Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Unlocking the potential of plant phenotyping data through integration and data-driven approaches

Frederik Coppens[1,2], Nathalie Wuyts[1,2], Dirk Inzé[1,2] and Stijn Dhondt[1,2]

Abstract

Plant phenotyping has emerged as a comprehensive field of research as the result of significant advancements in the application of imaging sensors for high-throughput data collection. The flip side is the risk of drowning in the massive amounts of data generated by automated phenotyping systems. Currently, the major challenge lies in data management, on the level of data annotation and proper metadata collection, and in progressing towards synergism across data collection and analyses. Progress in data analyses includes efforts towards the integration of phenotypic and -omics data resources for bridging the phenotype–genotype gap and obtaining in-depth insights into fundamental plant processes.

Addresses

[1] Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, B-9052, Ghent, Belgium

[2] Center for Plant Systems Biology, VIB, Technologiepark 927, B-9052, Ghent, Belgium

Corresponding author: Inzé Dirk (dirk.inze@ugent.vib.be)

Current Opinion in Systems Biology 2017, **4**:58–63

This review comes from a themed issue on **Big data acquisition and analysis (2017)**

Edited by **Pascal Falter-Braun and Michael A. Calderwood**

For a complete overview see the Issue and the Editorial

Available online 8 July 2017

http://dx.doi.org/10.1016/j.coisb.2017.07.002

Keywords

Plant phenotyping, Data management, Data integration, Data-driven analysis.

Introduction

During the past decade, plant phenomics has evolved from an emerging niche to a thriving research field, both in academia and industry. This can be largely attributed to the use of imaging for the non-invasive analysis of structural, physiological and performance-related plant traits [1]. Automated image analysis procedures allow substantial increases in the throughput of trait measurements, thereby countering the so-called phenotyping bottleneck, which considers phenotypic measurements the rate-limiting factor in the functional analysis of specific genotypes or the assessment of genotype performance in plant breeding [2]. Improvements in plant imaging have been accompanied by technological advancements in plant handling and camera positioning to keep up with the speed of image acquisition. Plant-to-sensor systems, utilizing conveyors and grippers to present the plant to the camera, and sensor-to-plant systems, which move the camera to the plants, have been developed in growth cabinets, chambers and greenhouses [3]. While the vast majority of the phenotyping is still done manually under field conditions, automated image acquisition always occurs in a sensor-to-plant fashion, assisted by manual or engine-driven 'phenomobiles', gantry systems on the ground, or unmanned aerial vehicles (UAVs) [4].

Undoubtedly, it is the development of digital image sensors that underlies this remarkable evolution in plant phenotyping. Sensitivity of the sensor for a specific part of the electromagnetic spectrum, in combination with appropriate filters, defines which traits can be extracted. Typical Red Green Blue (RGB) color sensors are sensitive to wavelengths in a range from 400 to 1000 nm. Most color cameras provide an infrared (IR) cut-off filter for imaging specifically in the visible spectrum, but without this filter, they allow near-IR imaging, and as such image acquisition of plants in the dark [5,6]. Indium gallium arsenide (InGaAs) sensors show a spectral response to a range from approximately 900 to 1700 nm. These sensors are used in Short Wave InfraRed (SWIR) cameras, which can be adopted for the measurement of water content in plants [7]. Long Wave Infrared (LWIR) sensors with a spectral range of 3–14 μm, on the other hand, are used for thermal imaging of shoots as a proxy for stomatal conductance or water use behavior in general [8].

The use of advanced imaging systems has drastically increased the volume of data from a couple of bytes, e.g. manually scored traits in a spreadsheet, to several megabytes (MB) or sometimes more than 100 MB, e.g. in the case of hyperspectral imaging or scene characterization by means of video capture. Data are also stored in a myriad of formats on diverse types of media ranging from a researcher's hard drive to local server stations or in "the cloud". Proper annotation of data to ensure their continued relevance after acquisition is thus essential. Furthermore, because the plant's phenotype is the result of a strong interaction between its genotype and the environment in which it grows

(G × E) [9], plant phenotyping efforts should include the logging of environmental conditions, which in turn requires the collection of metadata on the sensors in use. Because of the tremendous amounts and diversity of data produced within the plant phenotyping research field, data management, storage and analysis are currently considered as the major challenges. On the other hand, large datasets may also create opportunities for data modeling and machine learning towards "Big Data" analyses.

Data management to enable data integration

The current technologies and methods used in plant phenotyping generate a huge amount of complex, unstructured "Big Data", which can give the impression that a lot of the phenotype data might not be retrieved anymore [10]. In first instance, phenotypic data management requires the use of ontology terms for the unique and repeatable annotation of data in order to ensure their persistence in view of traceability and reuse under the form of data sharing and meta-analyses. The use of ontologies therefore promotes synergism. Moreover, in contrast to repositories such as the European Nucleotide Archive (ENA) [11] or Sequence Read Archive (SRA) [12] for sequencing data, there is currently no central, structured repository for phenotyping data or metadata. Although data can be uploaded to general purpose repositories such as Zenodo (https://zenodo.org/), FigShare (https://figshare.com) and Dryad (http://datadryad.org), these do not provide services to facilitate the description of, access to and integration of data. As a consequence of the lack of a central repository, advanced data mining and discovery depends on the error-prone scavenging of scientific literature. As a consequence, a plethora of resources has been developed by individual research groups and consortia, ranging from resources dedicated to one species or one type of phenotyping system to more generic platforms allowing the integration of several data types. AraPheno provides a central repository of population-scale phenotypes for Arabidopsis accessions [13], whereas the Plant Genomics and Phenomics (PGP) research data repository is an infrastructure to comprehensively publish plant research data covering cross-domain datasets [14]. The Phenomics Ontology Driven Data (PODD) repository was developed to handle and distribute phenotyping data and metadata from Australian facilities [15]. ClearedLeaves DB functions as an online database of cleared plant leaf images [16]. Phenopsis DB is an information system for sharing data generated by the PHENOPSIS plant phenotyping platform [17] and PhenoFront is a web-server front end to the LemnaTec Phenotyper platform [18]. Whereas BreeDB hosts datasets of tomato and potato populations (https://www.eu-sol.wur.nl), Genoplante Information System (GnpIS) is a multispecies integrative information system dedicated to plant and fungi pests, bridging genetic and genomic data [19]. This non-exhaustive list illustrates the variety of available resources, which in some cases, provide the data for download and further analysis.

Many of these data resources have been built to organize a huge amount of collected phenotypic data. In the light of high-throughput phenotyping, there is a need for managing the data at the moment it is being generated (Figure 1). Besides data derived from experiments, provisions are made for metadata related to the environment sensors in use, and to the imaging sensors themselves, including the type of sensor, the camera systems and their optical properties. The latter are required for image analysis, whereas the whole ensures traceability and quality insurance. These functionalities are built-in in PIPPA, the PSB (Plant Systems Biology) Interface for Plant Phenotype Analysis (https://pippa.psb.ugent.be), a web-based framework for the analysis, visualization and management of phenotypic data, which enables biologists to perform dedicated image processing and (statistical) analyses of data generated by Weighing, Imaging and Watering Machine (WIWAM) phenotyping platforms or of externally imported data. Frameworks harboring comparable functionalities include Integrated Analysis Platform (IAP), and Plant Computer Vision (PlantCV) [18,20].

Image data extraction

The advanced development of imaging in plant phenotyping enables multi-dimensional, high-throughput monitoring of plants at an increasing pace. Although numerous image analysis software tools are available for the extraction of biologically meaningful phenotypic or physiological parameters from these images [21,22], they mainly focus on the analysis and often are disconnected from the data management part. To address this, dedicated analysis platforms have been developed: IAP [20], PlantCV [18], InfraPhenoGrid [23], OMERO [24], BisQue on CyVerse [25], and PIPPA (https://pippa.psb.ugent.be). These systems offer a user-friendly interface to a grid compute cluster that facilitates researchers without a computer science background to run image analysis pipelines. Moreover, they also cater for bioinformaticians as they are inherently flexible, allowing custom analysis pipelines through extensions or Application Programming Interfaces (APIs). These platforms ensure provenance through metadata and thus play an important role in data management. Data visualization is also an important aspect, both for reporting and interpretation, as well as for quality control of the input data (Figure 1). For example, PIPPA deploys several 'sanity check' algorithms to flag outliers for further inspection.

As our capacity to extract information from images increases, so do the size and complexity of the derived