediction process.

The process used to recognise genes at DNA level takes various different methods into consideration.
The next and final step is the characterisation of the predicted genes, in order to assign a possible biological function to them. This stage, referred to as annotation, requires considerable computing resources to search for matches against various different protein sequence databases, protein domains and known functional motifs. A preliminary list of the predicted genes has been drawn up, based on the reconstruction of the genome available today. This characterisation of the genes is incremental and proceeds as new elements are gradually added to the reconstruction of the genome.

Conclusion

The sequencing of the *Coffea arabica* genome forms part of an ambitious research programme that aims to reconstruct, for the first time, the entire genome of a tetraploid organism, in which the two ancestral genomes are extremely similar.
The aim, difficult though it may be, is to be able to distinguish the sequences that derive from the *C. eugenioides* progenitor from those deriving from the *C. canephora* progenitor. A good understanding of the genetic material is fundamental for plant, and therefore seed, quality because all their structures and functions are dictated by the sequence of DNA nucleotides, which determines the sequence of the amino acids in the protein molecules.
These considerations are the starting point for the project, which has so far achieved a good level of decoding of the genetic base of the *Arabica* coffee plant.
BAC clones have been sequenced and an initial reconstruction of the genome has been completed. Transcripts of the leaves and roots of several varieties of Arabica have been sequenced in order to identify the encoding

genes, or in other words the sections of effectively active DNA on which the characteristics of the plant and therefore of the coffee bean depend. Powerful bioinformatic techniques are being used to identify the possible functions of the encoding genes. The sequencing of the Arabica genome is expected to have knock-on effects in various sectors, not least economically, with possible applications in agriculture and industry. In addition, structural and functional genomic analysis may explain how the two genomes of *C. eugenioides* and *C. canephora* combine and interact in the C. arabica genome to give it its distinctive characteristics, which do not represent the simple sum of the characteristics of the two ancestral species, and which make it so sought after for the production of coffee to drink.