1 Omics: Opening up the “Black Box” of the Phenotype

Roberto Fritsche-Neto\textsuperscript{a} and Aluízio Borém\textsuperscript{b}
\textsuperscript{a}Department of Genetics, University of São Paulo/ESALQ, Piracicaba, SP, Brazil
\textsuperscript{b}Department of Crop Science, Federal University of Viçosa, Viçosa, MG, Brazil

From the time that is believed agriculture began, in approximately 10,000 BC, people have consciously or instinctively selected plants with improved characteristics for cultivation of subsequent generations. However, there is disagreement as to when plant breeding became a science. Plant breeding became a science only after the rediscovery of Mendel’s laws in 1900. However, some scientists disagree with this view. It was only in the late 19th century that the monk Gregor Mendel, working in Brno, Czech Republic, uncovered the secrets of heredity, thus giving rise to genetics, the fundamental science of plant breeding.

Scientists added a few more pieces to the puzzle that was becoming this new science in the first half of the 20th century by concluding that something inside the cells was responsible for heredity. This hypothesis generated answers and thus consequent new hypotheses, leading to the continuing accumulation of knowledge and progress in the field. For example, the double helix structure of DNA was elucidated in 1953 (Table 1.1). Twenty years later, in 1973, the first experience with genetic engineering opened the doors of molecular biology to scientists. The first transgenic plant, in which a bacterial gene was inserted stably into a plant genome, was produced in 1983. Based on these advances, futuristic predictions about the contribution of biotechnology were published in the media, both by laypeople and scientists themselves, creating great expectations for its applications. Euphoria was the tone of the scientific community. Many companies, both large and small, were created, encouraged by the prevailing enthusiasm of the time (Borém and Miranda, 2013).

Many earlier predictions have now become reality (Table 1.1), leading to the consensus that each year the benefits of biotechnology will have a greater impact on breeding programs. Consequently, new companies...
Table 1.1 Chronology of major advances in genetics and biotechnology relevant to plant breeding. Adapted from Borém and Fritsche-Neto (2013).

<table>
<thead>
<tr>
<th>Year</th>
<th>Historical landmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>1809–1882</td>
<td>Charles Darwin develops the theory of natural selection: those individuals most adapted to their environment are selected, survive, and produce more offspring.</td>
</tr>
<tr>
<td>1865</td>
<td>Gregor Mendel establishes the first statistical methodologies applicable to plant breeding, giving rise to the “era of genetics,” with his studies on the traits of pea seeds.</td>
</tr>
<tr>
<td>1910</td>
<td>Thomas Morgan, studying the effects of genetic recombination in <em>D. melanogaster</em>, demonstrates that genetic factors (genes) are located on chromosomes.</td>
</tr>
<tr>
<td>1941</td>
<td>George Beadle and Edward Tatum demonstrate that a gene produces a protein.</td>
</tr>
<tr>
<td>1944</td>
<td>Barbara McClintock elucidates the process of genetic recombination by studying satellite chromosomes and genetic crossing-over related to linkage groups in chromosomes 8 and 9 of maize.</td>
</tr>
<tr>
<td>1944</td>
<td>Hunter and Markert develop biochemical markers based on the expression of enzymes (isoenzymes).</td>
</tr>
<tr>
<td>1947</td>
<td>Herbert Boyer discovers restriction enzymes, opening new perspectives for DNA fingerprinting and the cloning of specific regions.</td>
</tr>
<tr>
<td>1953</td>
<td>James Watson and Francis Crick, using X-ray diffraction, propose the double helix structure of the DNA molecule.</td>
</tr>
<tr>
<td>1957</td>
<td>Recombinant DNA technology begins with the first cloning of a DNA fragment.</td>
</tr>
<tr>
<td>1972</td>
<td>Stanley Cohen and Herbert Boyer perform the first genetic engineering experiment on a microorganism, the bacterium <em>Escherichia coli</em>. The result was considered to be the first genetically modified organism (GMO).</td>
</tr>
<tr>
<td>1975</td>
<td>Sanger develops DNA sequencing by the enzymatic method; in 1984, the method was improved, and the first automatic sequencers were built in the 1980s.</td>
</tr>
<tr>
<td>1977</td>
<td>Maxam and Gilbert develop DNA sequencing by chemical degradation.</td>
</tr>
<tr>
<td>1980</td>
<td>Botstein et al. develop the RFLP (Restriction Fragment Length Polymorphism) technique for genotypic selection.</td>
</tr>
<tr>
<td>1983</td>
<td>The first transgenic plant is produced, a variety of tobacco into which a group of Belgian scientists introduced kanamycin antibiotic resistance genes.</td>
</tr>
<tr>
<td>1985</td>
<td>Genentech becomes the first biotech company to launch its own biopharmaceutical product, human insulin produced in cultures of <em>E. coli</em> transformed with a functional human gene.</td>
</tr>
<tr>
<td>1985</td>
<td>The first plant with a resistance gene against Lepidoptera is produced.</td>
</tr>
<tr>
<td>1986</td>
<td>The first field trial of transgenic plants is conducted in Ghent, Belgium.</td>
</tr>
<tr>
<td>1987</td>
<td>The first plant tolerant to a herbicide, glyphosate, is created.</td>
</tr>
<tr>
<td>1987</td>
<td>Mullis and Faloona identify thermostable Taq DNA polymerase enzyme, which enabled the automation of PCR.</td>
</tr>
<tr>
<td>1988</td>
<td>The first transgenic cereal crop, Bt maize, is developed.</td>
</tr>
<tr>
<td>1990</td>
<td>Rafalski et al. (1990) develop the first genotyping technique using PCR, RAPD (Random Amplified Polymorphism DNA).</td>
</tr>
</tbody>
</table>
Table 1.1 (continued)

<table>
<thead>
<tr>
<th>Year</th>
<th>Historical landmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>New tools for NCBI sequence alignment are created (BLAST – Basic Local Alignment Search Tool) (National Center of Biotechnology and Information – <a href="http://www.ncbi.gov">www.ncbi.gov</a>).</td>
</tr>
<tr>
<td>1994</td>
<td>The first permit is issued for the commercial cultivation and consumption of a GMO, the Flavr Savr tomato.</td>
</tr>
<tr>
<td>1997</td>
<td>The first plant containing a human gene, the human protein c-producing tobacco, is produced.</td>
</tr>
<tr>
<td>2000</td>
<td>The first complete sequencing of a prokaryotic organism, the bacterium <em>E. coli</em>, is conducted.</td>
</tr>
<tr>
<td>2003</td>
<td>The first eukaryotic genome sequence, that of the human, is released by two major independent research groups in the United States.</td>
</tr>
<tr>
<td>2005</td>
<td>Large-scale sequencing (NGS – Next Generation Sequencing) is used as a tool to unravel whole genomes quickly.</td>
</tr>
<tr>
<td>2011</td>
<td>Second- and third-generation large-scale sequencing systems are developed; eukaryotic genomes are sequenced in just a few hours.</td>
</tr>
<tr>
<td>2012</td>
<td>Technologies that control the temporal and spatial expression of genes are used in genetic transformation and the exclusion of auxiliary genes.</td>
</tr>
<tr>
<td>2014</td>
<td>Large-scale sequencing, macro- and microsynteny, associative mapping, molecular markers for genomic selection, QTL (quantitative trait loci) cloning, and large-scale phenotyping are widely used, the use of “omics” and specific, multiple GMO phenotypic traits is expanded and bioinformatics is used intensively.</td>
</tr>
</tbody>
</table>

have been established to take advantage of innovative, highly promising business opportunities.

The Post-Genomics Era

In the late 20th and early 21st centuries, genome sequencing studies developed rapidly. Gene sequences are now available for entire organisms, including humans. After these DNA base sequences are determined, it is necessary to organize them and identify the coding regions and their functions in the organism.

In this context, with a huge range of sequences being deposited in databases, geneticists are faced with a challenge as great as that which propelled the “genomics era”: correlating structure with function. This challenge has given rise to functional genomics, the science of the “era of omics.”

Omic is the neologism used to refer to the fields of biotechnology with the suffix omics: genomics, proteomics, transcriptomics, metabolomics, and physiognomics, among others. These new tools are helping to
develop superior cultivars for food production or even allowing plants to function as biofactories. The focal point for the 21st century will be the technological development of large-scale molecular studies and their integration into systems biology. These studies aim to understand the relationship between the genome of an organism and its phenotype, that is, to open up the "black box" that contains the path between codons and yield or resistance to biotic or abiotic stresses (Figure 1.1). Thus, systems biology is a science whose objectives are to discover, understand, model, and design the dynamic relationships between the biological molecules that make up living beings to unravel the mechanisms controlling these parts.

The Omics in Plant Breeding

In recent years, genetics and omics tools have revolutionized plant breeding, greatly increasing the available knowledge of the genetic factors responsible for complex traits and developing a large amount of resources (molecular markers and high-density maps) that can be used in the selection of superior genotypes. Among the existing omics tools, global transcriptome, proteome, and metabolome profiles created using EST, SAGE, microarray, and, more recently, RNA-seq libraries have been the most commonly used techniques to investigate the molecular basis of the responses of plants, tissues/organs or developmental stages to experimental conditions (Kumpatla et al., 2012). However, regardless of the omics used, the aid of bioinformatics is required for the analysis and interpretation of the data obtained.
Given the importance of these fields, subsequent chapters will discuss the various tools currently in use, or with great potential for future use, in plant breeding. The roles of these fields, the relationships between them and their corresponding biological processes (as well as their presentation in this book) can be visualized by the “trail” of omics, as shown in Figure 1.2.

The initial draft of the genome of the first plant to be sequenced (Arabidopsis thaliana) took approximately ten years to be developed. Today, with the use of the next generation of DNA sequencing technology (NGS, Next-Generation Sequencing) (e.g., Oxford Nanopore, PacBio RS, Ion Torrent, and Ion Proton, among others) and powerful bioinformatics and computational modeling programs, genomes can be sequenced, assembled, and related to the phenotypic traits specific to each genotype within a few weeks. This capability, combined with the drastic reduction in the cost of sequencing, has enabled the generation of an ever-increasing volume of data, thus enabling the comprehensive study of genomes and the development of informative molecular markers.

**Genomics, Precision Genomics, and RNA Interference**

All of this information has inspired the development of new strategies for genetic engineering. However, until recently, the available genetic engineering tools could only introduce changes into larger blocks of DNA sequences, which could subsequently be inserted only at random in the genome of a target species. Recent advances in this field have made it possible to obtain new variations from site-directed modifications, including specific mutations, insertions, and substitutions of genes and/or blocks of genes, making genetic engineering a precise and powerful alternative for the development of new cultivars.
These modifications to specific DNA sequences are initiated by generating a break on the double-stranded target DNA (Double Stranded DNA Break, DSB). Genetically modified nucleases are designed to identify the specific site of the target genome and catalyze the creation of the DSB, enabling the desired DNA modifications to occur at the specific break site or close to it.

To access specific sites, three enzymes have been genetically modified or constructed: zinc finger nucleases (ZFNs) (Figure 1.3), transcription activator-like effector nucleases (TALENs) (Figure 1.4) and meganucleases, also known as LAGLIDADG hormone endonucleases (LHEs).

Another widely used technique is post-transcriptional gene silencing (PTGS), or RNA interference (RNAi). This technique has assisted the development of transgenic plants capable of suppressing the expression of endogenous genes and foreign nucleic acids (Aragão and Figueiredo, 2008).

Knowledge about the mechanisms involved in RNA-mediated gene silencing has been important in the understanding of the biological function of genes, the interaction between organisms, and the development of new cultivars, among other applications.

The RNAi pathway begins with the presence of double-stranded RNA (dsRNA) in the cytoplasm, which may vary in origin and size (Figure 1.5). These dsRNAs are cleaved by the Dicer enzyme, a member of the RNase III nuclease family. After the processing of the dsRNA, small interfering RNAs (siRNAs) are formed, which are then integrated into an RNA-induced silencing complex (RISC). The RISC is responsible for the cleavage of a specific mRNA target sequence.
Figure 1.4 Transcription activator-like effector nucleases (TALENs). (See color figure in color plate section).

Figure 1.5 Pathways of gene silencing in plant cells. (Source: Based on Souza et al., 2007). (See color figure in color plate section).
Omics in Plant Breeding

Transcriptomics and Proteomics

Transcriptomics is the study of the transcriptome, defined as the set of transcripts (RNAs), including messenger RNAs (mRNAs) and non-coding RNAs (ncRNAs), produced by a given cell, tissue or organism (Morozova et al., 2009).

A single organism can have multiple transcriptomes. An organism’s transcriptome varies depending on several factors: different tissues or organs and developmental stages of the same individual may have different transcriptomes, and different environmental stimuli may also induce differences. Transcriptomics is currently one of the main platforms for the study of an organism’s biology. The methods of the differential expression analysis of transcripts have spread to almost every field of biological studies, from genetics and biochemistry to ecology and evolution (Kliebestein, 2012). Thus, numerous genes, alleles and alternative splices have been identified in various organisms.

In the same way, proteomics is the study of the proteome, which includes the entire set of proteins expressed by the genome of a cell, tissue or organism. However, this study can be directed only to those proteins that are expressed differentially under specific conditions (Meireles, 2007). Thus, proteomics involves the functional analysis of gene products, including the large-scale identification, localization, and compartmentalization of proteins, in addition to the study and construction of protein interaction networks (Aebersold and Mann, 2003).

Proteomics searches for a holistic view of an individual by understanding its response after a stimulus, with the end goal of predicting some biological event. This field has developed primarily through the separation of proteins by two-dimensional gel electrophoresis and chromatographic techniques (Eberlin, 2005).

Metabolomics and Physiognomics

Along with the advancement of research in the fields of genomics and proteomics, another area has gained prominence since the year 2000: metabolomics. This science seeks to identify the metabolites involved in the different biological processes related to the genotypic and phenotypic characteristics of a particular individual.

Plants metabolize more than 200,000 different molecules involved in the structure, assembly, and maintenance of tissues and organs, as well as in the physiological processes related to growth, development, and reproduction. Metabolic pathways are complex and interconnected, and they are, to some degree, dependent on and regulated by their own products or substrates, as well as by their genetic components and different levels of gene regulation.
This observation shows the great capacity for modulation or plasticity of the physiological response networks of plants under the same hierarchical control (DNA). Through the combined and simultaneous analysis of more than one regulatory level, such as the association of molecular markers and metabolic comparisons, a complex set of data can be generated, that is, the physiognomy. This science, in turn, generates systemic models aiming to understand and predict plant responses to certain stimuli and/or environmental conditions.

**Phenomics**

The field of phenomics employs a series of “high-throughput” techniques to enhance and automate the ability of scientists to accurately evaluate phenotypes, as well as to eventually reduce the determinants of phenotype to genes, transcripts, proteins, and metabolites (Tisné et al., 2013).

The phenome of an organism is dynamic and uncertain, representing a set of complex responses to endogenous and exogenous multidimensional signals that have been integrated during both the evolutionary process and the developmental history of the individual. This phenotypic information can be understood as a set of continuous data that change during the development of the species, the population, and the individual in response to different environmental conditions.

The emphasis of phenomics is phenotyping in an accurate (able to effectively measure characteristics and/or performance), precise (little variance associated with repeated measurements), and relevant manner within acceptable costs. This focus is important because phenotyping is currently the main limiting factor in genetic analysis. Unlike genotyping, which is highly automated and essentially uniform across different organisms, phenotyping is still a manual, organism-specific activity that is labor intensive and is also very sensitive to environmental variation.

The following are examples of phenomics approaches: (i) the use of digital cameras to take zenithal images for the automatic analysis of leaf area and rosette growth and the measurement of the characteristics of tissues, organs or individuals (Tisné et al., 2013); (ii) the use of infrared cameras to visualize temperature gradients, which can indicate the degree of energy dissipation (Munns et al., 2010) and have implications for responsiveness to drought stress and photosynthetic rate; (iii) the use of images generated by fluorescence detectors to identify the differential responses of populations of seedlings, fruits or seeds to a stressor (Jansen et al., 2009); (iv) the use of noninvasive methods to visualize subterranean systems (Nagel et al., 2012); and (v) the use of LIDAR (Light Detection and Ranging) technology to measure growth rate through differences between small distances measured using a laser (Hosoi and Omasa, 2009).
All these instruments generate objective digital data that can be transmitted to remote servers, many of which are connected to the Internet, for storage and further analysis, which is also often automated. The prospects for these technologies are very promising for breeding programs, which increasingly evaluate greater numbers of individuals.

**Bioinformatics**

The exponential increase in the volume of both molecular and phenotypic data requires increased computational capacity for its storage, processing, and analysis. To this end, numerous computers and analytical tools have been developed to address the massive volume of data originating from genomics, proteomics, metagenomics, and metabolomics, among other omics.

Biological data are relatively complex compared with those from other scientific fields, given their diversity and their interrelationships. All this information can only be organized, analyzed, and interpreted with the support of bioinformatics.

Bioinformatics can be defined as the field that covers all aspects of the acquisition, processing, storage, distribution, analysis, and interpretation of biological information. A number of tools that aid in understanding the biological significance of omics data have been developed through the combination of procedures and techniques from mathematics, statistics, and computer science. In addition, the creation of databases with previously processed information will accelerate research in other biological fields, such as medicine, biotechnology, and agronomy.

**Prospects**

Plant breeding is an art, science, and business that is little more than a century old. Using methods developed mainly in the 20th century, breeders have developed agronomically superior cultivars. Because of the constantly increasing challenge of agricultural food production, plant breeding must evolve and use new knowledge. Therefore, the omics will gradually assume greater relevance and be incorporated into the routines of breeding programs, making them more accurate, fast, and efficient. Although the challenges are great, the prospects are even greater.

**References**


