Roberto Fritsche-Neto · Aluízio Borém
Editors

Phenomics

How Next-Generation Phenotyping is Revolutionizing Plant Breeding

Springer

Phenomics

Roberto Fritsche-Neto · Aluízio Borém
Editors

Phenomics

How Next-Generation Phenotyping is Revolutionizing Plant Breeding

Springer

Editors

Roberto Fritsche-Neto
Departamento de Genética
University of São Paulo
Piracicaba, SP
Brazil

Aluízio Borém
Departamento de Fitotecnia
Federal University of Viçosa
Viçosa, MG
Brazil

Translation from the Portuguese language edition published by Suprema Editora, Visconde do Rio Branco.

ISBN 978-3-319-13676-9 ISBN 978-3-319-13677-6 (eBook)
DOI 10.1007/978-3-319-13677-6

Library of Congress Control Number: 2014956365

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

In recent years, plant breeding has experienced a revolution. Because of a reduction in genotyping costs and single-nucleotide polymorphisms, it is possible to obtain a large amount of genotypic data in a short time. This flood of genomic information has triggered the development of new strategies for the integration of molecular information in breeding programs. However, there is still a need for quality phenotypic data. This will not only foster efforts in mapping initiatives, but also in genomic selection and direct phenotypic selection. Tuberosa (2012) addressed this issue by saying that "phenotyping is now king, and has taken heritability as queen."

The objective now is phenomics—that is, phenotyping a large number of individuals for a great amount of traits throughout the development of the plants, in a nondestructive manner and with good accuracy. However, the development of high-throughput phenotyping platforms is still a bottleneck. Thus, several initiatives involving many species and several traits are underway to develop automation and robotics for the next generation of phenotyping in the field, greenhouses, and laboratories. Many of those technologies have shown promising results for practical applications in breeding programs.

This book aims to describe the new technologies for high-throughput phenotyping as applied to plant breeding. Written in an easy-to-understand style, this book can serve as a reference for students, educators, and researchers who are interested in innovative technologies in plant breeding. Enjoy it!

Roberto Fritsche-Neto
Aluízio Borém

Reference

Tuberosa R (2012) Phenotyping for drought tolerance of crops in the genomics era. Front Physiol 3:347

Contents

Chapter 1
New Technologies for Phenotyping

José Luis Araus, Abdelhalim Elazab, Omar Vergara, Llorenç Cabrera-Bosquet, Maria Dolors Serret, Mainassara Zaman-Allah and Jill E. Cairns

Abstract Improvements in agronomical practices and crop breeding are paramount responses to the present and future challenges imposed by water stress and heat (Lobell et al. 2011a, b; Cairns et al. 2013; Hawkins et al. 2013). On what concerns breeding, constraints in field phenotyping capability currently limit our ability to dissect the genetics of quantitative traits, especially those related to yield and water stress tolerance. Progress in sensors, aeronautics and high-performance computing is paving the way. Field high throughput platforms will combine non-invasive remote-sensing methods, together with automated environmental data collection. In addition, laboratory analyses of key plant parts may complement direct phenotyping under field conditions (Araus and Cairns 2014). Moreover, these phenotyping techniques may also help to cope with spatial variability inherent to phenotyping in the field.

Water stress is the main factor limiting agricultural productivity worldwide. Global change scenarios for the coming decades suggest an increase in water stress in many regions of the world, either directly due to a lower precipitation or as a response to increases in air temperature. As a consequence, crop yields will be affected, even for crops such as maize (Lobell et al. 2011a, b; Cairns et al. 2013; Hawkins et al. 2013). Improvements in agronomical practices and crop breeding are paramount responses to the present and future challenges imposed by water stress. Constraints in field phenotyping capability currently limit our ability to dissect the genetics of

J.L. Araus (✉) · A. Elazab · O. Vergara · M.D. Serret
Unit of Plant Physiology, Department of Plant Biology, University of Barcelona,
08028 Barcelona, Spain
e-mail: jaraus@ub.edu

L. Cabrera-Bosquet
INRA, UMR759 Laboratoire d'Ecophysiologie des Plantes sous Stress Environnementaux,
Place Viala, 34060 Montpellier, France

M. Zaman-Allah · J.E. Cairns
International Maize and Wheat Improvement Center, CIMMYT Southern Africa Regional
Office, Harare, Zimbabwe

R. Fritsche-Neto and A. Borém (eds.), *Phenomics*,
DOI 10.1007/978-3-319-13677-6_1

quantitative traits, especially those related to yield and water stress tolerance. Progress in sensors, aeronautics and high-performance computing are paving the way. Field high throughput platforms will combine non-invasive remote-sensing methods, together with automated environmental data collection. In addition laboratory analyses of key plant parts may complement direct phenotyping under field conditions (Araus and Cairns 2014). Moreover these phenotyping techniques may also help to cope spatial variability inherent to phenotyping in the field.

1.1 Field Phenotyping

Crop management has benefited strongly from the adoption of techniques to monitor crop water status and growth, as well as to predict yield through the fast development of fields, such as precision agriculture or deficit irrigation schedule. These agronomical approaches are helping to reduce the gap between the actual (farmer's) yield and the yield potential. In the case of crop breeding, genetic advances in yield and stress resistance have decreased in recent decades despite the increased adoption of molecular approaches (e.g. marker-assisted selection, transformation). Increased evidence shows that phenotyping, particularly at the field level, is actually limiting the efficiency of conventional breeding as well as preventing molecular breeding from delivering all its potential (Araus et al. 2008; Cabrera-Bosquet et al. 2012; Cairns et al. 2012; Cobb et al. 2013). Constraints in field phenotyping capability limit our ability to dissect the genetics of quantitative traits, particularly those related to stress tolerance. The development of effective field-based high-throughput phenotyping platforms (HTPPs) remains a bottleneck for future breeding advances (Araus and Cairns 2014). However, progress in sensors, aeronautics, and high-performance computing are paving the way. Some of these technologies have been successfully implemented in precision agriculture, but their use for breeding requires more accuracy and high throughput because the range of genotypic variability is usually far smaller than that caused by changing environmental conditions, and the target is to assess a large number of genotypes.

Field conditions are notoriously heterogeneous, and the inability to control environmental factors makes results difficult to interpret. However, results from controlled environments are far removed from the situation plants will experience in the field; therefore, they are difficult to extrapolate to the field (Fig. 1.1). For example, the volume of soil available to roots within a pot is considerably smaller than in the field, thereby reducing the amount of water and nutrients available to plants (Passioura 2006; Porter 2012). The soil environment plays a crucial role in plant growth and development and is difficult to simulate under controlled conditions (Whitmore and Whalley 2009). Drought stress phenotyping is particularly challenging because declining soil moisture content is associated with increased mechanical impedance in the field, which is an effect that is hard to replicate within pots (Cairns et al. 2011).

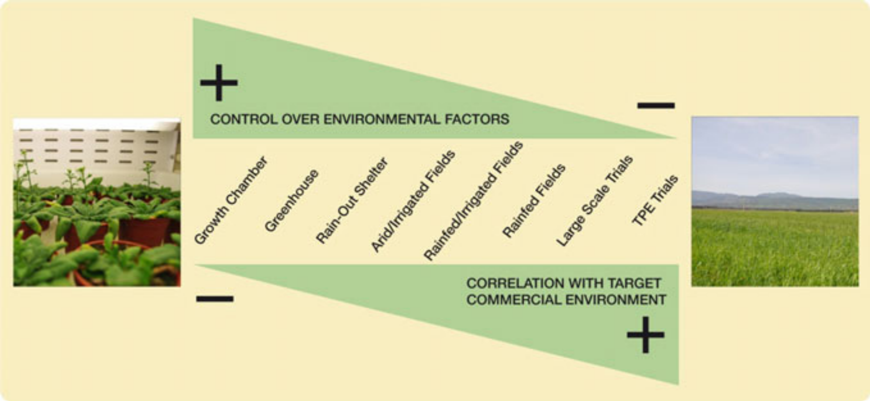


Fig. 1.1 Continuum of environments for drought resistance screening. The control over environmental factors decreases from the use of growth chambers to the target population environment (*TPE*) while the correlation of performance with the target commercial environments increases. Figure redrawn from Passioura (2006)

The most successful traits for field phenotyping integrate in time (throughout the crop cycle) and space (at the canopy level) the performance of the crop in terms of capturing resources (e.g. radiation, water, nutrients) and how efficiently these resources are used (Araus et al. 2002, 2008). Different methodological approaches have been proposed to evaluate these traits in the field (Fig. 1.2). They can be summarized into three categories: (i) proximal (remote) sensing and imaging, (ii) laboratory analyses of samples, and (iii) near-infrared reflectance spectroscopy (NIRS) analysis in the harvestable part of the crop (White et al. 2012). In practical terms, the second and third categories of traits may be considered within the same group of traits because NIRS may be eventually applicable to many of the traits usually analyzed in the laboratory.

1.2 Phenotypic Traits: Remote Sensing

Ground-based HTPPs allow data to be captured at the plot level, thus requiring little postprocessing. Moreover, this approach allows the implementation of closed multispectral imaging systems, which shut out wind and sunlight to ensure the highest possible precision and accuracy (Svensgaard et al. 2014). However, this also limits the scale at which ground-based HTTPs can be used. Furthermore, ground-based platforms do not allow simultaneous measurements of all plots within a trial (Busemeyer et al. 2013). Also, in the case of maize, its use is not very feasible, except for early stages of the crop (Montes et al. 2011).

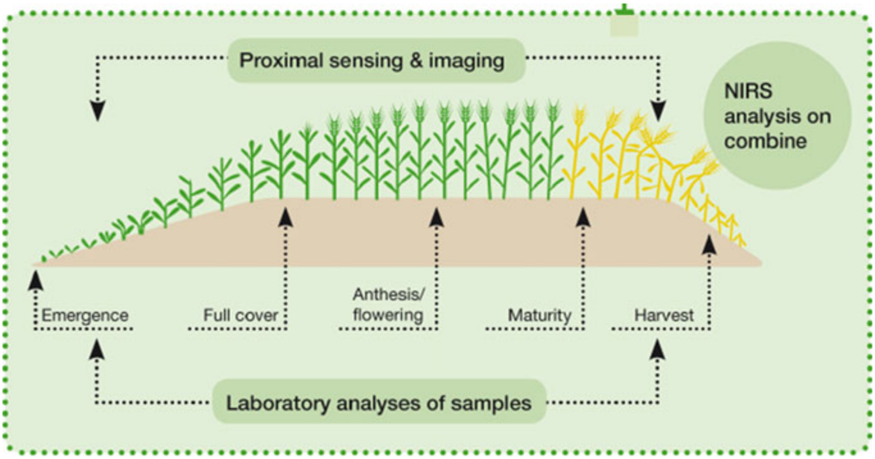


Fig. 1.2 Diagram of the main categories of phenotyping techniques deployed over the lifecycle of an annual seed crop, such as a cereal. Types of data acquisition include proximal sensing and imaging at frequent intervals, laboratory analyses of samples taken at specific intervals, and near-infrared spectroscopy (*NIRS*) on leaf matter or seeds to assess phenotypic traits potentially related with cereal performance under water stress, such as mineral content, stable carbon and oxygen isotope composition, or total nitrogen content (Cabrera-Bosquet et al. 2009a, b, 2011b). Redrawn from White et al. (2012) and Araus and Cairns (2014)

Field HTPPs should combine, at an affordable cost, a high capacity for data recording or scoring and processing and noninvasive remote-sensing methods, together with automated environmental data collection. Laboratory analyses of key plant parts may complement direct phenotyping under field conditions.

For almost any of the remote techniques, the use of imaging allows upscaling of the measurements—for example, from a single plot basis to dissecting an entire trial composed of different plots—provided that the image has enough resolution (pixels). There are different categories of sensors. RGB/CIR cameras combine color infrared (CIR) and red, green and blue light (called visible or RGB) imagery (Fig. 1.3A). It allows the estimation of green biomass, through a vegetation indices such as the normalized difference vegetation index (NDVI). Estimating the green leaf area index (GLAI, the ratio of green photosynthetic leaf area per ground area) is the proper way to assess the effect of drought (or any other stress that accelerates senescence) on potential canopy photosynthesis and thus grain yield (Lopes et al. 2011; Nguy-Robertson et al. 2012). For example, the ADC Lite (http://www.tetracam.com/adc_lite.html) and the ADC Micro (http://fieldofviewllc.com/tetracam-adc-micro) have spectral range bands in red, green, and near infrared (NIR), with the latter model having a weight of 100 g. Multispectral cameras are widely used for crop monitoring via remote sensing (Fig. 1.3B). They can acquire a limited number of spectral bands at once in the visible (VIS)–NIR regions.

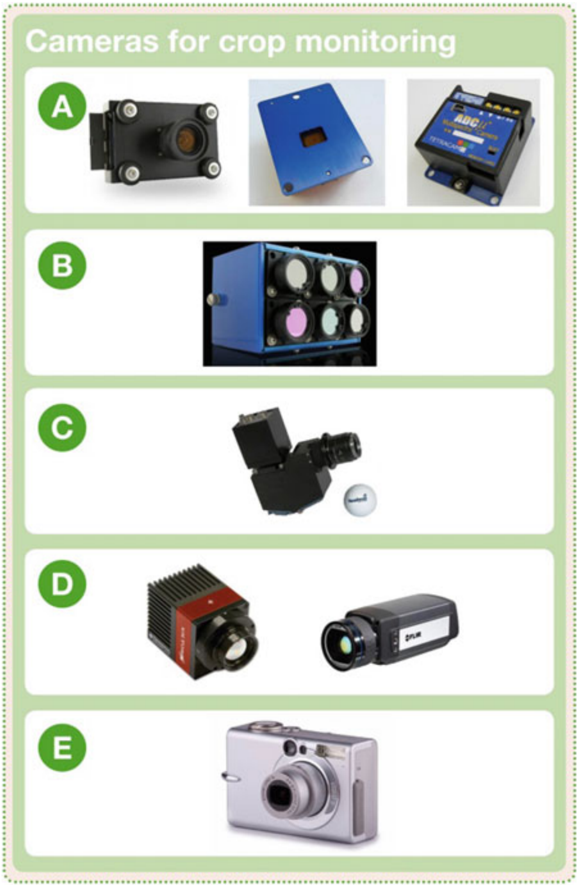


Fig. 1.3 Different categories of imaging systems for remote-sensing evaluation of vegetation. These include RGB/CIR **a** multispectral; **b** hyperspectral; **c** thermal; **d** conventional RGB; **e** cameras

Besides vegetation indices for evaluating green biomass, multispectral imagers can be formulated to other different spectral indices targeting senescence evaluation, nutrient status, pigment degradation, photosynthetic efficiency, or water content (Gutierrez et al. 2010). An example of a widely used camera is the Tetracam MCA (http://www.tetracam.com/Products-Mini_MCA.htm). Hyperspectral VIS–visible near-infrared (VNIR) imagers (Fig. 1.3C) allow the acquisition of hundreds of images at once, covering the entire electromagnetic spectrum between the VIS and the NIR regions in a continuous mode (wavelengths ranging from 400 to 900 nm). Other configurations cover the range from 1,000 to 2,500 nm. Therefore, it is possible to run empirical calibrations (like in a "NIRS-mode") against a wide and miscellaneous set of traits.

Figure 1.3C depicts the Micro-Hyperspec VNIR model (http://www.headwallphotonics.com/Portals/) which measures up to 260 bands of 5–7 nm full-width half-maximum in the 400–885 nm spectral region. This is a particularly promising approach given the possibility for multispectral information to predict complex traits, such as grain yield (Weber et al. 2012). Longwave infrared cameras or thermal imaging cameras render infrared radiation in the range of micrometers as visible light (Fig. 1.3D). The potential use of thermal imaging in phenotyping includes predicting water stress in crops. Thermal sensing has been used to assess maize response to drought (Romano et al. 2011, Winterhalter et al. 2011; Zhia et al. 2013). Low resolution may represent a limitation to the use of such cameras from aerial platforms. Examples of light thermal cameras are the FLIR Tau 640 LWIR with a 640 × 512 resolution (http://www.flir.com/cvs/cores/view/?id=51374) and the Thermoteknix Miricle camera with a 640 × 480 resolution (http://www.thermoteknix.com/products/oem-thermal-imaging/miricle-thermal-imaging-modules/). Due to their small size and weight, these cameras are not thermostabilized. Conventional digital RGB cameras (Fig. 1.3E) are very low-cost instruments that allow estimating plant cover (green biomass), senescence, and yield (Casadesús et al. 2014). At the leaf level, it allows one to assess chlorophyll and nitrogen content from digital images (Rorie et al. 2011). They can eventually replace portable chlorophyll meters, which cost several thousands of dollars. Moreover, the software needed is usually freely available (Casadesús et al. 2007).

Other remote-sensing techniques are starting to be adopted for field phenotyping, such as the use of laser imaging detection and ranging (Lidar). This is an active remote sensing technique that uses Lidar sensors to directly measure the three-dimensional distribution of plant canopies as well as subcanopy topography, thus providing high-resolution topographic maps and highly accurate estimates of vegetation height, cover, and canopy structure (Weiss and Biber 2011; Comar et al. 2012; Deery et al. 2014).

In the case of maize, its height prevents (or at least makes difficult) the use of growth-based platforms, such as phenomobiles (Deery et al. 2014), except for in the early phases of the crop. In these crops, the use of aerial HTPPs becomes a need. Considering cost and versatility, the use of unmanned aerial vehicles (UAVs) is the most promising alternative, compared with the use of cranes, tethered balloons, or manned aircrafts, to install remote-sensing approaches (Fig. 1.4). On the other hand, research on affordable technologies also should be a priority if the adoption of quality field high-throughput phenotyping is pursued for small companies and national agricultural systems from developing countries. These low-cost technologies include remote-sensing approaches, such as the use of RGB imaging and the implementation of NIRS calibrations of key analytical components.

In any case, improvements in user-friendly data management, together with a more powerful interpretation of results, should increase the use of field HTPP. Overall field high-throughput precise phenotyping needs to be placed in its right context as a one of the components that integrates advanced crop breeding, together with molecular biology, quantitative genetics, and even modelling (Cabrera-Bosquet et al. 2012; Araus and Cairns 2014; Cooper et al. 2014).

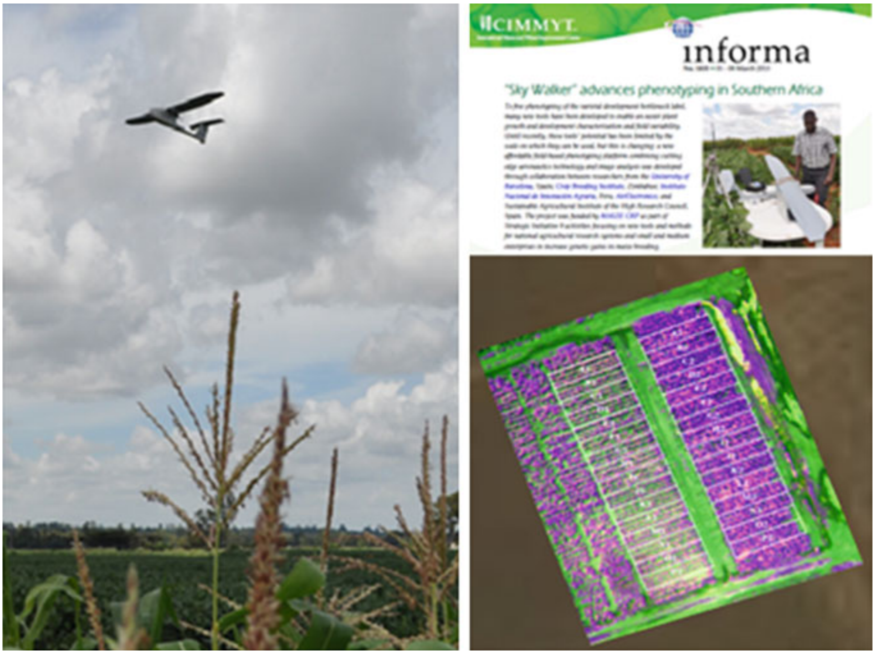


Fig. 1.4 Example of an aerial platform developed by the University of Barcelona in collaboration with Airelectronics and the Instituto de Agricultura Sostenible (Spain), sponsored by the Global Maize Program of CIMMYT (International Maize and Wheat Improvement Center)

1.3 Phenotypic Traits: Laboratory Analyses

In addition to proximal sensing approaches, the analysis of plant samples may complement direct phenotyping under field conditions. This is the case, for example, with stable isotopes (Yousfi et al. 2012). When breeding for yield potential and adaption to abiotic stresses such as drought, carbon isotope composition ($\delta^{13}C$) in dry matter, frequently expressed as a discrimination ($\Delta^{13}C$) against the source (i.e. atmospheric) CO_2, is a very promising tool that frequently exhibits high heritability and genetic correlation with yield (Condon et al. 2002, 2004; Araus et al. 2013); it has already been applied to breeding programs for C_3 cereals such as wheat (Rebetzke et al. 2008). However, its use as a phenotypic trait for crops such as maize (as well as sorghum, sugar cane, pearl millet, and others) appears to be limited because the specific characteristics of their photosynthetic C_4 metabolism makes the range of response of $\delta^{13}C$ to varying water conditions far smaller (and in the case of maize, in a opposite direction) than for crops with C_3 metabolism (Farquhar 1983; Henderson et al. 1992). Even so, $\delta^{13}C$ still allows one to differentiate between growing water conditions in maize (Cabrera-Bosquet et al. 2009a), as well as between hybrids and inbred lines (Araus et al. 2010) and highly

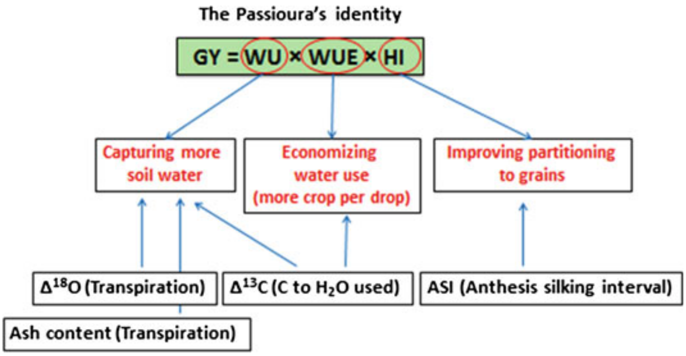
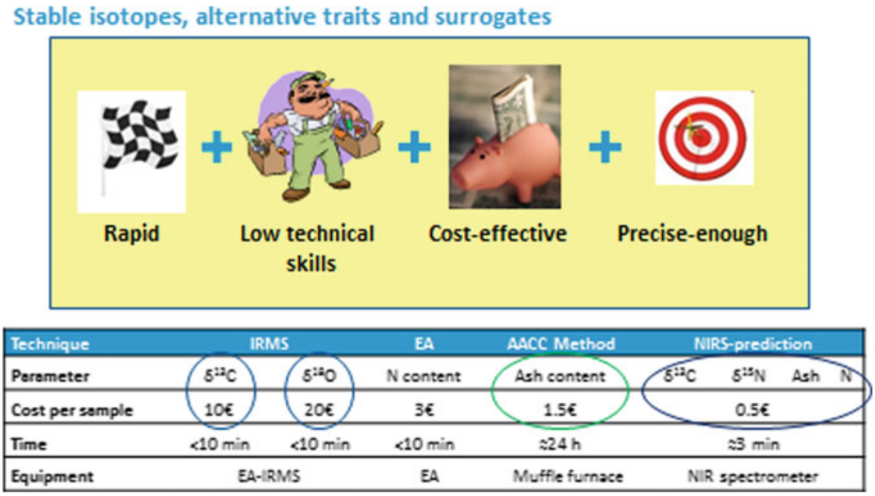

The Passioura's identity
GY = WU × WUE × HI
Capturing more soil water
Economizing water use (more crop per drop)
Improving partitioning to grains
Δ18O (Transpiration)
Δ13C (C to H2O used)
ASI (Anthesis silking interval)
Ash content (Transpiration)

Fig. 1.5 Potential analytical traits to phenotype for crop performance under water-limited environments. The physiological meaning of traits is placed in the context of the Passioura's identity (Passioura 1977). δ^{13}C, carbon isotope composition or Δ^{13}C, carbon isotope discrimination; δ^{18}O, oxygen isotope composition; Δ^{18}O, oxygen isotope enrichment with regard to water. ASI is placed here as example of successful trait related with HI, in this case for maize. For this and other crops phenological traits such as date of flowering may be also relevant

heritable significant genetic variation for Δ13C has been detected under field and greenhouse conditions (Gresset et al. 2014).

As for C_3 species, in maize, δ^{13}C (or Δ^{13}C) is an indicator of water use efficiency (Farquhar 1983; Henderson et al. 1992) but it also informs indirectly on water use (Cabrera-Bosquet et al. 2009a) (Fig. 1.5). Oxygen isotope composition (δ^{18}O) on dry matter (sometimes expressed as enrichment from the source water, Δ^{18}O) is an indicator of transpiration and therefore water used by the plant (Barbour et al. 2000; Farquhar et al. 2007; Cabrera et al. 2011a). Moreover, it is independent of the kind of photosynthetic metabolism that makes at first its use in feasible for maize (Cabrera-Bosquet et al. 2009b; Araus et al. 2010). However, to date, the use of δ^{18}O for breeding has been less promising than initially expected, probably due to a set of miscellaneous factors that affects ^{18}O isotopic signature, such as the plant's source (s) of water (irrigation, rainfall, water table may have different δ^{18}O) or the kind of tissue analyzed (^{18}O fractionation in the assimilates moving from the photosynthetic to the reproductive tissues probably exists). A relatively low-cost trait with low technical demands to assess plant transpiration and thus water used in an integrated manner is the total amount of minerals accumulated in transpiring organs, which in its simplest approach consists of analyzing the ash content (Cabrera-Bosquet et al. 2009a).

NIRS is regularly used to analyze in (intact) seeds the protein, nitrogen, starch, and oil content, as well as grain texture and grain weight, among others (Montes et al. 2007; Hacisalihoglu et al. 2010; Mir et al. 2012; White et al. 2012). In any case, the NIR spectrum captures physical and chemical characteristics of the samples, either of vegetative plant tissues or harvested seeds. By using calibration models, several traits can be determined on the basis of a single spectrum. However, the same spectrum may be used to develop prediction models for analyzing traits of potential



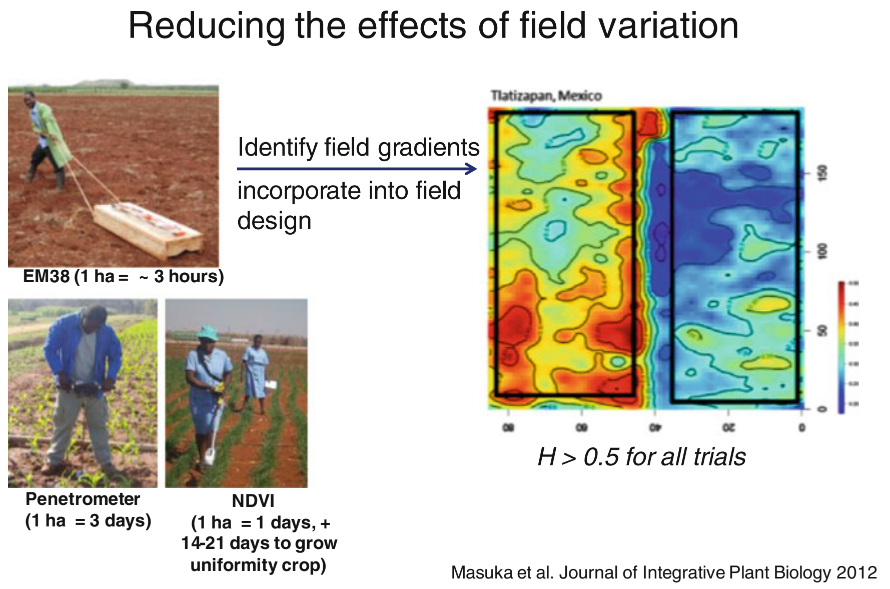
Technique	IRMS		EA	AACC Method	NIRS-prediction			
Parameter	$\delta^{13}C$	$\delta^{18}O$	N content	Ash content	$\delta^{13}C$	$\delta^{15}N$	Ash	N
Cost per sample	10€	20€	3€	1.5€	0.5€			
Time	<10 min	<10 min	<10 min	≈24 h	≈3 min			
Equipment	EA-IRMS		EA	Muffle furnace	NIR spectrometer			

Fig. 1.6 Different traits, analyzed in plant dry matter, that are potentially useful for maize phenotyping under water stress. Analytic methodologies used as well as the indicative cost per sample are included. $\delta^{18}O$, oxygen stable isotope composition; $\delta^{13}C$, stable carbon isotope composition; *EA* elemental analyzer; *IRMS* isotope ratio mass spectrometer; *NIRS* near-infrared reflectance spectroscopy

interest for phenotyping for stress adaptation, such as $\Delta^{13}C$, mineral content, or the composition of other stable isotopes (Ferrio et al. 2001; Kleinebecker et al. 2009; Cabrera-Bosquet 2011b). While the precision of these indirect estimations may be lower than those of direct analysis, the fast, low-cost, and nondestructive nature of NIRS may justify its use, at least in the early generations of a breeding program as a first screening approach when thousands of genotypes need to be evaluated (Fig. 1.6).

1.4 Phenotyping Tools Help to Cope with Spatial Variability

Field variation increases error variances, thereby masking important genetic variations for key traits and reducing repeatability, regardless of the cost and precision of a phenotyping platform (Masuka et al. 2012). Spatial variation can be caused by a number of factors, including the soil, which is inherently heterogeneous even in relatively uniform experimental sites. Earlier measurements of field variation relied on direct (i.e. destructive soil sampling) and subsequent laboratory analysis. Advances in proximal and remote-sensing technologies allow high-resolution mapping of spatial variability (Gebbers and Adamchuk 2010) (Fig. 1.7). Proximal



Reducing the effects of field variation

Identify field gradients → incorporate into field design

EM38 (1 ha = ~ 3 hours)

Penetrometer (1 ha = 3 days)

NDVI (1 ha = 1 days, + 14-21 days to grow uniformity crop)

H > 0.5 for all trials

Masuka et al. Journal of Integrative Plant Biology 2012

Fig. 1.7 Different strategies to mapping and further reducing the effects of field variation. These approaches include analyzing the naked soil using penetrometers and conductimeters; the measurement, from a single variety planted, of spatial variability in plant biomass or canopy temperature through the measurements of vegetation indices; and using spectroradiometric or multispectral imagery, or canopy temperature, through infrared thermometry or thermal imaging, respectively. Vegetation indices or canopy temperature can be measured at the field levels or from an aerial platform

sensors include the measurement of electrical conductivity, which is closely related to clay, water, and ionic content; electromagnetic surveys can be used to determine field gradients in soil texture (Cairns et al. 2012; Rebetzke et al. 2013). Alternatively, aerial HTPPs that allow fast non-destructive global position system–linked measurements of biomass using NDVI can be used to measure field variability, either on a single variety planted in the off season to develop subsequent planting maps or within experiments to build up performance maps to guide the next season's planting. Moreover, in HTTPs, the implementation of environmental characterization is essential to facilitate data interpretation, meta-data analysis, and, in the case of drought phenotyping, understanding patterns of water availability (Masuka et al. 2012). The need for environmental data is particularly pertinent in drought screening, where knowledge of soil moisture availability is necessary to ensure that the field environment and the type of drought imposed are representative of the target environment (Rebetzke et al. 2013).

1.5 Root phenotyping

Although there are considerable advances for evaluating the aerial parts of plants, roots are notoriously difficult to phenotype in the field (Reynolds et al. 2012: Leitner et al. 2014; see also DoVale and Fritsche-Neto 2014, Chap. 4 in this book). Traditional studies have focused on excavation techniques, from which root depth and root length density can be determined. Trenching is labor intensive and slow, which means that it is not really feasible for a large-scale evaluation in crops such as maize. An approach that was recently proposed for maize and less intensive in terms of resources deployed is "shovelomics" (Trachsel et al. 2011). Values for root architectural traits are derived from the visual scoring of roots, which for maize includes numbers, angles, and branching patterns of crown and brace roots. However, for the moment, this technique has delivered less than initially expected.

1.6 Concluding remarks

To conclude, field phenotyping must go hand-in-hand with methods to characterize and control field site variations (for improving repeatability), adopting appropriate experimental designs, selection of the right traits, and finally, proper integration of heterogeneous data sets, analysis, and applications, including prediction models (Araus et al. 2012; Araus and Cairns 2014; Prasanna et al. 2013; White et al. 2012). In the near future, what will pave the way for adoption of field HTPP is the efficient integration of all the components of the system. This includes a more user-friendly data management combined with data gathering and processing.

Acknowledgments The preparation of this chapter was supported by a grant (Affordable Field High Throughput Phenotyping Platform) from the MAIZE CGIAR Research Program and the Spanish project AGL2013-44147-R.

References

Araus JL, Slafer GA, Reynolds MP, Royo C (2002) Plant breeding and water stress in C_3 cereals: what to breed for? Ann Bot 89:925–940

Araus JL, Slafer GA, Royo C, Serret MD (2008) Breeding for yield potential and stress adaptation in cereals. Crit Rev Plant Sci 27:1–36

Araus JL, Cabrera-Bosquet L, Sánchez C (2010) Is heterosis in maize mediated through better water use? New Phytol 187:392–406

Araus JL, Serret MD, Edmeades GO (2012) Phenotyping maize for adaptation to drought. Front Physiol 3:305

Araus JL, Cabrera-Bosquet L, Serret MD, Bort J, Nieto-Taladriz MT (2013) Comparative performance of δ^{13}C, δ^{18}O and δ^{15}N for phenotyping durum wheat adaptation to a dryland environment. Funct Plant Biol 40:595–608

Araus JL, Cairns J (2014) Field high-throughput phenotyping—the new crop breeding frontier. Trends Plant Sci 19:52–61

Barbour MM, Fischer RH, Sayre KD, Farquhar GD (2000) Oxygen isotope ratio of leaf and grain material correlates with stomatal conductance and grain yield in irrigated wheat. Aust J Plant Physiol 27:625–637

Busemeyer L, Mentrup D, Möller K, Wunder E, Alheit K, Hahn V, Maurer HP, Reif JC, Würschum T, Müller J, Rahe F, Ruckelshausen A (2013) BreedVision—a multi-sensor platform for non-destructive field-based phenotyping in plant breeding. Sensors 13:2830–2847

Cabrera-Bosquet L, Sánchez C, Araus JL (2009a) How yield relates to ash content, Δ^{13}C and Δ^{18}O in maize grown under different water regimes. Ann Bot 104:1207–1216

Cabrera-Bosquet L, Sanchez C, Araus JL (2009b) Oxygen isotope enrichment (Δ^{18}O) reflects yield potential and drought resistance in maize. Plant Cell Environ 32:1487–1499

Cabrera-Bosquet L, Albrizio R, Nogués S, Araus JL (2011a) Dual Δ^{13}C/δ^{18}O response to water and nitrogen availability and its relationship with yield in field-grown durum wheat. Plant Cell Environ 34:418–433

Cabrera-Bosquet L, Sánchez C, Rosales A, Palacios-Rojas N, Araus JL (2011b) NIRS-assessment of δ^{18}O, nitrogen and ash content for improved yield potential and drought adaptation in maize. J Agric Food Chem 59:467–474

Cabrera-Bosquet L, Crossa J, von Zitzewitz J, Serret MD, Araus JL (2012) High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. J Integr Plant Biol 54:312–320

Cairns JE, Impa SM, O'Toole JC, Jagadish SVK, Price AH (2011) Influence of the soil physical environment on drought stress and its implications for drought research. Field Crop Res 121:303–310

Cairns JE, Sanchez C, Vargas M, Ordoñez RA, Araus JL (2012) Dissecting maize productivity: ideotypes associated with grain yield under drought stress and well-watered conditions. J Integr Plant Biol 54:1007–1020

Cairns J, Hellin J, Sonder K, Araus JL, MacRobert JF, Thierfelder C, Prasanna BP (2013) Adapting maize to climate change in sub-saharan Africa. Food Secur 5:345–360

Casadesús J, Kaya Y, Bort J, Nachit MM, Araus JL, Amor S, Ferrazzano G, Maalouf F, Maccaferri M, Martos V, Ouabbou H, Villegas D (2007) Using vegetation indices derived from conventional digital cameras as selection criteria for wheat breeding in water-limited environments. Ann Appl Biol 150:227–236

Casadesús J, Villegas D (2014) Conventional digital cameras as a tool for assessing leaf area index and biomass for cereal breeding. J Integr Plant Biol 56:7–14

Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. Theor Appl Genet 126:867–887

Comar A, Burger P, Benoit de Solan C, Baret F, Daumard F, Hanocq J-F (2012) A semi-automatic system for high throughput phenotyping wheat cultivars in-field conditions: description and first results. Funct Plant Biol 39:914–924

Condon AG, Richards RA, Rebetkke GJ, Farquhar GD (2002) Improving intrinsic water-use efficiency and crop yield. Crop Sci 42:122–131

Condon AG, Richards RA, Rebetkke GJ, Farquhar GD (2004) Breeding for high water-use efficiency. J Exp Bot 55:2447–2460

Cooper M, Messina CD, Podlich D, Totir LR, Baumgarten A, Hausmann NJ, Wright D, Graham G (2014) Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. Crop Pasture Sci 65:311–336

Deery D, Jimenez-Berni J, Jones H, Sirault X, Furbank R (2014) Proximal remote sensing buggies and potential applications for field-based phenotyping. Agronomy 5:349–379

Farquhar GD (1983) On the nature of carbon isotope discrimination in C_4 species. Aust J Plant Physiol 10:205–226

Farquhar GD, Cernusak LA, Barnes B (2007) Heavy water fractionation during transpiration. Plant Physiol 143:11–18

Ferrio JP, Bertran E, Nachit M, Royo C, Araus JL (2001) Near infrared reflectance spectroscopy as a new surrogate analysis for Δ^{13}C in mature kernels of durum wheat. Aust J Agric Res 52:809–816

Gebbers R, Adamchuk VI (2010) Precision agriculture and food security. Science 327:828–831

Gresset S1, Westermeier P, Rademacher S, Ouzunova M, Presterl T, Westhoff P, Schön CC (2014) Stable carbon isotope discrimination is under genetic control in the C4 species maize with several genomic regions influencing trait expression. Plant Physiol 164(1):131–143. doi: 10.1104/pp.113.224816. Epub 2013 Nov 26

Gutierrez M, Reynolds MP, Klatt AR (2010) Association of water spectral indices with plant and soil water relations in contrasting wheat genotypes. J Exp Bot 12:3291–3303

Hacisalihoglu G, Larbi B, Settles AM (2010) Near-infrared reflectance spectroscopy predicts protein, starch, and seed weight in intact seeds of common bean (*Phaseolus vulgaris* L.). J Agric Food Chem 58:702–706

Hawkins E, Fricker TE, Challinor AJ, Ferro CAT, Ho CK, Osborne TM (2013) Increasing influence of heat stress on French maize yields from the 1960s to the 2030s. Glob Change Biol 19:937–947

Henderson S, Caemmerer S, Farquhar G (1992) Short-term measurements of carbon isotope discrimination in several C_4 species. Funct Plant Biol 19:263–285

Kleinebecker T, Schmidt SR, Fritz C, Smolders AJP, Hölze N (2009) Prediction of δ^{13}C and δ^{15}N in plant tissues with near-infrared reflectance spectroscopy. New Phytol 184:732–739

Leitner D, Meunier F, Bodner G, Javaux M, Schnepf A (2014) Impact of contrasted maize root traits at flowering on water stress tolerance—a simulation study. Field Crops Res 165:125–137

Lobell DB, Schlenker W, Costa-Roberts J (2011a) Climate trends and global crop production since 1980. Science 333:616–620

Lobell DB, Bänziger M, Magorokosho C, Vivek B (2011b) Nonlinear heat effects on African maize as evidenced by historical yield trials. Nat Glob Change 1:42–45

Lopes MS, Araus JL, van Heerden PDR, Foyer CH (2011) Enhancing drought tolerance in C_4 crops. J Exp Bot 62:3135–3153

Masuka B, Araus JL, Sonder K, Das B, Cairns JE (2012) Deciphering the code: successful abiotic stress phenotyping for molecular breeding. J Integr Plant Biol 54:238–249

Mir RR, Zaman-Allah M, Sreenivasulu N, Trethowan R, Varshney RK (2012) Integrated genomics, physiology and breeding approaches for improving drought tolerance in crops. Theor Appl Genet 125:625–645

Montes JM, Melchinger AE, Reif JC (2007) Novel throughput phenotyping platforms in plant genetic studies. Trends Plant Sci 12:433–436

Montes JM, Technow F, Dhillon B, Mauch F, Melchinger A (2011) High-throughput non-destructive biomass determination during early plant development in maize under field conditions. Field Crops Res 121:268–273

Nguy-Robertson A, Gitelson AA, Peng Y, Viña A, Arkebauer T, Rundquist D (2012) Green leaf area index estimation in maize and soybean: combining vegetation indices to achieve maximal sensitivity. Agron J 104:1336–1347

Passioura JB (1977) Grain yield, harvest index, and water use of wheat. J Aust I Agr Sci 43:117–120

Passioura JB (2006) The perils of pot experiments. Funct Plant Biol 33:1075–1079

Poorter H, Bühler J, van Dusschoten D, Climent J, Postma JA (2012) Pot size matters: a meta-analysis of the effects of rooting volume on plant growth. Funct Plant Biol 39:839–850

Prasanna BP, Araus JL, Crossa J, Cairns JE, Palacios N, Das B, Magotokosho C (2013) High-throughput and precision phenotyping for cereal breeding programs. In: Gupta PK, Varshney RK (eds) Cereal genomics II. Springer, Dordrecht, pp 341–374 (Chapter 13)

Rebetzke GJ, Condon AG, Farquhar GD, Appels R, Richards RA (2008) Quantitative trait loci for carbon isotope discrimination are repeatable across environments and wheat mapping populations. Theor Appl Genet 118:123–137

Rebetzke GJ, Chenu K, Biddulph B, Moeller C, Deery DM, Rattey AR, Bennett D, Barrett-Lennard EG, Mayer JE (2013) A multisite managed environment facility for targeted trait and germplasm phenotyping. Funct Plant Biol 40:1–13

Reynolds MP, Pask AJD, Mullan DM (eds) (2012) Physiological breeding I: interdisciplinary approaches to improve crop adaptation. CIMMYT, Mexico

Romano G, Zia S, Spreer W, Cairns J, Araus JL, Müller J (2011) Use of thermography for high throughput phenotyping of tropical maize adaptation in water stress. Comput Electron Agric 79:67–74

Rorie RL, Purcell LC, Karcher DE, King CA (2011) The assessment of leaf nitrogen in corn from digital images. Crop Sci 51:2174–2180

Svensgaard J, Roitsch Y, Christensen S (2014) Development of a mobile multispectral imaging platform for precise field phenotyping. Agronomy 4:322–336

Trachsel S, Kaeppler SM, Brown KM, Lynch JP (2011) Shovelomics: high throughput phenotyping of maize (*Zea mays* L.) root architecture in the field. Plant Soil 341:75–87

Weber VS, Araus JL, Cairns JE, Sanchez C, Melchinger AE, Orsini E (2012) Prediction of grain yield using reflectance spectra of canopy and leaves in maize plants grown under different water regimes. Field Crop Res 128:82–90

Weiss U, Biber P (2011) Plant detection and mapping for agricultural robots using a 3D LIDAR sensor. Robot Auton Syst 59:266–273

White JW, Andrade-Sanchez P, Gore MA, Bronson KF, Coffelt TA, Conley MM, Feldmann KA, French AN, Heun JT, Hunsaker DJ, Jenks MA, Kimball BA, Roth RL, Strand RJ, Thorp KR, Wall GW, Wang G (2012) Field-based phenomics for plant genetics research. Field Crops Res 133:101–112

Whitmore AP, Whalley WR (2009) Physical effects of soil drying on roots and crop growth. J Exp Bot 60:2845–2857

Winterhalter L, Mistele B, Jampatong S, Schmidhalter U (2011) High throughput phenotyping of canopy water mass and canopy temperature in well-watered and drought stressed tropical maize hybrids in the vegetative stage. Eur J Agron 35:22–32

Yousfi S, Serret MD, Márquez AJ, Voltas J, Araus JL (2012) Combined use of δ^{13}C, δ^{18}O and δ^{15}N tracks nitrogen metabolism and genotypic adaptation of durum wheat to salinity and water deficit. New Phytol 194:230–244

Zia S, Romano G, Spreer W, Sanchez C, Cairns J, Araus JL, Müller J (2013) Infrared thermal imaging as a rapid tool for identifying water stress tolerant maize genotypes of different phenology. J Agron Crop Sci 199:75–84

Chapter 2
Experimental Designs for Next Generation Phenotyping

Luiz Alexandre Peternelli and Marcos Deon Vilela de Resende

Abstract The increase in popularity of high-throughput genotyping in breeding programs is associated with recent advances in DNA sequencing technology and large decreases in genotyping costs. However, the limits of using genotyping for making predictions and, therefore, identifying potential candidate materials for selection thus reside in the quality of the phenotyping. High-throughput phenotyping technologies have been developed and implemented prior to planting and during cultivation. Much of this phenotyping has occurred in relatively small and restricted environments where many influential factors in the quality of phenotype can be adequately controlled. In many situations, however, it is necessary to perform phenotyping under field conditions. In this case, depending on the characteristic of interest to be collected, the influence of factors difficult to be controlled in such adverse conditions can cause the need for use of alternatives that can ensure a sufficiently accurate and precise phenotyping. In this sense, the science of Statistics contributes with an important role, either in the use of traditional basic concepts, in the planning of controlled experiments, or in modeling and developing appropriate analyzes. This chapter will discuss several experimental designs that can potentially be used for phenotyping under variable conditions, describing their various characteristics. Also it will address on topics related to the problem of obtaining accurate and precise phenotypic information, and the role of statistics in the success of this venture so fashionable today.

L.A. Peternelli (✉)
Universidade Federal de Viçosa, Viçosa, Brazil
e-mail: peternelli@ufv.br

M.D.V. de Resende
Embrapa Florestas, Universidade Federal de Viçosa, Viçosa, Brazil
e-mail: marcos.deon@ufv.br

R. Fritsche-Neto and A. Borém (eds.), *Phenomics*,
DOI 10.1007/978-3-319-13677-6_2

2.1 Introduction

Genotyping is becoming increasingly routine and more widely accepted in breeding programs. This increase in popularity is associated with recent advances in DNA sequencing technology and large decreases in genotyping costs. As high-throughput genotyping can be performed with satisfactory quality, the limits of using genotyping for making predictions and, therefore, identifying potential candidate materials for selection thus reside in the quality of the phenotyping (Lado et al. 2013). Genotyping is now highly mechanized and uniform across organisms, but phenotyping methods still vary by species, are laborious, and are sensitive to environmental variation (Cobb et al. 2013). The ideal situation would be a phenotypic characterization that does not have any errors and therefore reproduces the true population or individual phenotypic value, at least for the conditions under which it is measured. However, to obtain an accurate predictive model, the genetic differences among the materials and the experimental conditions that affect the precision of the phenotypic value should be taken into consideration.

Regarding field experiments in which the breeder will select the materials, detailed knowledge of the field conditions and the material being selected is essential for a successful breeding program. To obtain this information, high-throughput phenotyping technologies have been developed and implemented prior to planting and during cultivation. When possible, characterization of the experiment before planting provides better information on the heterogeneity of the field and therefore allows one to define experimental strategies for subsequent, more accurate phenotyping studies. In addition, the measurements made during crop growth seek to reduce the variance caused by any nongenetic factor to which the material may still be subjected (Cabrera-Bosquet et al. 2012; Masuka et al. 2012; Crossa et al. 2006).

If the researcher is unable to use advanced phenotyping technologies or can only use a limited aspect of these technologies, traditional methods can be applied to studies, including effective experimental designs that can capture a large portion of the field variance, as well as correction methods employed during modeling and data analysis. In this context, various strategies can be employed, including strategies for spatial analysis that involve modeling the covariance matrix of the errors and the polynomial functions of rows and columns for fitting spatial trends.

Genetic analysis of field materials has two aims: (i) to infer the genotypic values of the materials and (ii) to rank the genetic materials by their genotypic values. Clearly, there is no interest in estimating the phenotypic means of the genetic materials in the experiments aimed at estimating the genetic means, also known as the genotypic values. In other words, the researcher is interested in future means, when the materials are planted again on commercial farms after the selection process. When planted commercially, even when planted at the same site or in the same region as the experiment, the effects of blocks and plots and the random environmental effects will not be repeated. As these effects are included to an extent in the phenotypic means, they are not sufficient to draw conclusions concerning the genotypic values of the genetic materials. Thus, utilizing phenotypic means for predicting results of subsequent studies is not desirable or recommended. On the contrary, the breeder is interested in the genotypic values free of environmental effects. These should be the values used for analyses of future outcomes (e.g., subsequent analyses based on molecular marker linkage, genomic selection, quantitative trait loci identification, and differential gene expression analysis by RNA-seq) based on genotyping data, thus allowing for better model predictions and ensuring conclusive results.

However, phenotyping via field experiments is generally associated with unbalanced data for several reasons, including plant and plot losses, unequal numbers of seeds and seedlings available for treatments, experimental networks with different numbers of replicates and different experimental designs, and non-evaluation of all combinations of genotypes and environments. In addition, when the automated collection of phenotypic data is impractical and a group of researchers analyzes the materials in the field, researcher bias can decrease accuracy. Thus, statistical models should include all of the sources of variation and noise to better "correct" the measured phenotypic values. Therefore, the optimal procedure for genetic analysis is restricted or residual maximum likelihood/best linear unbiased prediction, also generically called *mixed linear models*. Mixed-model data analysis allows for various sources of variation to be included in the model, without impeding analysis. In addition, these models seamlessly handle unbalanced data, leading to more precise estimations and predictions of genetic parameters and genetic values, respectively.

Currently, the development of effective phenotyping methods requires multidisciplinary collaboration involving biologists, agronomists, computer scientists, engineers, and statisticians (Cobb et al. 2013). The level of expertise required is related to the use and development of equipment for automated and efficient data collection, the definitions of the variable of interest to be collected, appropriate field conditions for plant growth and analysis, volume of data to be collected, stored and analyzed and, finally, the planning of experiments to better control for systematic variations. For data analysis, the wide availability of computer resources (software and computational power) has facilitated the work of statisticians during experimental planning. In the past, it was common to have restrictions for implementing various experimental and field data collection designs because the theoretical knowledge and available computational power were limiting factors. Now, it is possible to obtain more accurate means (or effects) for complex experimental designs in the context of mixed linear models, therefore ensuring greater effectiveness of subsequent analyses that require sufficiently accurate phenotypic values.

In addition to mixed models, Bayesian analyses have facilitated data analysis and have increasingly ensured that one can obtain adjusted data with the desired quality. Bayesian analysis provides more precise estimates of variance components, genetic parameters, genetic values, and genetic gains, in addition to allowing for accurate analyses of samples with finite sizes. The informational richness provided by this approach allows for the determination of point estimates and probability intervals for the posterior distributions of the parameters. The great advantage is that it is a modeling approach whereby, via the prior distributions of the effects and model parameters, a researcher can incorporate future knowledge regarding the problem in question. More details on this data analysis approach can be found in the literature (Silva et al. 2013; Resende 2002).

Although there is consensus in the numerous publications that address the importance of increasing the accuracy of phenotyping in field studies, little has been noted regarding the experimental designs that would be the most appropriate for experiments in which large-scale phenotyping is desired. This chapter will discuss several experimental designs that can potentially be used for this purpose, describing their various characteristics and results to the extent that the reader can implement them satisfactorily.

2.2 Basic Principles of the Experiments

Experiments differ among studies. However, all experiments are guided by several basic principles established at the beginning of the twentieth century by Fisher in several of his publications (Fisher 1926, 1935). The use of these principles (replication, randomization, and local control) is necessary to obtain valid conclusions.

The principle of replication consists of applying the same treatment to several plots within the same experiment for estimating the experimental error, or residual variance.

The principle of randomization provides all of the experimental units the same chance of receiving any of the treatments, thus preventing one of the treatments from being systematically favored or disfavored by external factors. A great benefit of randomization is to provide reliability for the estimates of the experimental error of the means for the treatments. By allowing the experimental error to be validly estimated, this principle ensures the use of significance tests (e.g., comparisons of treatment means) by making the experimental errors independent.

Finally, local control is a commonly applied principle, but it is not obligatory because experiments can be conducted without it. The goal of local control (or blocking) is to divide a heterogeneous environment into homogenous sub-environments. Treatments are distributed within the sub-environment, making the experimental design more efficient by reducing experimental error.

2.3 Experimental Design

There are no explicit citations for the experimental designs most commonly applied for large-scale phenotyping. Several studies (Araus and Cairns 2014; Fiorani and Schurr 2013; Cobb et al. 2013; Poorter et al. 2012) have noted the importance of organizing experiments according to an experimental design that allows for increasing the accuracy of the phenotypic information, but few studies name these designs (Lado et al. 2013; Auer and Doerge 2010).

Because the main interest of the researcher when evaluating various phenotypic characteristics is to better characterize the material under analysis, by destructive means or not, it is expected that collecting the most accurate data is of utmost importance. In this context, the term *accuracy* should be well understood. Figure 2.1 illustrates the concepts of accuracy and precision.

Accuracy is defined as the correlation between the true genotypic value and the value estimated from the genotypic and phenotypic data from the experiments. An accurate estimator has a small difference between the true and estimated values, that is, it has a small mean squared error (MSE). An optimal estimation/prediction method should minimize the MSE, given by MSE = bias^2 + precision = bias^2 + PEV, where PEV is the prediction error variance. Thus, a minimum MSE estimator has little or no bias and high precision (low PEV). With no bias, MSE = PEV.

The concepts of bias, precision, and accuracy are illustrated in Fig. 2.1. High accuracy (the capacity to hit the target) is a combination of high precision (low variance in the various attempts; i.e., low PEV) and low bias (mean of the various attempts equal to the prediction target). Thus, accuracy is the ability to identify the truth, and precision is the ability to always obtain the same answer but not necessarily the truth.

Designs recognized as having potential to improve the effectiveness (less prediction variance) of phenotyping in field experiments include the randomized complete block design (RCBD), the augmented block design (ABD), and the incomplete block design (IBD), with their possible variations.

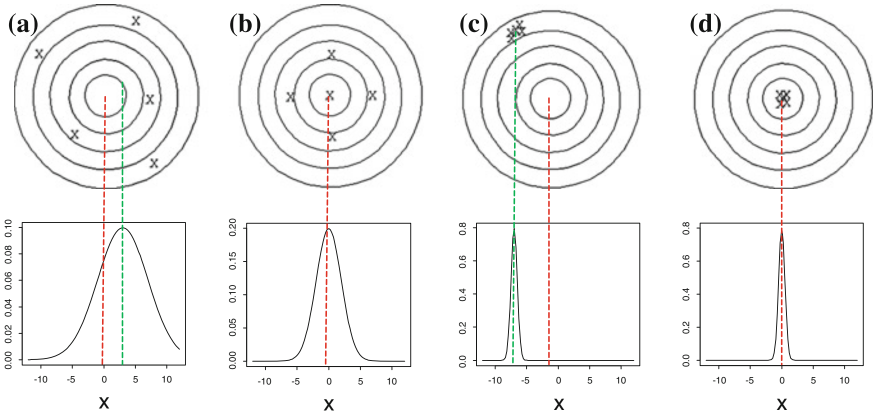


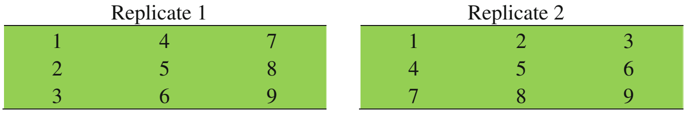
Fig. 2.1 Illustration of the concepts of accuracy, precision, and bias. **a** High bias, low precision ⇒ low accuracy; **b** low bias, low precision ⇒ low accuracy; **c** high bias, high precision ⇒ low accuracy; **d** low bias, high precision ⇒ high accuracy. The *vertical red line* shows the true value (target value). The *vertical green line* shows the prediction bias. The shape of the *curve* shows the precision: a curve more concentrated at the mean implies higher precision (low PEV), while a curve less concentrated at the mean implies less precision (high PEV). PEV, prediction error variance

ABD is most commonly used in the initial steps of a breeding program when there is still a substantial amount of material to be analyzed and, mainly, when there is little propagation material (and thus, the replication of treatments is difficult or impossible). One advantage of ABD is the ease of establishing the experiments, which is particularly useful for sugarcane breeding programs, for example (Souza et al. 2006; Peternelli et al. 2009). RCBD, in turn, is more commonly used in the later stages of breeding programs when, in addition to possessing sufficient propagation material to perform several replicates, more reliable conclusions concerning the analyzed treatments are desired. In contrast, RCBD is unviable when the number of treatments is large. Under this scenario, the block will be very large and will likely encompass heterogeneous conditions, thus limiting the efficiency. IBD, in turn, is employed when the block size is smaller than the number of treatments. If the researcher purposefully creates the blocks according to the number of treatments, the blocks may become very large, which will result in environmental heterogeneity within the block, thus leading to high prediction error. For this reason, IBD is preferred when homogeneity within the block is needed. This homogeneity can be guaranteed, for example, when the block to be homogeneous could only contain 20 plots, but the researcher must phenotype more than 20 different materials (treatments).

Several other aspects regarding these designs will be discussed. Theoretical and practical details of the analysis of these designs can be found in Resende (2007), Faraway (2005), Ramalho et al. (2005), Hinkelmann and Kempthorne (1994, 2005), Barbin (2003), Storck et al. (2000), Steel et al. (1997), Scott and Milliken (1993), Cochran and Cox (1992), Banzato and Kronka (1989), and Cox (1958).

2.3.1 Randomized Complete Block Design

RCBD is the most widely used of all of the experimental designs. It is suitable when there is complete homogeneity in the experimental conditions. In this case, the experimental area or material is divided into blocks (or groups), maintaining homogeneity within each block, and each block contains at least one replicate of each treatment distributed randomly within each block (Fig. 2.2).



Replicate 1

1	4	7
2	5	8
3	6	9

Replicate 2

1	2	3
4	5	6
7	8	9

Fig. 2.2 Layout of an experiment employing a RCBD with nine treatments. There are two replicates in this arrangement (often called blocks). Treatments are numbered *1–9*. In an RCBD, if one wants to add control treatments, the controls are allocated to new plots within each replicate. Within each replicate (or block), the treatments are allocated randomly

In experiments with this design, the blocks should be defined in a layout that confers homogeneity to each block. Theoretically, it does not matter if the experimental conditions in one block differ from the experimental condition of another block because these differences do not cause treatment × block interactions. This lack of interaction means that comparisons between pairs of treatments, for example, are not affected by the block in which these treatments are established.

It is important to emphasize that the use of an RCBD when it is not necessary results in a loss of efficiency and a decrease in the precision in the experiment. However, in general, it is necessary to divide the experimental area into homogeneous blocks that contain the treatments. Thus, this type of design is widely used for field conditions.

2.3.2 Augmented Block Design

ABD has been employed in various phases of breeding programs (e.g., for sugarcane). Initially proposed by Federer (1956), an ABD allows for genotypes to be analyzed without using replicates; only the controls are replicated (Fig. 2.3).

The experimental error can be estimated from the controls. This design is a type of IBD and is commonly called Federer blocks, in honor of its creator. This design is unbalanced and nonorthogonal. Thus, it should be analyzed using a mixed-model method.

The establishment of an ABD is very simple. It starts similarly to an RCBD with controls. However, the treatments, or new materials, are distributed among these blocks but not replicated between blocks. The statistical analysis of this design entails a fit of the "effects" attributed to each treatment, corresponding to a penalization of the treatments allocated to the best blocks and a bonus for the treatments located in the worst blocks.

In certain instances, two replicates are included when enough material is available to obtain a better estimate for the effects of each treatment, thus doubling the material requirements and operational costs of the program and reducing the area available for other goals or reducing the number of clones available in the area. However, by keeping the size of the experimental area constant, numerous studies have demonstrated that this practice of doubling the ABD does not necessarily

Block 1			Block 2		
1	4	7	10	11	12
2	5	8	13	14	15
3	6	9	16	17	18
A	B		A	B	

Fig. 2.3 Layout of an ABD, with 18 treatments and two controls. There are two blocks in the arrangement. *A* and *B* are the controls. Treatments are numbered *1–18*. All of the treatments and controls are randomly distributed across the blocks

result in gains in estimates of treatment effects (Peternelli et al. 2009). The greater difficulty is in defining the material that should be used as a control, thus providing the estimate of the experimental error. It is possible that this estimate is influenced by the choice of controls. Therefore, the researcher should use this design with care.

2.3.3 *Incomplete Block Design*

As mentioned previously, the heterogeneity within very large blocks will lead to a larger experimental error, which makes the phenotypic estimates of interest less precise by reducing the precision of the experiment. In the IBD design, the blocks are smaller, leading to less environmental heterogeneity within the blocks. The theory behind the planning and use of this design is extremely complex. Below, we briefly describe several concepts and peculiarities underlying IBD. However, there are several other important concepts and details of the analysis that must be addressed and are important to note. Valuable references on this topic are cited at the end of this chapter.

IBD designs can be classified into two categories: *resolvable designs*, in which the blocks can be grouped into replicates, and *nonresolvable designs*, in which the blocks cannot be grouped into replicates. Resolvable designs are preferred because analyses can be performed, when necessary and possible, using the completely randomized block design.

For the explanations below, the following definitions apply: v = number of treatments, k = size of the blocks or number of plots within each block, b = number of blocks, and r = number of replicates in the experiment.

Suppose that there are r replicates and v treatments. Additionally, suppose that within each replicate there are b blocks, each of size k. Figure 2.4 provides an example of this design.

In this layout, the blocks can be grouped into treatment replicates (*resolvable design*). Some authors (Williams and Matheson 1994) call this a *generalized lattice design*.

A balanced lattice square design is the most efficient IBD design if the aim is to compare two treatments. In this design, $v = k^2$, which may restrict its use in practice. To be balanced, $r = k + 1$. The high efficiency of the balanced IBD design is attributed to all of the treatment pairs occurring in at least one block of the experiment. However, for all of the treatment pairs to occur together at least once, a large number of

	Replicate 1			**Replicate 2**		
Block	1	2	3	1	2	3
	1	4	7	1	2	3
	2	5	8	4	5	6
	3	6	9	7	8	9

Fig. 2.4 Layout of an experiment using an IBD with nine treatments. There are three blocks (b = 3), two replicates (r = 2), nine treatments (v = 9) and the blocks have a size k = 3

replicates is needed ($r = k + 1$), which could prevent the implementation of a balanced IBD in practice. In this case, a partially balanced lattice square can be established. This design is obtained by considering a number of the $k + 1$ replicates from a balanced design. Therefore, another more practical design that can be implemented in the field is the alpha-lattice design (Patterson and Williams 1976), in which the v treatments are arranged in b blocks of size k, such that $v = ks$, for $s > 1$.

An advantage of alpha-lattice over the lattice square is the ability to use a large number of v values. That is, the relationship $v = ks$ in the alpha-lattice is less restrictive.

2.4 Modeling and Appropriate Analyses

There are various types of analyses for incomplete block experiments: (a) intrablock analysis, in which comparisons are only made between plots in the same block to estimate the treatment effects; and (b) analysis with recovery of interblock information, in which comparisons between blocks are also used to estimate treatment effects. Because it provides more precise results, the latter type of analysis is used by most computer programs in the context of mixed-model analyses.

If the researcher has additional available information that can contribute to a better fit (correction) of the phenotypic values collected in the field, additional analyses can be incorporated into the design model. Several examples are discussed in the following sections.

2.4.1 Covariance Analysis

If supplementary information that can somehow predict the performance of the experimental units is available, which would be the case when several variables are collected in the experiment, it is sometimes possible to estimate the extent to which the observations of interest were influenced by the variations in these supplementary measurements. The aim of these analyses would then be to adjust the mean response of each treatment to remove the experimental error from this external source, the covariate. Thus, the variance from the supplementary variable is removed from the experimental error, without having to include this variable in the experimental design. In summary, the usefulness is the removal of the experimental error that arises from external sources of variation, which would be impractical or very expensive to control for using more refined techniques. A typical covariate is the stand of plants per plot, which varies between plots and, therefore, should be controlled for during analysis and not by design.

Aulchenko et al. (2007) proposed fitting the model $y = Xb + Zg + e$, which yields $\hat{e} = y - X\hat{b} - Z\hat{g}$ after fitting, where g is a vector of polygenic effects. The model $\hat{e} = 1u + Wm_i + e$ is then fit to the residuals ($\hat{e}$) to identify the significant markers (m_i). This analysis seeks to capture only the effects associated with Mendelian segregation, which arise only from the linkage disequilibrium between markers and genes. Thus, this approach is applicable to genome-wide association studies (GWAS) and genome-wide selection (GWS) of advanced generations, as opposed to the training population. Conversely, the fitting of the model $y = Xb + e$ is applicable to GWS in the current generation and in the short term (a few generations after the training population) and contains both the genetic effects from Mendelian segregation and those explained by genealogy (which are contained in the residuals $\hat{e} = y - X\hat{b}$), which are used for genomic analyses. In both models, the data in y are adjusted for the effects of the covariates in $\hat{b}$.

2.4.2 Spatial Analysis

The researcher will often want to conduct his or her experiment in a new area and, therefore, does not have in-depth knowledge of the spatial heterogeneity of the site where the experiment is being implemented. Thus, when the heterogeneity is unknown a priori, the definition of blocks becomes arbitrary, which can result in strong heterogeneity within blocks, thus causing a decrease in the efficiency of the chosen design. When the program does not have the technology and resources required for the high-resolution collection of data on spatial variation in variables at the study site, one alternative is to randomly allocate plots of a single plant in the experimental field and then control for environmental heterogeneity by using covariance analysis to correlate a covariate with the studied variable (Papadakis method: Papadakis 1984) or by using regional or spatial variables (geostatistical methods). Potentially, a posteriori fitting of the environmental gradients in progeny tests may significantly increase the effectiveness of the selection of genetic parameters. Thus, establishing randomized plots of a plant (completely randomized design) is important. However, Gilmour (2000) advises that a posteriori blocking should not be based solely on the statistical significance of arbitrary contrasts. The researcher should identify the physical and environmental causes that lead to a given type of blocking. If the number of treatments and partitions allows the researcher to use a certain, efficient experimental design, they can reduce the need for a posteriori fitting techniques (e.g., spatial analysis; Resende 2002).

2.4.3 Polynomial Functions for Rows and Columns for Fitting Spatial Trends

This method is based on the procedure proposed by Federer et al. (2001), which basically involves the selection of a polynomial function for the rows and columns that refer to the coordinates of the experimental plots to better absorb the random variations inherent in the data, according to the model for the design implemented. The mean values associated with each treatment will thus be corrected by this function, providing adjusted values that are used in subsequent analyses.

2.5 Important Considerations

2.5.1 More on Accuracy and the Number of Replicates

Figure 2.5 illustrates the accuracy of the data collected for a given individual as a function of the number of replicates of that individual (pure line or clone; i.e., absent of genetic variability but with environmental variability). If the trait follows a normal distribution with a mean $\mu = 10$ and $\sigma^2 = 4$, sampling replicates (e.g., $r = 1$, 2, 3, 4, 5) produce the plots shown in Fig. 2.5. When the environmental variability can be removed by blocking, the precision of the estimate (even with only one replicate) is much larger; that is, the curve will be more concentrated around the true value μ.

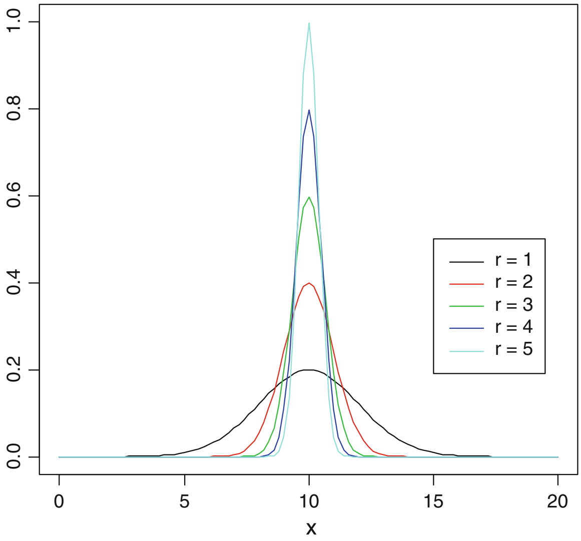


Fig. 2.5 Illustration of the range of values for trait *X* that can be obtained from various numbers of replicates of the experimental material. In this case, we are assuming $X \sim$ Normal (10, 4), which exhibits a CV = 20 %. Thus, with five replicates of the experimental material, it would be practically impossible to obtain a mean value greater than 12, but sampling only one replicate ($r = 1$) would likely yield values from 5 to 15. It is assumed that there is no genetic variability

2.5.2 Plot Size and the Number of Replicates

Several studies have confirmed that designs with a small number of plants per plot and numerous replicates are more efficient than those with numerous plants per plots and a small number of replicates (de Resende 2002). This relative superiority comes from the following: (a) the higher precision in the comparisons between treatments because of the greater number of replicates for a fixed-size experimental area; (b) the greater selective accuracy because of the greater number of replicates in a fixed-sized experimental area; (c) the greater individual heritability in the block because of the creation of more homogeneous blocks; (d) the lower overestimation (from any genotype × environment interaction) of heritability and genetic gain in a site because of the greater number of replicates analyzed (which may represent various environments); and (e) the smaller size and greater homogeneity of the block, reducing the need for spatial analysis of the experiments because local control is more effective.

As will be discussed, the plot should be considered the observational unit for data collection. For example, in sugarcane, the concept of plots of one plant should be interpreted as one furrow per plot for situations when all plants of a furrow are combined together in a composite sample to proceed with analyses. In this case, the individual heritability is defined as the heritability of a furrow and the number of replicates is determined as a function of the magnitude of this heritability at the furrow level.

The determination of sample sizes (in terms of numbers of replicates) for the estimation and prediction of various practical genetic breeding scenarios are explained by Resende (2002). To determine sample size, the criteria chosen was the maximization of the selective accuracy (the correlation between the true and estimated genetic values) as the number of replicates was increased (Table 2.1).

For example, with a heritability of 40 %, an accuracy of 90 % can be obtained with approximately seven replicates per clone. These are the recommended numbers per site. When there is a considerable genotype × environment interaction and a large planting area, experiments should be repeated in other sites before selection to minimize the adverse effects of the genotype × environment interaction.

When a researcher is conducting experiments with families, the genetic variability within the family contributes to the complexity of the problem. Thus, one must know the number of genotypes representing the families under study to determine the appropriate plot size and the number of replicates. In sugarcane breeding, for example, recommendations in the literature vary from 16 to 150 plants per family. In addition, the recommendations vary greatly depending on the parameter to be estimated and the type of trait to be analyzed (Peternelli et al. 2012; Leite et al. 2006, 2009).

A general approach for choosing sample size uses the confidence interval (CI; a 95 % CI is considered appropriate) for the sample mean ($\bar{y}$) of a normally distributed population. In this case, CI $= \bar{y} \pm 1.96\ \mathrm{s}(\bar{y})$, where $\mathrm{s}(\bar{y}) = \left(\hat{\sigma}^2 / \mathrm{n}\right)^{1/2}$ = the standard error of the mean. Thus, one can set a tolerance error (δ) in the estimate of the mean,

Table 2.1 Adequate number (N) of replicates per clone, in clonal tests, as a function of individual heritability (in one plot) (h_g^2), broadly speaking, to obtain an accuracy ($r_{\hat{g}g}$) of 90 and 95 %

h_g^2	N for $r_{\hat{g}g}$ = 90 %	N for $r_{\hat{g}g}$ = 95 %	h_g^2	N for $r_{\hat{g}g}$ = 90 %	N for $r_{\hat{g}g}$ = 95 %
0.05	81	176	0.40	7	14
0.10	39	84	0.45	6	12
0.15	25	53	0.50	5	10
0.20	18	38	0.60	3	7
0.25	13	28	0.70	2	4
0.30	10	21	0.80	2	3
0.35	8	17	0.90	1	2

given by $\delta = 1.96\ \mathrm{s}(\bar{y}) = (\hat{\sigma}^2/\mathrm{n})^{1/2}$. From this expression, $\mathrm{n} = (1.96^2\hat{\sigma}^2)/\delta^2$, which is the adequate sample size for an error tolerance of δ. The error tolerance is chosen by the researcher. If σ^2 is unknown, the estimated $\hat{\sigma}^2$ and t value (1.96) from Student's distribution is used in place of the z-score of the normal distribution.

Thus, to determine n, the δ for the estimate of the mean must be chosen and there must be an estimated $\hat{\sigma}^2$ for the phenotypic variability of the population. The error δ can be specified as a percentage of the mean (e.g., 10 %). In this case, δ is given by $0.10\bar{y}$.

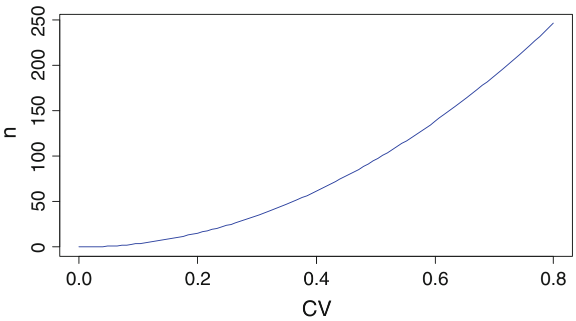
Thus:

$$n = \frac{1.96^2\hat{\sigma}^2}{(0.10\bar{y})^2} = \frac{1.96^2\hat{\sigma}^2}{0.10^2\bar{y}^2} = \frac{1.96^2}{0.10^2}\left(\frac{\hat{\sigma}}{\bar{y}}\right)^2 = \frac{1.96^2}{0.10^2}\mathrm{CV}^2;$$

where CV is the coefficient of variation of the trait in the population (Resende 2007). Using this approach, only an estimate or prior knowledge of the individual phenotypic CV in the population is required: the larger the CV, the larger the required sample size (Fig. 2.6).

For binomial variables, using the approximation to the normal distribution, the same expression for n can be solved by replacing $\hat{\sigma}^2$ with $p(1 - p)$ and $\bar{y}$ with p,

Fig. 2.6 Sample size (n) as a function of the phenotypic coefficient of variance (CV) of the trait in the population



where p refers to the observed proportion of the phenotypic class defined as "success." In the absence of information on p, one can use $p = 0.5$, which guarantees the largest variance possible.

2.5.3 Genetic Sampling and Effective Population Size

The genetic representativeness or effective size of a family is relevant to phenotyping in two aspects: (i) determining the size of the family in the experiment and (ii) obtaining adequate genetic representativeness of populations.

The effective size of full-sib families is given by $N_{ef} = (2n)/(n+1)$, where n is the number of individuals per family. The values of Ne for various values of n are listed in Table 2.2. This table also provides the results for half-sib and S1 families, which are discussed later.

Table 2.2 lists the number of individuals per family necessary to achieve a specific percent of the maximum N_{ef} of the family. An sample size of 100

Table 2.2 Effective size (N_{ef}) and the fractions of the maximum effective size (N_{efmax}) of a full-sib, half-sib, and S1 family as a function of the number (n) of individuals sampled per family

n	Full-sib		Half-sib		S1	
	N_{ef}	Fraction of N_{efmax}	N_{ef}	Fraction of N_{efmax}	N_{ef}	Fraction of N_{efmax}
1	1	0.500	1	0.250	0.670	0.670
5	1.667	0.833	2.5	0.625	0.909	0.909
7	1.750	0.875	2.8	0.700	0.933	0.933
10	1.818	0.910	3.1	0.775	0.952	0.952
12	1.846	0.923	3.2	0.800	0.960	0.960
15	1.875	0.938	3.3	0.825	0.968	0.968
18	1.895	0.947	3.4	0.850	0.973	0.973
20	1.905	0.952	3.5	0.875	0.976	0.976
25	1.923	0.962	3.6	0.90	0.980	0.980
30	1.935	0.968	3.64	0.91	0.984	0.984
40	1.951	0.976	3.72	0.93	0.987	0.987
50	1.961	0.980	3.77	0.94	**0.990**	**0.990**
60	1.967	0.984	3.88	0.97	0.992	0.992
100	**1.980**	**0.990**	3.88	0.97	0.995	0.995
150	1.987	0.993	3.92	0.98	0.996	0.996
200	1.990	0.995	3.94	0.985	0.997	0.997
250	1.992	0.996	3.95	0.988	0.998	0.998
300	1.993	0.997	**3.96**	**0.990**	0.998	0.998
∞	2.000	1.000	4.00	1.00	1.00	1.00

Source Adapted from Resende and Barbosa (2005) individuals will encompass 99 % of the maximum representativeness of the full-sib family. Therefore, increasing the sample size above 100 contributes almost nothing to increasing the representativeness for a family.

For half-sib families, the effective size of a family (N_{ef}) is given by $N_{ef} = (4n)/(n+3)$, where n is the number of individuals per family. For half-sib families, 300 individuals provide 99 % of the maximum representativeness of a family (Table 2.2). Because the ideal crossing design for selecting parents and clones within families assumes that three crosses are performed per parent, three full-sib families are associated with each parent. Thus, by adopting a family size of 100 for full-sib families, we obtain a size of exactly 300 for each half-sib family.

The effective size of an S1 family is given by $N_{ef} = (n)/(n+0.5)$ and the maximum equals 1, when n goes to infinity. However, with n = 1, the N_{ef} is already equal to 0.67. With n = 50, N_{ef} is already 0.99—that is, 99 % of the maximum N_{ef} (Table 2.2). One can say that the probability of adding an effectively different individual is less than 1 for each 100 individuals added after n = 50 (or exactly 0.67 for the first 100 after 50). Nevertheless, there would not be sufficient precision for including only this individual in the selection among 150 S1 individuals. Thus, it is believed that 50 individuals is an adequate size for selection in S1 families. The number n = 50 for S1 families is comparable to the numbers n = 100 and n = 300 for the progeny of full and half sibs, respectively. In other words, these numbers (50, 100, and 300) provide 99 % of the maximum representativeness of S1, full-sib, and half-sib families, respectively, and therefore would be adequate sizes of progeny for selection within said families.

In conclusion, 100 individuals per full-sib family and 300 per half-sib family are adequate sample sizes. The 100 individuals from each full-sib family can be divided into two or three environments at multiple sites.

2.5.4 Number of Experimental Sites

The appropriate number of experimental sites can be determined from the selection efficiency (E_f) for the mean of several environments (ℓ) relative to the selection in only one environment aiming to obtain gains in the mean of ℓ sites. This efficiency can be inferred (for heritability, at the level of the mean, similar and tending to 1 in various environments, similar to well-designed clonal tests) by the expression $E_f = [\ell/[1+(\ell-1)r_{gg}]]^{1/2}$, where, r_{gg} is the genetic correlation involving the performance of the germplasm in the environment (Table 2.3; Resende 2002).

The results shown in Table 2.3 demonstrate that when the genetic correlation is equal to or greater than 0.70, the gain in efficiency from analyzing more than one experimental site is less than 10 %. If the genetic correlation is greater than 0.80, the gain in efficiency is less than 5 %. Conversely, using three sites instead of two sites

Table 2.3 Efficiency (in terms of genetic gain in the mean of the sites) of using ℓ sites instead of one site for assessing genetic material, for various values of the genetic correlation (r_{gg}) involving the performance of the germplasm in the environment

r_{gg}	ℓ	E	r_{gg}	ℓ	E
0.90	2	1.03	0.55	2	1.14
	3	1.04		3	1.20
0.80	2	1.05	0.50	2	1.15
	3	1.07		3	1.22
0.70	2	1.08	0.40	2	1.20
	3	1.12		3	1.29
				4	1.35
0.60	2	1.12	0.30	2	1.24
	3	1.17		3	1.37
				4	1.45
				5	1.51

is recommended only when the correlation (estimated using three or more sites) is less than 0.5. Using four sites would be advantageous when the correlation is less than 0.40. Resende (2002) presented other approaches for various selection strategies for which the ideal numbers of sites are defined. The interested reader should refer to this reference.

The appropriate number of experimental sites for a fixed total number of individuals depends on the heritability of the trait and the intraclass genetic correlation across sites. Setting $n\ell$ as the total number of individuals per accession, where n refers to the number of individuals per site, and comparing the analysis of the $n\ell$ individuals in one environment or in several environments, the efficiency of selection based on various sites compared to selection based on one single site is given by

$$E = \left[\frac{1 + (n\ell - 1)\hat{h}_i^2}{1 + (n - 1)\hat{h}_i^2 + n(\ell - 1)\hat{r}_{gg}\hat{h}_i^2}\right]^{1/2}$$

where $\hat{h}_i^2$ is the estimated individual heritability within the site.

For example, for $h^2 = 0.20$, using 30 individuals per family will provide accuracy on the order of 90–95 % for the selection of individuals for propagation by seeds or clones for vegetative propagation (Resende 2002). For a total fixed number of individuals assessed, the author noted that it is advantageous (a gain of at least 6 %) to use four, three, and two sites for correlations with magnitudes of 0.30, 0.50, and 0.70, respectively. This result demonstrates that the interaction can be minimized without devoting additional resources but simply by dividing a large experiment across several sites.

References

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19(1):51–61

Auer PL, Doerge RW (2010) Statistical design and analysis of RNA sequencing data. Genetics 185:405–416

Aulchenko YS, Koning D, Haley C (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177:577–585

Banzato DA, Kronka SN (1989) Experimentação agrícola. FUNEP, Jaboticabal, 247 pp

Barbin D (2003) Planejamento e análise estatística de experimentos agronômicos. Midas, Arapongas, 208 pp

Cabrera-Bosquet LJ, Crossa J, von Zitzewitz MD, Serret J, Araus L (2012) High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. J Integr Plant Biol 54:312–320

Cobb JN, Declerck G, Greenbrg A, Clark R, McCouch S (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. Theor Appl Genet 126:867–887

Cochran WG, Cox GM (1992) Experimental designs, 2nd edn. Wiley, New York, 611 pp

Cox DR (1958) Planning of experiments. Wiley, New York, 308 pp

Crossa J, Burgueño J, Cornelius PL, McLaren G, Trethowan R et al (2006) Modeling genotype·environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Sci 46:1722–1733

Faraway JJ (2005) Linear models with R. Chapman & Hall/CRC, New York, 229 pp

Federer WT (1956) Augmented (hoonuiaku) designs. Hawaiian Planters' Rec 55:191–208 (Aica)

Federer WT, Reynolds M, Crossa J (2001) Combining results from augmented designs over sites. Agron J 93:389–395

Fiorani F, Schurr U (2013) Future scenarios for plant phenotyping. Ann Rev Plant Biol 64:267–291

Fisher RA (1926) The arrangement of field experiments. J Ministry Agric Great Brit 33:503–513

Fisher RA (1935) The design of experiments, 2nd edn. Oliver & Boyd, Edinburgh

Gilmour AR (2000) Post blocking gone too far! Recovery of information and spatial analysis in field experiments. Biometrics 56:944–946

Hinkelmann K, Kempthorne O (1994) Design and analysis of experiments—volume I: introduction to experimental design. Wiley, New York, 495 pp

Hinkelmann K, Kempthorne O (2005) Design and analysis of experiments—volume II: advanced experimental design. Wiley, New York 780 pp

Lado B, Matus I, Rodriguez A, Inostroza L, Poland J, Belzile F, del Pozo A, Quincke M, Castro M, von Zitzewitz J (2013) Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. G3 3:2105–2114

Leite MSO, Peternelli LA, Barbosa MHP (2006) Effects of plot size on the estimation of genetic parameters in sugarcane families. Crop Breed Appl Biotech 6(1):40–46

Leite MSO, Peternelli LA, Barbosa MHP, Cecon PR, Cruz CD (2009) Sample size for full-sib family evaluation in sugarcane. Pesquisa Agropecuária Bras 44:562–1574

Masuka BJ, Araus L, Das B, Sonder K, Cairns JE (2012) Phenotyping for abiotic stress tolerance in maize. J Integr Plant Biol 54:238–249

Papadakis J (1984) Advances in the analysis of field experiments. Communicationes dÁcademie dÁthenes 59:326–342

Patterson HD, Williams ER (1976) A new class of resolvable block designs. Biometrika 63:83–92

Peternelli LA, Souza EFM, Barbosa MHP, Carvalho MP (2009) Delineamentos aumentados no melhoramento de plantas em condições de restrições de recursos. Ciência Rural 39:2425–2430 (UFSM-Impresso)

Peternelli LA, Resende MDV, Mendes TO (2012) Experimentação e análise estatística em cana-de-açúcar. In: Santos F, Borém A, Caldas C (eds) Cana-de-açúcar: bioenergia, açúcar e etanol—Tecnologias e perspectivas, 2nd edn. Editora Folha de Viçosa Ltda., Viçosa, pp 333–353

Poorter H, Fiorani F, Stitt M, Schurr U, Finck A, Gibon Y, Usadel B, Munns R, Atkin OK, Tardieu F, Pons TL (2012) The art of growing plants for experimental purposes: a practical guide for the plant biologist. Funct Plant Biol 39:821–838

Ramalho MAP, Ferreira DF, Oliveira AC (2005) Experimentação em genética e melhoramento de plantas. UFLA, Lavras, 300 pp

Resende MDV (2002) Genética biométrica e estatística no melhoramento de plantas perenes. Embrapa Informação Tecnológica, Brasília, 975 pp

Resende MDV (2007) Matemática e estatística na análise de experimentos e no melhoramento genético. Embrapa Florestas, Colombo, 560 pp

Resende MDV, Barbosa MHP (2005) Melhoramento genético de plantas de propagação assexuada. Embrapa Florestas, Colombo, 130 pp

Scott RA, Milliken GA (1993) A SAS program for analyzing augmented randomized complete-block designs. Crop Sci 33:865–867

Silva MAG, Peternelli LA, Nascimento M, da Silva FL (2013) Modelos mistos na seleção de famílias de cana-de-açúcar aparentadas sob o enfoque clássico e bayesiano. Revista Brasileira de Biometria 31:1–12

Souza EFM, Peternelli LA, Barbosa MHP (2006) Designs and model effects definitions in the initial stage of a plant breeding program. Pesq Agropec Bras 41(3):369–375 (Brasília)

Steel RGD, Torrie JH, Dickey DA (1997) Principles and procedures of statistics: a biometrical approach, 3rd edn. McGraw-Hill Companies, New York, 666 pp

Storck L, Garcia DC, Lopes SJ, Estefanel V (2000) Experimentação vegetal. In: Santa Maria RS (ed) da Universidade Federal de Santa Maria, 199 pp

Williams ER, Matheson AC (1994) Experimental design and analysis for use in tree improvement. CSIRO Information Services, East Melbourne, 174 pp

Chapter 3
Statistical Analysis of Gene Expression and Genomic Data

Marcos Deon Vilela de Resende, Fabyano Fonseca e Silva, Moysés Nascimento, Camila Ferreira Azevedo and Luiz Alexandre Peternelli

Abstract This chapter deals with the statistical analysis of genomic, transcriptomic and proteomic data. Emphasis is given to the analysis of gene expression data including hints on experimental designs to sound data generation. Epigenetic variation is also addressed as a mean to enhance the analyses.

3.1 Statistical Analysis of Genomic Data

Genomic studies began with the mapping of quantitative trait loci (QTLs) by means of low-density genome scanning. Subsequently, marker-assisted selection (MAS) was proposed and implemented using a mixed inheritance model combining polygenic components with a component related to a major effect QTL (Fernando and Grossman 1989; Lande and Thompson 1990). With the advent of single nucleotide polymorphism (SNP) markers, association mapping has become common, and it is implemented via high-density genomic scanning to produce fine mapping. In addition, genomic selection (GS) or genome-wide selection (GWS) has

M.D.V. de Resende (✉)
Embrapa Florestas, Paraná, Brazil
e-mail: marcos.deon@ufv.br

M.D.V. de Resende · F.F. e Silva · M. Nascimento · C.F. Azevedo · L.A. Peternelli
Universidade Federal de Viçosa, Viçosa, Brazil
e-mail: fabyanofonseca@ufv.br

M. Nascimento
e-mail: moysesnascim@ufv.br

C.F. Azevedo
e-mail: camila.azevedo1504@gmail.com

L.A. Peternelli
e-mail: peternelli@ufv.br

R. Fritsche-Neto and A. Borém (eds.), *Phenomics*,
DOI 10.1007/978-3-319-13677-6_3

Table 3.1 Evolution of the type of data and analysis methods used in the selection of quantitative traits

Data types	Analysis method
Phenotypic and pedigree	Mixed models—BLUP
Phenotypic, pedigree and microsatellite markers	LA and LA-MAS mapping
Phenotypic (few) and SNP markers	LD, LD-MAS, and GWS mapping
Phenotypic (rare) and SNP markers	Mixed models combined with coalescence
Phenotypic (many) and SNP markers	GWAS
Phenotypic (many), SNP markers and methylation data	GWS and GWAS with epigenetic effects

LA linkage analysis; *LD* linkage disequilibrium

become possible, which is superior to MAS. Transcriptomics and proteomics have also emerged as new sources of information that can be used in genetic evaluation procedures.

The evolution of genetic evaluation procedures can be characterized according to Table 3.1, which is based on the report of Perez-Enciso (2007) and includes additional content.

When migrating to fine mapping (in which the confidence interval of the QTL localization is narrowed), pedigree information has become irrelevant, whereas other statistical techniques have become more useful.

Even with this technological evolution, phenotypic analysis remains crucial and is essential in the analysis prior to and/or concomitant with genomic, transcriptomic, and proteomic data.

Generally, an appropriate procedure for the prediction of genetic values that simultaneously uses information from molecular markers and phenotypes (y) can be obtained by means of the mixed models method.

The linear mixed model encompassing fixed (b), random genetic (g) and random environmental effects (e) is given by $y = Xb + Zg + e$.

Including the effects (q) of QTLs for each locus j, the model becomes $y = Xb + Zg + \sum_j W_j q_j + e$, where W_j is an incidence matrix that relates individuals with the alleles of locus j and where q contains the allelic effects for each locus. The incidence matrices W and their dimensions, which are produced by the number of alleles in each locus, are not known. The number of loci that affects the trait is also not known, which is in contrast with the first model, in which the incidence matrices for b and g (X and Z, respectively) are known. If W is known, the mixed model equations could be used without any changes and would constitute the model associated with the QTL analysis and MAS.

A better model is given by $y = Xb + \sum_j W_j q_j + e$, in which all loci are individualized and there is no need to include the polygenic genetic residual or infinitesimal component g. With W and q inferred from markers, this model characterizes GWS.

An ideal method for GWS should have three attributes: (i) **accommodate the genetic architecture** of the trait in terms of genes of small and large effects and their distributions; (ii) **regularize** the estimation process in the presence of multicollinearity and large numbers of markers using shrinkage estimators; and (iii) **select covariates** (markers) that affect the characteristic under analysis. The main problem associated with GWS is the estimation of a large number of effects from a limited number of observations and the collinearities derived from linkage disequilibrium between markers. Shrinkage estimators appropriately manage this issue and treat marker effects as random variables, estimating them simultaneously. The main methods for GWS can be divided into three large classes: explicit regression, implicit regression, and regression with dimensional reduction. The first class includes the methods of random regression–best linear unbiased prediction (RR-BLUP), least absolute shrinkage and selection operator (LASSO), elastic net (EN), BayesA, and BayesB. In the implicit regression class, the neural networks, reproducing kernel Hilbert spaces (RKHS) and nonparametric kernel regression methods via generalized additive models are included. The regression methods with dimension reduction include the independent components, partial least squares, and principal components. Additional details on GWS and GWAS were presented by Resende et al. (2014a, b).

3.2 Statistical Analysis of Transcriptomic Data

The transcriptome is the set of all transcripts in a cell under a given physiological condition; thus, inferences related to the transcriptome are fundamental to understanding cellular dynamics by allowing investigations of the mechanisms by which the genome interacts with its environment. Such inferences mainly reside in the mapping and quantification of the transcripts generated by genomic regions under different environmental conditions, which are usually referred to as treatments.

Since the 1990s, one of the primary techniques of transcriptome inference has been DNA microarrays. The data produced by this technology are based on DNA slides that simultaneously characterize the expression of thousands of genes subjected to different treatments. Laboratory procedures for the production of this type of data involve the extraction of messenger RNA (mRNA), reverse transcription to obtain complementary DNA (cDNA), fluorescent labeling, and cDNA hybridization with commercial DNA probes that are carefully placed on slides. The microarray technique provides an inference of the level of gene expression (i.e., the abundance of RNA transcripts) by the use of specific dyes that translate the level of expression into light intensity when irradiated with specific wavelengths.

The term *genetical genomics* was coined by Jansen and Nap (2001) to designate the combined study of transcriptome variability and polymorphism of DNA sequences. Genetical genomics involves two approaches: (i) determination of the genetic architecture of the transcriptome through the analysis of thousands of expression QTLs (eQTLs), where the phenotypes are cDNA levels associated with each gene; and (ii) use of gene expression data for the localization of candidate genes. For the latter approach to be successful, it is necessary that the levels of gene expression be under genetic control and that a number of the heritable expression levels be correlated with the trait of interest. Perez-Enciso et al. (2003) reported the combination of information from molecular markers and gene expressions for the mapping of quantitative traits.

Expression data are related to transcription (mRNA levels). Microarray-based technology is used to determine the differential gene expression of the entire genome in biological samples of specific tissues. Recently, large-scale sequencing technology has been used as an alternative for microarray-based techniques; such technology is referred to as *RNA-seq* and is based on the sequencing of a sample of all transcripts from an individual under certain conditions and in a particular tissue. The reading depth associated with each transcript is correlated with the expression level of the gene in question. Additional details on these approaches will be provided.

The genetic expression levels or mRNA amounts are then subjected to correlation analysis with quantitative traits in individuals of a segregating population with a focus on the detection of QTLs. As an example, differences in the mRNA amount produced by disease-resistant and susceptible plants may indicate that a particular mRNA is associated with a resistance gene. Using the amount of genetic expression to detect QTLs is more suitable for traits of resistance to stress caused by abiotic factors, such as drought and salinity, and for traits of resistance to diseases and pests.

In genetical genomics, the association between the mRNA level and DNA polymorphism (instead of the association between phenotype and DNA polymorphism) is justified by the greater proximity between RNA and DNA than between phenotype and DNA. However, a fundamental question is how to link the expression of a QTL with the phenotypic trait of interest. Direct analysis methods of gene expression and function are also essential to determine whether two close markers are detecting the same QTL or two close QTLs.

The combination of genetic data and gene expression data for the entire genome has provided information on the genetic basis of gene expression. In this case, the mRNA levels are phenotypic data subject to variations related to genetic and environmental causes, and the genomic regions that control the level of expression of the genes studied are identified. Basically, the regulation of the level of expression is divided into two classes, *cis* and *trans*. If the polymorphism associated with the differential expression level is close to the gene from which the mRNA was transcribed, the regulation is of the *cis* type. Otherwise, if the marker (and then the eQTL) is mapped in a position different from the transcript position, the gene is *trans*-regulated. The latter type of regulation is usually associated with a transcription factor that alters (or activates/deactivates) the level of mRNA expression in question. Studies have shown high genetic variation among genotypes for gene expression, and significant heritability estimates have been obtained. In humans, the heritability of gene expression levels is approximately equal to 30 %, which is important because the statistical power to detect genetic variants that affect gene expression depends on heritability. Thus, genes are expressed as a function of an environmental stimulus.

DNA microarray data (also referred to as slides) simultaneously involve the expression of thousands of genes at a certain age of the individual and under certain environmental conditions. The laboratory procedures for the production of such data involve the extraction of mRNA, reverse transcription to obtain cDNA, fluorescent labeling, and cDNA hybridization with commercial DNA probes. The microarray technique provides an inference of the level of gene expression via the abundance of transcribed RNAs. It also allows, in certain cases, the integration of genetics and physiology by determining networks among sets of genes associated with physiological characteristics. A disadvantage of using microarrays is the need for a priori knowledge of DNA sequences to develop the probes used in the hybridization. Thus, if a transcript is not previously known, it is not possible to build a probe and the expression of this gene will not be detected. In the RNA-seq method, a sample of all transcripts is sequenced and does not require a priori knowledge of the sequence of each gene.

The analysis of gene expression can be used to infer the function of genes and provide an understanding of the differential gene expression among tissues, developmental stages, responses to environmental stresses, and different genotypes. The analysis of such data is addressed in detail in the literature (Kerr et al. 2000; Wolfinger et al. 2001; Tempelman 2005; Rosa et al. 2007; Ayroles and Gibson 2006). In the case of microarrays, two types of array platforms may be used: (1) platforms based on a system of two colors that generate two samples per array (spotted cDNA); and (2) platforms based on a single color system (dye) or single channel array that generate one sample per array (Affymetrix). System 1 requires more complex designs (loop or circular and split plots) and analyses, and a large number of technical replicates. In contrast, system 2 allows multiple probes per gene and has a tendency to use a smaller number of replicates. Additionally, the design is simple, and there is no reference sample.

Two main approaches have been used in experiments with two-color arrays: (1) one color (green = cyanine 3 or Cy3) that is reserved for the reference or control sample, and another color (red = cyanine 5 or Cy5) that is used to evaluate the treatments; and (2) two colors that are used to evaluate treatments of interest. In approach 1, the Cy3/Cy5 ratio between the fluorescence intensities provides a measure of the intensity of gene expression. This approach is intuitive and suitable for situations in which there are a large number of treatments of the same factor with a small number of replicates. Approach 2 requires more refined designs to avoid confounding between factors (slides and samples of nucleic acid); in addition, the effects of the dyes are pronounced, and it is essential that each sample be represented by technical replicates of both dyes in equal proportions.

The loop design should be used in comparisons of the contributions of each factor. The split-plot design should be used to determine the effect of one factor through samples that include effects of another factor that is of less interest. For either of the two approaches, the array effect must be taken as random to consider that the two measurements in the same array are correlated and to adjust for the effect of common array environments. The experimental design guides the formulation of the appropriate linear model for analysis. Each slide or array is analogous to an incomplete block because they encompass only two of the various treatments. Additionally, each slide contains the effects of two dyes, and the design is of a row and column type with the dimensions 2 × s, where s is the number of slides or arrays.

For single-channel arrays, the experimental design is simplified and there is no need to consider the array and dye effects because there is no confounding; each sample is hybridized on a different array and measured independently. The reference or control sample is used to correct the data for the slide effects. In this case, the design is the incomplete block type with common treatments. The comparison between treatments is performed indirectly by means of the differences among treatments and the reference or control on each slide. In addition, in circular designs, the comparative effects of treatments are estimated through combinations of direct comparisons (among treatments within blocks or slides) and indirect comparisons (among treatments between blocks or slides).

For cDNA sequencing (RNA-seq), the experimental designs tend to be simpler. In general, the factors fitted in the model are the effects of different channels (lanes) and different runs (flow-cell), using a randomized complete block design (Auer and Doerge 2010). Gene expression is usually quantified by normalizing the reading coverage and sequence size. The most common method is normalization of the number of reads per kilobase (Kb) per million of sequences mapped to the reference (RPKM; Mortazavi et al. 2008).

In transcriptomics, a distinction is made between technical replication and biological replication. Technical replication refers to replicating the hybridization of the same RNA samples from the same common biological source. Thus, these replications are not completely independent from one another and are used to validate the accuracy of measurements of the level of transcripts and to model residual effects, such as the variation related to sequencing of the same sample on different channels. Therefore, these replications do not provide information on the level of variation in the population. Similarly, replicated probes within an array are used to reduce the need for technical replicates by increasing the reliability of abundance measures of transcripts for a specific target gene. With high-quality commercial arrays, technical errors are much smaller than the biological variance and more than two replicates per sample are generally not required.

Biological replication refers to replicating the hybridization of RNA samples from independent biological sources under the same conditions or treatment, such as samples taken from different individuals who received the same dose of a treatment, or even two replicas of the same genotype of a plant. These replicas are intended to provide information on biological variation between individuals (Ayroles and Gibson 2006). As for the required number of replicates, Wolfinger et al. (2001) and Tempelman (2005) recommend the use of at least four technical replicates for each biological replicate to detect 80 % of the genes expressed differentially among experimental groups.

The expression intensity data of each dye are initially converted to a base 2 logarithmic scale. The log transformation has the advantage of making the data approach a normal distribution, and it is more symmetric. With the data transformed at a logarithmic scale, the mean components associated with linear models can be used as appropriate statistical procedures. Therefore, additional statistics, such as the median, are not required. Without applying the logarithmic transformation, the statistical median is recommended because it is robust for outliers. After cleaning the data (e.g., removing nonexpressed genes, low intensity arrays), the data must be normalized to remove the global effects of arrays and dyes that do not reflect true genetic variation within and between arrays. These biases result from factors such as variations in the amount of DNA placed between arrays. Normalization methods, such as locally weighted regression scatterplot smoother (LOWESS; a robust nonparametric regression), can be used. LOWESS uses local regressions to remove general correlations between the intensity and intensity ratio. Another procedure is quantile normalization, which performs a nonlinear transformation that produces arrays with equal means, medians, and variances by obtaining the mean intensity of each quantile through the arrays.

However, such a global standardization can artificially remove true biological differences. Thus, alternative models can be used to remove the effects of slides and dyes, and the modeling of these effects in the statistical analysis can adjust the data for such effects. Wolfinger et al. (2001) proposed a two-model process. The first model fits the data (log transformed) to the overall effects (all genes simultaneously), slide or array effects (A), dyes effects (D), and their interactions (AD) using the model $\text{Log}_2(y) = u + \text{A} + \text{D} + \text{AD} + \text{Residual}(1)$. The second model uses the residual estimated by the previous model in a new analysis model designed for specific or individual genes. The first model that is designed for global normalization expresses fluorescence intensity as a deviation from the overall mean, and the second model can infer if these deviations differ among the factors (treatments, etc.) of the model and for individual genes.

The gene-specific model of Wolfinger et al. (2001) is given by $\text{Residual}(1) = u + \text{A} + \text{D} + \text{AD} + T + error$, where T is the treatment factor and *error* is a vector of errors specific to each gene. This model is fit separately for each gene in the array and considers specific variance components for each gene. The A and AD effects should be set as random, and the D factor effects should be set as fixed effects. The T factor effects should be considered fixed when referring to comparisons of different levels of stress to which a certain genotype is subjected and considered as random when referring to more than five genotypes taken from a population. Significance tests can be applied to fixed (F, Wald) and random (LRT or analysis of deviance) effect factors.

An alternative is to perform normalization simultaneously with the fitting of all other factors of the model and then to evaluate all of the effects of individual genes. According to Kerr et al. (2000), this is accomplished using the following model:

$y_{ijkm} = u + A_i + D_j + AD_{ij} + G_m + AG_{im} + DG_{jm} + T_k + TG_{km} + e_{ijkm}$, where y_{ijkm} is the variable abundance of transcription at a $\log_2$ scale and e_{ijkm} is a common residual to all genes. The effects of genes (G) and their interactions should be considered random. The interaction of major interest is TG_{km}, which depicts the effects of treatment k on the level of expression of gene m. More complex models that encompass the levels of biological variation (different genotypes) can also be used, and they allow the estimation of the components of variance and heritability of gene expression patterns. This model is relevant because it considers all of the effects simultaneously in a single analysis. However, it has the disadvantage of considering a residual variance common to all genes.

The use of the least squares method in the analysis of microarray data with all genes simultaneously has certain restrictions because of the large number of genes in relation to the number of slides. That is, the number of effects to be estimated is larger than the amount of data, which leads to estimation problems for modeling covariances among levels of expression of various genes because of the reduced number of degrees of freedom. The alternative to be adopted refers to the use of shrinkage estimators for the variance components (Cui et al. 2005).

Certain experiments may use multiple probes within an array, and the model at the observation level in each probe (P) can be fit for each gene. This model may be written as follows:

$y_{ijkm} = u + A_i + D_j + AD_{ij} + P_m + T_k + TP_{km} + e_{ijkm}$, where the probe effect and interaction term (TP_{km}) are random.

Models of this type were used by Drost et al. (2008) in *Eucalyptus*. In this genus, genetic markers have been generated from gene expression data. Thus, the expression intensity and detection of sequence polymorphisms are obtained simultaneously. Two classes of polymorphisms are obtained: (i) polymorphisms in sequences complementary to oligonucleotides of expressed genes (SFP; single-feature polymorphisms); and (ii) gene expression markers (GEM). The distinction between SFPs and GEMs by microarray data analysis allows SFP markers to be quickly obtained for use in association studies and genomic selection.

In significance tests of the model effects, the p values must be adjusted when multiple tests are conducted in an experiment, which is the case when thousands of genes are tested simultaneously. In such cases, a Bonferroni correction is applied, and the overall desired significance level α is specified and divided by the number (n) of tests to be performed. Thus, the corrected significance level $\alpha^* = \alpha/n$ is obtained, which is used as a significance threshold for each test. This approach is conservative and reduces the power of the tests. A more appropriate criterion for such cases is the false discovery rate (FDR), which is defined as the expected rate of false-positives among all significant tests (Rosa et al. 2007).

Gene expression studies and estimations of the effects of SNP markers or diversity arrays technology (DArT) (in the context of genome-wide selection) can characterize or determine the molecular or genetic signatures of the traits, which refers to the determination of the entire set of genes that affect a certain phenotypic trait.

3.2.1 Gene Expression Analysis in RNA-Seq Experiments

Because the extent of gene expression in RNA-Seq experiments is discrete (number of reads mapped to a particular transcript), statistical methods for the detection of differential gene expression are formulated based on discrete probability distributions, such as binomial, Poisson, and negative binomial (NB) distributions. It was initially assumed that the number of reads was distributed according to a binomial distribution; however, in the presence of a large number of reads (as is common in RNA-Seq data), the mapping probability is small, so these counts are characterized as rare events and can be described by a Poisson distribution. Thus, the Poisson distribution was adopted in early studies (Marioni et al. 2008; Wang et al. 2010) based on RNA-Seq data analysis. In subsequent studies, such as that of Anders and Huber (2010), RNA-Seq data were reported as having a greater variability than that presented by the Poisson distribution, in which it is assumed that the variance is equal to the mean. In count data modeling, when the variance exceeds the mean, the phenomenon called superdispersion is observed, which causes an underestimation of the sampling error when adopting the Poisson model. In contrast, the NB distribution assumes that the variance is greater than the mean, thus allowing a natural treatment for superdispersion by estimating the parameters of this distribution. In practical terms, this superdispersion is caused by the extra variation (heterogeneity of variances) inherent to the biological replicates (Anders and Huber 2010).

One of the statistical methods proposed for the detection of gene expression based on an NB distribution was presented by Robinson and Smyth (2008) and is implemented in the edgeR package (Robinson et al. 2010) of the free R/Bioconductor software. This method assumes a dispersion of extra variance in relation to the mean that is common to all genes. The estimation of the dispersion parameter is accurate because a large amount of data is used, as this parameter is the same for all genes considered. With Y_{ij} representing the number of reads of gene i in library j, the method assumes that $Y_{ij} \sim \mathrm{BN}\left(\mu_i = m_j\lambda_{ij}, \varphi\right)$, where μ_i is the distribution mean, which can be defined by the product of the scale factor (total number of reads) of library j (m_j) by the expression of gene i in this library j (λ_{ij}), and φ is the superdispersion parameter common to all genes.

In the absence of normalization (library standardization), the term m_j is the total number of reads in a library, and in its presence, m_j is the scale factor calculated according to Robinson and Oshlack (2010). Robinson and Smyth (2008) performed this estimation based on the maximum likelihood (ML) method, which consists of obtaining $\hat{\lambda}_{ij}$ *and* $\hat{\varphi}$ as the values that maximize the log-likelihood function:

$$\begin{aligned}\log L(\lambda_{ij}, \varphi|\mathbf{y}) &= \log\prod_{i=1}^{n} f(y_{ij}|\lambda_{ij}, \varphi)\\ &= \prod_{i=1}^{n}\frac{\Gamma(y_{ij}+\varphi^{-1})}{\Gamma(\varphi^{-1})\Gamma(y_{ij}+1)}\left(\frac{1}{1+m_j\lambda_{ij}\varphi}\right)\left(\frac{m_j\lambda_{ij}}{\varphi^{-1}+m_j\lambda_{ij}}\right)^{y_{ij}},\end{aligned}$$

where $f(Y_{ij}|\lambda_{ij}, \varphi)$ is the probability density function of the NB distribution.

Once the parameters related to the expression of gene i in libraries j ($\hat{\lambda}_{ij}$) and j' ($\hat{\lambda}_{ij'}$) are estimated under the superdispersion control provided by the estimate of φ, the next step is to test the hypothesis Ho: $\hat{\lambda}_{ij} = \hat{\lambda}_{ij'}$. Such tests can be based on asymptotic theory, such as Wald, likelihood score and ratio, and an accurate test was developed by Robinson and Smyth (2008); according to the authors, it has a better performance than the other tests. Another NB-based method was proposed by Anders and Huber (2010) and is implemented in the DESeq package of the free R/Bioconductor software.

In certain cases, high costs prevent the use of biological replicates, which implies low statistical power to differentially test the expressed genes from the perspective of frequentist statistics. Although the method developed by Anders and Huber (2010) does not require analyses with replicates (by means of the "blind" option of the DESeq package), the results cannot be considered in terms of statistical significance. To present an alternative method of analysis with small numbers of or without replicates, Hardcastle and Kelly (2010) proposed a Bayesian method to establish a posteriori likelihoods of various differential expression models. This method is implemented in the baySeq package of the free R/Bioconductor software.

Regardless of the technique adopted for gene expression analysis, microarrays or RNA-Seq, the genes identified as significant in different comparisons among treatments are usually subjected to a cluster analysis to facilitate the interpretation and visualization of groups of genes with similar expression patterns. Thus, inferences on the regulatory mechanisms and gene ontologies (GO) can be produced for specific groups of genes to identify related biological functions (Mukhopadhyay and Chatterjee 2007). When the treatments under study are characterized as different instances of time—meaning that they are from the viewpoint of longitudinal data—specific clustering methods must be used to consider the temporal dependency that occurs between the treatments (Nascimento et al. 2012).

3.3 Statistical Analysis of Proteomic Data

Proteomics describes the tools used to analyze the proteome, which in turn refers to the set of all proteins produced by the genome of an organism (Hollung et al. 2007). In general, the genome contains information related to the genes that are available and their possible locations, the transcriptome contains information related to the genes that are being expressed under certain conditions, and the proteome contains information related to the expressed genes that are being effectively translated into proteins of biological interest.

The main goal of proteomics is to identify new and potentially unexpected changes in the expression of proteins and their interactions or modifications as a result of an experimental treatment. In essence, proteomics can evaluate the complete scenario of cellular functions rather than a particular action of a protein (Lippolis and Reinhardt 2008).

The classical method for protein evaluation is two-dimensional (2D) gel electrophoresis, in which proteins are separated by the isoelectric point in the first dimension. The separation then occurs according to molecular weight in the second dimension, which is followed by identification by mass spectrometry (MS). The combination of 2D gel plus MS is known as gel-based proteomics, and it has been shown to be highly efficient in agriculture (Hochholdinger et al. 2006; Lametsch 2011). Although this method is the most frequently used, its efficiency depends on the characteristics of the biological and biochemical material as well as the level of expression of genes preselected for analysis (Lippolis and Reinhardt 2008).

In general, the analysis of gel-based proteomics data does not require significant statistical and/or computational costs because it includes a simple analysis process that is based on two distinct steps. The objective of the first step is to obtain a matrix with dimensions N by p, where N is the number of gel plates (assumed as treatments) containing profiles of p proteins (or their precursors, such as peptides). Thus, each row of the matrix corresponds to a treatment (or replicate of the same treatment), and each column corresponds to a different protein, whose expression is quantified by the intensity of the spot obtained from the gel run. Therefore, this first step is applied to correct or standardize the results of the spot images by means of an expression index (Morris et al. 2010). Specific software such as Image Master 2D Platinum use techniques to eliminate the distortions of spot images through alignments, normalizations, and scale corrections to condense such images in this index (usually referred to as a peak), which represents the intensity at which a certain protein (represented in the column of the matrix) was expressed in each treatment of interest (represented in the row of the matrix).

Once the matrix of results is obtained in the first step, it is possible to use different statistical methods to associate the expression intensity of each protein in the columns with the respective treatments in the rows. Thus, it is possible to infer on which protein a given treatment is more highly expressed and compare specific proteins between treatments. These inferences are made from either the univariate or multivariate viewpoint (Morris et al. 2010). In the univariate methods, once gel-run technical replicates are obtained (i.e., two or more matrix rows to represent the same treatment), simple tests, such as Student's t-test, or generalizations of analyses of variance (ANOVAs) can be used to compare treatments within a certain protein or to compare different proteins within the same treatment. In contrast, multivariate methods are used to outline a general pattern of all proteins and all treatments to generate visualizations that can provide conclusions on the differences between proteins and between treatments at once. Among the multivariate methods used in proteomic analysis, the principal components method and factor analysis are included.

3.4 Statistical Analysis with Epigenetic Models

Epigenetic variation refers to all reversible and heritable changes in the functional genome that do not alter the DNA nucleotide sequence. There are three main mechanisms of epigenetic changes: DNA methylation, histone modifications, and the action of noncoding RNAs. Of these, DNA methylation patterns are the most important.

Methylation affects the construction of the W matrix of incidence of the number of marker alleles. Data related to the probability of methylation in specific portions of DNA are available for genetic analysis in conjunction with phenotypic, pedigree, and genetic markers data. The variation in methylation patterns among individuals contributes to phenotypic variability, even if these individuals are genotypically identical. This bias should be removed from the overall expression of the phenotypic variation decomposition according to the infinitesimal genetic model, to obtain more accurate estimates of genetic values.

The next step in the evolution of genetic methods of genomic association and prediction is the simultaneous incorporation of phenotypes, pedigree, SNPs, indels, and methylation data in statistical estimation methods. In this context, certain important definitions are presented below.

- **Epigenetic inheritance**: transmission of phenotypic variation among generations that does not occur by variation in DNA sequences
- **Epigenetic transmissibility**: probability of transmission of ancestral phenotypes
- **Reversal or reset coefficient** (***v***): probability of changes in epigenetic states during gametogenesis or initial developmental stage
- **Epigenetic transmissibility coefficients(*1 – v*)**: the complement of the reset, return, or reversal coefficient
- **Inducing environment**: environmental signal or stress agent that causes changes to the epigenetic state.

In terms of quantitative genetics, the following terms and equations are important:

(a) **Covariance between relatives for sexual reproduction systems** (Tal et al. 2010)

Phenotypic model (*y*) in the presence of epigenetic variation (σ^2_ξ)

$$y = Xb + Zg + Z\xi + e$$

$\sigma^2_y = \sigma^2_g + \sigma^2_\xi + \sigma^2_e$: total phenotypic variance.

Covariance between relatives with epigenetic variation

$$\mathrm{COV}(P, F) = (1/2)\sigma^2_a + (1/2)(1 - v)\sigma^2_\xi$$

$$\mathrm{COV}(MI) = (1/4)\sigma^2_a + (1/4)(1 - v)^2\sigma^2_\xi$$

$$\mathrm{COV}(TS) = (1/4)\sigma_a^2 + (1/4)(1-v)^3\sigma_\xi^2$$

Epigenetic variation has been found to inflate the genetic covariances between relatives.

Estimators of variance components

$$(1-v) = \frac{2[\mathrm{COV}(MI) - \mathrm{COV}(TS)]}{\mathrm{COV}(P,F) - 2\mathrm{COV}(MI)}$$

$$\sigma_\xi^2 = \frac{2[\mathrm{COV}(P,F) - 2\mathrm{COV}(MI)]}{v\,(1-v)}$$

$$\sigma_g^2 = 2\mathrm{COV}(P,F) - (1-v)\sigma_\xi^2$$

Epigenetic heritability: $h_\xi^2 = \frac{\sigma_\xi^2}{\sigma_y^2}$

The model can be fitted by means of the mixed-model equations: $\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z + A^{-1}\frac{\sigma_e^2}{\sigma_g^2} & Z'Z \\ Z'X & Z'Z & Z'Z + \Lambda^{-1}\frac{\sigma_e^2}{\sigma_\xi^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \\ \hat{\xi} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \end{bmatrix}$, where A is the additive genetic correlation matrix among individuals and Λ is the epigenetic transmissibility matrix $(1-v)$.

The residual maximum likelihood (REML)/best linear unbiased prediction (BLUP) procedure can estimate the variance components:

σ_g^2: additive genetic variance; σ_ξ^2: epigenetic variance; σ_e^2: residual variance; h_ξ^2: epigenetic heritability.

(b) Covariance between relatives for asexual reproduction systems (Tal et al. 2010)

$\sigma_y^2 = \sigma_{gt}^2 + \sigma_\xi^2 + \sigma_e^2$: total phenotypic variance.

Covariance between relatives

$$\mathrm{COV}(P,F) = \sigma_{gt}^2 + (1-v)\sigma_\xi^2$$

$$\mathrm{COV}(RAM) = \sigma_{gt}^2 + (1-v)^2\sigma_\xi^2$$

$$\mathrm{COV}(TSC) = \sigma_{gt}^2 + (1-v)^3\sigma_\xi^2$$

Estimation of variance components

$$(1 - v) = \frac{\mathrm{COV}(RAM) - \mathrm{COV}(TSC)}{\mathrm{COV}(P, F) - \mathrm{COV}(RAM)}$$

$$\sigma^2_{\xi} = \frac{\mathrm{COV}(P, F) - \mathrm{COV}(RAM)}{v(1 - v)}$$

$$\sigma^2_{gt} = \mathrm{COV}(P, F) - (1 - v)\sigma^2_{\xi}$$

σ^2_{gt}: total genotypic variance.

COV(P, F), COV(MI), COV(TS), COV(TSC), COV(RAM): covariances between progeny-father; half-sibs; uncle-nephew; cloned uncle-nephew and between ramets, respectively.

References

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106

Auer PL, Doerge RW (2010) Statistical design an analysis of RNA sequencing data. Genetics 185:405–416

Ayroles JF, Gibson G (2006) Analysis of variance of microarray data. Methods Enzymol 411:214–233

Cui X, Hwang JTG, Qiu J et al (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics 6:59–75

Drost DR, Novaes E, Boaventura-Novaes C, Benedict CI, Brown RS, Yin T, Tuskan GA, Kirst M (2008) A microarray-based genotyping and genetic mapping approach for highly heterozygous outcrossing species enables localization of a large fraction of the unassembled Populus trichocarpa genome sequence. Plant J 58:1054–1067

Fernando RL, Grossman M (1989) Marker-assisted selection using best linear unbiased prediction. Genet Sel Evol 21:467–477

Hardcastle TJ, Kelly K (2010) baySeq: empirical bayesian methods for identifying differential expression in sequence count data. BMC Bioinform 11:422

Hochholdinger F, Sauer M, Dembinsky D, Hoecker N, Muthreich N, Saleem M, Liu Y (2006) Proteomic dissection of plant development. Proteomics 6:4076–4083

Hollung K, Veiseth E, Jia X, Faergestad EM, Hildrum KI (2007) Application of proteomics to understand the molecular mechanisms behind meat quality. Meat Sci 77:97–104

Jansen RC, Nap J (2001) Genetical genomics: the added value from segregation. Trends Genet 17:388–391

Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Comput Biol 7:819–837

Lametsch R (2011) Proteomics in muscle-to-meat conversion. In: 64th annual reciprocal meat conference. American Meat Science Association

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756

Lippolis JD, Reinhardt TA (2008) Centennial paper: proteomics in animal science. J Anim Sci 86:2430–2441

Marioni JC, Mason CE et al (2008) RNA-seq : an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517

Morris JS, Baggerly KA, Gutstein HB, Coombes KR (2010) Statistical contributions to proteomic research. Methods Mol Biol 641:143–166

Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by rna-seq. Nat Methods 5:621–628

Mukhopadhyay ND, Chatterjee S (2007) Causality and pathway search in microarray time series experiment. Bioinformatics 23:442–449

Nascimento M, Safadi T, Fonseca FS, Nascimento ACC (2012) Bayesian model-based clustering of temporal gene expression using autoregressive panel data approach. Bioinformatics 4:1–5

Perez-Enciso M (2007) Emerging tools in quantitative trait loci detection. Acta Agric Scand A —Animal Sci 57(4):202–207

Perez-Enciso M, Toro MA, Tenenhaus M, Gianola D (2003) Combining gene expression and molecular marker information for mapping complex trait genes: a simulation study. Genetics 164:1597–1606

Resende MDV, Silva FF, Resende Jr MFR, Azevedo CF (2014a) Genome-wide association studies (GWAS). In: Borem A, Fritsche-Neto R (Org.) Biotechnology and plant breeding, 1st edn. Elsevier, Dordrecht, pp 83–104

Resende MDV, Silva FF, Resende Jr MFR, Azevedo CF (2014b) Genome-wide selection (GWS). In: Borem A, Fritsche-Neto R (Org.) Biotechnology and plant breeding, 1st edn. Elsevier, Dordrecht, pp 105–134

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11:R25

Robinson MD, Oshilack A, Smyth GK (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11:R25

Robinson MD, Smyth GK (2008) Small sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9:321–332

Rosa GJM, Rocha LB, Furlan LR (2007) Estudos de expressão gênica utilizando-se microarrays: delineamento, análise, e aplicações na pesquisa zootécnica. Revista Brasileria de Zootecnia 36:185–209

Tal O, Kisdi E, Jablonka E (2010) Epigenetic contribution to covariance between relatives. Genetics 184:1037–1050

Tempelman RJ (2005) Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. Vet Immunol Immunopathol 105:175–186

Wang Z, Gerstein M, Snyder M, (2010) RNA-Seq: a revolutionary tool for transcriptomics. Nature 10:57–63

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 8:625–637

Chapter 4
Root Phenomics

Júlio César DoVale and Roberto Fritsche-Neto

Abstract Root systems have several functions that go beyond of plant support. Several metabolic reactions in plant roots are initiated that adjust them to a stress that can be instantaneous or permanent. Thus, many breeders believe that the key to success for obtaining genotypes tolerant to many types of abiotics is situated below the soil surface. Because of this, for decades much efforts have been invested in trying to develop tools that enable to analyze precisely the growth and development of roots under undesirable conditions. In the early 2000s were created some hardwares and softwares that enabled evaluation of several root parameters such as length, volume, surface area, projected area, among others. However, most of them have the disadvantage of destroying the sample to be evaluated. Recently, others methods have been developed which enable large-scale phenotyping, such as computed tomography-based. They are important for breeding programs because they allow evaluation of hundreds of genotypes in an easy and fast way. Moreover, they are not destructive methods and they permit to follow the root development in several phenological phases of the plant and in real- time. Given the above, the aim of this chapter is present the most used methods of root phenotyping for plant breeding. For that, we present some procedures and their computational basis, followed by their advantages and limitations.

J.C. DoVale (✉)
Federal University of Ceará (Universidade Federal do Ceará – UFC),
Fortaleza, Brazil
e-mail: juliodovale@ufc.br

R. Fritsche-Neto
Departamento de Genética, University of São Paulo,
Piracicaba, SP, Brazil
e-mail: roberto.neto@usp.br

R. Fritsche-Neto and A. Borém (eds.), *Phenomics*,
DOI 10.1007/978-3-319-13677-6_4

4.1 Introduction

In addition to providing support, roots determine a plant's ability to absorb water and nutrients present in the soil, synthesize and provide active biomolecules, and identify/signal stresses, among several other functions that are important for plant establishment in a given environment (Hodge et al. 2009). However, due to high soil resistance, detailed studies on root growth and development were lacking for a long time, especially when compared to other plant compartments, such as the leaves and stem.

The way roots develop in soil may have a critical effect on plant growth and, consequently, crop productivity. Since the 1990s, this has spurred innovative tools that have allowed roots to be studied in more detail. Currently, a wide variety of techniques are available for this purpose, ranging from the invasive (i.e., allowing the genotype to be differentiated by destructive sampling) to the noninvasive (i.e., allowing plant growth and development to be followed under desired conditions). These methods are extensively employed in ecology and physiology, especially in plant breeding programs aiming to select genotypes with efficient water and nutrient use or those tolerant to water and nutrient scarcity (Chun et al. 2005; Fritsche-Neto et al. 2012).

When plants are grown in an environment with a nutritional or water deficit, they usually exhibit greater carbohydrate allocation to the root system (Nielsen et al. 2001). With this modification, there is increased root length and density but reduced root diameter, which allows roots to have more contact with the soil (Ma et al. 2001; López-Bucio et al. 2002). These adaptations also include obtaining nutrients or water with minimal carbon cost, and this is only possible due to increased root growth and changes in branching pattern, total root length, root hair elongation, lateral root formation, and root architecture (Lynch and Brown 2001; Fan et al. 2003). In a quick survey of papers published and indexed in the Web of Science, it appears that most studies addressing the root system, especially those related to stress, have focused on root architecture. Root architecture refers to the spatial configuration of the root system—that is, the geometric arrangement of the root axes within the soil portion.

With the development of modern methods, it has been possible to initiate studies associating root architecture with other root attributes in young plants, thereby providing a basis for rapid phenotypic characterization (Singh et al. 2010). In this context, in addition to better understanding root system development and the mechanisms of tolerance to certain stresses, these modern methods allow breeders to undergo early and efficient selection among the thousands of genotypes that arise during every cycle in breeding programs. In this sense, this chapter aims to address the main methods for large-scale root phenotyping, their applications in plant breeding, and prospects for the future.

4.2 Large-Scale Root Phenotyping Methods

Due to technological advances and rapid dissemination of information, much work has been performed to automate plant phenotyping. In this context, several platforms and computer programs for collecting and analyzing root images have been developed in recent years. These platforms are used to accurately characterize root systems regarding both quantitative and qualitative aspects.

Roots are notoriously difficult to phenotype under field conditions. In addition to technical considerations, characterization under these conditions is limited by genotype–environment interactions, which are usually significant (Gregory et al. 2009). Traditional methods employed to study roots have emphasized root excavation techniques in which root system length and density may be determined (Araus and Cairns 2014). However, the excavation process is laborious and slow. Because of this, an Australian industrial research organization recently implemented a high-yield soil sampling system. This system consists of a hydraulic press that compresses up to 200 cm soil depth per day. Currently, this system is used to evaluate the effect of root architecture on water absorption, and especially to characterize more drought-tolerant genotypes (Gregory et al. 2009).

Some noninvasive techniques, such as those based on electrical capacitance, and other more innovative ones, such as magnetic resonance and three-dimensional (3D) computed tomography, have been proposed mainly for annual (herbaceous) crops. Trachsel et al. (2011) proposed a less costly and rapidly executed method designed a priori for grasses. This method was named *Shovelomics*, and it assigns grades (scores) to root architectural traits by visually inspecting the roots of individuals. In maize, for example, the numbers, angles, and patterns of nodulated and adventitious root branches are considered. Thus, at the end of the procedure, each individual has a final score that allows for classifying it as adapted or nonadapted to a marginal condition of cultivation. However, there are still problems related to limited resolution when working with tree species (Wasson et al. 2012).

Typically, these phenotyping platforms are divided into two main groups: ex situ analysis-based (using samples or the entire root system outside of the growth environment) and in situ analysis-based, which are also named noninvasive (evaluating the entire root system and in situ). Next, some of the most widely used methods in plant breeding programs for phenotyping roots of genotypes are presented.

4.2.1 Ex Situ Evaluations

In many situations, it is only possible to observe the effect of a given phenomenon on crop growth and developmental dynamics under controlled conditions (i.e., artificial environments). In fact, this requires avoiding the action of other factors not considered in the experiment, and it facilitates visualizing and capturing images.

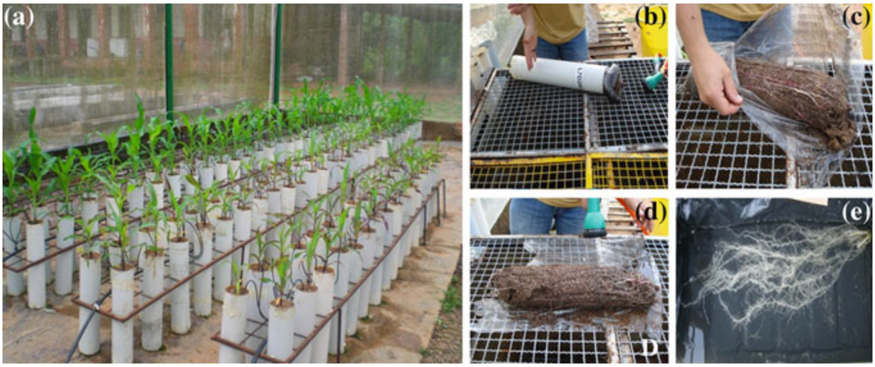


Fig. 4.1 Planting of maize strains in PVC pipes wrapped in plastic bags under high and low nitrogen availability (**a**); root washing process (**b**, **c**, and **d**); and the root system obtained using this technique (**e**)

The most commonly used techniques for this purpose are hydroponics, aeroponics, culture medium in agar, pots, and even polyvinyl chloride (PVC) pipes. Alternatively, rhizotrons and minirhizotrons have been employed, which allow roots to be studied while still within soil. However, they artificially restrict the direction of root system growth to two dimensions. Moreover, they do not allow for phenotyping a large number of individuals, which slows the selection step in breeding programs.

Planting in pots or PVC pipes (Fig. 4.1a) are methods that require washing the root system, which often leads to underestimating fine roots due to breakage during the washing process. To minimize errors from losing these roots, it is recommended that the containers used in the experiment (pots or pipes) be wrapped in a plastic bag (Fig. 4.1b). However, in addition to being a very laborious process, the spatial configuration of the roots can be lost and the inferences about root architecture can be limited (Mairhofer et al. 2013). However, there are methods that allow for phenotyping a reasonable number of individuals, and they are currently the most employed.

4.2.2 Scanning or Digital Scanning

Scanning combined with computerized image analysis is a fast method of evaluating root morphological patterns, such as length, diameter, topology, and branching. Computerized scanning complements manual estimates and those obtained using cameras. Scanning can be performed on small root samples or whole root systems obtained from hydroponic crops. Digital output from an image is stored on a computer as a TIFF file and then analyzed with the appropriate software. However, accurate image scanning and analysis depends not only on the software used, but also on the sample preparation and the scanning protocol (Polomsky and Kuhn 2002).

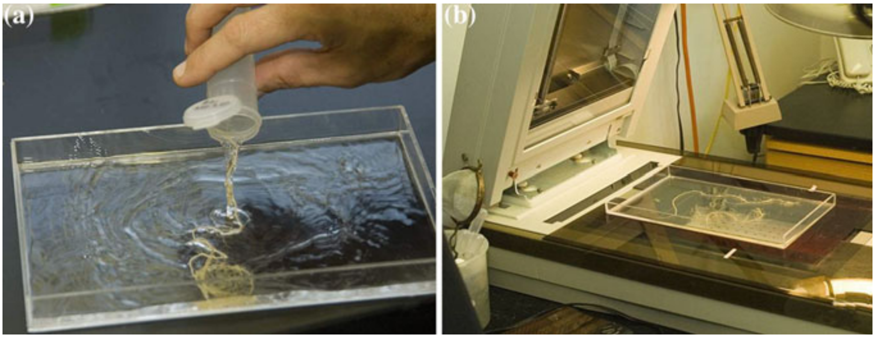


Fig. 4.2 Root samples placed in a transparent acrylic box with a predefined volume of water (**a**). The box is placed on the scanner table to start the scanning process (**b**)

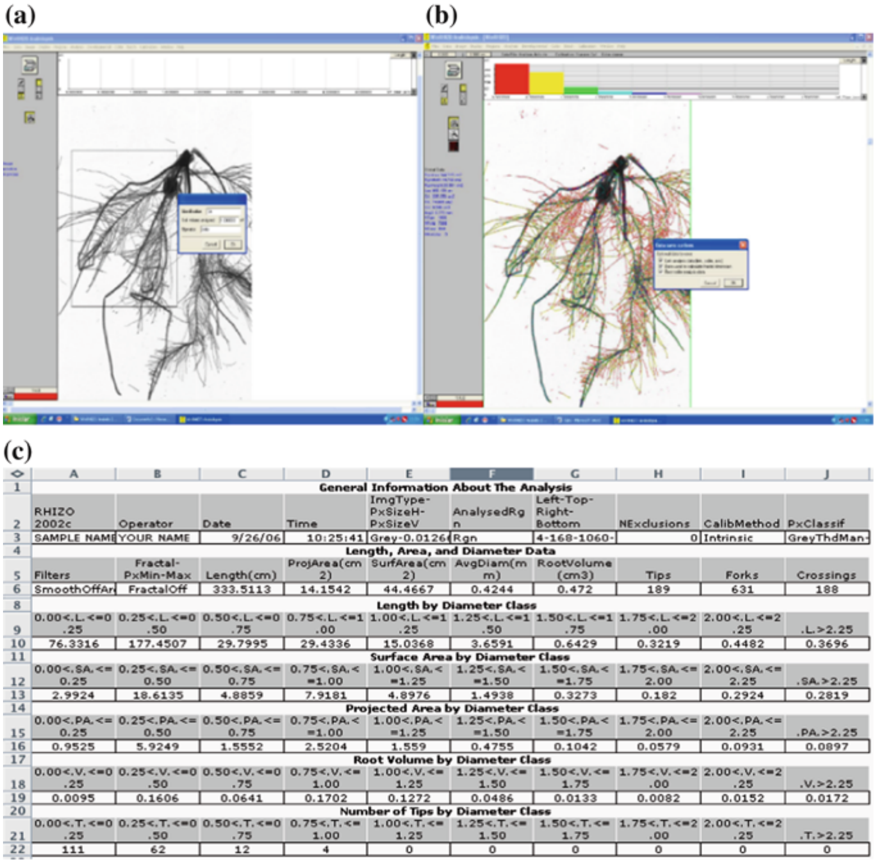
4.2.2.1 Sample Preparation

Whole root systems or samples with root segments are washed to remove soil particles from their surface. Next, they are placed and spread onto an acrylic box (transparent) with a predefined volume of water (usually covered with 2–5 mm of water; Fig. 4.2). The box is then placed on the table of the scanner and the roots are scanned. Samples with large volumes of roots should be divided into subsamples to minimize overlay, which is one of the main sources of error in such estimates (Bouma et al. 2000).

4.2.2.2 Scanning Protocol

Scanning resolution and initial transformations are important parameters that should be detailed in the study methods because they allow for possible comparisons of the results (Polomsky and Kuhn 2002). Software such as WinRHIZO and Delta T-Scan recommend a resolution of 400 dpi (Bouma et al. 2000).

The original images, which are obtained in grayscale in most of the scanning procedures, are transformed into binary versions (black and white). The highest pixel values in grayscale from the initial procedure are considered in only a portion of the image and are defined as black (value of 1). Conversely, the lowest pixel values of the gray scale represent the background of the image and are defined as white (value of zero; Polomsky and Kuhn 2002). According to these authors, the subsequent step is the skeletonization process of the axial roots (larger diameter), which is obtained by repeatedly removing pixels from the edge of the image until only a single chain of pixels represents a line in the center of the sample.



	A	B	C	D	E	F	G	H	I	J
1	General Information About The Analysis									
2	RHIZO 2002c	Operator	Date	Time	ImgType-PxSizeH-PxSizeV	AnalysedRgn	Left-Top-Right-Bottom	NExclusions	CalibMethod	PxClassif
3	SAMPLE NAMI	YOUR NAME	9/26/06	10:25:41	Grey-0.0126	Rgn	4-168-1060-	0	Intrinsic	GreyThdMan-
4	Length, Area, and Diameter Data									
5	Filters	Fractal-PxMin-Max	Length(cm)	ProjArea(cm2)	SurfArea(cm2)	AvgDiam(mm)	RootVolume (cm3)	Tips	Forks	Crossings
6	SmoothOffAr	FractalOff	333.5113	14.1542	44.4667	0.4244	0.472	189	631	188
8	Length by Diameter Class									
9	0.00<.L.<=0.25	0.25<.L.<=0.50	0.50<.L.<=0.75	0.75<.L.<=1.00	1.00<.L.<=1.25	1.25<.L.<=1.50	1.50<.L.<=1.75	1.75<.L.<=2.00	2.00<.L.<=2.25	.L.>2.25
10	76.3316	177.4507	29.7995	29.4336	15.0368	3.6591	0.6429	0.3219	0.4482	0.3696
11	Surface Area by Diameter Class									
12	0.00<.SA.<= 0.25	0.25<.SA.<= 0.50	0.50<.SA.<= 0.75	0.75<.SA.<=1.00	1.00<.SA.<=1.25	1.25<.SA.<=1.50	1.50<.SA.<=1.75	1.75<.SA.<= 2.00	2.00<.SA.<= 2.25	.SA.>2.25
13	2.9924	18.6135	4.8859	7.9181	4.8976	1.4938	0.3273	0.182	0.2924	0.2819
14	Projected Area by Diameter Class									
15	0.00<.PA.<= 0.25	0.25<.PA.<= 0.50	0.50<.PA.<= 0.75	0.75<.PA.<=1.00	1.00<.PA.<=1.25	1.25<.PA.<=1.50	1.50<.PA.<=1.75	1.75<.PA.<= 2.00	2.00<.PA.<= 2.25	.PA.>2.25
16	0.9525	5.9249	1.5552	2.5204	1.559	0.4755	0.1042	0.0579	0.0931	0.0897
17	Root Volume by Diameter Class									
18	0.00<.V.<=0.25	0.25<.V.<=0.50	0.50<.V.<=0.75	0.75<.V.<= 1.00	1.00<.V.<= 1.25	1.25<.V.<= 1.50	1.50<.V.<= 1.75	1.75<.V.<=2.00	2.00<.V.<=2.25	.V.>2.25
19	0.0095	0.1606	0.0641	0.1702	0.1272	0.0486	0.0133	0.0082	0.0152	0.0172
20	Number of Tips by Diameter Class									
21	0.00<.T.<=0.25	0.25<.T.<=0.50	0.50<.T.<=0.75	0.75<.T.<= 1.00	1.00<.T.<= 1.25	1.25<.T.<= 1.50	1.50<.T.<= 1.75	1.75<.T.<=2.00	2.00<.T.<=2.25	.T.>2.25
22	111	62	12	4	0	0	0	0	0	0

Fig. 4.3 a Acquisition (scanning) of an image of a maize sample using an EPSON Expression 10000 XL scanner equipped with an additional light (TPU). **b** Image analyzed using WinRHIZO Pro 2009c software. **c** Spreadsheet with data from each sample generated using this software. *Source* Basic, Reg, Pro & Arabidopsis for Root Measurement

4.2.2.3 Evaluations Using WinRHIZO Software

WinRHIZO software allows for more flexible and automatic selection of variation generated using the initial image capturing procedure. Estimating root diameter combined with different colors and configurations makes this procedure very precise (Fig. 4.3a, b).

The measurements involve total root length, mean root diameter, root projection and surface areas, root volume, and number of root types as a function of ten diameter classes. These classes vary from roots with diameter smaller than or equal to 0.5 mm up to roots with diameter larger than or equal to 4.5 mm. All of the information can be saved to an XLS file and then worked on in Excel (Fig. 4.3c).

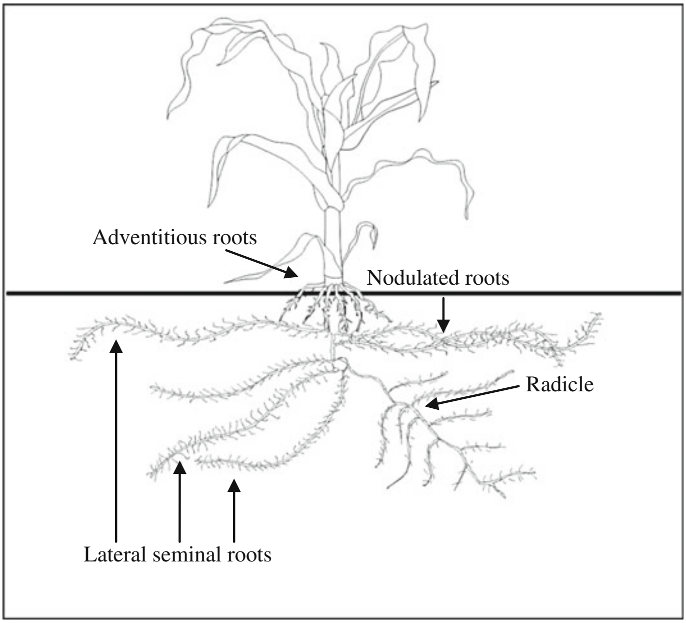


Fig. 4.4 Shoot and root system of a maize plant with six fully expanded leaves (V6)

Additionally, the software detects portions of overlapping roots and considers them in estimating the root parameters (Himmelbauer et al. 2004).

Aiming to facilitate the process of characterizing genotypes in plant breeding programs for abiotic stress conditions, some authors have suggested simplifying the ten diameter classes in maize provided by WinRHIZO into only two. Thus, fragments with diameter smaller than or equal to 0.5 mm are considered for parameters related to lateral roots and fragments with diameter larger than 0.5 mm are considered for parameters related to axial roots (Hund et al. 2009; Trachsel et al. 2009).

One of the goals of the study conducted by DoVale and Fritsche-Neto (2013) was to determine the role of the root system in efficient use of phosphorus in maize. For this, experimental hybrids were used at V6 stage—that is, with six fully expanded leaves (Fig. 4.4). The root systems of all the individuals were simplified as mentioned above. These authors found significant positive correlation coefficients ($p < 0.01$) between axial roots and phosphorus absorption efficiency both under high and low phosphorus availability conditions. This study allowed for the conclusion that this simplification is valid in the process of identifying genotypes with more efficient phosphorus use.

Even though root scanning allows for a large number of genotypes to be phenotyped in a breeding program, it usually has low yield. This is because one sample is evaluated at a time. Due to the slowness of this procedure, techniques that allow for high-yield phenotyping have been developed.

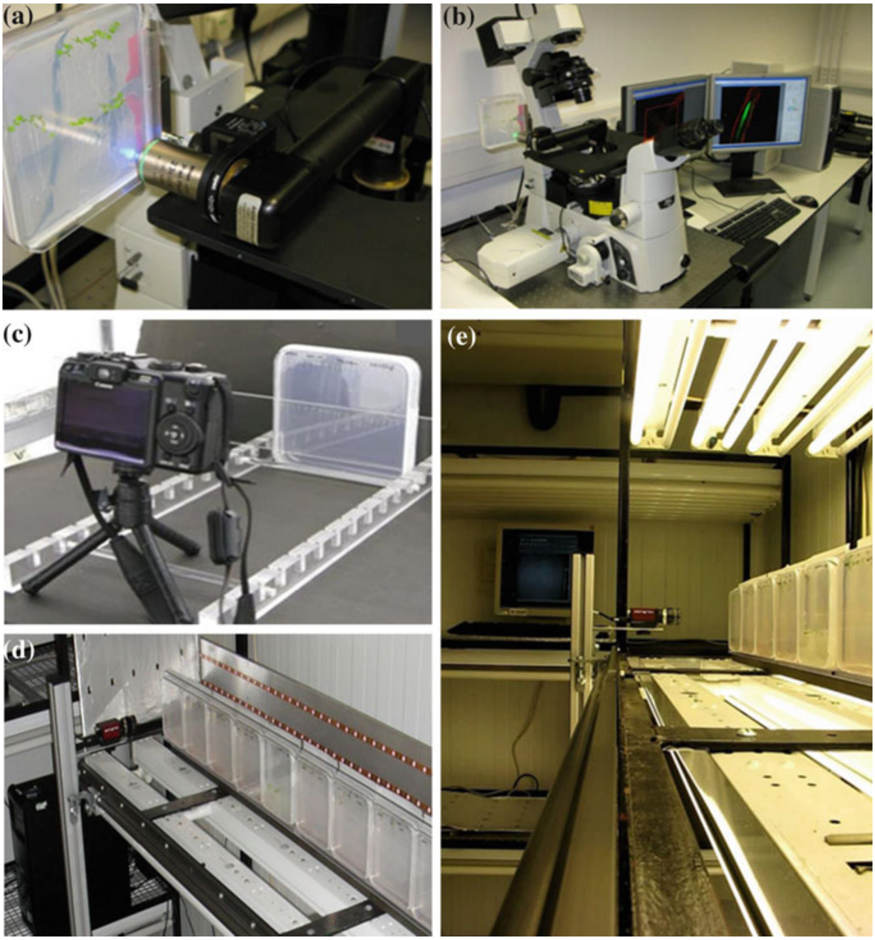


Fig. 4.5 Microscope for acquiring images in vertical plates (**a** and **b**), digital camera (**c**), and hardware for automated capture of root images (**d** and **e**). *Source* French et al. (2012)

4.2.2.4 Other Hardware and Software

In addition to scanners, images of roots can be captured using other devices that have higher yield. Researchers from the Center for Plant Integrative Biology, University of Nottingham, United Kingdom currently acquire information for their studies using microscopes with vertical plates (Fig. 4.5a, b), digital cameras (Fig. 4.5c), and hardware for acquiring automated images (Fig. 4.5d, e). The latter are able to phenotype up to 500 genotypes at a time at a maximum speed of 60 mm/s with an accuracy of approximately 187 μm (French et al. 2012). The major limitation of these devices is that evaluation must occur at a very early stage of development (seedling stage). However, they appear to be useful enough for characterizing genotypes tolerant to the presence of aluminum in soil.

Images captured using these devices are stored in an image database and can then be evaluated using different software, similarly to WinRHIZO. RootTrace version 1 (RT1), RootTrace version 2 (RT2), and RootNav are some examples of software compatible with the automated devices most commonly used by research groups that conduct studies involving root systems.

The RootTrace application analyzes the image from top to bottom based on a starting point predefined by the user. Thus, it is possible to follow plant growth and monitor the changes that occur in the root system using the growth rate of primary roots, angulation, and branching, among other parameters. RT1 considers root growth in the direction of gravity. The model developed for following these growth dynamics moves one pixel (~ 0.05 mm) every step, reflecting the effect of gravity on root growth. However, this monitoring is only possible if the roots exhibit curvature less than or equal to 90° at the root tip (Naeem et al. 2011).

The user can adjust the model with a type of multidirectional bar. When this bar is configured further to the left, the gravity-dependent RT1 model is employed. In contrast, when the bar is moved to the right, more or less points are fitted to predict the model (Fig. 4.6a, b). This procedure allows for more reliable monitoring of root curvature (Fig. 4.6g–j). RT2 users are able to calculate (predict) a monitoring model to analyze the data referring to root growth without needing to input a numerical parameter (Naeem et al. 2011).

The RootNav application is another new tool that allows root system architecture to be quantified for a range of crop species. An automatic component of this software is also based on a top-down approach, and it uses a powerful algorithm of maximum classification to analyze regions of the input image and then calculate the probability that certain pixels correspond to roots (Pound et al. 2013). According to these authors, this information is used as the basis for optimized approximation in detecting and quantifying roots.

Thus, like RootTrace, RootNav makes an optimized estimate from the seed to the root apices (Fig. 4.7). However, it also allows the user to easily and intuitively refine the results by visual inspection. Moreover, it provides supporting information necessary for extracting a variety of biologically relevant measurements. This is because there is a separate viewer tool in the center of the application that allows a rich set of traits related to root architecture to be retrieved from the original image.

4.2.3 In Situ or Nondestructive Evaluation

When growing plants in pots, the roots quickly fill the container, bending, distorting, and consequently substantially modifying their growth and development compared to what would usually be observed in the field. Therefore, to observe a more realistic root distribution, pots with volume much greater than the estimated volume of the roots should be used for each plant (which usually makes the study

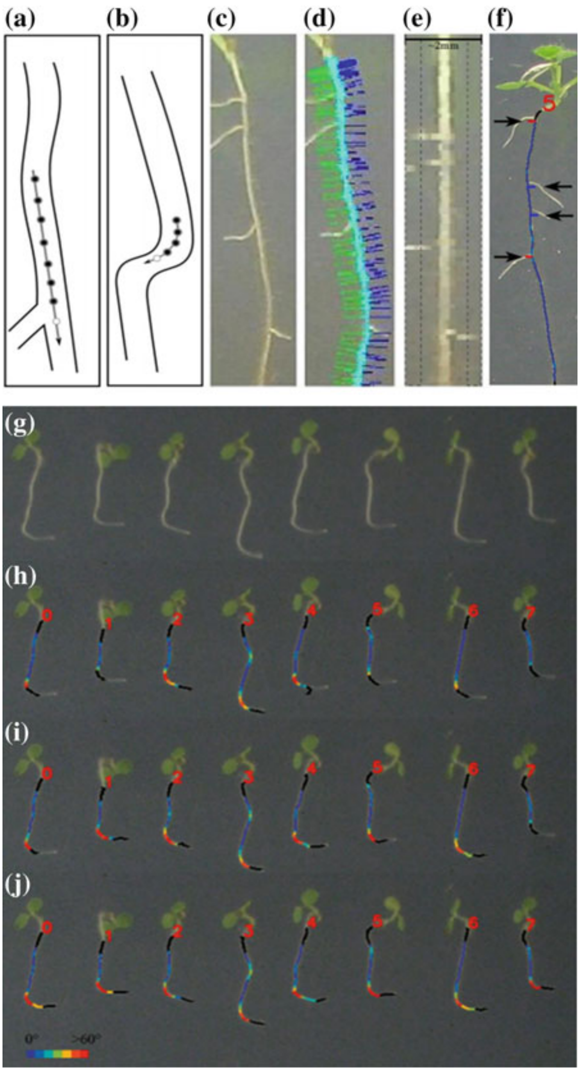


Fig. 4.6 Models for monitoring growth in roots with low curvature (**a**) and high curvature (**b**) obtained using RootTrace. (**c**) Original root image. (**d**) 40 pixels based on growth angles. (**e**) Area selected to detect the marked lateral regions. (**f**) Lateral detection. (**g**) Original image showing a pronounced gravitropic response. (**h**–**j**) Results of applying monitoring models using the multidirectional bar; colors indicate root curvature intensity (*red* for high and *blue* for low). *Source* Naeem et al. (2011)

infeasible), or the plants should be evaluated in their natural environment (in the field). Under these conditions, roots can grow deeper without their spatial distribution being affected. However, evaluation in these “open systems” has drawbacks, such as the sheer volume of the soil to be analyzed and the difficulty of sampling roots for further analysis.

The analysis of root images using samples or even whole roots is an arduous task. Additionally, as aforementioned, in the process of obtaining the samples, significant information regarding root system distribution in the soil can be lost. In this context, many researchers have sought to develop faster and more reliable methods for analyzing root images. Among these, the following two are the most innovative.

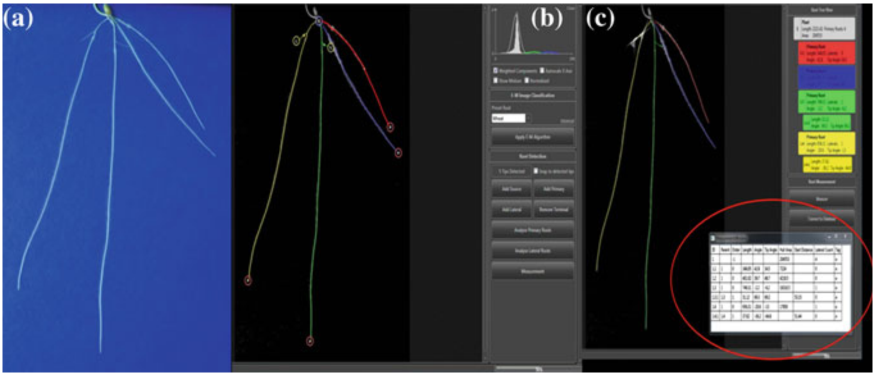
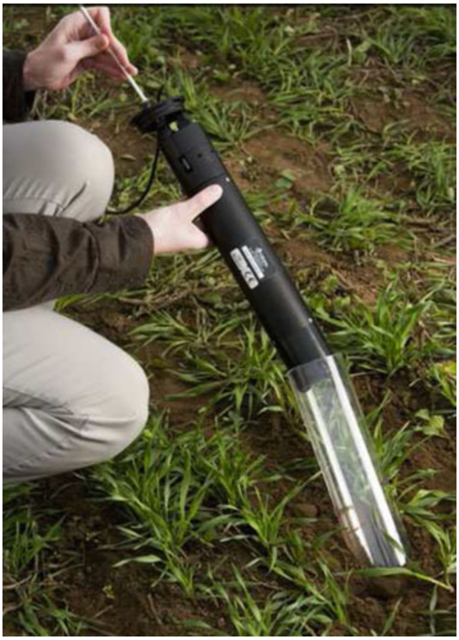


Fig. 4.7 **a** Original image of the wheat root system. **b** Partial analysis using RootNav software with the option of redirecting the analysis to selected areas. **c** Output of the application with data stored in spreadsheets. *Source* French et al. (2012)

4.2.3.1 “CI-600 RootSnap Scanner and Software” System

This system is used to analyze root growth, development, and function in adapting to a given environment. In this, the CI-600 scanner nondestructively captures high-resolution digital images (Fig. 4.8).

Fig. 4.8 CI-600 root scanner and acrylic tubes used in capturing the images. *Source* CID Bio-Science (2014) http://www.cid-inc.com/

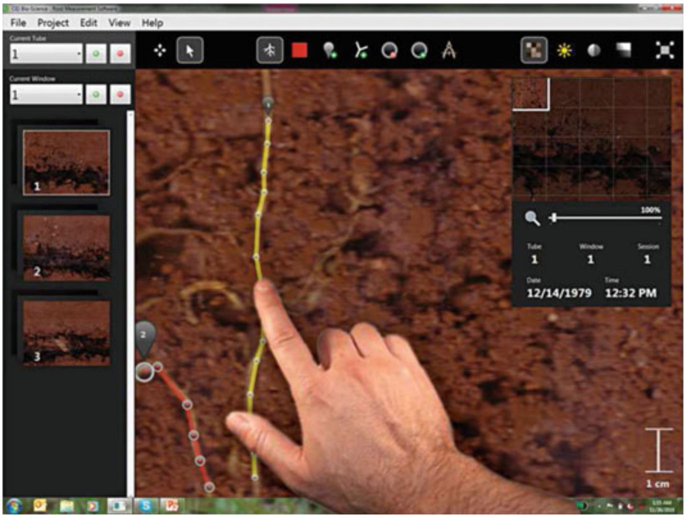


This phenotyping system is designed for long-term studies on living plants in the field where each plant can be evaluated several times during their growth cycle. For this, before or during planting, acrylic tubes are installed within the study area (in the plots). When the plants begin to construct their root “networks” around the tube, images of the structure and behavior of the roots can be obtained with the scanner and analyzed using CI-690 RootSnap software.

To evaluate images of the roots in the plots, it is necessary to insert the CI-600 reading device into a transparent acrylic tube preinstalled underground and start the scanning program on a computer (Fig. 4.8). The reading device automatically rotates approximately 360°, creating images of the soil and roots of approximately 21.59 × 19.56 cm, in color, and in high resolution (188 million pixels). Regarding the reading depth, it is possible to easily move the device to different depths, and from tube to tube, choosing an ideal image according to the goal of the study and the species.

The equipment is extremely portable (750 g) and fast handling (5–15 s per reading depending on the resolution). Additionally, it allows for viewing root growth and behavior during an entire growth season or for even longer periods.

To interpret the images and store them, the equipment has a USB interface that allows for connection to mobile devices, such as tablets and laptops. However, it is necessary to have software that processes the images and estimates the phenotypic values of the individuals such as length, volume, and surface area of the roots. RootSnap is such a root image analysis package. When installed on equipment with a multi-touch LCD screen, the software allows users to quickly and easily track roots using their fingers (Fig. 4.9).



Current Tube: 1

Current Window: 1

Tube 1 Window 1 Session 1

Date 12/14/1979 Time 12:32 PM

Fig. 4.9 Output of the RootSnap software, an automated root image analyzer. *Source* CID Bio-Science (2014) http://www.cid-inc.com/

The software also automatically overlays different tracing points on the root system. Additionally, the files are stored in the common open XML format and data export to applications such as Excel, WinRHIZO, RootTrace, and RootNav is supported.

4.2.3.2 X-ray Computed Tomography for Obtaining Noninvasive 3D Images

Methods based on this technology are well described in the review by Mooney et al. (2012). Usually, such methods seek to observe the roots in their natural state in the soil, both in space and time, maintaining their complex 3D morphology throughout their growth and development (four-dimensional, 4D). Several energy sources may be used to generate tomographic images. The X-ray technique is the most widely adopted because it is noninvasive and allows viewing inside objects in 2D or 3D based on the principle of attenuation of electromagnetic waves.

In this context, medical scanners have been used the most to investigate macroscopic characteristics of roots (Heeraman et al. 1997). These scanners are advantageous because several images can be easily obtained in a relatively short period of time. However, their resolution is usually limited to a slice thickness of 0.5 mm. Thus, if the goal of the study is to analyze fine roots, industrial X-ray devices are necessary such as synchrotron scanners or others specific for this purpose. There are already advanced systems of this type for animals, such as in vivo *X-treme* used for analyses in mice, and they can be adapted to plants. This system captures images in 3D and with high sensitivity for luminescence, fluorescence, X-rays, and radioisotopes (Bruker 2014; http://www.bruker.com/).

Although many studies have successfully visualized roots in situ, few have been able to extract the volumetric descriptions of material necessary to produce 3D models of their architecture. In this sense, two automated root tracing approaches were recently proposed, one based on assigning probability functions named RootViz (Tracy et al. 2012; www.rootviz3d.org) and another based on level set methods named Rootrak (Mairhofer et al. 2011).

The first assigns a probability function to determine which specific pixels of an image represent root material and if they can be used to provide a 3D view of root distribution in the soil. Kaestner et al. (2006) used this technique and successfully characterized the root architecture of speckled alder (*Alnus incana*) (Fig. 4.10).

Another example of RootViz application was presented by Tracy et al. (2012) in wheat (*Triticum aestivum*), where it was possible to view the root architecture at the initial growth phase (Fig. 4.11).

The second approach processes a set of "slices" of 2D images to construct a gray scale related to the known root structure. These initial grayscale values allow a simple model to be constructed. Thus, a search for connectivity between the images obtained is initiated to construct overall and 3D images of the root system. This technique detects both thick roots and fine roots that grow vertically. However, any disconnected roots or those that grow in irregular directions are not recorded.

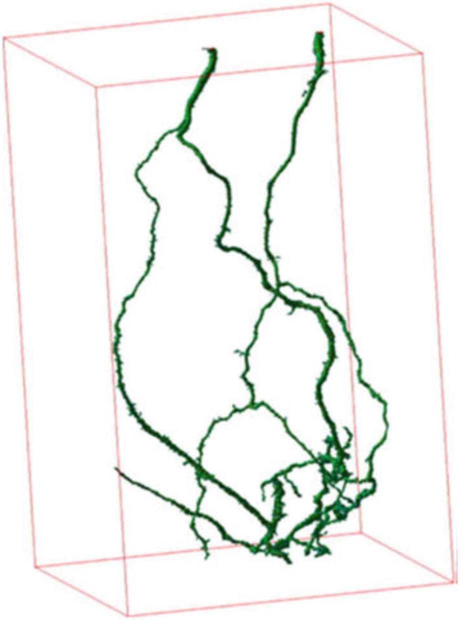


Fig. 4.10 A 3D view of the *Alnus incana* root system, with resolution of 36 µm and sample spatial dimensions of 36.9 × 36.9 × 59.15 mm^3. *Source* Kaestner et al. (2006)

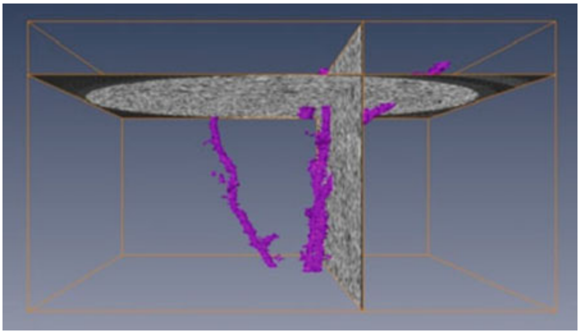


Fig. 4.11 A 3D view of roots of wheat grown in sandy soil with resolution of 18 µm, obtained using RootViz. Sample dimensions = 91 mm high × 29 mm wide. *Source* Mooney et al. (2012)

A risk in applying this type of segmentation is its imprecision, which may introduce attenuation values into the appearance model that do not derive from the root. Subsequent segmentations can thus undergo greater imprecision, leading to higher distortion of the model. To avoid this situation, the shapes of root sections extracted from adjacent images are compared, and if they differ significantly then

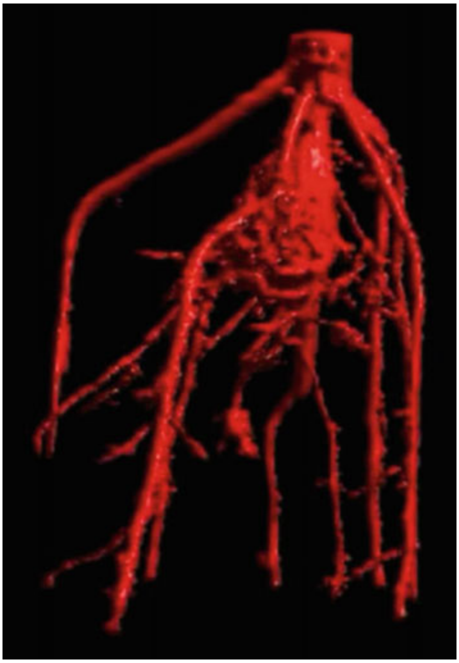


Fig. 4.12 A 3D view of maize roots grown in sandy soil with 30-µm resolution obtained using Rootrak, with sample dimensions of 50 × 120 mm. *Source* Mooney et al. (2012)

the model is discarded. Despite some limitations, the method has been successfully applied in tomography of maize (Fig. 4.12), wheat, and tomato grown in a variety of contrasting soil textures.

The use of relatively small samples (e.g., 25 mm wide) and higher resolution (e.g., voxel size of 100 mm) has been suggested for obtaining images, aiming to ensure that the fine roots can be accurately viewed (Jenneson et al. 2003). When the goal is to analyze thick roots (or main), larger samples (e.g., 150 mm wide; 500 mm tall) and relatively low resolution (e.g., >1 mm) may be used (Johnson et al. 2004).

Image quality is also strongly affected by the type of container for the samples. In this sense, thinner (<3 mm) pots made of low-density plastic material are preferred compared to metal cylinders (Lontoc-Roy et al. 2006).

The soil moisture content of the sample is another key issue. Soil moisture conditions below field capacity produce better quality images than those obtained from soil closer to saturation. This is most likely due to the moisture content in the roots (Mooney et al. 2012).

Other points that still deserve more studies are related to obtaining 4D images—that is, repeated images of the same plant over time to evaluate root growth and development. Repeated exposure to X-rays potentially has deleterious effects on the plant, which can lead to errors in obtaining and interpreting results. Additionally, researchers should be aware of the position of the sample inside the scanner, always seeking to put it in the same position as for the previous readings, because this prevents errors in reading the dimensional axes (Mooney et al. 2012).

4.2.3.3 3D Views of Roots Using Other Nondestructive Methods

A series of other imaging techniques has been developed to view and quantify root properties in situ, such as nuclear magnetic resonance (Jennette et al. 2001), magnetic resonance (Pohlmeier et al. 2008), thermal neutron tomography (Tumlinson et al. 2008), and neutron radiography (Carminati et al. 2010). Similarly to X-ray tomography, each of these approaches has a series of advantages and some limitations when used to view root system architecture directly in soil.

Magnetic resonance can be used alone or together with other techniques to view root morphology, volume, and length. However, this technique is particularly sensitive to the moisture content of the samples. Additionally, the use of this technique is limited to studies in soil of root diameters greater than 1 mm due to the presence of paramagnetic ions, such as Cu^{2+}, Fe^{2+}, Fe^{3+}, and Mn^{2+}.

This technique has been combined with positron emission tomography to quantify carbon allocation and storage as sugars in beet and maize. The other aforementioned techniques, and even X-ray tomography, when combined with magnetic resonance allow for performing other types of studies, such as identifying the water status of roots throughout the growth cycle. However, in contrast to computed X-ray tomography, for which benchtop systems have become extensively available, limited access to magnetic resonance facilities limits its use. Additionally, there are other significant disadvantages compared to X-ray techniques regarding most soil containing iron and/or manganese ions in large quantities, which negatively affects image quality (Heeraman et al. 1997).

4.3 Prospects

Despite major advances in root phenotyping, there are still large drawbacks to be resolved, especially related to the quality of data collection, adequate image resolution, and accurate analysis.

Regarding collection, it is noteworthy that the medium (substrate) and the cultivation container (pot or natural soil) have significant effects on image quality obtained and on root growth and development. This may lead to serious experimental errors and, consequently, misleading results and conclusions. In this sense, it is still necessary to improve and standardize protocols for conducting and evaluating experiments with this goal so that the real growth conditions reliably represent the site for which the new genotypes to be developed will be recommended. Additionally, it is necessary to improve the image capturing equipment to maximize reproducibility and minimize interference of the medium in image quality and resolution.

Regarding aspects after obtaining the images, it is necessary to develop statistical-mathematical algorithms and models that better describe the 4D structure of roots and transform the image data into quantitative variables that can be analyzed as such. This will facilitate the use not only of simple variables such as length, diameter, and volume, but also of routinely employing complex traits, such as angles between roots, growth rate, and their spatial distribution as a function of changes in soil factors.

Finally, the study of roots is a relatively new subject among geneticists and breeders. However, the results already observed and the equipment and techniques developed (or under development) give this field a very exciting outlook.

References

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19:52–61

Bouma TJ, Nielsen KL, Koutstaal B (2000) Sample preparation and scanning protocol for computerised analysis of root length and diameter. Plant Soil 218:185–196

Bruker (2014) http://www.bruker.com/

Carminati A, Moradi AB, Vetterlein D, Vontobel P, Lehmann E, Weller U, Vogel HJ, Oswald SE (2010) Dynamics of soil water content in the rhizosphere. Plant Soil 332:163–176

Chun L, Mi G, Li J, Chen F, Zhang F (2005) Genetic analysis of maize root characteristics in response to low nitrogen stress. Plant Soil 276:369–382

CID Bio-Science (2014) http://www.cid-inc.com/

DoVale JC, Fritsche-Neto R (2013) Genetic control of traits associated with phosphorus use efficiency in maize by REML/BLUP. Revista Ciência Agronômica 44:554–563

Fan M, Zhu J, Richards C, Brown KM, Lynch JP (2003) Physiological roles aerenchyma in phosphorus-stressed roots. Funct Plant Biol 30:493–506

French A, Wells D, Everitt N, Pridmore T (2012) High-throughput quantification of root growth. In: Macuso S (ed) Measuring roots: an updated approach. Springer, Heidelberg, 382 pp

Fritsche-Neto R, DoVale JC, Lanes ECM, Resende MDV, Miranda GV (2012) Genome-wide selection for tropical maize root traits under conditions of nitrogen and phosphorus stress. Acta Scientiarum Agron 34:389–395

Gregory PJ, Bengough AG, Grinev D, Schmidt S, Thomas WTB, Wojciechowski T, Young IM (2009) Roots phenomics of crops: opportunities and challenges. Funct Plant Biol 36:922–929

Heeraman DA, Hopkins JW, Clausnitzer V (1997) Three dimensional imaging of plant roots in situ with X-ray computed tomography. Plant Soil 189:167–179

Himmelbauer ML, Loiskandll W, Kastanek F (2004) Estimating length, average diameter, and surface area of roots using two different image analysis systems. Plant Soil 260:111–120

Hodge A, Berta G, Doussan C, Merchan F, Crespi M (2009) Plant root growth, architecture and function. Plant Soil 321:153–187

Hund A, Trachsel S, Stamp P (2009) Growth of axile and lateral roots of maize: I development of a phenotying platform. Plant Soil 325:335–349

Jenneson PM, Gilboy WB, Morton EJ, Gregory PJ (2003) Na X-ray micro-tomography system optimized for the low dose study of living organisms. App Rad Isotopes 58:177–181

Jennette MW, Rufty JR. TW, MacFall JS (2001) Visualization of soybean root morphology using magnetic resonance imaging. In: Zhou X, Luo X (eds) Advances in non-destructive measurement and 3D visualization methods for plant root based on machine vision. Key Laboratory of Key Technology on Agricultural Machine and Equipment, Ministry of Education, South China Agricultural University

Johnson SN, Read DB, Gregory PJ (2004) Tracking larval insect movement within soil using high resolution X-ray microtomography. Ecol Entom 29:117–122

Kaestner A, Schneebeli M, Graf F (2006) Visualising threedimensional root networks using computed tomography. Geoderma 136:459–469

Lontoc-Roy M, Dutilleul P, Prasher SO, Han L, Brouillet T, Smith DL (2006) Advances in the acquisition and analysis of CT scan data to isolate a crop root system from the soil medium and quantify root system complexity in 3-D space. Geoderma 137:231–241

López-Bucio JL, Hernandéz-Abreu E, Sánchez-Calderón L, Nieto-Jacobo MF, Simpson J, Herrera-Estrella L (2002) Phosphate availability alters architecture and causes changes in hormone sensitivity in the Arabidopsis root system. Plant Physiol 129:244–252

Lynch JP, Brown KM (2001) Topsoil foraging: an architectural adaptation to low phosphorus availability. Plant Soil 237:225–237

Ma Z, Bielenberger DF, Brown KM, Lynch JP (2001) Regulation of root hair density by phosphorus availability in Arabidopsis thaliana. Plant Cell Environ 24:459–467

Mairhofer S, Zappala S, Tracy SR, Sturrock C, Bennett MJ, Mooney SJ, Pridmore TP (2011) Automated recovery of 3D plant root architecture in soil from X-ray micro computed tomography using object tracking. Plant Physiol 158:561–569

Mairhofer S, Zappala S, Tracy S, Sturrock C, Bennett MJ, Mooney SJ, Pridmore TP (2013) Recovering complete plant root system architectures from soil via X-ray μ-computed tomography. Plant Methods 9:1–7

Mooney SJ, Pridmore TP, Helliwell J, Bennett MJ (2012) Developing X-ray computed tomography to non-invasively image 3-D root systems architecture in soil. Plant Soil 352:1–22

Naeem A, French AP, Darren MW, Pridmore TP (2011) High-throughput feature counting and measurement of roots. Bioinform Appl Note 27:1337–1338

Nielsen KL, Eshel A, Lynch JP (2001) The effect of P availability on the carbon economy of contrasting common bean (*Phaseolus vulgaris* L.) genotypes. J Exp Bot 52:329–339

Pohlmeier A, Oros-Peusquens A, Javaux M, Menzel MI, Vanderborght J, Kaffanke J, Romanzetti S, Lindenmair J, Vereecken H, Shah NJ (2008) Changes in soil water content resulting from Ricinus root uptake monitored by magnetic resonance imaging. Vad Zone J 7:1010–1017

Polomsky J, Kuhn N (2002) Root research methods. In: Waisel Y, Eshel A, Kafkafi U (eds) Plant roots: the hidden half, 3rd edn. Marcel Dekker, Inc., New York, pp 447–487

Pound MP, French A, Atkinsin J, Wells DM, Bennett MJ, Pridmore TP (2013) RootNav: navigation images of complex roots architectures. Plant Physiol 162:1802–1814.

Singh V, Oosterom EJ, Jordan DR, Messina CD, Cooper M, Hammer GL (2010) Morphological and architectural developmental of root systems in sorghum and maize. Plant Soil 333:287–299

Trachsel S, Messmer R, Stamp P, Hund A (2009) Mapping of QTLs for lateral and axile root growth of tropical maize. Theor Appl Genet 119:1413–1424

Trachsel S, Kaeppler SM, Brown KM, Lynch JP (2011) Shovelomics: high throughput phenotyping of maize (Zea mays L.) root architecture in the field. Plant Soil 341:75–87

Tracy SR, Black CR, Roberts JR, McNeill A, Davidson R, Tester M, Samec M, Korošak D, Sturrock C, Mooney SJ (2012) Quantifying the effect of soil compaction on three varieties of wheat (*Triticum aestivum* L.) with differing root architecture using X-ray micro computed tomography (CT). Plant Soil 353:195–208. doi:10.1007/s11104-011-1022-5

Tumlinson LG, Liu HY, Silk WK, Hopmans JW (2008) Thermal neutron computed tomography of soil water and plant roots. Soil Sci Soc Am J 72:1234–1242

Wasson AP, Richards RA, Chatrath R, Misra SC, Prasad SV, Rebetzke GJ, Kirkegaard JA, Christopher J, Watt M (2012) Traits and selection strategies to improve root systems and water uptake in water-limited wheat crops. J Exp Bot 63:3485–3489

Chapter 5
Seed Phenomics

Jeffrey L. Gustin and A. Mark Settles

Abstract Plant seeds present complex phenotypes that can be difficult to assess quantitatively. The burgeoning field of phenomics seeks to describe phenotypes in high-throughput and with quantitative descriptors that allow computational methods for analysis. This chapter summarizes technology platforms for collecting information-rich seed phenotypes that can also be scaled to high-throughput. Seed phenotypes can be assessed using imaging, spectroscopy, transcriptomes, proteomes, metabolomes, and ionomes. We focus on how these technologies have been developed and applied to maize seeds to define genotype-phenotype relationships.

5.1 Introduction

Determining phenotype-to-genotype relationships is critical for predictive modeling of breeding goals. Next-generation sequencing has greatly increased genotyping depth to obtain millions of polymorphism data points for individual plant breeding lines (reviewed in Huang and Han 2014). However, these genotyping data are usually associated with only a few phenotypes due to the expense of collecting phenotypic data. It is critical that more phenotypes are related to high-density genotypes in order to obtain better predictive models for the expected phenome of a genotype. There has been significant progress in developing high-throughput and high-density phenotyping platforms for many aspects of plant growth and development. These phenotyping approaches are collectively known as phenomics (Houle et al. 2010). Phenomics covers many technologies at multiple scales as well

J.L. Gustin · A.M. Settles (✉)
Horticultural Sciences Department, University of Florida, Gainesville, FL, USA
e-mail: settles@ufl.edu

J.L. Gustin
e-mail: jgustin@ufl.edu

R. Fritsche-Neto and A. Borém (eds.), *Phenomics*,
DOI 10.1007/978-3-319-13677-6_5

as a myriad of growth and development stages in plant life cycles (Dhondt et al. 2013). In this chapter, we review phenomics approaches that can be used to characterize maize seeds. We illustrate the complexity of describing a phenotype, but also the promise of multiple omics and imaging technologies to be able to more rapidly and completely characterize seed phenotypes and relate these to crop traits of interest.

Mature kernels are the primary product of the maize crop, and there is intense interest in improving yield, while maintaining an acceptable kernel phenotype (Ranum et al. 2014). The kernel is composed of the maternal pericarp, the triploid endosperm, and the diploid embryo (Kiesselbach 1949). The endosperm represents the bulk of grain yield and is composed primarily of starch and storage proteins (reviewed in Gwirtz et al. 2014). The embryo accumulates the vast majority of the kernel oil and is also rich in essential macro- and micronutrients, cellular protein, as well as some starch. The pericarp is mostly composed of fiber. Only the embryo and outer aleurone layer of the endosperm are desiccation tolerant, whereas the pericarp and starchy endosperm undergo programmed cell death at the end of development (Dominguez and Cejudo 2014). All of these tissues show variation in storage molecule composition and growth. In addition, the maternal parent has a strong influence over kernel development (Fig. 5.1). Thus, the phenome of a mature kernel is complex with many genes as well as environmental and developmental processes interacting to produce a grain of maize.

Kernel phenotypes traditionally are described based on color, grain-fill, and grain qualities related to end uses, such as popping ability in popcorn, flavor and sugar levels in sweet corn, or hardness and protein or oil quantity in field corn. Through traditional molecular genetics, genes controlling kernel characteristics have been identified for carotenoid, starch, oil, and storage protein traits. Phenomic analysis provides the opportunity to evaluate both mutant and natural variant alleles for their impact on all of these kernel traits.

5.2 Imaging

Trait evaluation based on visual inspection is one of the oldest means of phenotying. For example, archeological evidence suggests that grain size was one of the first traits to experience selective pressures by early agrarian cultures (Purugganan and Fuller 2009). This practice continues with maize breeders relying on visual scoring and, consciously or not, selecting lines based on the appearance of the plant or seed. The relatively recent introduction of molecular markers and genomic selection complements visual screening in plants and, given the complex nature of genotype by environment interactions, genotyping is not likely to fully supplant visual screening. Even though visual screening has been and remains an immensely powerful method of selecting lines and scoring phenotypes, it is low resolution,

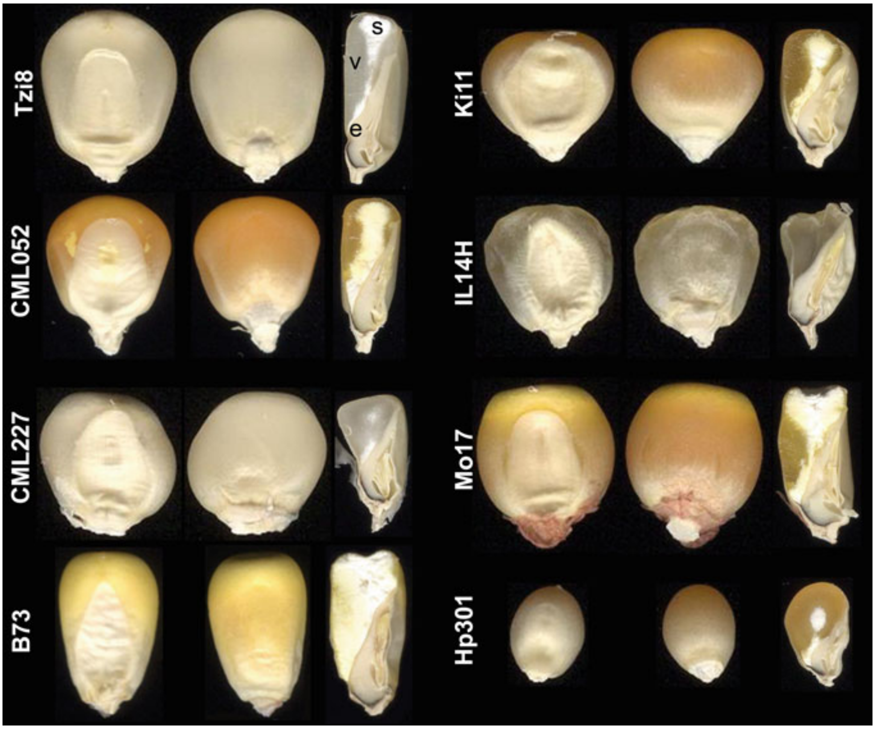


Fig. 5.1 Examples of diverse kernel phenotypes found in maize inbred lines. These lines display a range of structure and composition variation in kernel phenotypes. Each kernel was imaged for the germinal face, the abgerminal face, and a sagittal hand section. The sagittal section shows the virteous endosperm (*v*), starchy endosperm (*s*), and embryo (*e*)

subjective, and difficult to document. Capturing an image or video of a phenotype can provide high-resolution, objective measures of a phenotype. For traits that can be imaged in a high-throughput manner, image-based phenotyping methods greatly contribute to the description of the plant phenome.

Optical sorting is the highest throughput method of optical imaging and can separate kernels by color or size (Liao et al. 1994). Optical sorting machines typically process kernels in parallel using two wavelengths of light from visible, near infrared, or ultraviolet light. Reflectance data are processed in real time, and a compressed air ejector sorts rejected kernels into a separate bin (Fig. 5.2a). Optical sorting can identify and remove discolored grain, including discolorations associated with fungal contamination of the kernel (Pearson et al. 2004, 2010). Although these machines typically do not give quantitative data about individual kernels, they can be used to characterize frequencies of unacceptable kernels within a population and thereby determine the sensitivity of a line to disease or mechanical damage.

Grain shape is an important trait in several crop species due to preferences of consumers and grain processors. For example, long-grained rice is generally

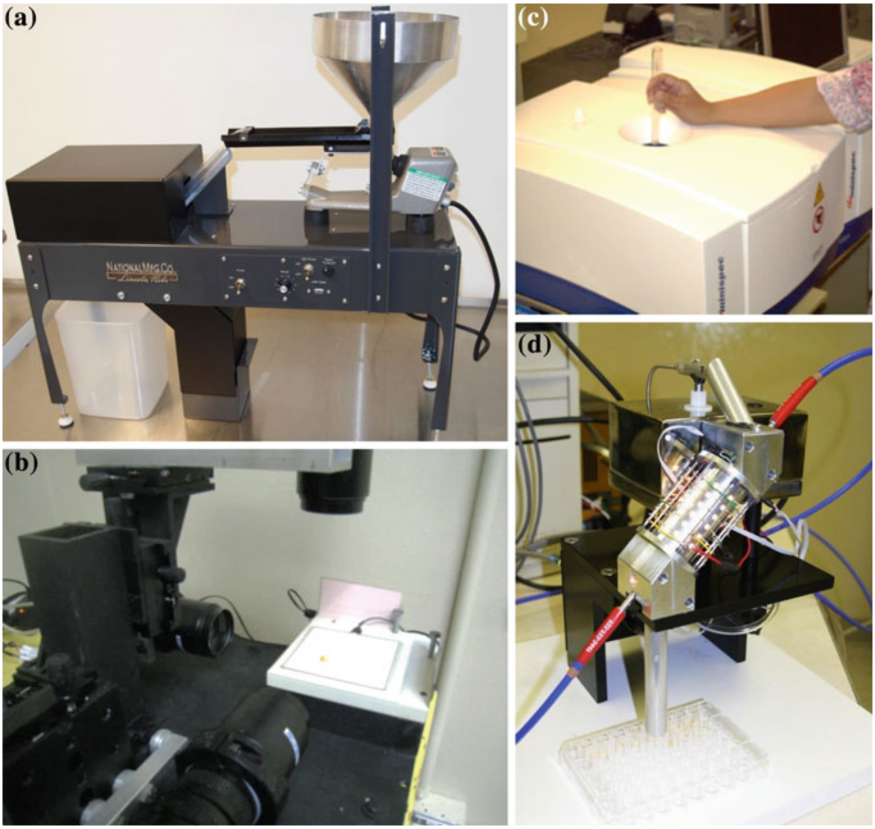


Fig. 5.2 Machine vision phenotyping platforms for maize kernels. **a** Optical sorting machines can use a few wavelengths of vision or NIR light to classify kernels for decoloration or damage. **b** Single-kernel optical imaging system reported by Sekhon et al. (2014). Three digital cameras capture silhoutte images of the kernel from multiple angles simultaneously. **c** Single-kernel NMR can quantify water and oil levels from mature maize kernels. **d** Single-kernel NIR platform reported by Spielbauer et al. (2009). A microbalance collects kernel weight. An NIR spectrum is captured while kernel falls through an illuminated glass tube

preferred by consumers (Juliano and Villareal 1993). To maximize the commercial value of rice, both yield and grain shape must be selected simultaneously as important breeding targets (Miura et al. 2011). In wheat, grain shape is a breeding target for milling performance with large spherical grains thought to produce the highest milling yield (Evers et al. 1990). Unlike other grains, maize kernels develop on a cob in close proximity to sibling kernels, and the degree of sibling crowding influences final shape. Ultimately, individual kernel shape is primarily determined by an interaction between the final volume of the kernel and proximity to its neighbors. Therefore, the shape of individual kernels mostly encodes information about the cob on which the kernels developed. For example, a study using two quantitative trait loci (QTL) mapping populations found that kernel number per plant is positively correlated with the tip-to-cap depth and negatively correlated with the abgerminal-to-germinal thickness of the kernel (Peng et al. 2011). In these populations, increased kernel numbers per ear are most likely due to packing more kernels within a row without changing cob size, which would have the effect of making the growing kernels taller and thinner.

Sekhon et al. (2014) reported a high-throughput imaging system to characterize the Krug long-term selections for small and large kernel size. The system captures single kernel images from multiple angles and processes kernel contours to describe kernel size (Fig. 5.2b). In the Krug selections, increased abgerminal area correlates positively with increased kernel size, reduced seed number, decreased ear row number, and increased biomass accumulation in the endosperm. Thus, high-throughput imaging of kernels can be correlated not only to grain quality, but also ear morphology and other yield traits.

A key quality trait in maize is hard, vitreous kernels, which provide fungal and insect resistance and have desired milling characteristics (reviewed in Fox and Manley 2009). However, vitreousness is a trait normally scored by eye using transmitted visible light. There are several examples where digital imaging of light attenuation through kernels has been used to generate a more objective vitreousness score (Eramus and Taylor 2004; Holding et al. 2008, 2011). These scores have been used to map recombinant inbred lines for *opaque2* modifiers from Quality Protein Maize.

Recently, we found that individual kernel density correlates well with endosperm vitreousness by using microcomputed tomography (microCT) (Gustin et al. 2013). microCT uses high energy X-ray irradiation to image "slices" of the kernel, which can be reconstructed into a three-dimensional model of the kernel. Attenuation of the X-rays as they pass through the sample is directly related to the density of the tissue, and microCT images can be used to calculate the density of the kernel or endosperm tissue. The difference in density between kernel tissues allows microCT to distinguish between embryo, endosperm, and pericarp in maize kernels (Takhar et al. 2011; Gustin et al. 2013). However, data collection is slow, limiting microCT application for high-throughput phenotyping of kernels.

5.3 Spectroscopy

Rapid, nondestructive methods for measuring the chemical composition of seeds has great value for commodity evaluation, breeding line selection, and fundamental research. Near-infrared (NIR) spectroscopy and nuclear magnetic resonance (NMR) spectroscopy are two commonly used, nondestructive techniques that have largely replaced more expensive analytical methods. Both of these techniques use the energy absorption properties of chemicals within the kernel to estimate traits, and the number of traits for which these technologies can be applied is continuing to grow making spectroscopy approaches very useful for seed phenomics.

NMR spectroscopy and the related technique of magnetic resonance imaging (MRI) have been extensively developed by their use in medical imaging. Consequently, there are many different devices and techniques that leverage magnetic resonance to probe biological systems. We will focus only on those few that have been used for phenotyping the chemical and structural variation in seeds. Magnetic resonance techniques make use of elemental isotopes that have magnetic properties due to an uneven number of protons and neutron in the nucleus (for a detailed review, see Chary and Govil 2008; Marion 2013). The primary magnetic isotope used in NMR and MRI studies in seeds is ^{1}H due to its high abundance. Strong magnetic fields are used to force ^{1}H nuclei into a high energy state. When the magnetic field is removed, the nuclei relax to a lower energy state and release radio waves that can be detected and converted into signal peaks. The intensity and wave frequency information are used to identify the chemical constituent and estimate its quantity. Movement of nuclei within a solid matrix is restricted, which causes broadening of the signal peaks and reduction in peak intensity. Consequently, NMR and MRI can only be applied to detect molecules in a liquid phase, such as measuring water and oil in seeds (Conway and Earle 1963). NMR spectroscopy averages the total signal from a sample while MRI uses stronger magnetic fields and higher resolution detectors to scan a sample and produce one-dimensional (1D), two-dimensional (2D), three-dimensional (3D), or even four-dimensional (4D) representations of where the water or oil is within a sample (Fig. 5.2c).

In maize, NMR and MRI are primarily used to characterize oil content. For example, the Illinois long-term selections for maize grain oil content have long used both bulk and single-seed NMR to make the divergent selections for high and low oil (Alexander et al. 1967; Alexander 1982). More recent examples of NMR applications in kernel phenotyping include identifying QTL affecting total oil content (Song et al. 2004), mapping and cloning a major oil QTL (Zheng et al. 2008), and long-term selection to increase oil content in breeding lines (Song and Chen 2004). MRI has been adapted by Monsanto to screen for high-oil kernels in breeding programs. For this application, an array of 24-well microtiter plates holds 2,592 individual kernels and is scanned by a clinical MRI in less than 40 min (Kotyk et al. 2005). Custom software identifies each kernel and measures relative oil content with a throughput that was estimated to be at least one order of magnitude faster than conventional NMR methods.

NIR spectroscopy uses light from the region of the electromagnetic spectrum between 780 and 2,500 nm to infer chemical composition. NIR light is absorbed predominantly by vibrations of organic bonds, including C–H, O–H, and N–H (Siesler 2008). The fundamental resonance frequency of these bonds occur in the mid-infrared (IR) range (2,500–50,000 nm), but chemical bonds also have overtone and combination vibrations that absorb in the NIR spectrum. NIR absorption intensities are 10–100 times weaker than absorption of mid-IR light, which allows NIR light to penetrate thicker samples, such as seeds. Biological samples contain diverse chemical components, and NIR spectra contain multiple, overlapping signal peaks that represent the major constituents of the sample. Multivariate regression models are used to predict chemical composition from the spectra.

Most commercial NIR spectroscopy platforms measure bulk grain composition. These devices may be designed to collect reflectance or transmittance spectra from either ground or whole grain batches of kernels. In addition, NIR spectrometers differ in the spectral range that they can collect data. Unfortunately, each of these data collection formats fundamentally change the spectra and separate calibrations need to be developed to estimate traits for each combination of spectral acquisition and sample preparation. However, the value of NIR becomes apparent because multiple seed traits can be evaluated in the matter of seconds with minimal sample preparation. Indeed, many calibrations have been developed for maize grain constituents, including protein, oil, starch, and moisture (Orman and Schuman 1991; Berardo et al. 2009); fatty acids (Yang et al. 2009); amino acids (Fontaine et al. 2002); carotenoids (Brenna and Berardo 2004); and amylose/amylopectin ratios (Campbell et al. 1997, 1999, 2002). For chemical composition, calibrations are generally more accurate for constituents that accumulate to significant levels within the kernel such as starch, protein, and oil. NIR absorption is also sensitive to light scatter, which is influenced by starch or storage protein packing within the kernel. Thus, NIR spectra can also be related to kernel traits that are not directly related to chemical composition of the kernel. Calibrations have been developed for properties such as wet milling starch yields (Paulsen et al. 2003a, b, 2004), kernel hardness (Robutti 1995; Correa et al. 2002; Ngonyamo-Majee et al. 2008), and fungal infection (Pearson et al. 2001; Dowell et al. 2002).

High-throughput characterization of kernel composition traits has facilitated large-scale genetic studies to identify QTL and, in some cases, genes involved in these important traits. For example, Cook et al. (2012) measured oil, protein, and starch levels using a commercial bulk grain NIR analyzer from 26,000 lines representing seven environmental replications of the Nested Association Mapping (NAM) population and 282 inbred lines of an association mapping panel. Studies of this size would be practically impossible if the traits were measured with analytical chemistry. Importantly, if the spectra for each sample are saved, the data can be reanalyzed for new traits as new calibrations become available.

In contrast to NMR, single-kernel NIR has been far more challenging to develop for maize. This is primarily due to the shape and internal structure of maize kernels. Flattened kernel shapes tend to present the germinal or abgerminal side of the kernel randomly to the NIR spectrometer. The two sides of the kernel have different reflectance spectra due to a higher concentration of oil at the germinal face and higher concentration of starch and protein at the abgerminal face (Orman and Schumann 1992; Weinstock et al. 2006; Baye et al. 2006; Janni et al. 2008). NIR transmittance spectroscopy has been attempted to account for these difference, but transmittance has not proven to be robust for calibrating on multiple kernel composition traits (Finney and Norris 1978; Orman and Schumann 1992; Codgill et al. 2004).

Single-kernel NIR reflectance has shown to be more robust when the spectrum is averaged over the surface of the kernel (Armstrong 2006; Janni et al. 2008). The highest throughput design for a single-kernel device captures an NIR spectrum from a kernel as it tumbles down an illuminated glass tube (Armstrong 2006) (Fig. 5.2d).

This glass-tube device has been used to develop multiple calibrations for maize kernel traits, including starch, protein, oil, weight, density, and kernel volume (Armstrong 2006; Spielbauer et al. 2009; Tallada et al. 2009; Armstrong and Tallada 2012; Gustin et al. 2013). These calibrations have been completed with either a pooling strategy to obtain average analytical values for individual genotypes of maize versus true single-kernel calibrations where the spectra are regressed onto analytical data from individual kernels. The latter approach appears to give reduced error in predictions. In the process of developing these calibrations, we found that at least 90 % of the variance in single-kernel spectra can be attributed to the volume and density of the kernel (Gustin et al. 2013). Interestingly, single-kernel density appears to correlate well with test weight in hybrid seeds, suggesting that single-kernel selection could be used to predict or improve test weight in maize (Gustin et al. 2013).

5.4 Transcriptome

As the most proximate readout of the genome, RNA transcript levels have been intensely studied as a method of inferring genotypes and associating transcript levels with gene functions (reviewed in Gault and Settles 2014). However, transcript levels themselves can be considered a phenotype, and transcriptome profiling provides thousands of phenotypic data points from individual samples. Transcript levels from individuals in a mapping population can be used as quantitative traits to map expression QTL (eQTL). The loci controlling gene expression can be separated into *cis*- and *trans*-acting factors. Strong *cis*-acting eQTL identify loci in which contrasting alleles show large differences in gene expression. When *trans*-acting eQTL for multiple genes cluster at a single locus, these are referred to as hotspots and potentially regulate related biochemical processes. The few maize eQTL studies have focused on seedling gene expression in B73xMo17 recombinant inbred lines (Shi et al. 2007; Swanson-Wagner et al. 2009; Holloway et al. 2011; Li et al. 2013) (Fig. 5.1).

There have not been any eQTL studies reported for maize seed development, but the major insights from seedling analysis indicate that most genes in RILs show parental levels of gene expression (Li et al. 2013), with *trans*-eQTL generally showing small effects (Swanson-Wagner et al. 2009; Holloway et al. 2011). Halloway et al. (2011) also reported that incomplete genome assembly or errors in assembly give significant false positive signals for *trans*-eQTL and that most strong effect *trans*-eQTL result in misplacement of the specific gene within the physical map. Most of the maize studies used independent microarray transcript profiling and single nucleotide polymorphism genotyping to associate transcript levels with genotypes. With RNA-seq, there is the potential to score the data both for transcript levels and for sequence polymorphisms to generate genetic maps for the QTL analysis. This would obviate the need to integrate the physical genome sequence with the QTL mapping. Swanson-Wagner et al. (2009) used reciprocal crosses in their eQTL mapping to determine that the paternal allele generally shows dominance for expression level in hybrid seedlings, suggesting that genomic imprinting can extend from seed to plant development. For maize seeds, eQTL analysis may provide great insights into the control and biological function of imprinted gene expression.

5.5 Proteome

Like the transcriptome, the proteome can also be viewed as a phenotype. Most genes express as proteins, and proteins determine much of the structural and metabolic activity of the cell. Like transcriptomics, proteomics technologies are still relatively expensive to examine hundreds of samples. Most maize kernel proteomics studies focus on two-dimensional polyacrylamide gel electrophoresis (PAGE) followed by directed identification of specific protein spots using mass spectroscopy. This technique is limited to detecting the most abundant 500–700 proteins in a sample.

Two-dimensional PAGE analysis of maize kernels has mainly focused on associating protein levels with physiological responses to either different environments or a mutant perturbation. For example, Silva-Sanchez et al. (2014) compared glycoprotein spots in normal and *minature1* mutant kernels, while Jin et al. (2013) identified ~40 proteins that showed significant abundance changes as grain-filling rates change in different hybrid genotypes. However, this latter study could not account for the protein differences as being caused by genotype or by changes in physiological state. A study designed to dissect genotype and environment components of the kernel proteome found that about 50 % of the variation in kernel proteomes can be accounted for by genotype, while the remaining variation can be explained by growing season or location of the plant (Anttonen et al. 2010). Proteomics analysis of F1 hybrid embryos found that nearly 25 % of protein spots accumulated to levels closer to one of the parents rather than fitting an additive model with many protein spots accumulating to the low-parent level (Marcon et al. 2010). These observations suggest that proteome phenotypes are influenced heavily by environment and the parent-of-origin of the haplotypes used in a cross.

Finally, two-dimensional PAGE has been used to characterize protein changes resulting from transgenic events, with about 40 proteins showing differences between transgenic and isogenic, nontransgenic controls (Zolla et al. 2008). However, the majority of the proteins found to be different in this study correspond to many of the proteins that have been subsequently shown to be influenced by environment or grain-filling rates (Anttonen et al. 2010; Jin et al. 2013). A more complete proteome of developing kernels has been characterized using liquid chromatography coupled with mass spectroscopy (Walley et al. 2013). As proteomics technologies reduce in cost, it is likely that the proteome will become a more robust phenotype for seed phenomics applications.

5.6 Metabolome

Metabolites represent a snapshot of cellular metabolism for a tissue sampled. Profiling even a limited number of metabolites can give significant insight into kernel phenotypes. For example, analysis of only a handful of sugars and sugar phosphates in maize starch biosynthetic mutants was able to group the *shrunken2* and *brittle1* mutants as having related phenotypes (Tobias et al. 1992). Only after cloning and biochemical characterization did it become clear that *brittle1* encodes the ADP-glucose transporter, which transports the product of *shrunken2* from the cytosol into the amyloplast for starch synthesis (Kirchberger et al. 2007). Targeted metabolite profiles have also been used as biomarkers to study the development of hybrid kernels relative to parental inbreds (Romisch-Margl et al. 2010).

Probably the most significant progress in using targeted metabolite profiling in maize kernels has been to identify alleles for breeding improved carotenoid accumulation. High-performance liquid chromatography (HPLC) was used to profile association mapping populations to find favorable alleles in both the *lycopene epsilon-cyclase* and *hydroxylase3* genes (Harjes et al. 2008; Vallabhaneni et al. 2009; Yan et al. 2010). These genes encode enzymes that determine the flux to the optimal provitamin A carotenoid, beta-carotene, as well as the catabolism of beta-carotene to nonprovitamin A xanthophylls. The variants identified from metabolite profiling have been used to increase provitamin A content in tropical maize lines for biofortification of the kernel (Azmach et al. 2013).

There have been tremendous advances in the ability to profile a broader spectrum of plant metabolites using liquid and gas chromatography to separate derivatized metabolites coupled with mass spectrometry to identify individual compounds (reviewed in Fernie and Schauer 2009). These technologies can quantify approximately 100 known metabolites and track signatures for more than 1,000 metabolite features. Profiling known metabolites can provide insight into the biochemical status of kernel tissues. For example, metabolite profiling of a mutant in the maize *6-phosphogluconate dehydrogenase3* locus found that the kernel showed metabolite shifts consistent not only with known roles in NADPH production and fatty acid biosynthesis but also an increase in reducing sugars consistent with reduced starch synthesis (Spielbauer et al. 2013).

This more comprehensive metabolomics has only recently been applied to maize kernel phenotyping. Shen et al. (2013) completed association analysis on thousands of metabolite features from cooked maize. Using network analysis, approximately half of the metabolite features could be grouped into 48 networks. Surprisingly, one network revealed polymorphism in the alpha-zein storage protein being associated with levels of the C-terminal peptide of this protein, suggesting the peptide was differentially released after cooking in diverse germplasm. By contrast, Wen et al. (2014) completed profiling of mature maize kernels to use the identified metabolites for QTL and association analysis. Nearly 1,500 individual locus to trait associations were detected in this study, with multiple examples of association analysis identifying enzymes that influence the accumulation of specific metabolites.

These studies illustrate the potential of using metabolomics to obtain high-density phenotypes for kernels to both uncover unexpected variations and to associate genes with specific biochemical functions.

5.7 Ionome

The ionome is a term used to encompass the entire mineral nutrient and trace element constituent of biological samples (Salt et al. 2008). The ionome of plants is a complex mixture of elements whose concentration in the plant is dependent upon the genotype of the plant, the environment in which it is growing, and the tissue being sampled. In cereal grains, elemental accumulation is heterogeneous. For example, the embryo and aleurone layer have high concentration of important mineral nutrients Zn and Fe, whereas Cu preferentially accumulates in the starchy endosperm (Ozturk et al. 2006; Cakmak et al. 2010; Iwai et al. 2012). Therefore, selection for altered seed composition or size directly and inadvertently impacts grain mineral nutrition.

The methods of choice for measuring the elemental profile of biological samples are inductively coupled plasma spectroscopy using either a mass spectrometer (ICP-MS) to directly measure ions based on mass or optical emission spectroscopy (ICP-OES) to indirectly measure ions based on their characteristic light emissions. These instruments can analyze multiple ions simultaneously in a matter of minutes, making them a useful tool for high-throughput phenotyping applications, as has been recently demonstrated in large-scale forward and reverse genetic screens in *Arabidopsis* (Salt et al. 2008) and QTL studies and association mapping studies in rice and *Arabidopsis* (Zhang et al. 2014; Norton 2014).

An ICP-MS pipeline has been reported for measuring 20 elements from individual maize kernels (Baxter et al. 2014). This pipeline was used to screen bi-parental RIL populations for QTL affecting grain nutritional quality. Screening the intermated B73xMo17 population identified 27 QTL from 9 elements (Baxter et al. 2013). Several QTL colocalized with loci identified in rice for Fe and Zn accumulation in the grain. In a separate study, 31 QTL for 13 elements were identified in a stiff stock by sweet corn (B73xIL14H) RIL population (Baxter et al. 2014). No overlapping QTL were found between the studies, suggesting that the extensive diversity of maize germplasm has a large impact on the elemental profile of maize grain.

5.8 Future Directions

Objectively describing kernel phenotypes in high-throughput is a serious challenge. The development of multiple phenotyping platforms is shifting seed phenotyping from a data-limited science to being computationally limiting. Imaging and spectroscopy techniques can provide nondestructive measures of shape, size, color, chemical composition, density, and even pathogen infection in just a few seconds. Although requiring a bit more time to assay, omics technologies provide even more data points to describe gene, protein, metabolite, or elemental composition of kernels. All of these tools have been used successfully to connect genes to phenotypes of interest for breeders. However, there are many challenges and open questions about the value of different phenomic technologies. It is not clear whether the transcriptome, proteome, metabolome, and ionome are highly correlated phenotypes in which measuring one might be able to predict the other analyses. It is possible that each omic technology will give clearly different phenotypic insights, but there are not many examples integrating multiple omics platforms in a single analysis of seed phenotypes. Integrating phenomics platforms also raises the challenge of analyzing and storing many different data types, such as digital images, spectra, and sequence read counts. In addition, unlike genomic data, phenotypes are impacted by environment, which raises the difficult issue of collecting and associating environmental metadata with phenomic data points. Robust computational skills will be required not just to analyze next-generation genotyping but also to handle these new generations of phenotyping technologies.

Acknowledgments We apologize to the authors of relevant research articles that were not highlighted in this chapter due to space constraints. The authors' research on seed phenomics is supported by grants from the National Science Foundation (awards IOS-1031416 and MCB-1412218), the National Institute of Food and Agriculture (awards 2010-04228 and 2011-67013-30032), and the Vasil-Monsanto Endowment.

References

Alexander DE (1982) The use of wide-line NMR in breeding high-oil corn. J Am Oil Chem Soc 59:A284

Alexander DE, Silvela L, Collins FI, Rodgers RC (1967) Analysis of oil content of maize by wide-line NMR. J Am Oil Chem Soc 44:555–558

Anttonen MJ, Lehesranta S, Auriola S, Röhlig RM, Engel KH, Kärenlampi SO (2010) Genetic and environmental influence on maize kernel proteome. J Proteome Res 9:6160–6168

Armstrong PR (2006) Rapid single-kernel NIR measurement of grain and oil-seed attributes. Appl Eng Agric 22:767–772

Armstrong PR, Tallada JG (2012) Prediction of kernel density of corn using single-kernel near infrared spectroscopy. Appl Eng Agric 28:569–574

Azmach G, Gedil M, Menkir A, Spillane C (2013) Marker-trait association analysis of functional gene markers for provitamin A levels across diverse tropical yellow maize inbred lines. BMC Plant Biol 13:227

Baxter IR, Gustin JL, Settles AM, Hoekenga OA (2013) Ionomic characterization of maize kernels in the intermated B73xMo17 population. Crop Sci 53:208–220

Baxter IR, Ziegler G, Lahner B, Mickelbart MV, Foley R, Danku J, Armstrong P, Salt DE, Hoekenga OA (2014) Single-kernel ionomic profiles are highly heritable indicators of genetic and environmental influences on elemental accumulation in maize grain (Zea mays). PLoS ONE 9:e87628

Baye TM, Pearson TC, Settles AM (2006) Development of a calibration to predict maize seed composition using single kernel near infrared spectroscopy. J Cereal Sci 43:236–243

Berardo N, Mazzinelli G, Valoti P, Lagana P, Redaelli R (2009) Characterization of maize germplasm for the chemical composition of the grain. J Agric Food Chem 57:2378–2384

Brenna OV, Berardo N (2004) Application of near-infrared reflectance spectroscopy (NIRS) to the evaluation of carotenoids content in maize. J Agric Food Chem 52:5577–5582

Cakmak I, Pfeiffer W, McClafferty B (2010) Biofortification of durum wheat with zinc and iron. Cereal Chem 87:10–20

Campbell KG, Bergman CJ, Gualberto DG, Anderson JA, Giroux MJ, Hareland G, Fulcher RG, Sorrells ME, Finney PL (1999) Quantitative trait loci associated with kernel traits in a soft x hard wheat cross. Crop Sci 39:1184–1195

Campbell MR, Brumm TJ, Glover DV (1997) Whole grain amylose analysis in maize using near-infrared transmittance spectroscopy. Cereal Chem 74:300–303

Campbell MR, Yeager H, Abdubek N, Pollak LM, Glover DV (2002) Comparison of methods for amylose screening among amylose-extender (ae) maize starches from exotic backgrounds. Cereal Chem 79:317–321

Chary KV, Govil G (2008) NMR in biological systems: from molecules to humans. In: Kapein R (ed) Series: focus on structural biology, vol 6. Dordrecht, The Netherlands

Cogdill RP, Hurburgh CR, Rippke GR (2004) Single-kernel maize analysis by near-infrared hyperspectral imaging. Trans ASAE 47:311–320

Conway TF, Earle FR (1963) Nuclear magnetic resonance for determining oil content of seeds. J Am Oil Chem Soc 40:265–268

Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Ross-Ibarra J, Buckler ES, Flint-Garcia SA (2012) Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant Physiol 158:824–834

Correa CES, Shaver RD, Pereira MN, Lauer JG, Kohn K (2002) Relationship between corn vitreousness and ruminal in situ starch degradability. J Dairy Sci 85:3008–3012

Dhondt S, Wuyts N, Inzé D (2013) Cell to whole-plant phenotyping: the best is yet to come. Trends Plant Sci 18:428–439

Domínguez F, Cejudo FJ (2014) Programmed cell death (PCD): an essential process of cereal seed development and germination. Front Plant Sci 5:366

Dowell FE, Pearson TC, Maghirang EB, Xie F, Wicklow DT (2002) Reflectance and transmittance spectroscopy applied to detecting fumonisin in single corn kernels infected with Fusarium verticillioides. Cereal Chem 79:222–226

Erasmus C, Taylor JRN (2004) Optimising the determination of maize endosperm vitreousness by a rapid non-destructive image analysis technique. J Sci Food Agric 84:920–930

Evers AD, Cox RI, Shaheedullah MZ, Withey RP (1990) Predicting milling extraction rate by image analysis of wheat grains. Asp Appl Biol 25:417–426

Fernie AR, Schauer N (2009) Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet 25:39–48

Finney EE, Norris KH (1978) Determination of moisture in corn kernels by near-infrared transmittance measurements. Trans ASAE 21:581–584

Fontaine J, Schirmer B, Horr J (2002) Near-infrared reflectance spectroscopy (NIRS) enables the fast and accurate prediction of essential amino acid contents. J Agric Food Chem 50:3902–3911

Fox G, Manley M (2009) Hardness methods for testing maize kernels. J Agric Food Chem 57:5647–5657

Gault CM, Settles AM (2014) Functional genomics. In: Wusirika R, Bohn M, Lai J (eds) Genetics, genomics, and breeding of maize. CRC Press, Boca Raton, pp 131–154

Gustin JL, Settles AM (2013) Machine vision for seed phenomics. In: Becraft PW (ed) Seed genomics. Wiley, Hoboken, pp 237–251

Gwirtz JA, Garcia-Casal MN (2014) Processing maize flour and corn meal food products. Ann NY Acad Sci 1312:66–75

Harjes CE, Rocheford TR, Bai L, Brutnell TP, Kandianis CB, Sowinski SG, Stapleton AE, Vallabhaneni R, Williams M, Wurtzel ET, Yan J, Buckler ES (2008) Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science 319:330–333

Holding DR, Hunter BG, Chung T, Gibbon BC, Ford CF, Bharti AK, Messing J, Hamaker BR, Larkins BA (2008) Genetic analysis of opaque2 modifier loci in quality protein maize. Theor Appl Genet 117:157–170

Holding DR, Hunter BG, Klingler JP, Wu S, Guo X, Gibbon BC, Wu R, Schulze JM, Jung R, Larkins BA (2011) Characterization of opaque2 modifier QTLs and candidate genes in recombinant inbred lines derived from the K0326Y quality protein maize inbred. Theor Appl Genet 122:783–794

Holloway B, Luck S, Beatty M, Rafalski JA, Li B (2011) Genome-wide expression quantitative trait loci (eQTL) analysis in maize. BMC Genom 12:336

Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. Nat Rev Genet 11:855–866

Huang X, Han B (2014) Natural variations and genome-wide association studies in crop plants. Annu Rev Plant Biol 65:531–551

Iwai T, Takahashi M, Oda K, Terada Y, Yoshida KT (2012) Dynamic changes in the distribution of minerals in relation to phytic acid accumulation during rice seed development. Plant Physiol 160:2007–2014

Janni J, Weinstock BA, Hagen L, Wright S (2008) Novel near-infrared sampling apparatus for single kernel analysis of oil content in maize. Appl Spectrosc 62:423–426

Jin X, Fu Z, Ding D, Li W, Liu Z, Tang J (2013) Proteomic identification of genes associated with maize grain-filling rate. PLoS ONE 8:e59353

Juliano BO, Villareal CP (1993) Grain quality evaluation of world rices. International Rice Research Institute, Manila

Kiesselbach TA (1949) The structure and reproduction of corn. Nebr Agric Exp Stn Ann Rep 161:1–96

Kirchberger S, Leroch M, Huynen MA, Wahl M, Neuhaus HE, Tjaden J (2007) Molecular and biochemical analysis of the plastidic ADP-glucose transporter (ZmBT1) from Zea mays. J Biol Chem 282:22481–22491

Kotyk JJ, Pagel MD, Deppermann KL, Colletti RF, Hoffman NG, Yannakakis EJ, Das PK, Ackerman JJ (2005) High-throughput determination of oil content in corn kernels using nuclear magnetic resonance imaging. J Am Oil Chem Soc 82(855):862

Li L, Petsch K, Shimizu R, Liu S, Xu WW, Ying K, Yu J, Scanlon MJ, Schnable PS, Timmermans MC, Springer NM, Muehlbauer GJ (2013) Mendelian and non-mendelian regulation of gene expression in maize. PLoS Genet 9:e1003202

Liao K, Paulsen MR, Reid JF (1994) Real-time detection of colour and surface defects of maize kernels using machine vision. J Agric Eng Res 59:263–271

Marcon C, Schützenmeister A, Schütz W, Madlung J, Piepho HP, Hochholdinger F (2010) Nonadditive protein accumulation patterns in Maize (Zea mays L.) hybrids during embryo development. J Proteome Res 9(12):6511–6522

Marion D (2013) An introduction to biological NMR spectroscopy. Mol Cell Proteomics 12:3006–3025

Miura K, Ashikari M, Matsuoka M (2011) The role of QTLs in the breeding of high-yielding rice. Trends Plant Sci 16:319–326

Ngonyamo-Majee D, Shaver RD, Coors JG, Sapienza D, Correa CES, Lauer JG, Berzaghi P (2008) Relationships between kernel vitreousness and dry matter degradability for diverse corn germplasm I. Development of near-infrared reflectance spectroscopy calibrations. Anim Feed Sci Tech 142:247–258

Norton GJ, Douglas A, Lahner B, Yakubova E, Guerinot ML, Pinson SRM, Tarpley L, Eizenga GC, McGrath SP, Zhao FJ, Islam MR, Islam S, Duan G, Zhu Y, Salt DE, Meharg AA, Price AH (2014) Genome wide association mapping of grain arsenic, copper, molybdenum and zinc in rice (Oryza sativa L.) grown at four international field sites. PLoS ONE 9:e89685

Orman BA, Schumann RA (1991) Comparison of near-infrared spectroscopy calibration methods for the prediction of protein, oil, and starch in maize grain. J Agric Food Chem 39:883–886

Orman BA, Schumann RA (1992) Nondestructive single-kernel oil determination of maize by near-infrared transmission spectroscopy. J Am Oil Chem Soc 69:1036–1038

Ozturk L, Yazici MA, Yucel C, Torun A, Cekic C, Bagci A, Ozkan H, Braun HJ, Sayers Z, Cakmak I (2006) Concentration and localization of zinc during seed development and germination in wheat. Physiol Plant 128:144–152

Paulsen MR, Mbuvi SW, Haken AE, Ye B, Stewart RK (2003a) Extractable starch as a quality measurement of dried corn. Appl Eng Agric 19:211–217

Paulsen MR, Pordesimo LO, Singh M, Mbuvi SW, Ye BY (2003b) Maize starch yield calibrations with near infrared reflectance. Biosyst Eng 85:455–460

Paulsen MR, Singh M (2004) Calibration of a near-infrared transmission grain analyzer for extractable starch in maize. Biosyst Eng 89:79–83

Pearson TC, Wicklow DT, Brabec DL (2010) Characteristics and sorting of white food corn contaminated with mycotoxins. Appl Eng Agric 26:109–113

Pearson TC, Wicklow DT, Maghirang EB, Xie F, Dowell FE (2001) Detecting aflatoxin in single corn kernels by transmittance and reflectance spectroscopy. Trans ASAE 44:1247–1254

Pearson TC, Wicklow DT, Pasikatan MC (2004) Reduction of aflatoxin and fumonisin contamination in yellow corn by high-speed dual-wavelength sorting. Cereal Chem 81:490–498

Peng B, Li Y, Wang Y, Liu C, Liu Z, Tan W, Zhang Y, Wang D, Shi Y, Sun B, Song Y, Wang T, Li Y (2011) QTL analysis for yield components and kernel-related traits in maize across multi-environments. Theor Appl Genet 122:1305–1320

Purugganan MD, Fuller DQ (2009) The nature of selection during plant domestication. Nature 457:843–848

Ranum P, Peña-Rosas JP, Garcia-Casal MN (2014) Global maize production, utilization, and consumption. Ann NY Acad Sci 1312:105–112

Robutti JL (1995) Maize kernel hardness estimation in breeding by near-infrared transmission analysis. Cereal Chem 72:632–636

Römisch-Margl L, Spielbauer G, Schützenmeister A, Schwab W, Piepho HP, Genschel U, Gierl A (2010) Heterotic patterns of sugar and amino acid components in developing maize kernels. Theor Appl Genet 120:369–381

Salt DE, Baxter I, Lahner B (2008) Ionomics and the study of the plant ionome. Annu Rev Plant Biol 59:709–733

Sekhon RS, Hirsch CN, Childs KL, Breitzman MW, Kell P, Duvick S, Spalding EP, Buell CR, de Leon N, Kaeppler SM (2014) Phenotypic and transcriptional analysis of divergently selected maize populations reveals the role of developmental timing in seed size determination. Plant Physiol 165:658–669

Shen M, Broeckling CD, Chu EY, Ziegler G, Baxter IR, Prenni JE, Hoekenga OA (2013) Leveraging non-targeted metabolite profiling via statistical genomics. PLoS ONE 8:e57667

Shi C, Uzarowska A, Ouzunova M, Landbeck M, Wenzel G, Lübberstedt T (2007) Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint x Flint maize recombinant inbred line population. BMC Genom 8:22

Siesler HW (2008) Basic principles of vibrational spectroscopy. In: Burns DA, Ciurczak EW (eds) Handbook of near-infrared analysis. CRC Press, Boca Raton, Florida, pp 7–18

Silva-Sanchez C, Chen S, Li J, Chourey PS (2014) A comparative glycoproteome study of developing endosperm in the hexose-deficient miniature1 (mn1) seed mutant and its wild type Mn1 in maize. Front Plant Sci 5:63

Song TM, Chen SJ (2004) Long term selection for oil concentration in five maize populations. Maydica 49:9–14

Song XF, Song TM, Dai JR, Rocheford T, Li JS (2004) QTL mapping of kernel oil concentration with high-oil maize by SSR markers. Maydica 49:41–48

Spielbauer G, Armstrong P, Baier JW, Allen WB, Richardson K, Shen B, Settles AM (2009) High-throughput near-infrared reflectance spectroscopy for predicting quantitative and qualitative composition phenotypes of individual maize kernels. Cereal Chem 86:556–564

Spielbauer G, Li L, Römisch-Margl L, Do PT, Fouquet R, Fernie AR, Eisenreich W, Gierl A, Settles AM (2013) Chloroplast-localized 6-phosphogluconate dehydrogenase is critical for maize endosperm starch accumulation. J Exp Bot 64:2231–2242

Swanson-Wagner RA, DeCook R, Jia Y, Bancroft T, Ji T, Zhao X, Nettleton D, Schnable PS (2009) Paternal dominance of trans-eQTL influences gene expression patterns in maize hybrids. Science 326:1118–1120

Takhar PS, Maier DE, Campanella OH, Chen G (2011) Hybrid mixture theory based moisture transport and stress development in corn kernels during drying: validation and simulation results. J Food Eng 106:275–282

Tallada JG, Palacios-Rojas N, Armstrong PR (2009) Prediction of maize seed attributes using a rapid single kernel near infrared instrument. J Cereal Sci 50:381–387

Tobias RB, Boyer CD, Shannon JC (1992) Alterations in carbohydrate intermediates in the endosperm of starch-deficient maize (Zea mays L.) genotypes. Plant Physiol 99:146–152

Vallabhaneni R, Gallagher CE, Licciardello N, Cuttriss AJ, Quinlan RF, Wurtzel ET (2009) Metabolite sorting of a germplasm collection reveals the hydroxylase3 locus as a new target for maize provitamin A biofortification. Plant Physiol 151:1635–1645

Walley JW, Shen Z, Sartor R, Wu KJ, Osborn J, Smith LG, Briggs SP (2013) Reconstruction of protein networks from an atlas of maize seed proteotypes. Proc Natl Acad Sci USA 110: E4808–E4817

Weinstock BA, Janni J, Hagen L, Wright S (2006) Prediction of oil and oleic acid concentrations in individual corn (Zea mays L.) kernels using near-infrared reflectance hyperspectral imaging and multivariate analysis. Appl Spectrosc 60:9–16

Wen W, Li D, Li X, Gao Y, Li W, Li H, Liu J, Liu H, Chen W, Luo J, Yan J (2014) Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat Commun 5:3438

Yan J, Kandianis CB, Harjes CE, Bai L, Kim EH, Yang X, Skinner DJ, Fu Z, Mitchell S, Li Q, Fernandez MG, Zaharieva M, Babu R, Fu Y, Palacios N, Li J, Dellapenna D, Brutnell T, Buckler ES, Warburton ML, Rocheford T (2010) Rare genetic variation at Zea mays crtRB1 increases beta-carotene in maize grain. Nat Genet 42:322–327

Yang XH, Guo YQ, Fu Y, Hu JY, Chai YC, Zhang YR, Li JS (2009) Measuring fatty acid concentration in maize grain by near-infrared reflectance spectroscopy. Spectrosc Spectral Anal 29:106–109

Zhang M, Pinson SR, Tarpley L, Huang XY, Lahner B, Yakubova E, Baxter I, Guerinot ML, Salt DE (2014) Mapping and validation of quantitative trait loci associated with concentrations of 16 elements in unmilled rice grain. Theor Appl Genet 127:137–165

Zheng P, Allen WB, Roesler K, Williams ME, Zhang S, Li J, Glassman K, Ranch J, Nubel D, Solawetz W, Bhattramakki D, Llaca V, Deschamps S, Zhong GY, Tarczynski MC, Shen B (2008) A phenylalanine in DGAT is a key determinant of oil content and composition in maize. Nat Genet 40:367–372

Zolla L, Rinalducci S, Antonioli P, Righetti PG (2008) Proteomics as a complementary tool for identifying unintended side effects occurring in transgenic maize seeds as a result of genetic modifications. J Proteome Res 7:1850–1861

Chapter 6
Agronomic Field Trait Phenomics

Dhyaneswaran Palanichamy and Joshua N. Cobb

Abstract Recent advances in high-throughput phenotyping allow breeders to collect phenotypic data with a level of accuracy that was impossible to achieve previously. However, many of these technologies depend on leveraging-controlled environments like green houses or growth chambers. While these controlled phenotypes can have strategic value for gene discovery, their relevance for breeding and understanding genotype x environment interactions to predict field performance is an active field of study and currently limited, at best. This chapter deals with various technologies that have empowered the collection of phenotypic data directly under field conditions and the relative advantages and disadvantages of using them to collect agronomic phenotypes. Important considerations to be aware of before planning a high-throughput phenotyping experiment that use technologies like field spectroscopy and remote sensing are also discussed including a review of various publically available and/or commercial aerial, ground-based and root phenotyping platforms.

6.1 Introduction

The human population is predicted to increase to 9.6 billion from the current world population of 7.2 billion by the year 2050 (United Nations 2013). Not the least of the global challenges associated with this population growth is the issue of supplying sufficient food, fuel, and fiber to feed a global economy driven by a population of this magnitude. The difficulty of meeting this demand is further compounded by a 5–10 million hectare reduction in agricultural land area annually due to land degradation (GEF/UNCCD 2011), as well as declines in agricultural

D. Palanichamy
School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA

J.N. Cobb (✉)
DuPont Pioneer, Johnston, IA 50131, USA
e-mail: joshua.cobb@pioneer.com

R. Fritsche-Neto and A. Borém (eds.), *Phenomics*,
DOI 10.1007/978-3-319-13677-6_6

productivity associated with predicted patterns of climate change (Melillo et al. 2014). Plant breeding and genetics sits at the center of the multidisciplinary solutions that will form to meet the challenge of feeding the future. Innovations in plant breeding will lead to the de novo generation of novel genetic material suitable to perform under such demanding environmental, economic, and political constraints.

Despite thousands of years of technological advancement since the origin of plant breeding, genetic recombination has remained the primary driver of genetic gain from selection. Selection in plant breeding can be defined as the science of discriminating among biological variants in a population to identify and pick desirable recombinants. However, identifying the recombinants that lead to superior phenotypes can be challenging as the phenotype is driven both by the genetic constituency of an organism as well as the influence of the environment within which it is grown. Complicating this further is the principle of genotype–environment interactions that favor genotypes differentially according to the environmental conditions. For the purposes of this discussion, the phenotype is defined as the observable effect of the genotype and its interaction with a given environment (Acquaah 2007).

The advent of genetic marker technologies has allowed breeders to visualize the effects of selection on the genome and in some measure predict recombinants that have a higher probability of producing superior phenotypes. Because this has been most easily accomplished via marker-assisted selection (MAS) in cases where the desirable phenotype is controlled by a single dominant gene, it has driven the development of diversity panels and mapping populations in several species, which attempt to explore existing genetic diversity in search of simple traits (McCouch et al. 2012).

In the more complex cases where the infinitesimal model of gene action applies (i.e., many genes all contributing small effects on a quantitative phenotype), this has proven more difficult. In these cases, genomic selection has emerged as one strategy to predict desirable recombinants based on a smaller training population that is carefully phenotyped and genotyped (Heffner et al. 2009; Asoro et al. 2011).

The reduction in the cost of genotyping that has come about with improvements in next-generation sequencing and single nucleotide polymorphism genotyping technology has resulted in the generation of an abundant amount of genotypic data in a very short time period. This deluge of genomic information has driven a strong focus on strategies for the integration of molecular information in the breeding pipeline.

However, despite these advances, there is still tremendous need for high-quality phenotype data—not only to empower genetic mapping efforts and genomic selection models, but also for direct phenotypic selection. No matter how much genetic or genomic information is used to predict phenotypic information, the phenotype is still driven by both the genome and the environment. To select desirable recombinants, the environment must either be sampled or controlled in order to separate genetic ‘signal’ from environmental ‘noise.’ This is a challenge because phenotyping is more expensive, labor intensive, and prone to error than genotyping. However, unlike with the evolution of genotyping technology, cost per phenotypic data point is unlikely to change dramatically. Tuberosa (2012) put it best by stating that “phenotyping is [now] king, and [has taken] heritability [a]s queen.”

Increasing heritability (the proportion of phenotypic variance explained by the genotype) of any phenotyping activity is partly a function of managing the influence of the environment. This can be done either by controlling the environmental variation directly (as with the use of a growth chamber or other managed environments) or by sampling the environmental variation by replicating observations of a genotype across a target population of environments. In this chapter, we primarily treat field-based phenotyping systems. For a more detailed comparison of field-based versus controlled phenotyping systems, see Cobb et al. (2013). Suffice it to mention that one of the major challenges of the field-based high-throughput phenotyping systems when compared to phenotyping in controlled environments is that field-based phenotyping systems require technology suitable for varied field conditions and data collection strategies that are simultaneously cheap and high throughput, but also avoid increasing error.

6.2 Major Considerations Before Planning a Phenotyping Experiment

Before conducting any study that requires high-throughput phenotyping (HTP), it is vital to understand the objective of the experiment. HTP for selection of desirable recombinants at a 50 % selection intensity will require a different level of accuracy than phenotyping a mapping population with the intention of identifying a marker linked to an important gene. Cost should be the primary driver of any analysis; if all a researcher needs to do is throw away the bottom half of a population, the cost per phenotypic data point can be substantially reduced compared to the approach that would be needed for more sophisticated genetic analyses. The trait(s) that the researcher is looking to phenotype should be well articulated beforehand, and relevance to selection targets should be established conclusively. For example, in a naturally self-pollinating organism, phenotyping a population for stigma exertion in order to increase cross pollination is only useful if stigma exertion has been demonstrated to contribute to increased cross pollination a priori.

Even though the application of HTP data is broad and is used in various experiments, such as genomic selection, gene discovery, germplasm characterization, QTL mapping, and phenotypic selection of breeding material, it is important to have clearly defined objectives prior to the collection of phenotypic data (Araus and Cairns 2014). Information on the genetic architecture of a trait and the heritability of the trait under various phenotyping protocols is also necessary to help the researcher prioritize phenotyping activities and plan costs before using a HTP platform.

A detailed description of many of the factors that need to be considered collecting HTP data have been described by Cobb et al. (2013), as shown in Fig. 6.1. One often overlooked detail is the use of standardized ontologies and careful

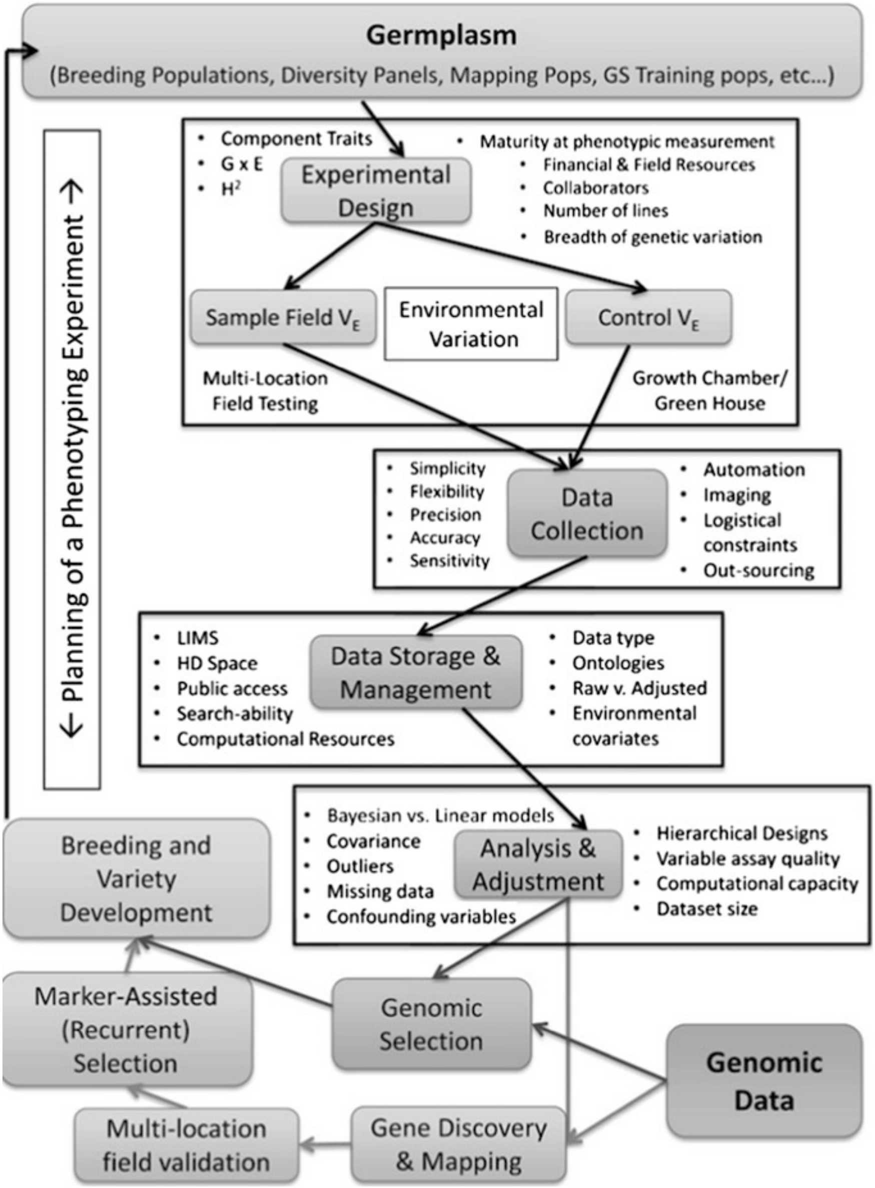


Fig. 6.1 Factors to be considered before obtaining data from a high-throughput phenotyping experiment (Cobb et al. 2013; used with permission)

management of phenotypic databases. If done properly and in alignment across groups, such strategies integrate results across experiments and can drive powerful meta-analyses, which can leverage information across a diversity of environments in order to enhance genetic signal.

6.3 Major Phenotypes Measured in Field Conditions

Agronomically relevant phenotypes that are also most amenable to HTP tend to be observable at the canopy level, are measurable at different points throughout the crop lifecycle, and also discriminate plants based on their ability to capture natural resources and use them efficiently (Araus et al. 2008). Examples of such traits would be canopy temperature, leaf area, normalized difference vegetation index (NDVI), and chlorophyll density. Each of these traits serves as a proxy for more relevant agronomic metrics, such as water use efficiency, biomass, harvest index, and disease pressure. Researchers have developed a variety of inexpensive and novel field-based HTP approaches for a number of these traits, with the goal of automating the phenotyping process without sacrificing the power of prediction (Araus and Cairns 2014). In recent years, improvements in various technologies such as remote sensing, digital image analysis, infrared thermography, robotics, data management, and farm machinery have accelerated the rate at which HTP data is collected and catalogued in a database. The challenge moving forward, however, is to address the question of how this data can be translated into actionable information upon which breeding decisions can be made.

The literature contains a few good examples of this, including Munns et al. (2010), who recorded the temperature gradients in wheat and barley fields using an infrared camera. This indicated the levels of energy dissipation in plants and allowed the researchers to visualize stress response capacity and photosynthetic rates, which serve as reliable indicators for drought tolerance (Munns et al. 2010). Additionally, estimation of leaf area density (LAD) by using a three-dimensional portable LIDAR (light detection and ranging) enabled the more precise phenotyping of the vertical canopy structure in Japanese *Zelkova* trees (Hosoi and Omasa 2007). Furthermore, drought tolerance in cotton (*Gossypium barbadense* L.) was estimated by evaluating canopy height, NDVI, and temperature using four sets of sensors mounted on a tractor (Andrade-Sanchez et al. 2013). This technique allowed the researchers to collect time series data from the field in order to understand physiological and developmental traits that are typically harder to phenotype under relevant growing conditions.

Advances in field-based HTP technology have also opened areas of research to the below-ground parts of an agro-ecosystem, which have been historically hard to phenotype. One good example comes from Zenone et al. (2008), who used ground penetrating radar and electrical resistivity tomography as an indirect nondestructive technique for estimating the shape and behavior of pine tree roots in the subsoil. Perhaps the earliest foray into this space is the rapid root biomass detection in maize, done using a portable capacitance meter by Van Beem et al. (1998).

Each of these examples highlights three key questions that need to be asked before funding an effort to collect and manage HTP data:

(1) What data are going to be collected?
(2) What mechanism is going to be used to collect that data?
(3) How is that data going to be analyzed and applied to meet the objectives of a breeding or genetics program?

6.4 Field Spectroscopy

In terms of what data should be collected, there are many options. Field spectroscopy is the most common technique used in field-based phenotyping systems due to the speed with which the data can be collected and the quantitative nature of the information itself. Spectroscopy is the study of the interaction between electromagnetic radiation and matter. Field Spectroscopy, specifically, is the measurement of the reflectance properties of soils, vegetation, rocks, and bodies of water under natural light. It is used for quantitative measurement of radiance, irradiance, reflectance, or transmission in the field; the resulting image-based measurements are converted to estimations of relevant phenotypes through an established calibration. Some the examples are RGB/CIR cameras, multispectral cameras, hyperspectral imagers, thermal imaging cameras, and even conventional handheld digital cameras (Araus and Cairns 2014). However, there are challenges associated with the collecting of spectroscopy data that, if not addressed, pose barriers to the precision and accuracy of the information a researcher collects.

Milton (1987) proposed several guidelines to help a researcher ensure quality data from a spectroscopy experiment. To increase the accuracy and consistency of field data, several methodologies have been suggested by various authors (Jackson and Robinson 1985; Kimes and Kirchner 1982). The following are the guidelines provided by Milton (1987) to standardize field spectroscopy experiments across different research groups:

(1) To ensure a fixed geometry between the sensor, the standard panel, and the target, a mast or tripod should be used. The variable geometry and the proximity of the operator to the target and of the target to the radiometer leads to the reduced precision of the handheld measurements.
(2) The sensor should be at least 1–2 m above the upper surface of the target.
(3) The sensor support should be pointed directly towards the sun. This is to be consistent with orienting sensor horizontal support and positioning other field equipment in same positions relative to the sun.
(4) The standard panels should be placed in such a way that it fills the field-of-view of all the bands of the sensor.
(5) While taking the measurements, use of a continuously recording solarimeter within the field area provides three benefits:

- All data from the primary sensor will be screened and any anomalies in them will be corrected.
- It helps to quantify the atmospheric variability on a range of time scales.
- The effects of variable cloud cover can be corrected using the data obtained from solarimeter.

(6) Operators are advised to wear dark clothing while taking measurements. Errors in the measurement of radiance in red and near-infrared wavelengths can be as high as 10 and 12 %, respectively, if an operator with white clothing approaches the target at 0.5 m. However, if the operator wears black clothing, the errors in both bands were reduced to less than 2 % (Kimes and Kirchner 1982). The distance from the target and the vehicles should be more than 3 m for the same reason.

Of course, as a general guideline, Milton (1987)'s principles are useful, but it would behoove any researcher who is thinking of seriously investing in an HTP experiment to conduct pilot studies to optimize the collection of the phenotype data to the specific conditions of the activity. Such pilot studies form the basis of the establishment of a truly empowering HTP platform.

An essential part of the phenotypic optimization process will be to determine the optimum wavelength of light to use as the spectroscopic metric. The most widely used field spectroscopy-based techniques are remote sensing and imaging, infrared thermometry and thermal imaging, and near infrared spectroscopy (NIRS) analysis (Araus and Cairns 2014).

6.4.1 Remote Sensing and Imaging

remote sensing refers to collection of data about an object or phenomenon without any physical contact with the object. It usually refers to the data collected through satellites using different sensors of various platforms with a wide range of spatio-temporal, radiometric, and spectral resolutions (Khorram et al. 2012). Remote sensing instruments function as either passive or active sensors.

Passive sensors are used to measure the natural radiation emitted or reflected by objects. The most common source of radiation detected by Passive sensors is reflected sunlight. Most of the Passive sensors operate in the visible, infrared, thermal infrared, and microwave portions of the electromagnetic spectrum. Some examples of Passive sensors include accelerometers, radiometers, imaging radiometers, spectrometers, spectro-radiometers, and hyperspectral radiometers.

Unlike Passive sensors, an active sensor emits its own radiation towards the target of interest. The reflected or backscattered radiation from the target is detected by the active sensor to predict the properties of the target. Most of the active sensors operate in the microwave portion of the electromagnetic spectrum. Some of the examples of active sensors are RADAR (sound), LIDAR (light), a scatterometer, or a laser altimeter (Allen 1998).

The electromagnetic waves emitted by water sources, vegetation, and fields from earth can be captured by these sensors embedded in satellites and used to detect their characteristics. Plants are easily distinguishable in these systems because they exhibit stronger radiation in the electromagnetic spectra of visible-near infrared (VIS-NIR). Field-based HTP approaches using remote sensing use this VIS-NIR data to infer attributes of plants or field conditions. This technique can be particularly valuable because it is possible to phenotype the same individuals multiple times across the growing season, as the methods are both noninvasive and nondestructive. remote sensing can measure a wide array of traits including green biomass, photosynthetic transpirative gas exchange, and even grain yield (Araus et al. 2008). Phenotypic data generated from remote sensing can be used for a broad range of breeding experiments related to increasing yield potential or tolerance to abiotic and biotic stress factors.

6.4.2 Infrared Thermometry and Thermal Imaging

Infrared thermometers can be used to measure the temperature of an object by capturing the thermal radiation it emits. They are typically attached to the vehicles used in HTP systems as sensors (Furbank and Tester 2011). Long-wave infrared cameras and thermal imaging cameras are used to predict a variety of traits in crop plants, including water stress, disease and pathogen detection, and the detection of bruises on fruits and vegetables. These cameras can be attached either to aerial platforms or proximal sensing platforms, depending on the needs and funds of the researcher. High-resolution infrared cameras can enhance the quality of data and also reduce the time required to collect that data compared to lower resolution cameras (Araus and Cairns 2014).

6.4.3 Near-Infrared Spectroscopy

NIRS uses the waves emitted from the near-infrared region of the electromagnetic spectrum to analyze proteins, nitrogen, starch, and oil content, as well as grain texture and grain weight of individual seeds (Cabrera-Bosquet et al. 2011). This technique aids in evaluating large numbers of plot material and covers a broad distribution of measurements within plots. In addition, NIRS can be also used to measure more complex traits, such as stress adaptation, by using prediction models to correlate physiochemical properties to stresses of interest (Araus and Cairns 2014). Even though the precision of these estimates might be lower than the data points obtained from direct analysis, the lower cost of estimation often justifies its use. For example, in early-generation plant breeding material where thousands of genotypes need to be screened with relatively low selection intensity, this approach can help in identifying the bottom end of a phenotypic distribution in a reasonably cost-effective manner.

6.5 Data Collection Platforms

Even after the question of what data to collect is answered, the mechanism for collecting that data still remains a challenge for researchers. The efficiency and accuracy of collecting data from the field depends heavily on the mechanism for moving the sensors along the field. It is important to maintain consistency in data collection and—especially in the case of time series experiments—not to damage plants. There are two obvious and major methods used to collect data across fields: ground-based platforms and aerial platforms.

6.5.1 Ground-Based Phenotyping Platforms

Several ground-based high-throughput phenotyping systems have been developed over the past several years. These are usually modified vehicles equipped with global positioning systems (GPS) and sensors, often referred to as "phenomobiles" (Araus and Cairns 2014). In an effort to balance cost, accuracy, and relevance, researchers have strived to develop both expensive as well as very cost-effective phenomobiles. Some examples of data collection using phenomobiles are as follows:

- Ruixiu et al. (1989) used a three-wheeled cart with mounted multiple ultrasonic proximity sensors for measuring the morphological characters of bush-type plants (Fig. 6.2a).
- Rundquist et al. (2004) used an all-terrain motorized platform with a boom extendable to 12 m. The boom was mounted with a spectroradiometer that was used to estimate the biophysical characteristics of vegetation in a corn field (Fig. 6.2b).
- White and Conley (2013) used a hand cart that was built by welding two bicycle frames to a steel scaffold. It was capable of positioning two monochrome cameras and three infrared thermometers over two rows (Fig. 6.2c).
- Andrande-Sanchez et al. (2013) used a near infrared spectrometer sensor mounted on a tractor to measure canopy height, reflectance, and temperature (Fig. 6.2d).

One of the most commonly used phenomobiles is the combine/harvester. It is used to collect data on grain yield, moisture, and, in many cases, grain quality. Several other tractor- and harvester-mounted sensors are used for collecting data from field experiments. Multiple sensors are attached to these phenomobiles and are used in the detection of various traits, such as grain yield, crop canopy height, leaf area index (LAI), NDVI, multispectral imaging, and hyperspectral reflectance (Comar et al. 2012; Montes et al. 2011).

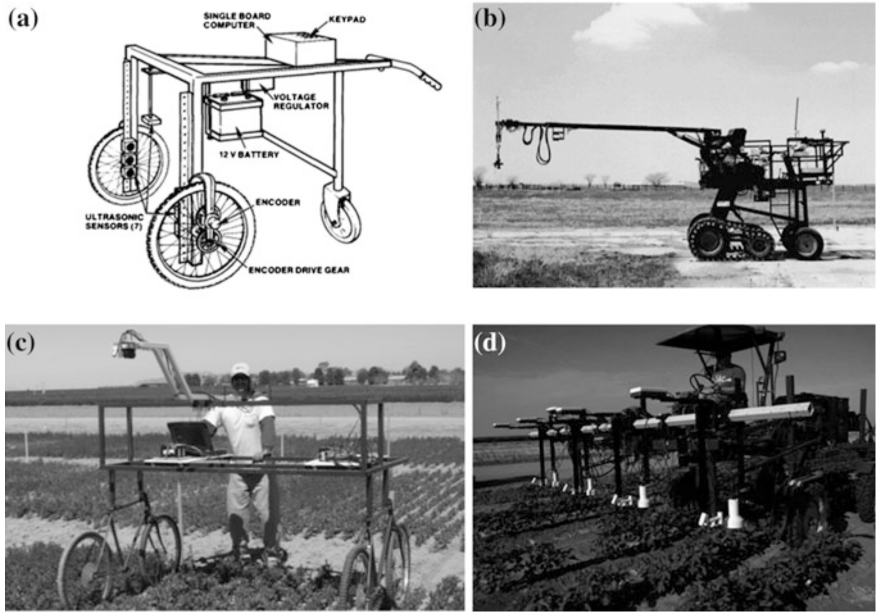


Fig. 6.2 Phenomobiles used in field-based phenotyping systems. **a** Three wheeled cart (Ruixiu et al. 1989); **b** All-terrain motorized platform (Rundquist et al. 2004); **c** Hand cart made by welding two bicycle frames (White and Conley 2013); Tractor mounted with sensors for high-throughput phenotyping (Andrande-Sanchez et al. 2013); (used with permission.)

6.5.2 Unmanned Aerial Platforms

Aerial platforms for field-based phenotyping are particularly useful for noninvasive measurements of a large number of phenotypes quickly and over large distances. However, the throughput offered by these platforms often comes at the cost of accuracy. Various types of carriers are used for aerial phenotyping, including unmanned polycopters, blimps, and small unmanned helicopters (Fig. 6.3).

Chapman et al. (2014) used a “pheno-copter” (Fig. 6.3a), a gas-powered robotic helicopter, to collect field data using three different cameras. The device was mounted with a visual camera, NIR camera, and a thermal camera. The three different types of images were captured using a GNC system and image capture computer, which allowed the pheno-copter to measure ground cover in sorghum, canopy temperature in sugarcane, and three-dimensional measures of crop lodging in wheat. One of the main advantages of using a pheno-copter is that it can fly low at a distance of up to 10–40 m above the field, which allows it to collect high-resolution image data for every plot and also lets the user collect data from many numbers of plots in a limited amount of time. Figure 6.3a depicts the pheno-copter platform, which can be flown with a radio transmitter or touchscreen control, and the ground station (Chapman et al. 2014).

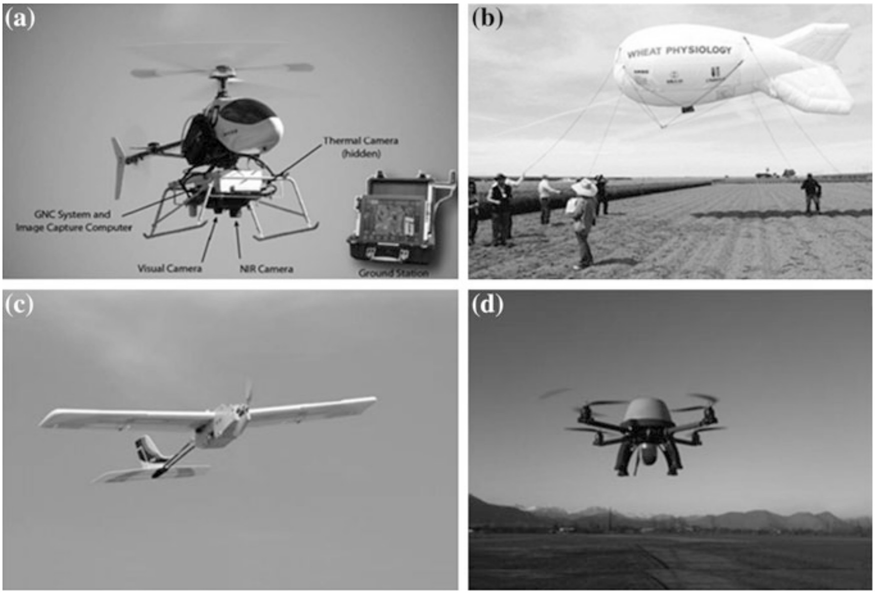


Fig. 6.3 Various aerial platforms used for collecting phenotypic data in field. **a** Pheno-Copter. **b** Helium filled blimp. **c** The Lancaster Hawkeye Mark II. **d** Octane multi-rotor system. (Chapman et al. 2014; Goth 2014; Hunting 2013; used with permission.)

The Centro Experimental de Norman E. Borlaug (CENEB) station, near Ciudad Obregon, associated with CIMMYT in Mexico, used an AB1100, 8-m-long helium-filled blimp (Fig. 6.3b) to evaluate wheat plants for abiotic stress tolerance during flowering time. The blimp was capable of floating up to 300 m above the fields while thermal infrared imagers and multispectral cameras were attached to the blimp for collecting data on canopy temperature, stomatal conductance, canopy water content, vegetation indices, and pigment indices (Goth 2014).

The Lancaster Hawkeye Mark II (Fig. 6.3c) is an unmanned aerial vehicle (UAV) built by the Precision Hawk Company. This vehicle is a completely autonomous fixed-wing system. It weighs only 3 pounds but can survey about 300 acres in 40 min. The company offers imagery, video, thermal, and multispectral imaging. Onboard sensors also help in adjusting the device for weather variability. The field information is programmed into the device before being tossed into the air by the user. The device then collects the phenotypic data and returns on its own. Additionally, it also supports an onboard Wi-Fi system for wireless data transmission (Hunting 2013).

The Octane multirotor system (Fig. 6.3d) is a bit heavier than Lancaster Hawkeye Mark II, but it is capable of taking high-resolution, GPS-referenced imagery and video. Manufactured by Volt Aerial Robotics, the company claims that the system is easier to operate under field conditions than the fixed-wing UAVs (Frey 2013). In addition to these examples, other commonly used UAVs include the

Astec Falcon 8 (an 8-rotor UAV), the Airelectronics Skywalker, the Yamaha RMAX, and the Penguin series UAVs and unmanned aircraft systems (Hunting 2013).

6.6 Root Phenotyping Platforms

Any treatment of HTP platforms would be incomplete without at least making mention of one of the greatest phenotyping challenges of the modern age. One of the most difficult parts of the plant to be phenotyped in the field is the roots. However, roots form often up to 50–60 % or more of total plant biomass and are essential to understand such agronomically important traits as drought tolerance, nutrient use efficiency, and lodging resistance. Scientists have increasingly used ground-penetrating radar (GPR) to study root characteristics in plants over the past few years. Figure 6.4 depicts a researcher using GPR to measure wheat root biomass (Tanikawa et al. 2013).

GPR is essentially a broadband electromagnetic pulse radar system that is used to predict the depth, size and position of matter buried in the soil using the timing and characteristics of the reflected waves. A transmitting antenna from the GPR transmits pulses of electromagnetic energy, which penetrate the soil. This energy is reflected by the boundary layers of the objects with different physical characteristics. The receiving antenna in the GPR intercepts the reflected waves and the variation in the reflected wavelength is caused due to the contrasts in the dielectric permittivity between the bulk medium and buried objects (Barton and Montagu 2004).

Both soil depth and root water content affect the reflected radar waves, making it harder to get accurate data. However collection of data in sandy soils rather than clay have been shown to be more accurate (Borem and Fritsche-Neto 2014; Hirano et al. 2009).

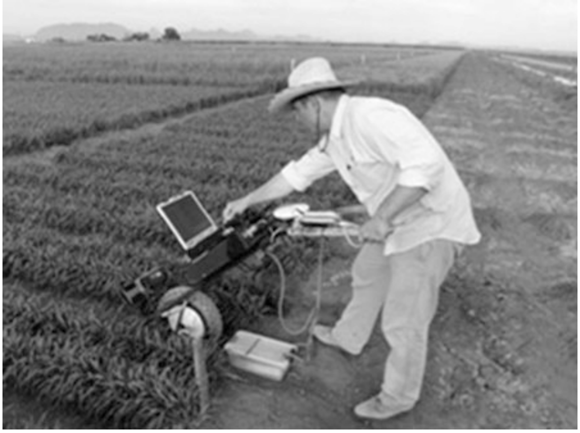


Fig. 6.4 This figure depicts a researcher using GPR to measure wheat root biomass (Thompson et al. 2013; used with permission)

6.7 Commercially Available Field-Based Phenotyping Systems

Apart from the various methodologies used by the public institutions to phenotype plants efficiently, several private companies have capitalized on the need to develop field-based phenotyping systems that are both powerful and accurate. The following sections discuss some of the commercially available field-based phenotyping platforms.

6.7.1 Scanalyzer^{Field}

Scanalyzer$^{\text{Field}}$ is a phenotyping platform built by the Lemna-Tech Company. The company has been in the business of building high-throughput phenotyping platforms since 1998, and most of their platforms have been dedicated to greenhouse and growth chamber-based phenotyping systems. However, due to the need to measure agronomically important traits under field conditions, they have developed the Scanalyzer$^{\text{Field}}$ platform, which is depicted in Fig. 6.5.

The Scanalyzer$^{\text{Field}}$ platform is built around a stationary portal crane system that acts as a support for various sensors in the field. The system allows the user to phenotype plants from a height of 3–6 m over several hundred square meters of area, with a maximum size of 10 × 40 × 6 m for the stationary portal crane (Scanalyzer$^{\text{Field}}$ 2014).

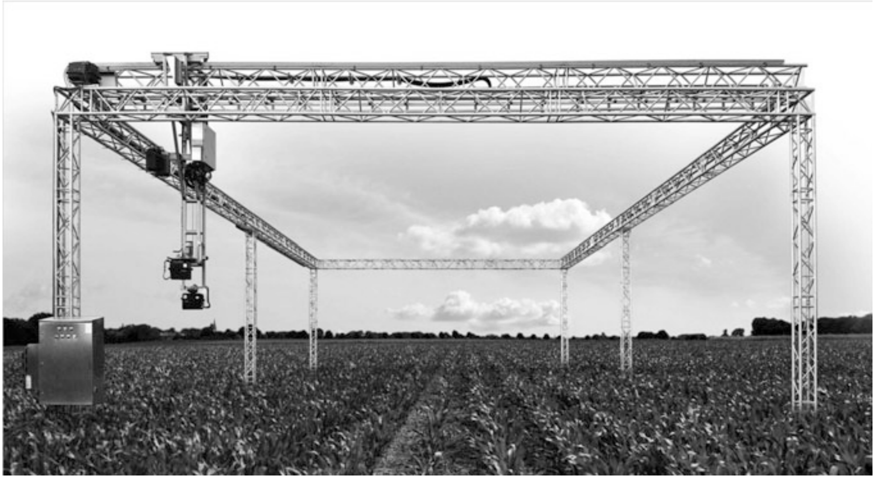


Fig. 6.5 The Lemnatech Scanalyzer$^{\text{Field}}$ platform for field-based phenotyping (Schwartz et al. 2013; used with permission.)

Even though the data collected from the system is more precise than the various aerial platforms, the portability and cost of the stationary portal crane makes the equipment less accessible for public research projects.

6.7.2 Smartfield ™

SmartField is a company that manufactures equipment designed to track relevant metrics of the field environment. This affords them the advantage of marketing both to researchers and to growers. Figure 6.6a shows a SmartProfile™ instrument loaded with four soil moisture sensors and four soil temperature sensors. Even though it is not used to measure field data on plants directly, it can be used to monitor stability and range of environmental conditions the plants are exposed to over time.

Figure 6.6b depicts an image of a SmartCrop® sensor that has an infrared thermometer to collect the canopy temperature of plants in a single field. The company's SmartCrop® sensor is capable of collecting canopy temperature readings every minute and averaging them at 15-min intervals and reports the data to a base station. Like SmartCrop®, the SmartWeather™ system collects field data on wind speed, wind direction, solar radiation, and barometric pressure in conjunction with the ambient temperature, relative humidity, and rainfall data collected by the base station (Smartfield™ 2014).

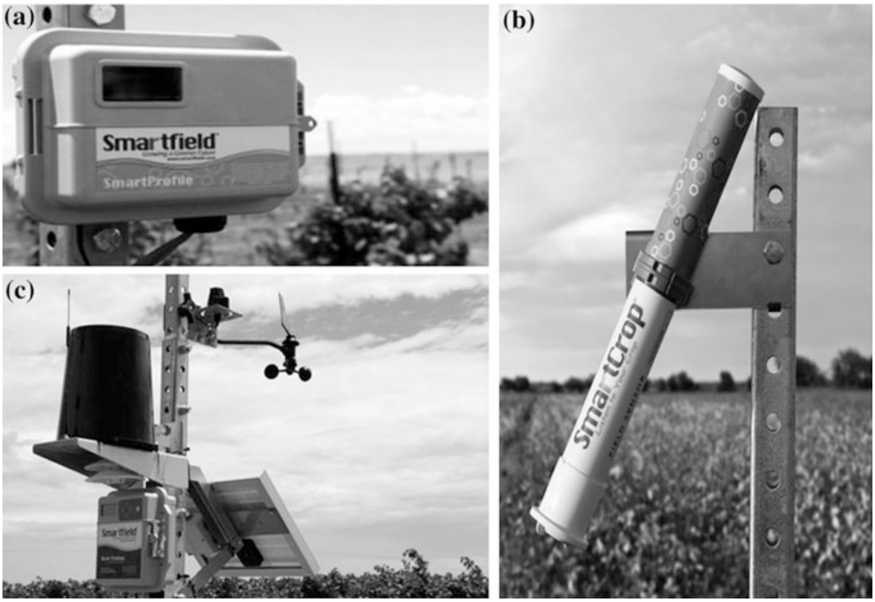


Fig. 6.6 Smart Crop® sensor **a** SmartProfile™ **b** SmartCrop® **c** SmartWeather™ (Smartfield 2014; used with permission.)

One of the most challenging components of phenotypic information for a researcher to understand is genotype–environment interaction. This is the phenomenon of adaptation that gives some genotypes favorable phenotypic expression in one environment, but allows it to be outcompeted by other genotypes when the environment changes. Critical to the understanding of genotype–environment interaction in any system is accurate and timely collection of environmental data that allows for the identification of the mechanisms that drive it. The Smartfield™ products are one of many market-based solutions for understanding a target population of environments.

6.8 Conclusion

In an age where genotype information is becoming cheaper, more reliable, and more readily available to smaller groups, and where the importance of phenotypic selection to plant breeding is often overlooked in favor of marker-based selection techniques, innovations in field-based HTP have the opportunity to accelerate the development of high-yielding, stress-tolerant, and disease-resistant plant varieties. Field-based HTP reduces the time and increases the accuracy of the parameters estimated from phenotypic data and allows for the selection of varieties and the identification of genetic loci with a precision not seen prior to the 21st century. However, just like in the case with high volumes of cheap genotype data, sorting through high volumes of cheap phenotype data will pose a significant logistical challenge to the scientific community. These challenges will be exacerbated by a dynamic and changing environment that is constantly shaping patterns of genotype–environment interaction and undermining the practices of selection and mapping. Meeting these challenges will take coordinated effort across groups and stable dedicated funding sources committed to the success of agriculture and plant breeding. Furthermore, these funds must be spent wisely to ensure that only the most relevant data are collected and employed for making genetic gain.

Looking to the future, once high volumes of genotype and phenotype data are available on genetically stable and agronomically important shared germplasm resources, high-throughput modeling of changes in the environment and their effects on the phenotype will be needed to complete the triangle and truly empower the mapping of genotype to phenotype.

References

Acquaah G (2007) Principles of plant genetics and breeding. Blackwell Publisher, Malden

Allen B (1998) Remote Sensing and Lasers. Accessed 17 June 2014 http://www.nasa.gov/centers/langley/news/factsheets/RemoteSensing.html

Andrade-Sanchez P, Gore M, Heun J, Thorp K, Carmo-Silva A, French A, Salvucci M, White J (2013) Development and evaluation of a field-based high-throughput phenotyping platform. Funct Plant Biol 41(1):68–79

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19(1):52–61

Araus J, Slafer G, Royo C, Serret M (2008) Breeding for yield potential and stress adaptation in cereals. Crit Rev Plant Sci 27(6):377–412

Asoro F, Newell M, Beavis W, Scott M, Jannink J (2011) Accuracy and training population design for genomic selection on quantitative traits in elite north american oats. Plant Genome 4 (2):132–144

Barton CVM, Montagu KD (2004) Detection of tree roots and determination of root diameters by ground penetrating radar under optimal conditions. Tree Physiol 24(12):1323–1331

Borém A, Fritsche-Neto R (2014) Omics in plant breeding. Ames: Wiley Blackwell

Cabrera-Bosquet L, Sanchez C, Rosales A, Palacios-Rojas N, Araus J (2011) Near-infrared reflectance spectroscopy (NIRS) assessment of delta O-18 and nitrogen and ash contents for improved yield potential and drought adaptation in maize. J Agric Food Chem 59(2):467–474

Chapman S, Merz T, Chan A, Jackway P, Hrabar S, Dreccer F, Holland E, Zheng B, Ling J, Jimenez-Berni J (2014) Pheno-copter: a low-altitude, autonomous remote-sensing robotic helicopter for high-throughput field-based phenotyping. Agronomy 4:279–301

Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. Theor Appl Genet 126(4):867–887

Comar A, Burger P, de Solan B, Baret F, Daumard F, Hanocq J (2012) A semi-automatic system for high throughput phenotyping wheat cultivars in-field conditions: description and first results. Funct Plant Biol 39(10–11):914–924

Frey T (2013) Agriculture, the new games of drones. The futurist—a magazine of forecasts, trends and ideas about the future

Furbank RT, Tester M (2011) Phenomics–technologies to relieve the phenotyping bottleneck. Trends Plant Sci 16(12):635–644

GEF/UNCCD (2011) Land for life: securing our common future. Global Environment Facility, Washington, DC

Goth B (2014) Behind the science: researcher helps remote sensing soar. Accessed 17 July 2014 http://blog.cimmyt.org/behind-the-science-researcher-helps-remote-sensing-soar/

Heffner EL, Sorrells ME, Jannink J (2009) Genomic selection for crop improvement. Crop Sci 49 (1):1–12

Hirano Y, Dannoura M, Aono K, Igarashi T, Ishii M, Yamase K, Makita N, Kanazawa Y (2009) Limiting factors in the detection of tree roots using ground-penetrating radar. Plant Soil 319 (1):15–24

Hosoi F, Omasa K (2007) Factors contributing to accuracy in the estimation of the woody canopy leaf area density profile using 3D portable lidar imaging. J Exp Bot 58(12):3463–3473 doi: 10.1093/jxb/erm203

Hunting K (2013) Do you really know your fields? http://farmindusrynews.com/site/farmindusrynews.com/files/uploads/2013/07/FIN_22-25_UAVs.pdf

Jackson RD, Robinson BF (1985) Field evaluation of the temperature stability of a multispectral radiometer. Remote Sens Environ 17(1):103–108

Khorram S, Koch FH, van der Wiele CF, Nelson SAC (2012) Remote sensing. Springer, Boston

Kimes DS, Kirchner JA (1982) Irradiance measurement errors due to the assumption of a lambertian reference panel. Remote Sens Environ 12(2):141–149

McCouch SR, McNally KL, Wang W, Hamilton RS (2012) Genomics of gene banks: a case study in rice. Am J Bot 99(2):407–423

Melillo JM, Richmond TC, Yohe GW (eds) (2014) Climate change impacts in the united states: the third national climate assessment. U.S. Global Change Research Program, Washington, DC

Milton EJ (1987) Review article principles of field spectroscopy. Int J Remote Sens 8(12):1807–1827

Montes JM, Technow F, Dhillon BS, Mauch F, Melchinger AE (2011) High-throughput non-destructive biomass determination during early plant development in maize under field conditions. Field Crops Res 121(2):268–273

Munns R, James RA, Sirault XRR, Furbank RT, Jones HG (2010) New phenotyping methods for screening wheat and barley for beneficial responses to water deficit. J Exp Bot 61(13):3499–3507

Ruixiu S, Wilkerson JB, Wilhelm LR, Tompkins FD (1989) A microcomputer-based morphometer for bush-type plants. Comput Electron Agric 4(1):43–58

Rundquist D, Perk R, Leavitt B, Keydan G, Gitelson A (2004) Collecting spectral data over cropland vegetation using machine-positioning versus hand-positioning of the sensor. Comput Electron Agric 43(2):173–178

Scanalyzer$^{\mathrm{Field}}$ (2014) A wide range of choices for the quantitative, non-destructive analysis of different crops or model plants in field conditions [Internet]. Accessed 7 July 2014 http://www.lemnatec.com/products/hardware-solutions/scanalyzer-field/#.U7r4yvldV8E

Smartfield™ (2014) Growing a greener future: equipment [Internet]. Accessed 7 July 2014 http://www.smartfield.com/smartfield-products/equipment/

Stefan S, Jörge V, Dirk V, Matthias E (2013) Digital phenotyping of field crops under field conditions. LemnaTec—plant phenomics workshop, 1/11/2013. Plant and Animal Genome XXI, San Diego

Tanikawa T, Hirano Y, Dannoura M, Yamase K, Aono K, Ishii M, Igarashi T, Ikeno H, Kanazawa Y (2013) Root orientation can affect detection accuracy of ground-penetrating radar. Plant Soil 373(1):317–327

Thompson SM, Cossani CM, Ibrahim AMH, Reynolds MP, Goodman D, Hays DB (2013) Estimating wheat root biomass using ground penetrating radar. ASA, CSSA & SSSA International Annual Meetings, 9/3, Tampa, Florida, USA. ASA, CSSA & SSSA International Annual Meetings, Tampa

Tuberosa R (2012) Phenotyping for drought tolerance of crops in the genomics era. Front Physiol 3:347

United Nations (2013) World population prospects: the 2012 revision, key findings and advance tables. Department of Economic and Social Affairs, New York

Van Beem J, Smith ME, Zobel RW (1998) Estimating root mass in maize using a portable capacitance meter. Agron J 90(4):566–570

White J, Conley M (2013) A flexible, low-cost cart for proximal sensing. Crop Sci 53(4):1646–1649

Zenone T, Seufert G, Morelli G, Teobaldelli M, Fischanger F, Matteucci M, Sordini M, Armani A, Ferre C, Chiti T (2008) Preliminary use of ground-penetrating radar and electrical resistivity tomography to study tree roots in pine forests and poplar plantations. Funct Plant Biol 35(10):1047–1058

Chapter 7
Disease Phenomics

Éder A. Giglioti, Ciro H. Sumida and Marcelo G. Canteri

Abstract Phenomics of plant disease defined as a large-scale study linking genomics with plant resistance and pathogenicity or pathogen aggressiveness. Performed by a set of new technologies such as automatic measurements, remote sensing, and image processing. Study of plant breeding disease resistant, the understanding of phenotypic characteristics and their change depending on genotype and environment interactions is essential to the studies success. Use of resistant cultivars is one of the best options among the strategies for disease control to be applicable over large areas and be economically viable. The future prospects in relation to Phenomics to improve plant resistance to disease are promising, promises to revolutionize plant breeding, accelerating the production of cultivars resistant to multiple diseases, whether through assisting the choice of parents, molecular marker-assisted selection, or the introduction of resistance of resistance genes by genetic engineering. Numerous automated tools have been developed and tested, which could be used for phenotyping on a large scale to assess disease under field conditions in order to optimize the growing demand of research related to the production of new cultivars, in order to increase yield and cultivars resistant to diseases, to support future agricultural production and ensure food security.

É.A. Giglioti
Adamantinenses Integrated College, Adamantina, Brazil
e-mail: edergiglioti@smartbiotecnologia.com.br

C.H. Sumida · M.G. Canteri (✉)
Agronomy Department, Londrina State University, Londrina, Brazil
e-mail: canteri@uel.br

C.H. Sumida
e-mail: cirosumida@hotmail.com

R. Fritsche-Neto and A. Borém (eds.), *Phenomics*,
DOI 10.1007/978-3-319-13677-6_7

7.1 Importance of Disease Resistance and Large-Scale Phenotyping

The constant need to improve and promote sustainable production to ensure a sufficient quantity of both human food and raw materials for the production of agroenergy is among the great challenges of modern agriculture. Given the limited area available for agricultural expansion, the production of new high-yielding cultivars that are adapted to various climate, soil, and management conditions and that exhibit resistance or tolerance to biotic and abiotic stresses will greatly contribute to overcoming this challenge. Productivity-inhibiting factors, including plant diseases, must be increasingly studied and controlled to prevent annual losses caused in various crops worldwide. The production and use of resistant cultivars represents the most rational and economical strategy for disease control because it is applicable to large areas, is highly effective against pathogens when well managed, and poses no risk to human health or harm to the environment. Thus, the use of resistant cultivars is always the first option among the various control measures and often suffices to control the losses caused by the main diseases that affect a specific crop and, in other cases, to assist with other complementary phytosanitary measures. Thus, resistance to disease is considered a “key agronomic trait” of new cultivars, and their development should be among the priorities of any breeding program.

Resistance is the rule and susceptibility the exception in natural pathosystems. Otherwise, a single pathogen could attack all plants of a species, which would succumb in its ecosystem. Susceptibility only occurs when parasitism is established, determining genetic compatibility between both through close relationships, which results from co-evolution between the pathogen and host. Consequently, genetic changes in the host population are accompanied by genetic changes in the pathogen population, and vice versa. In the struggle for survival, such genetic changes provide new pathogenic attack mechanisms and, conversely, the host responds with new defense mechanisms. Thus, both vertical and horizontal resistance (*plant phenotype*) may be expressed during various growth phases of a crop as a result of the complex and dynamic interactions among plant genetics (*plant genotype*), the genetics of pathogen virulence or aggressiveness (*pathogen genotype*), the physicochemical world within which both develop (*environment*), and the effect humans have on each of those three factors through their management activities (*management*; Fig. 7.1).

The success of genetic breeding therefore depends on a deep understanding of the complex interactions among plant, pathogen, host, and human. Hence, a technology race that has been occurring in recent years has resulted in the development of large-scale methods and equipment for gene sequencing and functional analyses. However, the effective use of this knowledge for realizing concrete benefits for genetic breeding towards disease resistance has been elusive, despite these recent advances and the large amount of information available in databases of nucleotides, genes, and partial or complete genome sequences of host plants and their respective pathogens.

What has happened? The answer is simple: while genomics has been evolving fast and continuously and phenotyping has lagged behind, *phenomics* has emerged.

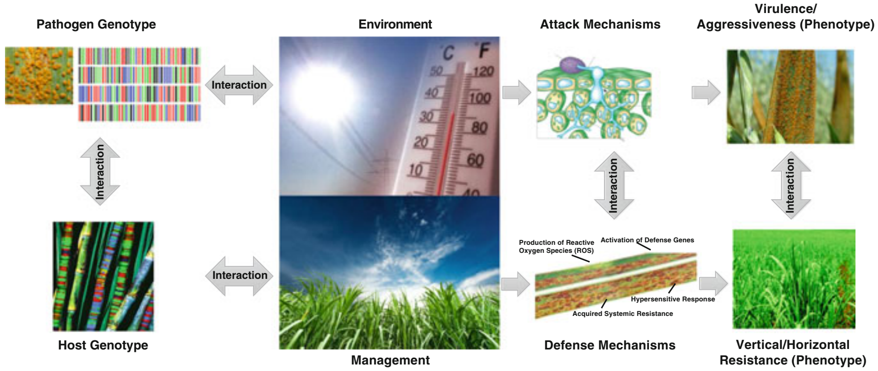
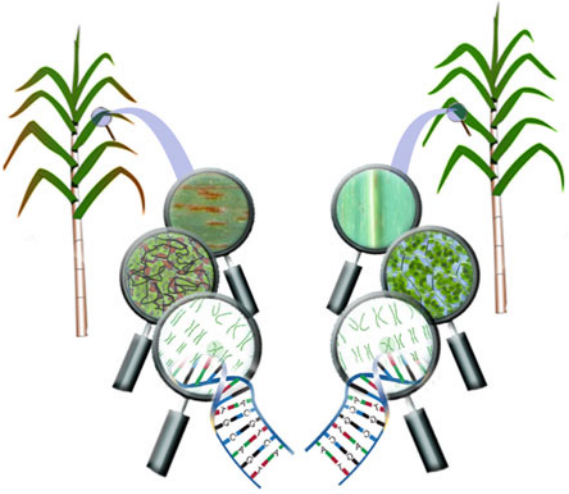


Fig. 7.1 Interactions resulting from the co-evolution between hosts and pathogens that determine the pathogenicity and resistance in the armed struggle they both wage for survival

Phenomics is the "area of science, which seeks to characterize [in large scale] phenotypes (***phenotyping***) in a rigorous and formal way, and link these traits to the associated genes and gene variants (alleles)" (Fig. 7.2). The missing piece to the puzzle is precisely that which seems easiest—namely, the development of methodologies, software, and platforms for large-scale phenotyping.

A new generation of phenotyping studies has benefited significantly from advances in platforms, methodologies, automation, georeferencing, control of experimental conditions, volume of collected data, software, precision and accuracy, reproducibility, complexity, and especially scalability, compared to the conventional method of using diagrammatic scales. Thus, fast, precise, and accurate evaluations of thousands of plants are possible. Hence, these new phenotyping capabilities promise to revolutionize plant breeding, accelerating the production of cultivars resistant to multiple diseases, whether through assisting in the choice of parents, molecular marker-assisted

Fig. 7.2 Phenomics of plant diseases, defined herein as large-scale studies linking genomics with plant resistance and pathogenicity or pathogen aggressiveness



selection, or the introduction of resistance genes by genetic engineering. Therefore, this chapter introduces the main methodologies, devices, software, and platforms available for the large-scale phenotyping of resistance to plant diseases.

7.2 Methodologies and Image Processing Tools for Large-Scale Phenotyping of Plant Diseases

The reflectance of sunlight in the visible (VIS, 400–700 nm), near-infrared (NIR, 700–1,100 nm) and shortwave infrared (SWIR, 1,100–2,500 nm) spectra are directed by multiple light–leaf interactions: (i) absorption of radiant energy by the leaf chemistry, (ii) light diffusion resulting from the leaf surface and internal cellular structures, and (iii) absorption of radiant energy induced by the leaf water content. These leaf characteristics are affected by plant diseases through changes in the physiology of diseased plants, including reduced photosynthesis, increased transpiration, reduced chlorophyll, increased anthocyanin content, decreased tissue water content, tissue death, onset of symptoms, and presence of pathogen structures, among others. Thus, there are considerable differences in the reflectance, absorption, diffusion, and transmission of light between healthy and diseased tissues (Fig. 7.3). Those differences enable the identification and quantification of

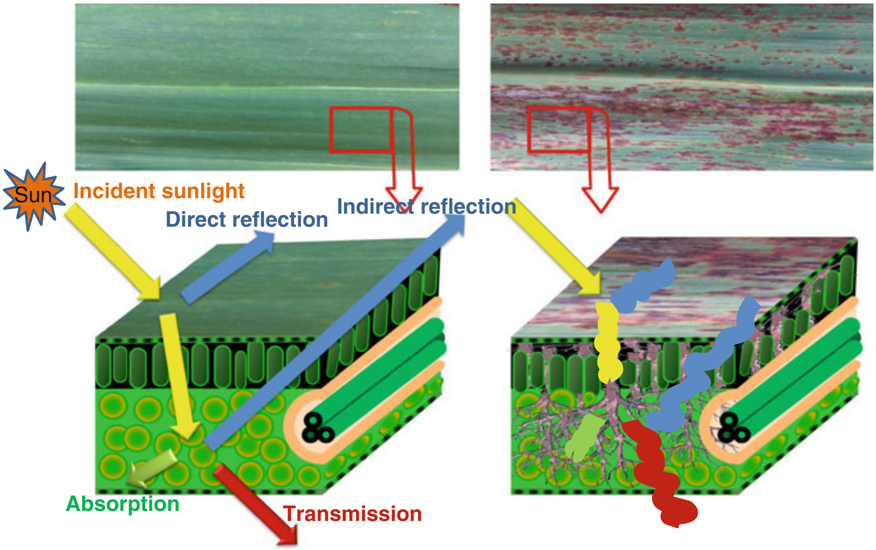


Fig. 7.3 Reflection, absorption, and transmission processes characteristic of the interaction between sunlight and either healthy sugarcane leaves or leaves with symptoms of Orange Rust (*Puccinia kuehnii*). *Left* Healthy leaf without compromised light–tissue interaction. *Right* Diseased tissue, indicating that the reflection of light will be altered

diseases through the assessment of the light–leaf interaction, usually conducted in the VIS, NIR, and SWIR spectral regions.

Based on this principle, phenotyping methods of plant diseases have been recently developed based on the acquisition, processing, and analysis of images that are generated from the referenced regions of the electromagnetic spectrum. Table 7.1 provides a brief summary of image types, their respective advantages and disadvantages, examples of their use for the evaluation of plant resistance to diseases, and software for the analysis and interpretation of imaging data. For further insight on the matter, please refer to general reviews of advanced methods for the detection and quantification of diseases (Bock et al. 2010; Sankaran et al. 2010), sensors (Naue et al. 2010), software (Bock et al. 2010), and algorithms (Barbedo 2013).

7.2.1 Analysis of Images in the Visible Spectrum

Color images in the visible spectrum are multichannel images formed by the additive color system red, green, and blue (RGB). RGB image processing (IP) systems are usually based on five stages: image acquisition, preprocessing, segmentation, representation and description, and recognition and interpretation. This IP process is the simplest, demanding the least complex software and cheaper sensors and devices, and is the most popular system. Thus, whenever possible, the first choice for large-scale phenotyping of plant resistance to pathogens in breeding programs is RGB imaging.

The images acquired to perform large-scale phenotyping may be stored or uploaded in real time to an image database. After removing the noise, adjusting the contrast or brightness, and smoothing specific image properties, the improved RGB image is then nonlinearly converted into the HSV color system formed by the hue (matrix), saturation, and value components, also known as HSB (B for brightness). Other color systems may be used, including HSL (L for luminosity) and HSI (I for intensity). Then, areas of interest contained in the image are identified and extracted to enable the representation, description, segmentation, and quantification of the ratio between healthy and diseased tissues—that is, the disease severity.

Figure 7.4 contains example plots of component *H* (matrix) values that were obtained from the sunlight–sugarcane leaf interaction for healthy tissues and diseased tissues with symptoms of Orange Rust (*Puccinia kuehnii*). Diseased tissues are characterized by *H* values ranging from 0 to 45 and 260 to 360, while the values of healthy tissues range from 60 to 190. The *S* values range from 0 to 40 and 8 to 60, while the B values range from 38 to 75 and 30 to 100 in the healthy and diseased tissues, respectively. Therefore, the HSB components may be used for phenotyping Orange Rust because they allow for the differentiation between healthy and diseased tissues. The *H* intervals of the healthy and diseased tissues can then form the basis for the analysis of RGB images of sugarcane plots in the field with varying levels of Orange Rust severity. Those images can be acquired and stored using a remote-controlled drone as a large-scale phenotyping platform.

Table 7.1 Types of images, respective advantages and disadvantages, examples of their application, and image processing software used in several studies for quantifying the severity of diseases in leaves, plants, and plots

Image	Advantages	Disadvantages	Pathosystems	Software[a]
RGB	• Fast, accurate and reliable when well calibrated • Powerful when automated • Technology • Low-cost acquisition and analysis devices • Existence of software adapted to applications and needs specific for disease quantification	• Low cost and popularity • Difficulty in dealing with the color variation from one plant to another and various artifacts and failures resulting from image acquisition • Not yet optimized to handle various diseases simultaneously and multiple diseases • Features precision and accuracy required for correct quantification of diseases	• Barley-*Erisyphe graminis* (Newton 1989) • Maize-*Fusarium* spp. (Todd and Kommendahl 1994) • Potato- (Niemira et al. 1999) • Maize-maize streak virus (Martin et al. 1999) • Oats-*Puccinia coronata* (Díaz-Lago et al. 2003). Vine-*Plamopara vitícola* (Boso et al. 2004) • Rye and triticale-*Magnaporthe oryzae* (Maciel et al. 2013) • Sugarcane-*Cercospora longipes* (Patil and Bodhe 2011) • Soybean-seed pathogens (2014)	• ASSESS • Image Pro Software • JLGenias • SigmaScan • Sigma Scan Pro software • Skye-Probetech • Soft Imaging Systems GmBH • Microsoft C compiler • Image09 • Visual C++ • QUANT • ER Mapper • EASI/PACE • ENVI • ERDAS Imagine • GRASS GIS • IDRISI • PG Steamer • TNT Mips • Image IntelligenceTM Suite • RemoteView
Hyperspectral	• Large amounts of image-generated data and information • Further possibilities for phenotyping several diseases simultaneously	• High cost • Extremely large size of the image database, hindering the acquisition and processing steps • Requires a high data storage capacity • Requires highly qualified and trained personnel to take advantage of the full potential	• Apple tree-*Venturia inaequalis* (Delalieux et al. 2007) • Rice-*Pyricularia oryzae* • Sweet potato-(*Cercospora, Erysiphe, Uromyces*) (Mahlein et al. 2012) • African oil palm-*Ganoderma boniense* (Lelong et al. 2010) • Wheat-*Puccinia striiformis* f. sp. *tritici* (Arora et al. 2013) • Wheat-*Fusarium* spp. (Bauriegel and Herppich 2014)	

(continued)

Table 7.1 (continued)

Image	Advantages	Disadvantages	Pathosystems	Software[a]
Thermographic	• Quantification possible even before the onset of visible symptoms	• Indicator of plant stress. May be nonspecific	• Pumpkin-*Pseudoperonospora cubensis* (Oerke et al. 2006) • Vine-*Plasmopara viticola* (Stoll et al. 2008) • Apple tree-*Venturia inaequalis* (Oerke et al. 2011) • Olive tree-*Verticillium* (Calderon et al. 2013)	
Chlorophyll fluorescence	• Further possibilities for phenotyping several diseases simultaneously • Quantification even before the onset of visible symptoms	• Several stress factors may affect chlorophyll fluorescence, and thus, it may be nonspecific	• Sweet potato-*Cercospora beticola* (Chaerle et al. 2007) • Millet-*Puccinia substriata* (Costa et al. 2009) • Bean plant-*Xanthomonas fuscans* pv. *fuscans* (Rousseau et al. 2013) • Wheat-*Fusarium* spp. (Bauriegel and Herppich 2014)	

[a] Cited by Bock et al. (2010)

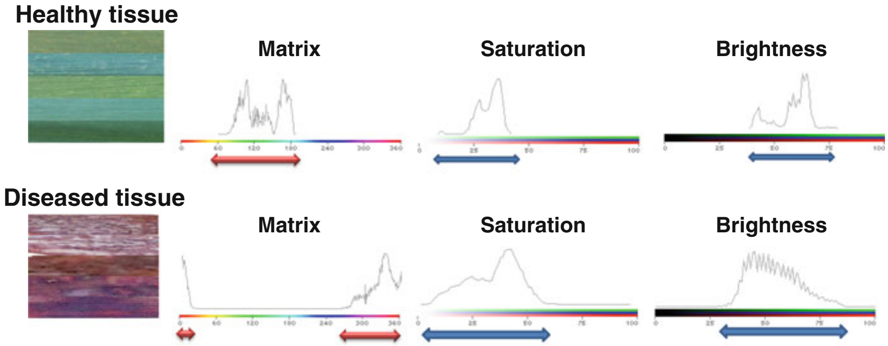


Fig. 7.4 Healthy (*top*) and diseased (*bottom*) sugarcane tissue attacked by *P. kuehnii*, and the respective HSB plots representing different intervals of the visible region

IP has enabled the separation among severity values corresponding to resistance levels of susceptible, intermediate, or moderately resistant genotypes (Fig. 7.5). Thus, the use of unmanned aircraft (drones or unmanned aerial vehicles) carrying image acquisition technology may provide an excellent large-scale phenotyping system at the field level, especially for taller crops. Fixed-land platforms may be used for the acquisition, storage, and even transmission of images of shorter crops, including soybean.

Figure 7.6a shows soybean plots with differences regarding the severity of defoliation caused by Asiatic Rust (*Phakopsora pachyrhizi*), corresponding to the six levels of the diagrammatic scale developed by Canteri et al. (2006): 5, 15, 45, 65, 85 and 100 %. Each image was processed and analyzed to generate an equation between the percentage of defoliation and the *Hue* value below 48, producing a

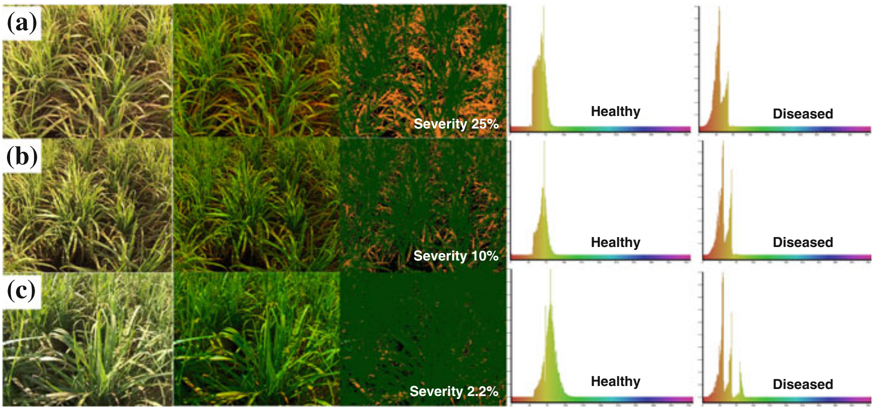


Fig. 7.5 Field images of sugarcane plots characterized by genotypes that are susceptible, intermediate, and moderately resistant to Orange Rust (*Puccinia kuehnii*). **a** Original RGB format; **b** segmented and transformed; and **c** after representation and description

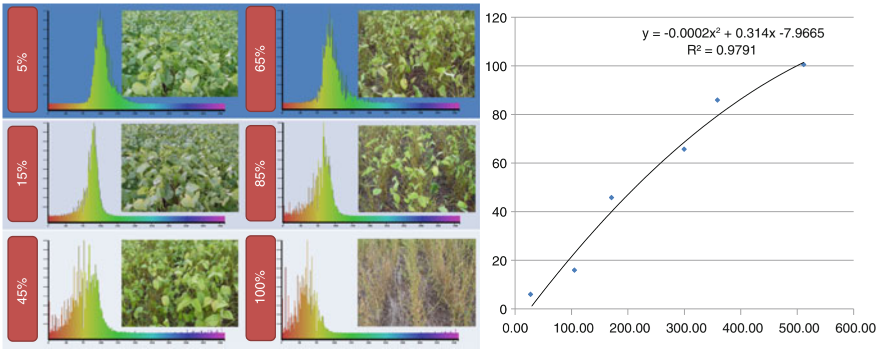


Fig. 7.6 Diagrammatic scale for assessing the severity of Asian Rust (*Phakopsora pachyrhizi*) and the respective plots of *H* values less than 48 (*left*). Plot of the number of pixels with *H* values below 48 (*x*-axis) against defoliation severity (*y*-axis); the data fit a polynomial equation with a high R^2 (*right*)

second-degree polynomial model ($-0.002X^2 + 0.314X - 7.9665$), with R^2 equal to 0.9791 (Fig. 7.6b). That confirms once again that RGB images adapt very well to the phenotyping of plant resistance to leaf diseases, even when using relatively simple algorithms.

7.2.2 Analysis of Hyperspectral Images

Hyperspectral images (HSI) are created from indices calculated from the reflectance of waves for the entire electromagnetic field and consist of hundreds of records of continuous spectral bands used to derive a complete reflectance spectrum for each pixel. Each disease, through its symptoms and/or signs, may be identified and quantified by the "spectral signature," also known as a spectral response, spectral curve, or simply spectrum. HSI have been widely used to identify and quantify plant diseases based on image acquisition at numerous scales, including leaves, plants, and plots in greenhouses, fields, or farms. Although HSI may use the entire electromagnetic spectrum, for the identification and quantification of plants, the regions from 400 to 700 nm and from 700 to 1,100 nm are more commonly used. These wavelength intervals enable a better differentiation between healthy and diseased tissue, especially for diseases that affect pigments, structures, and water contents of plants and leaves.

Hyperspectral acquisition, also known as imaging spectroscopy, may examine more than 100 bands, with widths that usually range from 0.01 to 0.02 μm. Spectrographs that are coupled to digital sensors (e.g., charge-coupled device, complementary metal-oxide semiconductor, indium gallium arsenide) are used for that purpose, enabling each photographic element (pixel) of the integrated circuit to be considered as a separate spectroradiometer. Therefore, a three-dimensional

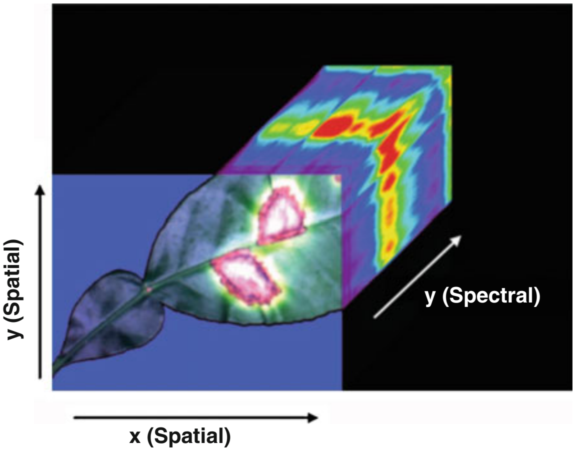


Fig. 7.7 Hyperspectral datacube of a grapefruit leaf with citrus canker lesions, indicating two spatial dimensions (*x* and *y*) and one spectral (*z*) dimension. *Source* Bock et al. (2010)

spectral image of data for processing and analysis is produced, also termed a hypercube or a datacube, with spatial *x*- and *y*- axes and a spectral *z*-axis (Fig. 7.7).

A large amount of data is generated for each hyperspectral image, producing extremely large files and therefore demanding high transmission, storage, and data processing capacities. Thus, specific software and algorithms are required for processing (Table 7.1). Consequently, the first step in the analysis of hyperspectral images following image acquisition is to reduce the dimensionality of the data. The removal of noise and reduction of image size may be performed in several ways. However, the two most commonly used routines are principal component analysis (PCA) and the minimum noise fraction (MNF) transformation.

The definition of the spectral library, which consists of a database composed by spectral signatures extracted from pixels of characteristic regions of each disease-specific symptom, is performed following the data reduction. The signatures of all the stages of symptom development for various genotypes of plants cultivated under representative soil, environmental, and management conditions must be included in the spectral library to represent the high symptomatological variation. Subsequently, the separation of pixels of a digital image into different groups or classes is conducted by classification. Several classification algorithms have already been validated, and the normalized difference vegetation index (NDVI) is among the most commonly used. The final step is the analysis of the hyperspectral data using algorithms that allow for differentiation between healthy and diseased tissues.

Such spectral signatures can enable the determination of the severity levels of sweet potato leaves inoculated with *Cercospora beticola* (Fig. 7.8). The reflectance in the visible spectrum, primarily in the spectral bands from 500 to 700 nm, increased with severity, indicating the power of using spectral signatures for the quantification and, therefore, phenotyping of plant diseases.

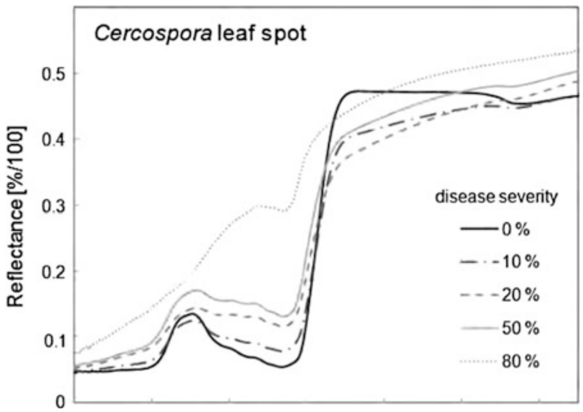


Fig. 7.8 Spectral signatures of sweet potato leaves with different severity levels of *Cercospora beticola* leaf spot. *Source* Mahlein (2010)

7.2.3 Analysis of Thermographic Images

Infrared thermography converts thermal radiation emitted by the leaf surface or canopy of one or many plants into detailed visual images of the temperature profile —that is, a thermogram. It enables the visualization of the temperature of plants through the detection of the emitted infrared radiation (far infrared, 8–14 μm), providing information on plant physiology, especially transpiration. The leaf temperature is subjected to the influence of environmental factors and the transpiration rate. Changes in the water status of a plant can cause changes in leaf transpiration, given the active regulation of stomatal conductance.

The analysis of thermographic data may be accomplished using a computer application that transforms data on emitted radiation into thermal images, within which the temperature levels are represented by a false-color gradient. Thus, thermography indirectly quantifies the transpiration rate and plant water content. Several thermography systems have been validated for use in breeding programs that aim to develop tolerance to water stress under laboratory, greenhouse, and field conditions. These systems may be validated for large-scale phenotyping of plant diseases, including those that affect transpiration or water transport in plants, such as wilting, stunting, and other diseases caused by pathogens that colonize the vascular system and affect the rates of transpiration and photosynthesis.

As shown in Fig. 7.9, apple leaves inoculated with conidia of *Venturia inaequalis* exhibited atypical concentric patches of low leaf temperature at 6 days after inoculation, even before the onset of visible symptoms of Apple scab. The affected leaf area and the difference between the temperatures of healthy and diseased tissues increased as the typical symptoms of disease became visible at 8 days after inoculation. Similarly, multiple leaf areas of lower temperature precede the onset of Apple scab lesions at the most advanced stages of the disease progression, seen when comparing the images from 8 and 11 days after inoculation. The leaf temperatures of noninfected apple trees exhibited a small temperature variation.

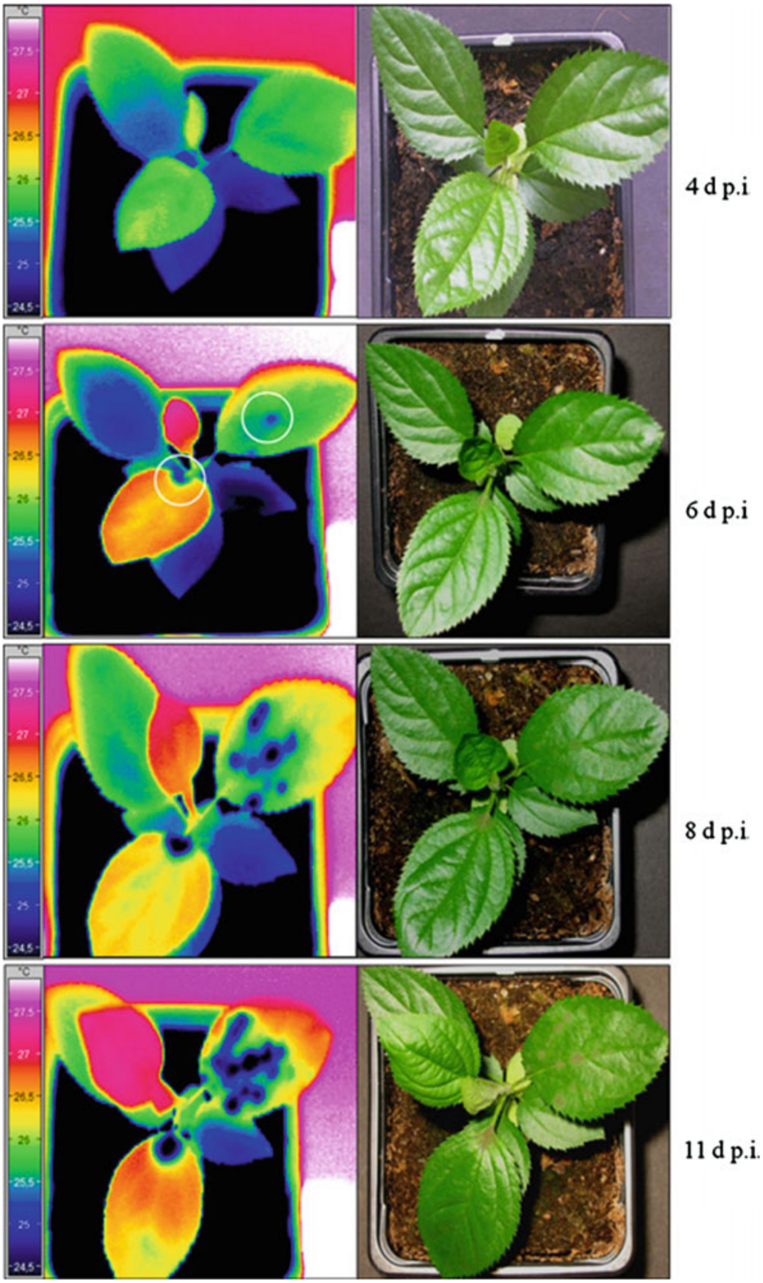


Fig. 7.9 Effects of *Venturia inaequalis* infection on the spatial heterogeneity of the leaf temperature of apple seedlings at 4, 6, 8, and 11 days after inoculation. Thermal images are displayed on the *left*, whereas RGB reflectance images are displayed on the *right*. *Source* Oerke et al. (2011)

These results provide further evidence that thermography enables the early detection and quantification of several plant diseases and is therefore very well adapted for the detection of nonvisible symptoms during their initial stages.

7.2.4 Analysis of Chlorophyll Fluorescence Images

The visible light energy absorbed by plant leaves may be used for conducting photochemical reactions and, therefore, for CO_2 assimilation or may be dissipated in the form of heat or fluorescence. Many pathogens, including fungi, bacteria, viruses, phytoplasmas, and mycoplasmas, mainly target the carbon metabolism and the photosynthetic apparatus, thus decreasing the production of chlorophyll in diseased tissues. In such instances, these diseases may be quantified and the selection of resistant cultivars may be promoted through the large-scale acquisition and analysis of chlorophyll fluorescence images.

Devices for fluorescence image acquisition have been developed for the microscopic, organ, leaf, plant, and small-plot scales. These devices consist of charge-coupled device cameras with light-emitting diode panels; they automatically estimate the maximum quantum yield of photosystem II (Fv/Fm = (Fm − Fo)/Fm). This parameter enables the detection and quantification of the difference in fluorescence between healthy and diseased tissues and is the most commonly used by plant breeders. Healthy tissues generate Fv/Fm values of approximately 0.84, while diseased tissues are significantly less efficient, with values ranging from 0.2 to 0.6, depending on the stage of symptom expression. Therefore, this nondestructive method exhibits great potential for use in in vivo phenomics studies of disease resistance. The range of the Fv/Fm interval enables the separation of disease symptoms into several stages of development for bean plants inoculated with *Xanthomonas fuscans* subsp. *fuscans*. Figure 7.10 illustrates the chlorophyll

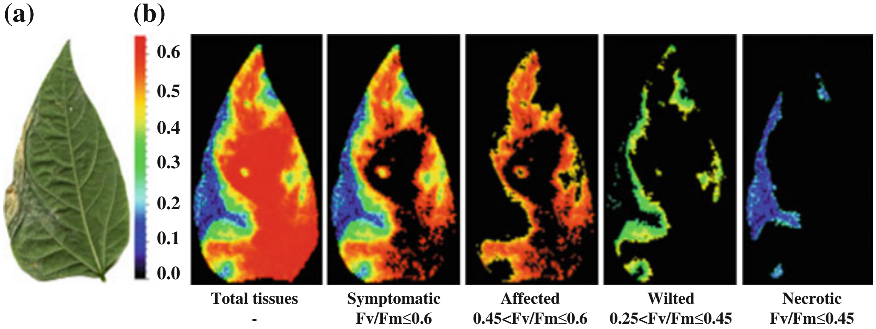


Fig. 7.10 Fv/Fm intervals during several stages of symptom development in bean plants (*Phaseolus vulgaris* cv. Flavert) inoculated with *Xanthomonas fuscans* subsp. *fuscans*. **a** RGB image with clearly visible necrosis in the *left margin* of a leaflet surrounded by wilted tissue. **b** Fv/Fm image generated by chlorophyll fluorescence. *Dark areas* represent pixels not selected with the interval used. The ratio of pixels in each segment was quantified after their separation. *Source* Rousseau et al. (2013)

fluorescence images acquired 11 days after inoculation and the respective Fv/Fm intervals used in the image processing.

Thermography is thus better suited for assessing diseases that affect the amount of chlorophyll, enabling the detection and quantification of infection even before the symptoms become visible. The great advantage of the method is that it enables the early detection of symptoms, that is, before symptoms become visible, and is thus an excellent tool for quantifying the area of infected tissue and the cultivar responses in terms of susceptibility or resistance to disease.

7.3 Platforms for Large-Scale Phenotyping of Disease Resistance

Considering the existence of various techniques, sensors, software, and algorithms for the quantification of diseases using IP methodology, it is necessary to define how all that technology may be combined to perform the large-scale acquisition, storage, and transmission of images. This knowledge will allow for advancements using the new generation of phenotyping studies of plant pathology, especially in the understanding of the genetics of the plant–pathogen interaction, therefore accelerating the production of resistant cultivars. Thousands of accessions and seedlings must be phenotyped in phenomics studies and routine breeding programs; thus, the use of platforms directed towards automated phenotyping at both the greenhouse and field levels should be prioritized. Diseases attack various plant parts, thus necessitating the acquisition of images at the levels of leaves, roots, the entire plant laterally, and the plant crown or canopy. Furthermore, platforms should be designed for nondestructive evaluations.

With the growing need and demand for phenomics studies, equipment manufacturing companies have recently developed several commercial phenotyping platforms for conducting large-scale screening and selection programs. Several are completely automated and versatile, enabling the characterization of several plant species through RGB, hyperspectral, thermographic, and chlorophyll fluorescence imaging. For example, the multi-sensor platforms Scanalyser^{3DHT} for greenhouses (Fig. 7.11a–c) and Scanalyser FIELD (Fig. 7.11d) for field studies, which are capable of analyzing thousands of plants, have been developed by the company LemnaTec (http://www.lemnatec.com). These platforms offer an enormous range of options for researchers, plant breeders, agrobiotechnology companies, and governmental organizations. The multinational companies Basf, Bayer, Syngenta, Dupont, Monsanto, and Pioneer, as well as institutes such as the National Institute of Agronomic Research in France (Institut National de la Recherche Agronomique, INRA), the Australian Plant Phenomics Facilities (APPF; http://www.plantphenomics.org.au), and the National Plant Phenomics Centre at Aberystwyth University (http://www.plant-phenomics.ac.uk/en), in the United Kingdom, and governmental organizations, stand out among the LemnaTec customers. Other institutes, universities, and companies have also entered the market and offer

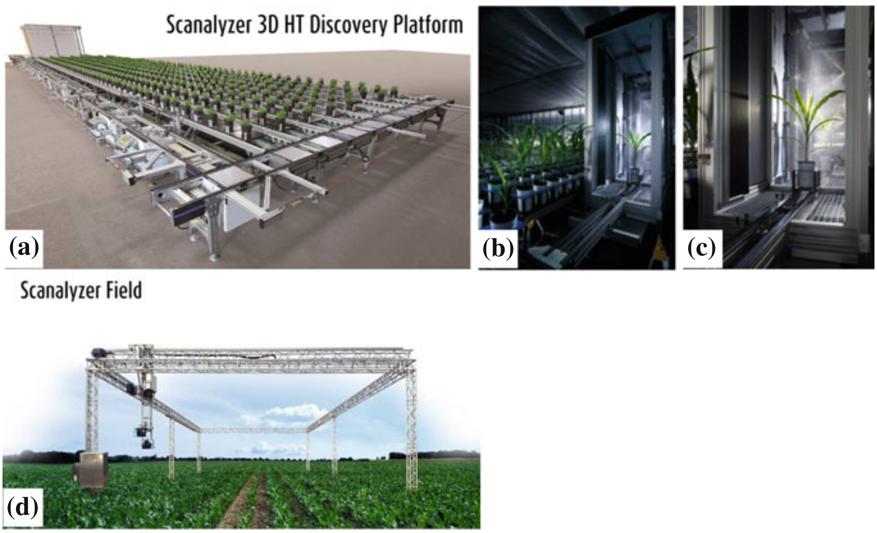


Fig. 7.11 Examples of large-scale phenotyping platforms for greenhouse and field studies. **a** Overview of the growth room. **b** Automated conveyor carrying potted plants to the imaging chamber. **c** Detail of the plant arriving at the imaging chamber. **d** Overview of the field platform. *Source* Equipment company, LemnaTec (http://www.lemnatec.com); accessed on 08/02/2014

excellent platforms for monitoring and quantifying several aspects of plant growth, development, and response to biotic and abiotic stresses, including the American company Qubti Systems (http://qubitsystems.com/portal/), which developed PlantScreen™ Conveyor Systems for greenhouses and PlantScreen™ Field Systems for field studies, and the Dutch company Phenopex, which offers FieldScan (with a total capacity for 12,800 plants and a throughput of 5.180 plants/h; http://phenospex.com). Several companies even include customized service, customizing the platform according to the customer's phenotyping demands.

In Brazil, the National Center for Agroenergy Research (Centro Nacional de Pesquisa em Agroenergia) of the Brazilian Agricultural Research Corporation (Empresa Brasileira de Pesquisa Agropecuária, EMBRAPA), in Brasília, has started assembling a structure for large-scale phenotyping studies of energy crops (Souza 2014). This structure will have five imaging modules, one for each standard, including RGB, spectral (UV/VIS), chlorophyll fluorescence, thermography, and NIR. Following its completion and validation, the goal is to adapt it to other species and conditions for which plant genetic breeding programs of EMBRAPA are conducted. The initiative is laudable, albeit still insufficient for an agricultural country.

Despite the several systems available in the market, researchers have developed their own systems at universities and centers for phenomics or for large-scale phenotyping studies, including automated platforms (Poland et al. 2012; Andrade-Sanchez et al. 2014) and a multispectral imaging box mounted on a tractor jib crane (Svensgaard et al. 2014), in efforts aimed at refining methodology, reducing costs, and meeting the specific requirements of each crop. The APPFC in Adelaide,

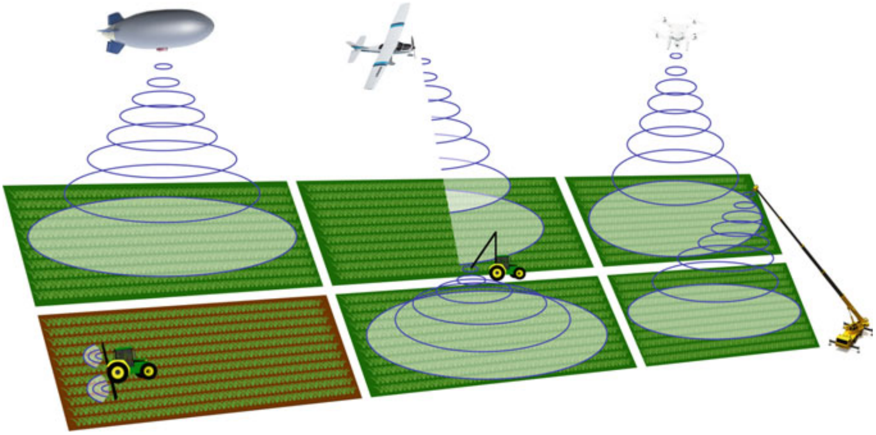


Fig. 7.12 Examples of platforms for carrying technology for the acquisition, storage, and transmission of images for phenotyping plant diseases in the field. **a** Inflatable; **b** Fixed-wing unmanned aerial vehicle (UAV); **c** Drone; **d** Fixed-backhoe loader; **e** High-backhoe loader; **f** Jib crane. *Source* Smartbio Desenvolvimento Tecnológico Ltda

Australia, is a reference center with facilities for plant phenotyping and has built the Plant Accelerator, with 3,500 m^2 of greenhouse area and 20 environmental control chambers. Furthermore, the Phenomobile, Imaging Tower, and Airborne Blimp (http://www.plantphenomics.org.au/) platforms have been developed for phenotyping in the field.

Figure 7.12 illustrates a few other platforms that have been used in the field. These are only a few examples. There are numerous possibilities for the development of customized systems that fall short of our immediate imagination and are thus up to the creativity of the inventor.

The vast majority of available platforms may be used for directly phenotyping disease resistance with small adaptations, providing that suitable methods and technologies are applied for data acquisition, analysis, and processing.

7.4 Application of Large-Scale Phenotyping to Plant Genetic Breeding for Producing Resistant Cultivars

The production of plant disease-resistant cultivars is the final objective of plant breeders. A significant effort should be directed towards accelerating the production of resistant cultivars through large-scale phenotyping, which will increasingly occur through the introduction of phenomics, molecular biology, and genetic engineering into classical breeding programs. This articulation will promote the acceleration of the production of plant disease-resistant cultivars, thus promoting sustainable agriculture.

7.4.1 General Flowchart of the Phenomics of Plant Diseases for the Identification of Genes and Quantitative Trait Loci Associated with Resistance and Pathogenicity

Plant disease phenomics, whether applied within or outside of breeding programs, enables the performance of large-scale assays linking genomics with plant resistance and pathogenicity of pathogen aggressiveness and, therefore, the identification of genes and quantitative trait loci (QTL) that control host resistance and pathogen virulence or aggressiveness. Then, the identified genes and QTL may be used for screening and molecular marker-assisted selection of genotypes with vertical or horizontal resistance, respectively. Furthermore, phenomics identifies a number of genes to correct the susceptibility of superior cultivars by introducing resistance genes into plants through genetic engineering. In pathogens, phenomics enables the identification of physiological strains and of the most aggressive isolates for their use as inocula in resistance tests. Virulence genes may also be knocked-out to produce avirulent isolates. Hence, the study of plant disease phenomics is not a simple task. The stress resulting from parasitism is abiotic and, thus, involves another organism: the pathogen. Hence, as shown in the general flowchart of activities of phenomics studies for disease resistance, the pathogen should be studied in the same stages applied to the plants, doubling the required workload, cost, and knowhow (Fig. 7.13).

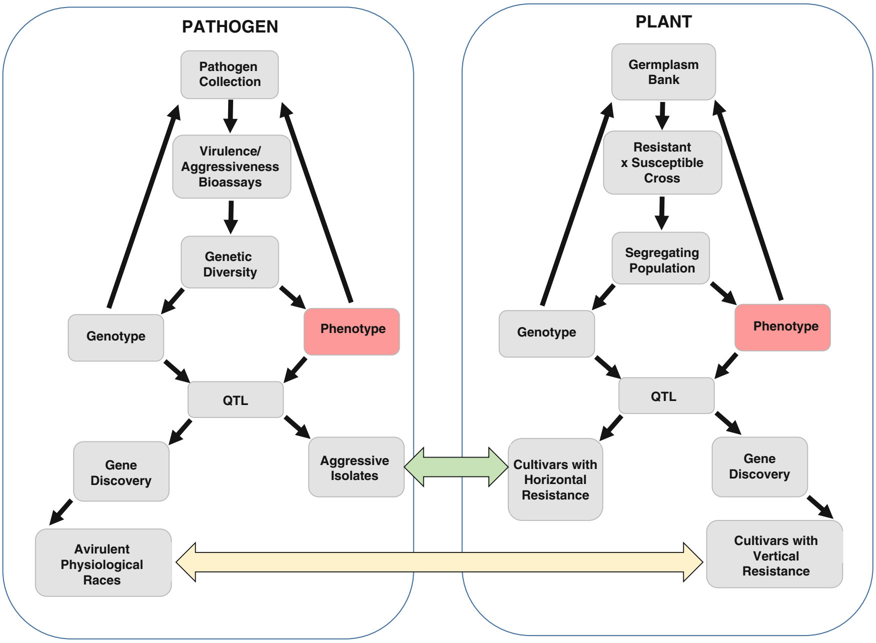


Fig. 7.13 General flowchart of the activities of a breeding program for phenomics and large-scale phenotyping studies aimed at producing resistant cultivars

In plants, the entire breeding process begins with the establishment of a germplasm bank with a fairly broad genetic base, containing sources of resistance and susceptibility to the main diseases affecting the crop in question. Segregating populations for resistance are produced from crosses between parents that are resistant and susceptible to one or more diseases. The genotypes of these segregating populations may be easily determined through genomics. Thereafter, the phenomics process proceeds only if the phenotype of each specimen of that segregating population—or, at least, of the resistant and susceptible individuals—is accurately, objectively, and rapidly identified, thus enabling the establishment of a link between plant genes (genotype) and resistance (phenotype).

The use of resistance phenomics knowledge for plant breeding should be accompanied by the use of pathogenicity phenomics knowledge. In other words, disease resistance phenotyping depends on plant inoculation with the correct pathogen isolate so that each gene or set of genes are expressed for the vertical and horizontal resistance of plant and pathogen, respectively. Thus, the establishment of a pathogen collection (of fungi and bacteria, among others) that encompasses a population's genetic diversity in a specific field, region, state, country, continent, or worldwide is the first step towards understanding the attack mechanisms of pathogens and, therefore, the type of plant resistance to be studied. Large-scale genomic and molecular studies enable the determination of the complete genome of each isolate of a population of a specific pathogen. However, the pathogen must also be inoculated in the host plant for the large-scale phenotyping of different cultivars, clones, and accessions of the plant to associate its genes with characteristics that confer virulence or aggressiveness. The isolates that exhibit pathogenicity are regulated by several genes (polygenic = quantitative) related to horizontal resistance in plants, which are identified by linking QTL with aggressiveness. In turn, establishing a link between genes and avirulence enables the identification of physiological races—that is, those isolates whose interaction with the plant follows the gene-for-gene relationship (for each gene conditioning resistance in the host, there is a corresponding gene conditioning virulence in the pathogen), interacting with the plant's monogenic vertical resistance, also termed *race-specific resistance.* Thus, similar to the manner in which large-scale phenotyping is key for the identification and quantification of plant resistance, it is also key for virulence and pathogen aggressiveness studies, thus further demonstrating its required use in breeding programs.

7.4.2 Phenotyping of Plant Resistance and Pathogen Virulence or Aggressiveness

Plant disease phenotyping requires that pathogens either naturally or artificially make contact with the host, thus enabling the pathogen and host to express pathogenicity and resistance, respectively. The inoculation may be performed by injection, coppicing, spraying, use of infective lines, or naturally (spores present in the air, soil and seedlings, among others), depending on the pathogen in question and whether the experiment is conducted in a greenhouse or in the field. In addition to the inoculation, which may occur at different experimental times, a series of activities should be conducted to quantify the pathogenicity and/or resistance. The general procedure for phenotyping, which is adaptable to any crop, disease, symptomatology, type of resistance, and greenhouse or field conditions, consists of the components presented in Fig. 7.14, which are briefly described here:

(1) *Choice of image and sensor type*: The first step is the choice of image type to be analyzed. RGB should be the first choice for diseases with visible symptoms on leaves. Spectral signatures will enable better disease quantification when diseases with symptoms that are difficult to distinguish are involved. Conversely, thermography is the most suitable method in the case of diseases that affect the plant water content, providing quantification of the infection in its early stages. The same condition applies with the chlorophyll fluorescence, albeit for the quantification of diseases affecting the photosynthetic apparatus rather than water content. These are the most appropriate indications for each type of image, but are not meant to limit the use of each type.

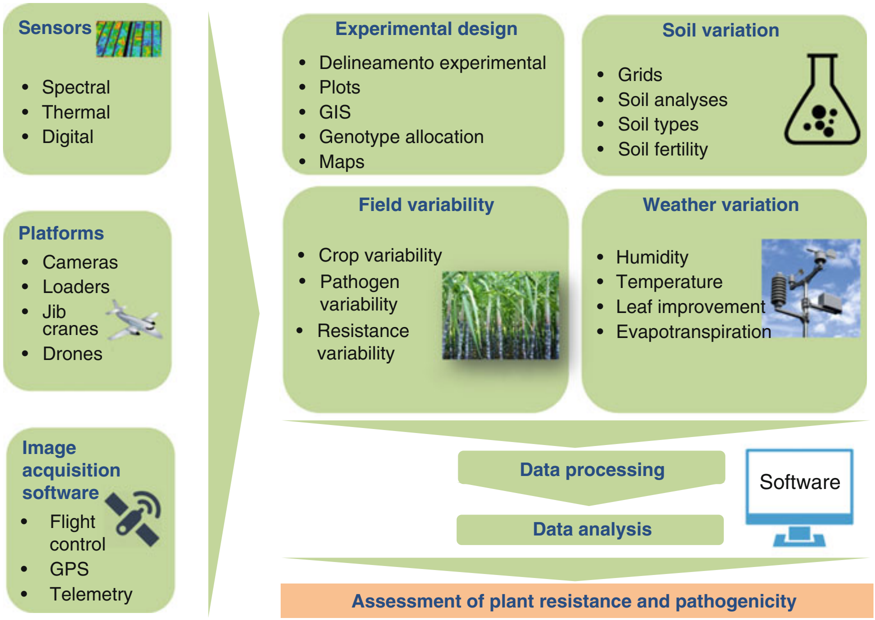


Fig. 7.14 Summary of the different components of the breeding process when conducting large-scale phenotyping for disease resistance. The overall process includes the evaluation of the most important diseases of a crop at the correct time, including the assessment of spatial variability, the characterization of the environment, and the integration of all the data. *Source* Adapted for use in plant diseases based on the general process introduced by Araus and Cairns (2014)

(2) *Definition of the platform*: The platform to be employed must be selected after the choice of sensor(s). The optimal platform will mainly depend on the crop, disease(s), and intended scale of the phenotyping. The use of a commercial automated platform with already existing models or with customizations for specific situations is recommended for greenhouse studies. The use of a conveyor belt and image acquisition software is essential to the process. An accurate global positioning system, geodata (GIS), and telemetry should be included in field experiments. In the case of drone use, which provides a platform with high potential for phenotyping, the development and use of flight systems are necessary and should be synchronized with the experimental design and the allocation of genotypes across the experimental area.

(3) *Choice of image acquisition, storage, and transmission software*: The choice and/or customized development of software to control the entire automation of a phenotyping platform must be appropriate to the scale and type of plant disease. Furthermore, image acquisition, storage, and transmission software are all necessary.

(4) *Experimental design*: The choice of the experimental design is critical both for identifying and quantifying diseases and, especially, for linking resistance to genes and QTL. The experimental area should be very well characterized through the design of allocation maps for plots and genotypes using a timer (greenhouse) and geo-referencing (field experiments). In the end, digital maps of the experimental design and the allocation of plots and genotypes should be developed. These maps serve as a guide to associate the image or portions of the image, which is also timed and/or geo-referenced, to each genotype and its replicate, when such exists. Variations of soil type and fertility among grids should also be monitored in the field, and the climactic variation must be recorded when using uncontrolled conditions. Thus, both the plant and the pathogen will be able to express the variability in resistance and pathogenicity, respectively. From the above-described information and the number of genotypes intended for phenotyping, the experimental design and map are developed, allocating each genotype to a georeferenced plot.

The field variability of both the crop and pathogen, if any exists, will be detected using the image acquisition platform after planting and inoculation, when necessary. The weather conditions during the crop development are also monitored. The image and weather data are stored in a database, and the images are processed and analyzed using specific software, designing maps and producing reports with the allocation of genotypes and their respective resistance levels. Finally, the quantification of the incidence or severity of disease(s) is determined by large-scale phenotyping, at various stages of crop development, both in greenhouse or field studies.

7.5 Future Prospects

The future is full of opportunities and challenges for phenomics studies of plant diseases, with an aim to obtain a better understanding of the complex plant–pathogen–environment–human interactions, to identify genes and QTL of pathogenicity and resistance, and, with that knowledge, to accelerate genetic breeding programs to produce new cultivars that are resistant to multiple crop diseases. Phenotyping plays an extremely significant role because there are already methodologies, sensors, devices, software, and platforms for the large-scale identification and quantification of diseases at the laboratory, greenhouse, and field levels, despite the associated difficulties. However, even greater levels of the dissemination, popularization, and use of these tools are required. Research centers in phenomics, including the National Plant Phenomic Center at Aberystwyth Center, the APPF, the Leibniz Institute of Plant Genetics, and the INRA should be used as models. Notwithstanding, the subject of plant diseases is still understudied, despite its importance. Plant breeders with holistic views and experience, especially in phytopathometry, epidemiology, genomics, and breeding, must be further engaged to realize the full benefits from employing the breadth of already developed phenomics technology for the advancement of the sustainability of food production and agroenergy. Brazil, for example, falls far behind other countries in which agriculture is considerably less important. Initiatives such as state and national virtual genome projects must be urgently developed to congregate various groups of excellence in Brazil to compensate for the delay caused by the lack of available researchers and structures for large-scale phenotyping programs, promoting breeding and phenomics programs for various crops of agricultural importance and, especially, for the development of various key features in one cultivar, including disease resistance.

References

Andrade-Sanchez P, Gore MA, Heun JT, Thorp KR, Carmo-Silva AE, French AN, Salvucci ME, White JW (2014) Development and evaluation of a field-based high-throughput phenotyping platform. Funct Plant Biol 41:68–79

Araus JL, Cairns JE (2014) Field high–throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19(1):52–61

Arora A, Sharma RK, Saharan MS, Venkatesh K, Dilbaghi N, Sharma I, Tiwari R (2013) Quantifying stripe rust reactions in wheat using a handheld NDVI remote sensor. In: Proceedings of BGRI2013 Technical Workshop, 19–22 de Agosto, New Delhi, India, pp 1–14

Bauriegel E, Herppich WB (2014) Hyperspectral and chlorophyll fluorescence imaging for early detection of plant diseases, with special reference to *Fusarium* spec. infections on wheat. Agriculture 4:32–57

Barbedo JGA (2013) Digital image processing techniques for detecting, quantifying and classifying plant diseases. Spinger Plus 2(660):1–12

Bock C, Poole GH, Parker PE, Gottwald TR (2010) Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. Crit Rev Plant Sci 29 (1–3):59–107

Boso S, Santiago JL, Martínez MC (2004) Resistance of eight different clones of the grape cultivar 'Albariño' to Plasmopara viticola. Plant Dis 88:741–744

Calderon R, Navas-Cortes JA, Lucena C, Zarco-Tejada PJ (2013) High-resolution airborne hyperspectral and thermal imagery for early detection of verticillium wilt of olive using fluorescence, temperature and narrow-band spectral indices. Remote Sens Environ 139:231–245

Canteri MG, Koga LJ, Godoy CV (2006) Escala diagramática para estimar desfolha provocada por doenças em soja [Diagrammatic scale for estimating disease-related defoliation in soybean]. In: IV Congresso Brasileiro de Soja, Londrina [IV Brazilian Congress of Soybean, 2006 Londrina]. Resumos do IV Congresso Brasileiro de Soja [Abstracts of the IV Brazilian Congress of Soybean]. Londrina, Embrapa, pp 106–106

Costa ACT, Oliveira LB, Carmo MGF, Pimentel C (2009) Avaliação visual e do potencial fotossintético para quantificação da ferrugem do milheto pérola e correlações com a produção [Rust quantification by visual and photosynthetic potential evaluation and its correlations with production in families of pearl millet]. Trop Plant Pathol 34(5):313–321

Chaerle L, Hagenbeek D, de Bruyne E (2007) Chlorophyll fluorescence imaging for disease-resistance screening of sugar beet. Plant Cell Tiss Organ Cult 91:97–106

de Souza CAF (2014) Fenotipagem de plantas: uma nova abordagem para um velho problema. Brasília: EMBRAPA Agroenergia, 9 pp. (Comunicado Técnico) [Plant phenotyping: a new approach to an old problem. Brasília: Brazilian Agricultural Research Corporation Agroenergy (Empresa Brasileira de Pesquisa Agropecuária, EMBRAPA Agroenergia), 9 pp. 2014. (Technical Report)]. http://www.infoteca.cnptia.embrapa.br/handle/doc/991030

Delalieux S, Van Aardt J, Keulemans W, Schrevens E, Coppin P (2007). Detection of biotic stress (Venturia inaequalis) in apple trees using hyperspectral data: non-parametric statistical approaches and physiological implications. Eur J Agron 27:130–143

Díaz-Lago JE, Stuthman DD, Leonard KJ (2003) Evaluation of components of partial resistance to oat crown rust using digital image analysis. Plant Dis 87:667–674

Lelong CCD, Roger JM, Brégand S, Dubertret F, Lanore M, Sitorus NA, Raharjo DA, Caliman JP (2010) Evalualtion of oil-palm fungal disease infestation with canopy hyperspectral reflectance data. Sensor 10:734–747

Maciel JLN, Nascimento A Jr, Boaretto C (2013) Estimation of blast severity on rye and triticale spikes by digital image analysis. Int J Agron 2013:1–8

Mahlein AK (2010) Detection, identification, and quantification of fungal diseases of sugar beet leaves using imaging and non-imaging hyperspectral techniques. 172 pp. Dissertation (Masters)–Institute of Crop Science and Rescource Conservation–Phytomedicine. Ansbach

Mahlein AK, Steiner U, Hillnhutter C, Dehne HW, Oerke EC (2012) Hyperspectral imaging for small-scale analysis of symptoms caused by different sugar beet diseases. Plant Methods 8 (3):1–13

Martin DP, Willment JA, Rybicki E (1999) Evaluation of Maize streak virus pathogenicity in differentially resistant Zea mays genotypes. Phytopathology 89:695–700

Naue CR, Marques MW, Lima NB, Galvíncio JD (2010) Sensoriamento remoto como ferramenta aos estudos de doenças de plantas agrícolas: uma revisão [Remote Sensing as a Tool for the Study of Plant Diseases on Agriculture: a Revision]. Revista Brasileira de Geografia Física 03:190–195

Newton AC (1989) Measuring the sterol content of barley leaves infected with powderymildew as a means of assessing partial resistance to Erysiphe graminis f. sp. hordei. Plant Pathol 38:534–540

Niemira BA, Kirk WW, Stein JM (1999) Screening for late blight susceptibility inpotato tubers by digital analysis of cut tuber surfaces. Plant Dis 83:469–473

Oerke EC, Steomer U, Dehne HW, Lindenthal M (2006) Thermal imaging of cucumber leaves affected by downy mildew and environmental conditions. J Exp Bot 57(9):2121–2132

Oerke EC, Frohling P, Steiner U (2011) Thermographic assessment of scab disease on apple leaves. Precision Agric 12(5):699–715

Patil SB, Bodhe SK (2011) Leaf disease severity measurement using image processing. Int J Eng Technol 3(5):297–301

Poland J, Price K, Gore M, Andrade-Sanchez P, Fritz A, Price R, White J, French A, Thorp K, Schapaugh W, Welch S, Zhang N TRPGR A field-based high-throughput phenotyping platform for plant genetics. http://www.wheatgenetics.org/downloads/Projects/HTP_ProjectNarrative_20130219.pdf

Rousseau C, Bove BE, Rousseau D, Fabre F, Berruyer R, Gyillaumès J, Manceau J, Jacques MA, Boureau T (2013) High throughput quantitative phenotyping of plant resistance using chlorophyll fluorescence image analysis. Plant Methods 9(17):1–13

Sankaran S, Mishra A, Ehsani R, Davis C (2010) A review of advanced techniques for detecting plant diseases. Comput Eletronics Agric 72(1):1–13

Stoll M, Schultz HR, Berkelmann-Loenhnertz B (2008) Exploring the sensitivity of thermal imaging for *Plasmopara viticola* pathogen detection in grapevines under different water status. Funct Plant Biol 35:281–288

Svensgaard J, Roitsch T, Christensen S (2014) Development of a mobile multispectral imaging plataforma for precise field phenotyping. Agronomy 4:322–336

Todd LR, Kommedahl T (1994) Image-analysis and visual estimates for evaluating disease reactions of corn to Fusarium stalk rot. Plant Dis 78:876–878

Yang W, Duan L, Chen G, Xiong L, Liu Q (2013) Plant phenomics and high-throughput phenotyping: accelerating rice functional genomics using multidisciplinary technologies. Curr Opin Plant Biol 16(2):180–187

Chapter 8
Proteomics and Metabolomics as Large-Scale Phenotyping Tools

Simone Guidetti-Gonzalez, Mônica T. Veneziano Labate, Janaina de Santana Borges, Ilara G. Frasson Budzinski, Felipe Garbelini Marques, Thaís Regiani, Andressa Peres Bini, Marisângela Rodrigues Santos and Carlos Alberto Labate

Abstract For centuries, plant breeders had a few choices of tools to evaluate the phenotypes and to choose and select the best gene combinations from the best genotypes. Basically, a balance and a rule were the main instruments used. In many cases these conventional phenotyping tools required destructive harvests at fixed times or at particular phenological stages which were slow and costly. Today researchers can combine novel technologies such as non-invasive imaging analysis, spectroscopy, robotics, high-performance computing and "omics" high throughput analyses, integrating all of this data more efficiently and in view of plant physiology and plant breeding. In this chapter we describe briefly the high-throughput analysis using proteomics and metabolomics and integrating these data with phenomics, the challenge of statistical analysis and bioinformatics, to deal with the large amount of data generated. We also provide an overview of proteomic associated with plant breeding programs; biotic and abiotic stress conditions; quality, nutritional value, and safety of food products; and physiology and biomass production. The use of metabolomics in large-scale phenotyping is also provided in this chapter.

8.1 Introduction

Phenomics is defined as the result of the acquisition of a set of information and/or phenotypic data for a specific organism, which may be used to chart a causal relation between its genotype, environment, and phenotype (Houle et al. 2010). The importance of phenomics has increased with the realization that sets of phenotypic data are needed to understand which, and how, genomic variables affect the

S. Guidetti-Gonzalez · M.T. Veneziano Labate · J. de Santana Borges · I.G. Frasson Budzinski · F.G. Marques · T. Regiani · A.P. Bini · M.R. Santos · C.A. Labate (✉)
Laboratório Max Feffer de Genética de Plantas, Departamento de Genética, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Av. Pádua Dias N°11, CEP 13400-970 Piracicaba, SP, Brazil
e-mail: calabate@usp.br

R. Fritsche-Neto and A. Borém (eds.), *Phenomics*,
DOI 10.1007/978-3-319-13677-6_8

phenotypes, or to understand questions involving pleiotropy or the supply of raw data necessary to decipher complex issues related to biotic and abiotic stress conditions that interfere with the productivity of plant cultures.

The expanding world population and the growing demand for renewable energy sources and cultivable land rely heavily on the increasing productivity of agriculturally important cultures. The genetic breeding programs employed to achieve this objective must be founded on studies that will allow the identification of molecular markers for specific characteristics by means of biotechnology and high-throughput techniques, such as proteomics and metabolomics.

Examples of high-throughput technology are nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry (MS). NMR is highly selective and nondestructive, but it has low sensitivity (Lindon and Nicholson 2008). The MS technique, on the other hand, offers a good combination of selectivity, sensitivity, and high precision (Lei et al. 2011); these essential characteristics for the analysis of complex samples are the reason why MS stands out and is now widely used in proteomic and metabolomic analyses (Barkla et al. 2013).

A landmark in the evolution of MS was the development of new sources of ionization at atmospheric pressure, such as electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI), which, unlike electron ionization (EI) and chemical ionization (CI), do not require a vacuum to generate ions. With the development of these new ionization methods and high-resolution analyzers, such as time of flight (TOF) instruments, a broad variety of chemical compounds, ranging from small polar molecules to macromolecules, began to be analyzed by MS. Also, coupling MS with certain separation systems, such as liquid and gas chromatography and electrophoresis, became a real possibility.

Therefore, with the use of these technologies involving proteomics and metabolomics in combination with data obtained from phenomics, high-productivity genotypes adapted to adverse environmental conditions are expected to be developed more efficiently, along with the results from research work on food safety and biofuel production from agricultural products.

8.1.1 Proteomics and Metabolomics as Large-Scale Phenotyping Tools

An organism's phenotype is the outcome of a combination of multiple interactions among different molecules (DNA, RNA, proteins, and metabolites) and the environment (Fig. 8.1), which is why large-scale studies require an accurate phenotypic description (Arbona et al. 2013) that will permit the use of automation and increase the capacity to evaluate a specific phenotype (Tisné et al. 2013).

A workflow of proteomic and metabolomic analyses (Fig. 8.2) shows that, after the extraction of the proteins and metabolites, the samples from different tissues (leaf, stem, or root) are analyzed by chromatography coupled to mass spectrometry

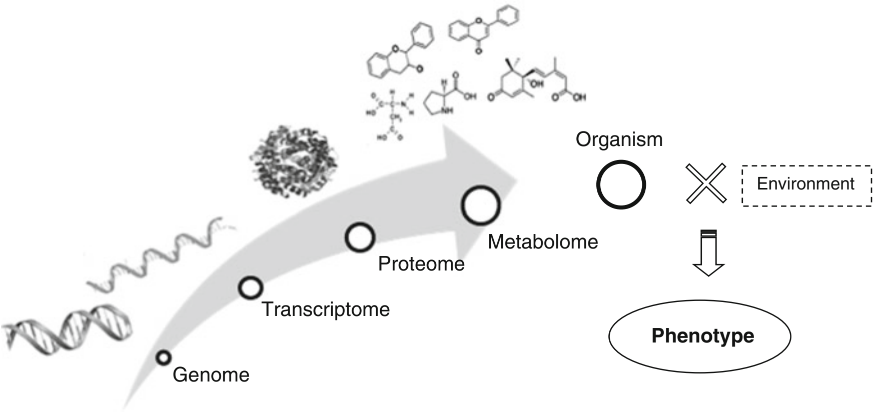


Fig. 8.1 The phenotype as the outcome of a combination of multiple interactions among genome, transcriptome, proteome, metabolome, and the environment

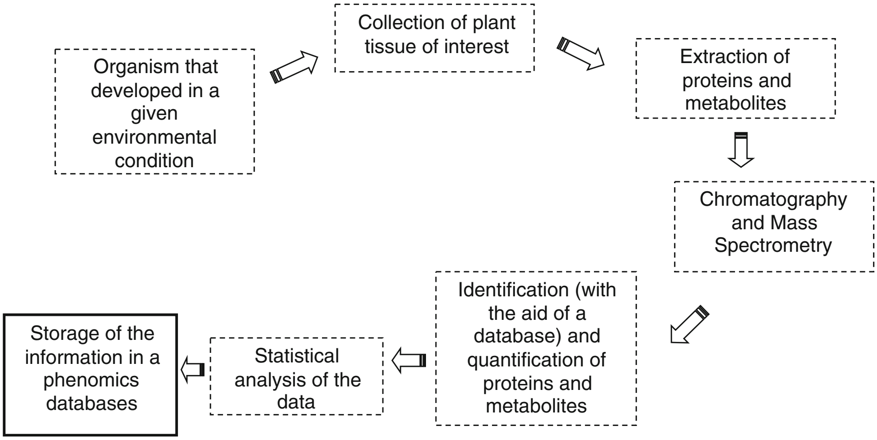

Organism that developed in a given environmental condition
Collection of plant tissue of interest
Extraction of proteins and metabolites
Chromatography and Mass Spectrometry
Identification (with the aid of a database) and quantification of proteins and metabolites
Statistical analysis of the data
Storage of the information in a phenomics databases

Fig. 8.2 Workflow of proteomic and metabolomic analyses

to identify and quantify these molecules, generating results that may subsequently become available in databases for reference and processing. The main characteristics available in such databases are the *m/z* value, retention time and index, similarity, score, and fragmentation spectrum. For proteins, gene ontology-related databases are also available (Ashburner et al. 2000) to provide information on biological processes, molecular functions, and cellular components in an organism. It is, however, very difficult to develop databases that compile all the information generated by the proteomics and metabolomics of a species developed in a given environmental condition, confirming the need for new studies related to systems biology.

8.1.2 Statistical Analysis and Bioinformatics

The large number of variables that can be visualized in plant research, as well as the large scale of phenomics-related data, represent a computational challenge; the program must create interactive methods for the correct visualization and interpretation of the data, and thus open the possibility of exploring them efficiently. Some software programs currently in use for these analyses are ProteinLynx (Waters®), MarkerLynx (Waters®), MetaboAnalyst (http://www.metaboanalyst.ca), Cytoscape (http://www.cytoscape.org/), Generic Model Organism Database (GMOD) (http://gmod.org/), and Gaggle (http://gaggle.systemsbiology.net/docs/).

To follow the exponential growth in the volume of data from high-throughput analyses, the computational capacity for data storage, processing, and analysis needs to increase (Fritsche-Neto and Borém 2013). Fortunately, technological development is advancing rapidly. The really substantial challenge, however, is to manage these data in a way that will allow the extraction of biologically and agronomically relevant information. Biological data and statistical analyses are complex because of their diversity and interrelationships, which is the reason why it is only possible to organize, analyze, and interpret all this information with the support of bioinformatics.

8.2 Proteomics Applied to Plant Breeding Programs

The ability of plants to adapt to climate changes and survive under conditions of biotic and abiotic stress is due to a reprogramming of their transcriptome, proteome, and metabolome (Qi et al. 2011; Sánchez-Rodríguez et al. 2011). These “omics” therefore allow discovering genes, gene products, and paths, which control a given phenotypical trait of agronomic interest and provide support for the development of prospection and analysis platforms as selection strategies for genetic improvement (Langridge and Fleury 2011).

Proteomics involves the analysis of the whole set of proteins of an organism in a given biological condition. It may involve posttranslational modifications (e.g., phosphorylation, methylation, and glycosylation) within the area known as *functional genomics*, which includes large-scale identification, location, and compartmentalization of the proteins, in addition to studies and construction of protein interaction networks (Aebersold and Mann 2003).

The purpose of this chapter is to provide a holistic view of a living organism to understand its responses to a given stimulus and, consequently, to predict a biological event.

The methodologies of proteomic analyses, described in detail by Regiani et al. (2013), include 2-DE gel (O’Farrell 1975), 2D-DIGE gel (Lilley and Friedman 2004), and shotgun proteomic analysis, with or without isotope labeling. The major isotope labeling techniques are stable isotope labeling by amino acids in cell culture (SILAC;

Ong et al. 2002) and stable isotope labeling in plants (SILIP; Schaff et al. 2008), as well as ^{15}N labeling (Kierszniowska et al. 2009), which includes the technique known as stable isotope labeling in arabidopsis (SILIA; Guo and Li 2011).

Most plant proteomic studies use different tissues and genotypes submitted to different biological conditions, which produce variables that complicate the task of making comparative analyses. Several functional proteomic projects were conducted in an attempt to circumvent this problem, including the international arabidopsis proteomic project (http://www.masc-Proteomics.org/), the maize proteomic project, (http://ppdb.tc.cornell.edu/dbsearch/searchcomp.aspx), the soybean proteomic project (http://proteome.dc.affrc.go.jp/Soybean/), and the "Organellome" project (http://podb.nibb.ac.jp/Organellome), among others (Jorrín-Novo et al. 2009). A more recent project is that of the International Plant Proteomics Organization (INPPO; http://www.inppo.com; Agrawal et al. 2012).

The results of functional genomics studies allowed the identification of molecular markers by reverse genetic analysis. The challenge, however, is to find effectively potential marker genes among such a complex set of data for use in plant breeding programs (Kirst et al. 2004).

The genic or proteic expression data of individuals in a segregating population may be analyzed by quantitative trait loci (QTLs) as a strategy to detect complex molecular markers of protein expression (pQTLs) in plants. Witzel et al. (2011) found pQTLs involved in metabolic processes related to the mechanism of response to the presence of a phytopathogen in barley; the identified markers include a disulfide isomerase, a BDA1-amilase inhibitor, a NADP-dependent malic enzyme, an adenosine kinase, and a BP1 peroxidase.

8.2.1 Proteomics Associated with Biotic and Abiotic Stress Conditions

One of the limiting factors of agricultural production is temperature oscillation. Comparative studies of tolerance to low temperatures (Neilson et al. 2011) and to high temperatures have identified heat shock proteins (HSPs), for example, among other proteins related to metabolic processes, as traits associated with the high temperature-tolerant phenotype in wheat (*Triticum durum*; Laino et al. 2010), tomato (Yang et al. 2006), rice (Jagadish et al. 2010), and soybean (Wang et al. 2012a) cultivars. Other proteins, such as calmodulin-binding proteins (CBPs), were identified in sorghum (Virdi et al. 2009) and arabidopsis (Zhang et al. 2009) with differential expression under high-temperature stress conditions, while the S-adenosylmethionine synthetase (SAM) protein identified in barley was proposed by Süle et al. (2004) as a potential molecular marker for the selection of high temperature-tolerant germplasms.

Another limiting factor of agricultural production is the availability of water for irrigation of the cultures. Drought stress leads to a variety of physiological changes that include, for example, a decrease in the photosynthetic rates of the plants, resulting in major productivity losses. The profile of proteins expressed in drought stress-tolerant cultivars was seen to undergo considerable change in susceptible lineages such as maize (Alvarez et al. 2008), sugarcane (Jangpromma et al. 2010), wheat (Ford et al. 2011), rice (Muthurajan et al. 2011; Mirzaei et al. 2014), eucalyptus (Borges 2013), and grape (Cramer et al. 2013) as a protective strategy against such adverse conditions. In these circumstances, as well as in the case of stress from excess salinity (Chaves et al. 2011; Gao et al. 2011), several metabolites, such as ascorbate and glutathione, and enzymes, such as peroxidases and superoxide dismutases, help to fight reactive oxygen species (ROS) that inflict chemical damage both to DNA and proteins, and they may produce severe effects on cell metabolism (Mittler 2002).

A further stress factor of importance is the attack of plants by pathogens, which leads to enormous productivity losses in economically important vegetable cultures. In this case as well, proteomics is very useful to elucidate molecular mechanisms involved in plant–pathogen interactions (Monavarfeshani et al. 2013), with the purpose of identifying potential resistance-related molecular markers that may be used in plant improvement.

The proteomic analysis conducted by Shah et al. (2012) simultaneously identified differential classes of proteins involved in defense mechanisms in the necrotrophic *Botrytis cinerea* fungus, which causes gray mold on the tomato leaf, and also in the tomato fruit in response to the presence of the pathogen. Among others, the protein classes identified in the plant include peroxidases, proteases, protease-inhibitor proteins, chitinases, endoglucanases, and carbohydrate metabolism-related enzymes (CAZy), which are expressed differentially in as an attempt to prevent the fungus from developing. These protein classes have also been found in response to *Puccinia striiformis* f. sp. *Tritici* and *P. hordei* fungi, which cause rust in wheat (Maytalman et al. 2013) and barley (Bernardo et al. 2012), respectively.

Among the classes of differentially expressed proteins with respect to pathogens, which are related to cell wall degradation, the following may serve as examples: pectin methylesterases, endo-polygalacturanases (Endo-PGs), *β*-galactosidases, cellulases, laccases, and CAZy enzymes, among others that were found in *Septoria tritici* (Yang et al. 2013), *Magnoporthe oryzae* (Franck et al. 2013), *Botrytis cinerea* (Shah et al. 2012), and *Candidatus Phytoplasma aurantifolia* (Monavarfeshani et al. 2013), which cause tritici blotch, blast, gray mold, and witches' broom disease, respectively, in wheat, rice, tomato, and the lime tree.

8.2.2 Proteomics Associated with Quality, Nutritional Value, and Safety of Food Products

The identification and quantification of proteins by highly advanced proteomic techniques have made it possible to evaluate the quality and nutritional value of food products in important cultures such as pea (Bourgeois et al. 2011), rice

(Shi et al. 2013), and soybean (Gomes et al. 2014). Proteomics has also been used in food safety analysis, such as the determination of the protein profile of rice to assess the effect of a genetic modification on the nutritional value of the grain (Wang et al. 2012b).

Taub et al. (2008) showed that high CO_2 (carbon dioxide) levels diminished the protein content in wheat, barley, rice, potato, and soybean. These results led the authors to conclude that the growing CO_2 emissions in the 21st century may result in lower protein concentration in a number of food crops, which would reduce their nutritional value and thus place human nutrition at risk.

AbdElgawad et al. (2014) also noted that high CO_2 concentrations, accompanied by water scarcity and high-temperature conditions, change the chemical composition of legumes and grasses that accumulate fructans; both are used as pasture. The extreme situation imposed by abiotic stress on legumes led to a reduction in their amounts of proteins, phosphorus (P), and magnesium (Mg), and to an increase in their phenol, lignin, tannin, carbon (C), and nitrogen (N) content as well as in their C:N, C:P, and N:P ratios. In contrast, the tissue composition of grasses was not affected to a similar extent. The authors concluded that the quality losses are less pronounced in grasses than in legumes in this type of environmental condition, because the fructans present in grasses contribute to their cell homeostasis.

An example of a proteomics study with a potential application in phenomics is the “Seeds in Chernobyl” database (http://www.chernobylproteomics.sav.sk), which contains information on the abundance of hundreds of proteins based on research studies with soybean and flax seeds collected from plants cultivated in the Chernobyl radioactive area. This database provides information on the proteomic profile of seeds permanently exposed to radioactive ionization, as well as the quantitative data for the proteins expressed in radioactive and nonradioactive areas (Klubicová et al. 2012).

Proteomics may also help to identify allergenic molecules that are the cause of serious health problems in sensitive individuals. Pechanova et al. (2013) discussed the evolution of mass spectrometry techniques for the accurate identification of allergenic proteins present in several types of grain and their respective products, such as flours. An interesting example is the work of Pastorello et al. (2000), who used immunological assays to identify a single allergenic protein, lipid transfer protein (LTP), in maize. With the development of mass spectrometry, however, six additional allergenic proteins—vicilin, globulin-2, 50 kDa gamma-zein, endochitinase, thioredoxin, and a tripsin inhibitor (Fasoli et al. 2009)—have been identified and enriched our knowledge of the allergenic potential of maize for the sensitive population.

In the wheat cultivar *Triticum durum*, using MALDI-TOF and MALDI-TOF/TOF, Laino et al. (2010), have also identified allergenic polypeptides that were induced in plants following a temperature increase. Other potentially allergenic proteins have been identified by mass spectrometry in a variety of cultures, such as maize, peanut, and tomato (revised by Nakamura and Teshima 2013) and in different varieties of lupin beans, a legume of high biological value that is being increasingly incorporated into the human diet (Islam et al. 2012).

8.2.3 Proteomics Associated with Physiology and Biomass Production

The relevance of proteomics involving physiology studies and biomass production cannot be overstated, because its objective is to identify proteins that play a crucial role in the adequate growth and development of plants.

In a study that involves an interdisciplinary approach to employ a variety of methodologies and tools to identify key proteins that respond to the *alf* (*abnormal leaf and flower*) gene, Chen et al. (2012) suggested that the proteins known as glutamate-1-semialdehyde 2,1-aminomutase and peptidyl-prolyl cis-trans isomerase perform an essential regulatory activity in the development of leaves and flowers, because both proteins exhibit a significant differential expression and their respective genes are found in the same chromosome, near the *alf* gene.

Proteomic approaches have also been successfully used for the detection of heterotic patterns at the level of protein expression in maize (Hoecker et al. 2008; Marcon et al. 2013) and rice (Wang et al. 2008). The phenotypic differences are commonly observed among reciprocal F1 maize hybrids due to parental imprinting. Comparative proteomic analyses of F1 hybrids and their parental lineages proved to be an excellent technique for the identification of genes associated with a uniparental domain, and may clear the way for better parent selection in plant breeding programs (Pechanova et al. 2013). Proteomics has also been used in a study of heterosis in sunflower (Mohayeji et al. 2014), in which the authors suggested that heterosis mechanisms may improve the energy balance of the plant, via carbon fixation and reduction of energy consumption, to produce superior hybrids.

Proteomics has also been applied in related research fields, particularly those involved in the production of second-generation biofuels, with the purpose of improving biomass production processes. Among the research work conducted with biomass-producing species, the studies with sorghum (Ngara and Ndimba 2011; Ngara et al. 2012), *Jatropha curcas* (Popluechai et al. 2011), *Miscanthus* (Straub et al. 2013), and sugarcane (Calderan-Rodrigues et al. 2014) may be cited as examples. In their study on *Miscanthus* genotypes, for instance, Straub et al. (2013) correlated the high biomass production rates of these genotypes with assimilation of primary carbon and reduction in secondary metabolism.

8.3 Use of Metabolomics in Large-Scale Phenotyping

The term *metabolomics* refers to the study of the entire set of metabolites in an organism in a given biological condition. Its application provides unique results, in that it allows the identification of compounds such as sugars, fatty acids, amino acids, hormones, alkaloids, and phenolic compounds, which have highly diverse chemical groups in their molecules and are characterized by a wide range of physicochemical properties such as size, polarity, and hydrophobicity

(Rasmussen et al. 2012). The predominant purpose of this approach is to provide an overall qualitative and/or quantitative view of an organism's metabolic profile.

The entire array of chemical reactions that take place continuously in a cell is defined as metabolism. Specific enzymes present in the cell guide such reactions and, in doing so, create the different metabolic pathways. The chemical compounds that are formed, degraded, or transformed by and during these processes are called metabolites. When produced by plants, such chemical compounds may be divided into two large groups that compose the primary and the secondary metabolism.

Primary metabolism is the set of metabolic processes that perform essential functions such as photosynthesis, respiration, and transport of solutes. Amino acids, nucleotides, lipids, and carbohydrates are examples of these metabolites. The secondary metabolism, on the other hand, is not always necessary for a plant to complete its life cycle, but it nevertheless plays an important role in the interaction of the plants with their environment. Secondary products also perform a protective function against abiotic stress forms, such as those associated with temperature changes, hydrological regime variations, luminosity levels, and mineral nutrient deficiencies. Terpenes, phenolic compounds, and alkaloids constitute the main secondary metabolite groups.

Metabolomics has become an important analytical strategy because changes in messenger RNA levels frequently fail to produce changes in protein levels and, once synthesized, a protein may or may not be functional. As a result, the changes observed in the transcriptome and the proteome may sometimes fail to correspond exclusively to changes in the phenotype. Given the abundance of metabolites, it is possible to infer molecular information about the function of the cell, and thereby define the phenotype of a cell or tissue in response to environmental or genetic changes, because the metabolites are a fundamental complement in functional genomics. A further advantage of the metabolomic approach is that it does not depend on the knowledge and availability of genetic information on the organism to be studied.

The high degree of diversity between primary and secondary metabolites (wide range of molecular mass values; presence of polar and nonpolar compounds) requires the use of different analytical tools for the study of the various classes of compounds (Dunn and Ellis 2005), as there is currently no available technology with the capability to analyze all of the existing metabolite classes at once. These techniques therefore need to be carefully selected according to the metabolites and metabolic pathways of interest (Pérez-Clemente et al. 2013).

Factors such as polyploidy, highly polygenic traits, and predicted epystatic and environmental effects limit the use of genetic markers. Thus, screening methods based on metabolic markers for the prediction of phenotypic traits have become more useful. Steinfath et al. (2010) published the first study to apply metabolomics to predict phenotypes of agronomical importance. The authors analyzed tubers from 20 potato cultivars in two contrasting environments, with respect to soil quality and climate, to predict the appearance of blotches on the tuber and a tendency for the potato to darken during frying. The metabolites tyrosine, threonine, valine, serine, and glutamine were found in all of the samples that proved to be susceptible to the appearance of spots; as a result, they came to be considered biomarkers for this phenotype. Glucose and fructose, on the other hand, were selected as biomarkers for potato quality.

A few desirable traits are related to the presence of specific primary metabolites, such as high sugar content in grasses due to the presence and concentration of fructans (Turner et al. 2006), as well as secondary metabolites, such as insect herbivority-resistant alkaloids (Clay and Schardl 2002). In certain studies, the analysis of such target metabolites is sufficient for it to be associated with QTLs (mQTLs) selection or analyses. However, there are traits, such as plant growth and drought tolerance, that involve complex networks of interaction between metabolites (Rasmussen et al. 2012). In this regard, Kliebenstein authored a manual to aid researchers in establishing and in interpreting the experimental data that intend to use metabolomics in the analyses of mQTLs (Kliebenstein 2010). Studies involving mQTL have been conducted for poplar (Morreel et al. 2006), rice (Gong et al. 2013), potato (Carreno-Quintero et al. 2012), arabidopsis (Kerwin et al. 2011), and apple (Khan et al. 2012).

Morreel et al. (2006) mapped four mQTLs associated with flavonoid biosynthesis in populus and concluded that the combination of metabolic profiling with QTLs analysis constitutes a valuable technique to identify control points in a metabolic pathway. In a more extensive study, Gong et al. (2013) related more than 2,800 mQTL to 900 different metabolites in rice. Carrero-Quintero et al. (2012) detected 139 primary metabolites in potato using GC-MS and found mQTLs in about 72 % of them, and Khan et al. (2012) investigated the genetic bases in apple to determine the quantitative variations in this species, detecting 669 mQTLs for potentially beneficial phenolic compounds.

Despite the high initial investment needed to conduct experiments in metabolomics, this technique provides a wealth of data that can be related to other "omics," as well as to characteristics selected for study in a given organism. In addition, the information generated by the analysis of a given crop can be extrapolated to others, and studies based on the integration of genomics and metabolomics data constitute an important strategy in the investigative work on the regulation between the metabolism and the phenotype of a given crop.

8.4 Conclusion

The integration of data generated by phenomics, genomics, proteomics, and metabolomics analyses can reduce the number of candidate genes in genetic studies (Fig. 8.3) and, as a result, greatly improve the efficiency for the discovery and use of new genes for genetic breeding programs (Chen et al. 2012).

The existing selection and improvement techniques are considered incapable of producing cultivars that will meet future demand for food crops and their products (Tester and Langridge 2010). In this respect, metabolomics and proteomics stand out as tools that can be employed in association, both with large-scale phenotyping

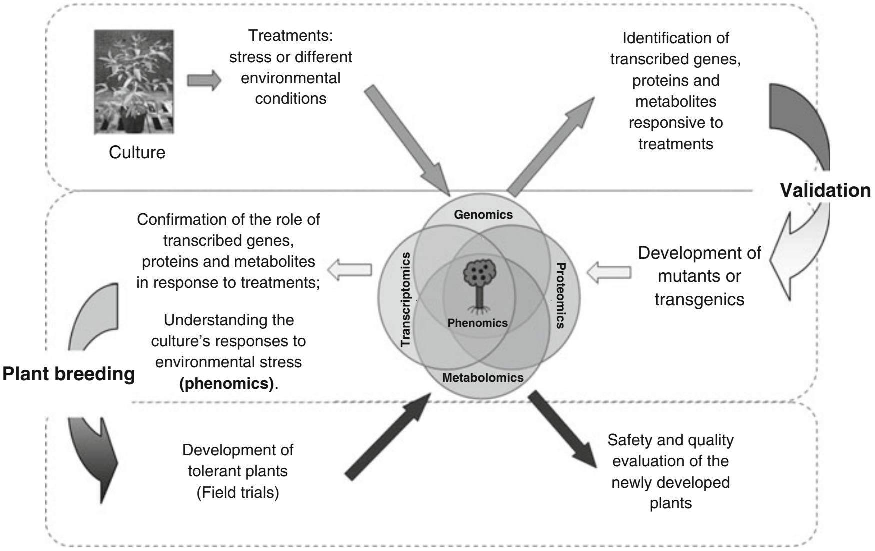


Fig. 8.3 Application of "omics" in plant breeding programs: identification of genes, proteins, and metabolites, and their validation relative to function and applicability. Adapted from Abreu et al. (2013)

and with the selection individuals/genotypes that are more productive or endowed with specific traits of interest. The identified proteins and/or metabolites may be selected for use in the development of biomarkers, which can help to conduct plant breeding programs, making it possible to predict the quality of a crop harvest, for example.

References

AbdElgawad H, Peshev D, Zinta G, Van den Ende W, Janssens IA, Han Asard H (2014) Climate extreme effects on the chemical composition of temperate grassland species under ambient and elevated CO_2: a comparison of fructan and non-fructan accumulators. PLoS ONE 9:e92044

Abreu IA, Farinha AP, Negrão S, Gonçalves N, Fonseca C, Rodrigues M, Batista R, Saibo NJM, Oliveira MM (2013) Coping with abiotic stress: proteome changes for crop improvement. J Proteomics 93:145–168

Aebersold RH, Mann M (2003) Mass spectrometry-based proteomics. Nature 422(6928):198–207

Agrawal GK, Sarkar A, Agrawal R, Ndimba BK, Tanou G, Dunn MJ et al (2012) Boosting the globalization of plant proteomics through INPPO: current developments and future prospects. Proteomics 12:359–368

Alvarez S, Marsh EL, Schroeder SG, Schachtman DP (2008) Metabolomic and proteomic changes in the xylem sap of maize under drought. Plant Cell Environ 31:325–340

Arbona V, Manzi M, Ollas C, Gómez-Cadenas A (2013) Metabolomics as a tool to investigate abiotic stress tolerance in plants. Int J Mol Sci 14:4885–4911

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29

Barkla BJ, Vera-Estrella R, Pantoja O (2013) Progress and challenges for abiotic stress proteomics of crop plants. Proteomics 13:1801–1815

Bernardo L, Prinsi B, Negri AS, Cattivelli L, Espen L, Valè G (2012) Proteomic characterization of *Rph15* barley resistance gene-mediated defence responses to leaf rust. BMC Genom 13:642

Borges JS (2013) Análise comparativa do proteoma e metaboloma de raízes de dois clones de *E. grandis* x *E.camaldulensis*, sendo um tolerante e um susceptível a condições de estresse hídrico. Dissertação de Mestrado [Comparative proteome and metabolome analysis of the roots of two *E. grandis* x *E.camaldulensis* clones, one of which is tolerant and the other one is susceptible to water stress. Master's degree dissertation]. University of São Paulo-ESALQ, Piracicaba, 199 pp

Bourgeois M, Jacquin F, Cassecuelle F, Savois V, Belghazi M, Aubert G, Huart LM, Marget P, Burstin J (2011) A PQL (Protein quantity loci) analysis of mature pea seed proteins identifies loci determining seed protein composition. Proteomics 11:1581–1594

Calderan-Rodrigues MJ, Jamet E, Bonassi MBC, Guidetti-Gonzalez S, Begossi AC, Setem LV, Franceschini LM, Fonseca JG, Labate CA (2014) Cell wall proteomics of sugarcane cell suspension cultures. Proteomics 14:738–749

Carreno-Quintero N, Acharjee A, Maliepaard C, Bachem CW, Mumm R, Bouwmeester H, Visser RG, Keurentjes JJ (2012) Untargeted metabolic quantitative trait loci analyses reveal a relationship between primary metabolism and potato tuber quality. Plant Physiol 158:1306–1318

Chaves MM, Costa JM, Saibo, NJM (2011) Recent advances in photosynthesis under drought and salinity. In: Turkan I (ed) Advances in botanical research-plant responses to drought and salinity stress: developments in a post-genomic era, vol 57. Academic Press, Elsevier, pp 49–104

Chen L, Xing G, Xu Y, Liu X, Zhao T, Gai J (2012) Identification of major responding proteins of *abnormal leaf and flower* in soybean with an integrative "omics" strategy. Comput Electr Eng 38:3–10

Clay K, Schardl C (2002) Evolutionary origins and ecological consequences of endophyte symbiosis with grasses. Am Nat 160:S99–S127

Cramer GR, Van Sluyter SC, Hopper DW, Pascovici D, Keighley T, Haynes PA (2013) Proteomic analysis indicates massive changes in metabolism prior to the inhibition of growth and photosynthesis of grapevine (*Vitis vinifera* L.) in response to water deficit. BMC Plant Biol 13:49

Dunn WB, Ellis DI (2005) Metabolomics: current analytical platforms and methodologies. Trends Anal Chem 24:285–294

Fasoli E, Pastorello EA, Farioli L, Scibilia J et al (2009) Searching for allergens in maize kernels via proteomic tools. J Proteomics 72:501–510

Ford KL, Cassin A, Bacic A (2011) Quantitative proteomic analysis of wheat cultivars with differing drought stress tolerance. Front Plant Sci 2:44

Franck WL, Gokce E, Oh Y, Muddiman DC, Dean RA (2013) Temporal analysis of the *Magnaporthe Oryzae* proteome during conidial germination and cyclic AMP (cAMP)-mediated appressorium formation. Mol Cell Proteomics 12:2249–2265

Fritsche-Neto R, Borém A (2013) Fenômica [Phenomics]. In: Fritsche-Neto R, Borém A (eds) Ômicas 360º—Aplicações e Estratégias para o Melhoramento de Plantas [360° Omics—applications and strategies for plant improvement, 1st edn.], 1a. ed., Viçosa: Suprema Gráfica e Editora Ltda, pp 243–265

Gao L, Yan X, Li X, Guo G, Hu Y, Ma W et al (2011) Proteome analysis of wheat leaf under salt stress by two-dimensional difference gel electrophoresis (2D-DIGE). Phytochemistry 72:1180–1191

Gomes LS, Senna R, Sandim V, Silva-Neto MAC, Perales JEA, Zingali RB, Soares MR, Fialho E (2014) Four conventional soybean [*Glycine max* (L.) Merrill] seeds exhibit different protein profiles as revealed by proteomic analysis. J Agric Food Chem 62:1283–1293

Gong L, Chen W, Gao Y, Liu X, Zhang H, Xu C, Yu S, Zhang Q, Luo J (2013) Genetic analysis of the metabolome exemplified using a rice population. PNAS 110:20320–20325

Guo G, Li N (2011) Relative and accurate measurement of protein abundance using ^{15}N stable isotope labeling in arabidopsis (SILIA). Phytochemistry 72:1028–1039

Hoecker N, Lamkemeyer T, Sarholz B, Paschold A, Fladerer C, Madlung J et al (2008) Analysis of nonadditive protein accumulation in young primary roots of a maize (*Zea mays* L.) F1-hybrid compared to its parental inbred lines. Proteomics 8:3882–3894

Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. Nat Rev/Genet 11:855–866

Islam S, Yan G, Appels R, Ma W (2012) Comparative proteome analysis of seed storage and allergenic proteins among four narrow-leafed lupin cultivars. Food Chem 135:1230–1238

Jagadish SVK, Muthurajan R, Oane R, Wheeler TR, Heuer S, Bennett J et al (2010) Physiological and proteomic approaches to address heat tolerance during anthesis in rice (*Oryza sativa* L.). J Exp Bot 61:143–156

Jangpromma N, Kitthaisong S, Lomthaisong K, Daduang S, Jaisil P, Thammasirirak S (2010) A proteomics analysis of drought stress-responsive proteins as biomarker for drought-tolerant sugarcane cultivars. Am J Biochem Biotechnol 6:89–102

Jorrín-Novo JV, Maldonado AM, Echevarría-Zomeño S, Valledor L, Castillejo MA, Curto M et al (2009) Plant proteomics update (2007–2008): second-generation proteomic techniques, an appropriate experimental design, and data analysis to fulfill MIAPE standards, increase plant proteome coverage and expand biological knowledge. J Proteomics 72:285–314

Kerwin RC, Jimenez-Gomez JM, Stacey LH, Maloof JN, Kliebestein DJ (2011) Network quantitative trait loci mapping of circadian clock outputs identifies metabolic pathway-to-clock linkages in Arabidopsis. Plant Cell 23:471–485

Khan SA, Chibon PY, De Vos RC, Schipper BA, Walraven E, Beekwilder J, Van Dijk T, Finkers R, Visser RGF, Van de Weg EW, Bovy A, Cestaro A, Velasco R, Jacobsen E, Schoute HS (2012) Genetic analysis of metabolites in apple fruits indicates an mQTL hotspot for phenolic compounds on linkage group 16. J Exp Bot 63:2895–2908

Kierszniowska S, Walther D, Schulze WX (2009) Ratio-dependent significance thresholds in reciprocal ^{15}N-labeling experiments as a robust tool in detection of candidate proteins responding to biological treatment. Proteomics 9:1916–1924

Kirst M, Myburg AA, De Léon JPG, Kirst ME, Scott J, Sederoff R (2004) Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. Plant Physiol 135:2368–2378

Kliebenstein DJ (2010) Metabolmics and plant quantitative trait locus analysis—the optimum genetical genomics plataform? In: Nikolau BJ, Wurtele ES (eds) Concepts in plant metabolomics. Springer, Dordrecht, pp 29–44

Klubicová K Vesel M Rashydov NM Hajduch M (2012) Seeds in chernobyl: the database on proteome response on radioactive environment. Front Plant Sci vol 3, artigo 231

Laino P, Shelton D, Finnie C, De Leonardis AM, Mastrangelo AM, Svensson B et al (2010) Comparative proteome analysis of metabolic proteins from seeds of durum wheat (cv. Svevo) subjected to heat stress. Proteomics 10:2359–2368

Langridge P, Fleury D (2011) Making the most of 'omics' for crop breeding. Trends Biotechnol 29:33–40

Lei Z, Huhman DV, Summer LW (2011) Mass spectrometry strategies in metabolomics. J Biol Chem 22:25435–25442

Lilley KS, Friedman DB (2004) All about DIGE: quantification technology for differential-display 2D-gel proteomics. Expert Rev Proteomics 1:401–409

Lindon JC, Nicholson JK (2008) Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. Trends Anal Chem 27:194–205

Marcon C, Lamkemeyer T, Malik WA, Ungrue D, Piepho HP, Hochholdinger F (2013) Heterosis-associated proteome analyses of maize (*Zea mays* L.) seminal roots by quantitative label-free LC–MS. J Proteomics 93:295–302

Maytalman D, Mert Z, Baykal AT, Inan C, Gunel A, Hasançebi S (2013) Proteomic analysis of early responsive resistance proteins of wheat (*Triticum aestivum)* to yellow rust (*Puccinia striiformis* f. sp. *tritici*) using ProteomeLab PF2D. Plant Omics J 6:24–35

Mirzaei M, Soltani N, Sarhadi E, George IS, Neilson KA, Pascovici D, Shahbazian S, Haynes PA, Atwell BJ, Salekdeh GH (2014) Manipulating root water supply elicits major shifts in the shoot proteome. J Proteome Res 13:517–526

Mittler R (2002) Oxidative stress, antioxidants and stress tolerance. Trends Plant Sci 7:405–410

Mohayeji M, Capriotti AL, Cavaliere C, Piovesana S, Samperi R, Stampachiacchiere S, Toorchi M, Lagana A (2014) Heterosis profile of sunflower leaves: a label free proteomics approach. J Proteomics 99:101–110

Monavarfeshani A, Mirzaei M, Sarhadi E, Ardeshir Amirkhani A, Nekouei MK, Haynes PA, Mardi M, Salekdeh GH (2013) Shotgun proteomic analysis of the mexican lime tree infected with *Candidatus Phytoplasma aurantifolia*. J Proteome Res 12:785–795

Morreel K, Goeminne G, Storme V, Sterck L, Ralph J, Coppieters W, Breyene P, Steenackers M, Georges M, Messens E, Boerjan W (2006) Genetical metabolomics of flavonoid biosynthesis in *Populus*: a case study. Plant J 47:224–237

Muthurajan R, Shobbar ZS, Jagadish S, Bruskiewich R, Ismail A, Leung H et al (2011). Physiological and proteomic responses of rice peduncles to drought stress. Mol Biotechnol 48:173–182

Nakamura R, Teshima R (2013) Proteomics-based allergen analysis in plants. J Proteomics 93:40–49

Neilson KA, Mariani M, Haynes PA (2011) Quantitative proteomic analysis of cold-responsive proteins in rice. Proteomics 11:1696–1706

Ngara R, Ndimba BK (2011) Mapping and characterisation of the sorghum cell suspension culture secretome. Afr J Biotechnol 10:253–266

Ngara R, Ndimba R, Borch-Jensen J, Jensen ON, Ndimba BK (2012) Identification and profiling of salinity stress-responsive proteins in *Sorghum bicolor* seedlings. J Proteomics 75:4139–4150

O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. J Biol Chem 250:4007–4021

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A et al (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 1:86–376

Pastorello EA, Farioli L, Pravettoni V, Ispano M et al (2000) The maize major allergen, which is responsible for food induced allergic reactions, is a lipid transfer protein. J Allergy Clin Immunol 106:744–751

Pechanova O, Takác T, Samajand J, Pechan T (2013) Maize proteomics: an insight into the biology of an important cereal crop. Proteomics 13:637–662

Pérez-Clemente RM, Vicente Vives V , Zandalinas SI, López-Climent MF, Muñoz V, Gómez-Cadenas A (2013) Biotechnological approaches to study plant responses to stress. BioMed Res Int, Article ID 654120

Popluechai S, Froissard M, Jolivet P, Breviario D, Gatehouse AMR, Donnell AGO, Chardot T, Kohli A (2011) *Jatrophus curcas* oil body proteome and oleosins: L-form *JeOle3* as a potential phylogenetic marker. Plant Physiol Biohem 49:352–356

Qi Y, Wang H, Zou Y, Liu C, Liu Y, Wang Y et al (2011) Over-expression of mito-chondrial heat shock protein70 suppresses programmed cell death in rice. FEBS Lett 585:231–239

Rasmussen S, Parsons AJ, Jones CS (2012) Metabolomics of forage plants: a review. Ann Bot 110:1281–1290

Regiani T, Budzinski IGF, Guidetti-Gonzalez S, Labate MTV, Cotinguiba F, Marques FG, Moraes FE, Labate CA (2013) Eletroforese, cromatografia e espectrometria de massas [Electrophoresis. chromatography and mass spectrometry] In: Borém A, Fritsche-Neto R (eds) Ômicas 360: Aplicações e Estratégias para o melhoramento de Plantas [360° Omics—applications and strategies for plant improvement]. Suprema Gráfica e Editora Ltda, Visconde do Rio Branco, Brasil, pp 123–148

Sánchez-Rodríguez E, Moreno DA, Ferreres F, Rubio-Wilhelmi MM, Ruiz JM (2011) Differential responses of five cherry tomato varieties to water stress: changes on phenolic metabolites and related enzymes. Phytochemistry 72:723–729

Schaff JE, Mbeunkui F, Blackburn K, Bird DM, Goshe MB (2008) SILIP: a novel stable isotope labeling method for in planta quantitative proteomic analysis. Plant J 56:840–854

Shah P, Powell ALT, Orlando R, Bergmann C, Gutierrez-Sanchez G (2012) Proteomic analysis of ripening tomato fruit infected by *Botrytis cinerea*. J Proteome Res 11:2178–2192

Shi W, Muthurajan R, Rahman H, Selvam J, Peng S, Zou Y, Jagadish KSV (2013) Source-sink dynamics and proteomic reprogramming under elevated night temperature and their impact on rice yield and grain quality. New Phytology 197:825–837

Steinfath M, Strehmel N, Peters R, Schauer N, Groth D, Hummel J, Steup M, Selbig J, Kopka J, Geigenberger P, Van Dongen J (2010) Discovering plant metabolic biomarkers for phenotype prediction using an untargeted approach. Plant Biotechnol J 8:900–911

Straub D, Yang H, Liu Y, Ludewig U (2013) Transcriptomic and proteomic comparison of two miscanthus genotypes: high biomass correlates with investment in primary carbon assimilation and decreased secondary metabolism. Plant Soil 372:151–165

Süle A, Vanrobaeys F, Hajós G, VanBeeumen J, Devreese B (2004) Proteomic analysis of small heat shock protein isoforms in barley shoots. Phytochemistry 65:1853–1863

Taub DR, Miller B, Allen H (2008) Effects of elevated CO_2 on the protein concentration of food crops: a meta-analysis. Glob Change Biol 14:75–565

Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science 327:818–822

Tisné S, Serrand Y, Bach L, Gilbault E, Ben Ameur R, Balasse H, Voisin R, Bouchez D, Durand-Tardif M, Guerche P, Chareyron G, Da Rugna J, Camilleri C, Loudet O (2013) Phenoscope: an automated large-scale phenotyping platform offering high spatial homogeneity. Plant J 74:534–544 (Oxford)

Turner LB, Cairns AJ, Armstead IP, Ashton J, Skot K, Whittaker D, Humphreys MO (2006) Dissecting the regulation of fructan metabolism in perennial ryegrass (Lolium perenne) with quantitative trait locus mapping. New Phytol 169:45–58

Virdi AS, Thakur A, Dutt S, Kumar S, Singh P (2009) A sorghum 85 kDa heat stress-modulated protein shows calmodulin-binding properties and cross-reactivity to anti-*Neurospora crassa* Hsp 80 antibodies. FEBS Lett 583:767–770

Wang W, Meng B, Ge X, Song S, Yang Y, Yu X et al (2008) Proteomic profiling of rice embryos from a hybrid rice cultivar and its parental lines. Proteomics 8:4808–4821

Wang L, Ma H, Song L, Shu Y, Gu W (2012a) Comparative proteomics analysis reveals the mechanism of pre-harvest seed deterioration of soybean under high temperature and humidity stress. J Proteomics 75:2109–2127

Wang Y, Xu W, Zhao W, Hao J, Luo Y, Tang X, Zhang Y, Huang K (2012b) Comparative analysis of the proteomic and nutritional composition of transgenic rice seeds with cry1ab/ac genes and their non-transgenic counterparts. J Cereal Sci 55:226–233

Witzel K, Pietsch C, Strickert M, Matros A, Roder MS, Weschke W, Wobus U, Mock HP (2011) Mapping of quantitative trait loci associated with protein expression variation in barley grains. Mol Breeding 27:301–314

Yang JY, Sun Y, Sun AQ, Yi SY, Qin J, Li MH et al (2006) The involvement of chloroplast HSP100/ClpB in the acquired thermo tolerance in tomato. Plant Mol Biol 62:385–395

Yang F, Melo-Braga MN, Larsen MR, Jorgensen HJL, Palmisano G (2013) Battle through signaling between wheat and the fungal pathogen *Septoria tritici* revealed by proteomics and phosphoproteomics. Mol Cell Proteomics 12:2497–2508

Zhang W, Zhou RG, Gao YJ, Zheng SZ, Xu P, Zhang SQ, Sun DY (2009) Molecular and genetic evidence for the key role of AtCaM3 in heat-shock signal transduction in arabidopsis. Plant Physiol 149:1773

Index

R. Fritsche-Neto and A. Borém (eds.), *Phenomics*,
DOI 10.1007/978-3-319-13677-6