The Next Generation of Training for Arabidopsis Researchers: Bioinformatics and Quantitative Biology

Joanna Friesner[1], Sarah Assmann[2], Ruth Bastow[3], Julia Bailey-Serres[4], Jim Beynon[5], Volker Brendel[6], Robin Buell[7], Alexander Buksch[8], Wolfgang Busch[9,10], Taku Demura[11,12], Jose R. Dinneny[13], Colleen J. Doherty[14], Andrea L. Eveland[15], Pascal Falter-Braun[16,17], Malia A. Gehan[15], Michael Gonzales[18], Erich Grotewold[19] Rodrigo Gutierrez[20], Ute Kramer[21], Gabriel Krouk[22], Shisong Ma[23], R.J. Cody Markelz[24], Molly Megraw[25], Blake C. Meyers[15], Jim Murray[26], Nicholas J. Provart[27], Sue Rhee[13], Roger Smith[28], Edgar Spalding[29], Crispin Taylor[30], Tracy Teal[31], Keiko U. Torii[32], Chris Town[33], Matthew Vaughn[34], Richard Vierstra[35], Doreen Ware[36,378], Olivia Wilkins[38], Cranos Williams[39], Siobhan M. Brady[40+]

[1] Agriculture Sustainability Institute and Department of Neurobiology, Physiology and Behavior, University of California, Davis, CA, USA 95616
[2] Biology Department, Penn State University, University Park, PA, USA, 16802
[3] GARNet, School of Biosciences, Cardiff University, Cardiff, UK
[4] Center for Plant Cell Biology and Botany and Plant Sciences Department, University of California, Riverside, CA, USA 92521
[5] School of Life Sciences, Gibbet Hill Campus, The University of Warwick, Coventry, CV4 7AL, UK
[6] Department of Biology and Department of Computer Science, Indiana University, Bloomington, IN, USA 47405
[7] Department of Plant Biology, Michigan State University, East Lansing, MI, USA 48824
[8] University of Georgia, Department of Plant Biology; Warnell School of Forestry and Natural Resources; and Institute of Bioinformatics, Athens, GA, USA 30602
[9] Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna Biocenter (VBC), Dr. Bohr-Gasse 3, 1030 Vienna, Austria
[10] Salk Institute for Biological Studies, Plant Molecular and Cellular Biology Laboratory, 10010 N Torrey Pines Rd, La Jolla, CA, USA 92037
[11] Graduate School of Biological Sciences, Nara Institute of Science and Technology, Ikoma, Nara, 630-0192, Japan
[12] RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, 230-0045, Japan
[13] Carnegie Institution for Science, Department of Plant Biology, Stanford, CA, USA 94305
[14] Department of Molecular and Structural Biochemistry, North Carolina State University, Raleigh, NC , USA 27695
[15] Donald Danforth Plant Science Center, St. Louis, MO, USA, 63132
[16] Institute of Network Biology (INET), Helmholtz Zentrum München (HMGU), German Research Center for Environmental Health, 85764 Neuherberg, Germany
[17] Department of Microbe-Host Interactions, Ludwig-Maximilians-Universität München (LMU Munich), Planegg-Martinsried, Germany
[18] Center for Applied Genetic Technologies (CAGT), 111 Riverbend Road, Athens, GA, USA 30602
[19] Center for Applied Plant Sciences and Dept. of Molecular Genetics, The Ohio State University, Columbus, OH, USA 43220.
[20] FONDAP Center for Genome Regulation. Millennium Nucleus Center for Plant Systems and Synthetic Biology. Departamento de Genética Molecular y Microbiología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Avenida Libertador Bernardo O'Higgins 340, Santiago, Chile 833115
[21] Molecular Genetics and Physiology of Plants, Faculty of Biology and Biotechnology, Ruhr University Bochum, 44801 Bochum, Germany
[22] Laboratoire de Biochimie et Physiologie Moléculaire des Plantes, UMR CNRS/INRA/SupAgro/UM, Institut de Biologie Intégrative des Plantes "Claude Grignon," Place Viala, 34060 Montpellier Cedex, France.

[23] School of Life Sciences, University of Science and Technology of China, 443 Huangshan Road, Hefei, Anhui, 230027, China
[24] Department of Plant Biology, University of California, Davis, CA, USA 95616
[25] Department of Botany and Plant Pathology; Department of Computer Science; and Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR, USA 97331
[26] School of Biosciences, Sir Martin Evans Building, Cardiff University, Museum Avenue, Cardiff CF10 3AX, Wales, UK.
[27] Department of Cell & Systems Biology / Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON. M5S 3B2, Canada
[28]Syngenta Biotechnology, Inc.PO Box 122573054 E. Cornwallis Road, Research Triangle Park, NC 27709, USA
[29] Department of Botany, University of Wisconsin, 430 Lincoln Drive, Madison, WI, USA 53706
[30] American Society of Plant Biologists, Rockville, MD, USA 20855
[31] Data Carpentry, Davis, CA, USA 95616
[32] Howard Hughes Medical Institute and Department of Biology, University of Washington, Seattle, WA, USA 98195
[33] J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD, USA 20850
[34] Life Sciences Computing, Texas Advanced Computing Center, 10100 Burnet Rd, Austin, TX, USA 78758
[35] Department of Biology, Washington University in St. Louis, St. Louis, MO, USA 63130
[36] Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA 11724
[37] US Department of Agriculture, Agricultural Research Service, Ithaca, New York, USA 14853
[38] Department of Plant Science, McGill University, Montreal, QC, Canada H9X 3V9
[39] Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, USA 27695
[40]Department of Plant Biology and Genome Center, University of California, Davis, CA, USA 95616
+Corresponding author – sbrady@ucdavis.edu

Abstract

Researchers studying plants using the model organism, *Arabidopsis thaliana*, can easily generate or access massive datasets using modern technologies. However, in order to best analyze such datasets to elucidate novel biological mechanisms, many individuals face critical deficiencies in their training. Ideally, these scientists will be able to, individually or in a team, integrate foundational concepts from biological science, chemistry, mathematics, statistics, computer science, bioinformatics and data science. Here, we provide examples of guidelines, skill sets, and core competencies that should be considered when developing curricula or training efforts at the undergraduate, graduate, postdoctoral and faculty levels. Discussion of specific training needs from the perspective of the agricultural biotechnology industry are also provided. Critical to "large-scale biology" is the formation of productive collaborations. Methods to identify the best collaborator, to define an effective collaboration on the part of all partners, and pedagogical methods to train students in the art of collaboration are also discussed. Finally, these challenges and potential solutions are addressed in a selected case study on high-throughput phenotyping.

Introduction

It has been over 50 years since *Arabidopsis thaliana* was first introduced as a model organism to understand basic processes in plant biology. A well-organized scientific community has used this small "reference" plant species to make numerous fundamental plant biology discoveries (Provart et al., 2016). Due to an extremely well annotated genome and advances in high-throughput sequencing, our understanding of this organism and other plant species has become ever more intricate and complex. Computational resources including CyVerse[1], Araport[2], TAIR[3] and BAR[4] have further facilitated novel findings with just the click of a mouse. As we move towards understanding biological systems, Arabidopsis researchers will need to use more quantitative and computational approaches in order to extract novel biological findings from these data. Here, we discuss guidelines, skill sets, and core competencies that should be considered when developing curricula or training undergraduate or graduate students, postdoctoral researchers, and faculty. A selected case study provides more specificity as to the concrete issues that plant biologists face and how best to address such challenges.

Transforming Education and Training – from Undergraduates to Faculty

An overhaul in training is necessary for plant biologists to make use of massive data sets and enabling technologies. This is not a novel idea in the life sciences. In fact, Bialek and Botstein (2004) articulated a concept for an integrated introductory quantitative science curriculum, primarily for undergraduates, in order to address this specific issue. Their publication has been highly cited and used as a foundational resource. They noted that biologists have too little education and experience in quantitative thinking and computation relative to what is needed for full participation in this new era of genomics research. Both then, and still now, many upper-level undergraduates in the life sciences versus quantitative sciences already speak "noticeably different languages". Bialek and Botstein (2004) proposed that instead of prerequisite courses in mathematics, physics, chemistry and computation, the fundamental ideas of each of these disciplines should be introduced at a high level of sophistication. Their point is that these ideas should be presented in context and with relevant biological problems for a "seamless" educational experience. This would also avoid the delivery of these quantitative science courses as a "service" for the life sciences students. In a "service course", students often exhibit a lack of enthusiasm due to the fact that they are required to take these courses. An additional issue is that many of the quantitative concepts presented are devoid of a biological perspective. Training at the graduate level must also necessarily

[1] http://www.cyverse.org/
[2] https://www.araport.org/
[3] http://www.arabidopsis.org/
[4] http://bar.utoronto.ca

integrate foundational concepts from biological science, chemistry, mathematics, statistics, computer science, bioinformatics and data science. We stress that this is more than simply an understanding of bioinformatics - that is, more than just using computation to extract knowledge from biological data. Instead, education in plant biology should be truly interdisciplinary, perhaps as exemplified by (i) theoretical biology whereby theoretical perspectives (often mathematical) are used to give insights into biological processes, (ii) quantitative biology whereby quantitative approaches and technologies are used to analyze and integrate biological systems or to construct and model engineered life systems, or (iii) computational biology whereby biological data are used to develop algorithms or models to understand relationships amongst various biological systems.

Implementation of Quantitative Training in the Life Sciences

Significant administrative, content, and logistical challenges often exist to impede the creation of new academic programs. Despite this, a growing number of institutions are developing undergraduate and graduate curricula in bioinformatics and computational biology for the life sciences, many of which incorporate the vision of Bialek and Botstein[5]. Practical strategies to overcome many of these challenges have been described for an overhaul in the graduate training program at Harvard Medical School (Gutlerner and Van Vactor, 2013). Our primary recommendation is to include in life sciences curricula the teaching of the skills and competencies described above, with the aim to develop students and future scientists that are adept at using transdisciplinary approaches to solve challenges in biology, and thus, well adapted to addressing current and future needs in modern plant biology research.

Minimal Skill Sets and Core Competencies

Over the last 15+ years a variety of meetings and task forces have been convened to determine the nature, extent, content, and available delivery tools for degree and training programs utilizing bioinformatics or computational biology in life sciences programs. Tan et al. (2009) proposed a generalized minimum set of competencies that the next generation of biologists will need to effectively cope with ever-increasing amounts of information and datasets, and the growth of importance in informatics in this genomics era; the following competencies have increased in relevance since they were first published, and thus could guide curricula development (or revisions of existing curricula):

1. Basic knowledge in the specific domains of computer science, statistics and mathematics that intersect with modern biology.

[5] http://www.bioinformatics.org/wiki/education_in_the_united_states

2. Expertise in communicating and representing biological knowledge and processes in mathematical, statistical and computing terms and concepts.
3. The ability to use or develop efficient bioinformatics and biocomputational tools and techniques for the acquisition, interpretation, analysis, prediction, modeling, simulation and visualization of experimental and other biological data.
4. Proficiency in the search, retrieval, processing, curation, organization, classification, management and dissemination of biological data and information in databases for deriving biological insights and knowledge discovery.
5. Critical thinking and problem-solving skills in quantitative aspects of biology.

As a community with expertise in quantitative and computational plant biology, and using these competencies as a guideline, we further propose a suite of minimal skill sets [adapted from (Rubinstein and Chor, 2014; Welch et al., 2014)] which will enable a plant biologist to generate and utilize multi-dimensional and scaled plant biology data in order to answer central biological questions (**Table 1**).

Table 1: Minimal Skill Sets Recommended for Plant Biology Students

Category	Specific Skills
Unix/Linux	Comfort/familiarity with using command line
Scripting Language	Perl or Python, for advanced students – C++, CUDA
Database creation and query	Mongo or MySQL, data mining
Software carpentry	Best practices, proper commenting, version control
Computation	Machine learning, algorithm design and analysis, distributed and high-performance computing
Statistical methods	Descriptive and inferential statistics, hypothesis testing, parameter estimation, power analysis, data transformations, meta-analysis, hierarchical clustering
Mathematical	Probability theory, differential equations, graph theory. Linear algebra, information theory
Statistical programming	R/Bioconductor (particularly for analysis of next-generation sequencing data)
Biological databases and resources	NCBI, EBI, Araport, TAIR, MaizeGDB, Gene Ontology etc.
Network Analysis	Cytoscape plugins
Data Visualization	Could include ggplot, visualization of genome-scale data in genome browsers, volcano plots, heat maps etc.

We suggest two possibilities to implement across diverse institutions this integrated paradigm for training in this suite of minimal skill sets and core competencies. So as not to reinvent the wheel, it may be fairly straightforward for a plant biology program to participate in an extant integrative biology/quantitative sciences program within their respective institution, if those programs fulfill this suite of core recommended competencies/skill sets, simply by augmenting existing programs with elective plant courses. Alternatively, a program could implement course curricula (both undergraduate and graduate) that have been described in the literature and for which resources are available. These include the Course Source Bioinformatics Learning framework, which has been developed and reviewed by members of the Genomics Education Partnership, the Network for Integrating Bioinformatics into Life Science Education, the Genome Consortium for Active Teaching of Next Gen Sequencing, and the Howard Hughes Medical Institute-sponsored Bioinformatics Workshop for Student/Scientist Partnerships (Rosenwald et al., 2016). Other curricula include a basic bioinformatics curriculum offered at the Free University of Berlin which emphasizes fundamentals in biology, mathematics and computer science (Koch et al., 2008), or a first-year graduate course in quantitative biology which emphasizes the integrated curriculum proposed by Bialek and Botstein (2004). The latter example uses breakthrough papers in diverse areas of biology, and that emphasize quantitative reasoning, theory, and experimentation, to convey the importance of quantitative knowledge to understand basic biological processes (Wingreen and Botstein, 2006). Similar curricula have been implemented in the UK and are considered requisite training for graduate students in plant biology[6] . A course entitled "Computational Approaches for Life Scientists[8]" has also been described which focuses on enriching the curriculum of life science students with abstract, algorithmic and logical thinking and exposes them to "computational culture" (Rubinstein and Chor, 2014). Such curricula should be followed by a more focused track in plant biology, again emphasizing the quantitative premises underlying plant biology. Finally, a capstone problem-solving course that integrates teamwork could provide practical examples of how to integrate these diverse and interdisciplinary subject materials to address unsolved questions in plant biology.

Bridge Programs, Bootcamps and Supportive Environments for Quantitative-Based Plant Biology Education

Even without creating new programs, supportive environments for students interested in both plant and computational biology could help lower the "intimidation" barrier. For example, this could involve creating quantitative biology interest groups. Additional vehicles to encourage peer-to-peer learning could include hackathons (events that bring people together in teams for collaborative computer programming efforts to creatively solve a problem) that would provide training, while encouraging interactions between plant biology and computational students.

Recently, organizations such as Software Carpentry[7] and Data Carpentry[8] (which are merging into one organization) and Amelieff[9] have been created to fill in some of the gaps in education for

[6] https://sysmic.ac.uk
[7] http://ca4ls.wikidot.com
[8] http://www.datacarpentry.org/

programming and data science skills. Since 2015, these organizations have held workshops at institutions across the world. Other short courses also exist globally which focus on training experimental biologists in bioinformatics, statistical genetics and mathematical modeling including the Summer Institute of Statistical Genetics (USA)[10], the Summer School for Statistical Genetics (Japan)[11], the Santa Barbara Advanced Summer School of Quantitative Biology (USA)[12], the BioComp training series (Austria), the Summer School (Germany)[13], the Saclay Plant Sciences summer schools (France)[14], the Integrative Database training course (Japan)[15], the Large Biological Data Analysis Course (Japan)[16], and the Cold Spring Harbor Laboratory courses[17] (USA) in "Frontiers and Techniques in Plant Science" and "Programming for Biology". However, access to these courses is limited, and the course fees and travel necessary to participate may present significant barriers. In order to enhance the flexibility and to minimize financial input, curricula could be complemented with short-courses or with certificates from online Massive Open Online Courses (MOOCs). As a community, developing a portal that provides reviews and ratings of these programs would be a valuable resource (Searls, 2012). It should be noted, however, that a recent report assessing boot camp programs (from 2 days to 2 weeks in length) typically designed to expose graduate students to data analysis techniques (amongst others) found a null difference when assessing research skill development, despite a statistically significant increase in perceived skill advancement (Feldon et al., 2017).

Funding

While many academic institutions recognize the importance of these training efforts, they need funding to come into existence. The United States National Science Foundation (NSF) Research Traineeship (NRT) Program[18] Traineeship Track specifically fosters interdisciplinary training. The German Research Foundation provides funding for International Research Training Groups dedicated to a focused "study abroad" research program and a structured training strategy. In France, local funding agencies named LABEX (for "Laboratoire d'Excellence") fund interdisciplinary interactions between local partners,

[9] http://amelieff.jp/english/
[10] https://www.biostat.washington.edu/suminst/sisg
[11] http://www.sg.med.osaka-u.ac.jp/school.html
[12] https://www.kitp.ucsb.edu/qbio
[13] GCBN/de.NBI
[14] https://www6.inra.fr/saclay-plant-sciences_eng/Teaching-and-training/Summer-schools/Summer-School-2016
[15] https://biosciencedbc.jp/en/
[16] https://biosciencedbc.jp/en/
[17] http://meetings.cshl.edu/courseshome.aspx
[18] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505015

an example being Numev[19], which promote interactions between computer and mathematical scientists and biologists with strong support of plant scientists. The CNRS (Centre National de la Recherche Scientifique) regularly promotes biology and math interactions through specific grant calls led by its Office for Interdisciplinary Research (PEPS). In many of these cases, however, proposals are granted only for specific areas deemed to be a 'high priority' to each funding organization, which may lower the success of proposals that do not fit easily into the chosen scope.

Additional Recommendations for Postdoctoral Scholars and Faculty

At the moment, there are no standardized modes of quantitative or interdisciplinary training for postdoctoral fellows in plant biology. Thus, postdoctoral scholars often need to identify their own opportunities for additional training, if they have not received such training during their undergraduate or graduate training. Many competitive postdoctoral scholar fellowships offer funds for additional training including NSF's Plant Genome Research Program Postdoctoral Research Fellowships in Biology (PGRP PRFB)[20], the USDA's AFRI Food, Agriculture, Natural Resources, and Human Sciences Education and Literacy Initiative Fellowship program (AFRI ELI)[21] and the National Institute of Health K99 grant program[22]. The Human Frontiers Science Program offers postdoctoral fellowships for citizens of many countries with a special category for cross-disciplinary fellowships to support training those in quantitative sciences in experimental biology[23]. Moreover, the European Union's Marie Skłodowska-Curie Actions Individual Fellowships offer funds for additional training and for short 3 to 6 month visits. The Plant Biology section of the General Program and the Young Scientists Fund of the National Natural Science Foundation of China (NSFC) encourages interdisciplinary research that combine methods from plant biology and other areas, such as mathematics, physics, and computer sciences[24].

However, these fellowships are quite competitive and can be restricted to postdoctoral scholars trained in certain disciplines. What if a postdoctoral scholar is unsuccessful at receiving such funds but still wishes to undergo interdisciplinary training? In Germany, there is a growing number of structured postdoctoral fellowship programs funded by individual research institutions that offer institutional support in identifying interdisciplinary training opportunities. The Postdoctoral Fellowship Program (PFP) by the HelmholtzZentrum Munich ensures that fellows are integrated into international and

[19] http://www.lirmm.fr/numev/
[20] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503622
[21] https://nifa.usda.gov/program/afri-education-and-literacy-initiative
[22] https://www.nhlbi.nih.gov/research/training/programs/postdoc/pathway-parent-k99-r00
[23] http://www.hfsp.org/funding/postdoctoral-fellowships
[24] http://www.nsfc.gov.cn/nsfc/cen/xmzn/2017xmzn/01/03sm/001.htm

interdisciplinary research groups, while the University Foundation Fellowship Program by the Technical University of Munich assists with the identification of interdisciplinary and collaborative research programs. Additional institutional solutions could provide the resources for postdoctoral participation (and instruction) in short courses that provide training in a particular competency, or could integrate postdoctoral scholars in existing courses provided for graduate students. At the mid- and senior-postdoctoral scholar level, perhaps the best way is to provide opportunities for senior “biologically-oriented” postdoctoral scholars to engage in dedicated training via short-term “residencies” (3 to 6 months) in a laboratory that specializes in quantitative, computational, or modeling analyses. Such longer-term dedicated learning programs would have the advantage of carrying out a distributed practice of learning, which has proven more beneficial in long-term retention of concepts, relative to the shorter mass “boot-camp”-type strategy. (Feldon et al., 2017).

Short-term or long-term sabbaticals in a computational lab are also a good solution for faculty members to acquire computational skills. The USA National Science Foundation’s Mid-Career Investigator Awards in Integrative Organismal Biology (MCA-IOS)[25] could be a source of funding for associated travel costs. The German Academic Exchange Service (DAAD) and the French AGreskills federal programs, as well as the local LabEx programs (mentioned previously) financially support sabbaticals for this purpose. Alternatively, it may be better for faculty to focus on how they can better assess and support research activities in their own lab and be able to better understand how to review papers or grants that contain research of an interdisciplinary nature. Short workshops could be developed to provide training to faculty on quantitative and computational methods and how to conduct high-quality computational/quantitative research.

Computational Training for Industry

The key attributes for researchers in industry with respect to projects involving computational approaches are strong interpersonal skills in teamwork, collaboration, communication, and project management. Industry requires individuals who are expert in one specific area but have the breadth of understanding that allows them to appreciate and respect the input of other disciplines to the overall project. This includes familiarity with biological databases and quantitative biology approaches. In addition, employees in industry benefit from training programs which expose workers in academia and industry to each other’s ways of working. The European funding model encourages partnerships between researchers and industry (e.g., the bread wheat initiative led by INRA[26]). Another model is to embed master's or doctoral students in industry placements for three to six months. Two UK-specific examples of this are the

[25] https://www.nsf.gov/pubs/2017/nsf17508/nsf17508.htm
[26] http://www.wheatinitiative.org

compulsory program of the UK Biotechnology and Biological Sciences Research Council (BBSRC), called "PIPS" (Professional Internships for PhD Students[27]), and the Flexible Interchange Program (FLIP[28]) which operates at the postdoctoral scholar and faculty level to promote training and exchange between industry and academic partners. An additional twist on this theme is provided by the Chilean scientific funding agency CONICYT that offers a post-graduate thesis in industry[29]. At the institutional level, research institutions dedicated to applied sciences and industrial cooperation, like the Fraunhofer Institutes in Germany, traditionally work in close cooperation with industry including master's and doctoral students.

Arabidopsis Training for Plant-Curious Data Scientists

A rapidly growing world population and a changing climate demand development of improved crop varieties that yield more with fewer inputs, as well as advances in renewable fuels, and biomaterials. Moving forward, a community-wide effort to promote the value of plant science research to data scientists is needed. Arabidopsis training for "plant-curious" data scientists should emphasize 1) how knowledge gained from Arabidopsis research is relevant to crop improvement, and 2) how to utilize Arabidopsis as a tool to rapidly test gene function and optimize emerging technologies prior to delivery to a crop system. The advent of gene editing technologies, such as CRISPR/Cas9-based approaches to specifically target loci for site-directed mutagenesis or sequence replacement, introduces a new paradigm. While these technologies create opportunities for targeted mutagenesis directly in crop species, significant bottlenecks in the transformation process limit the extent to which these experiments can be performed in crops. Therefore, Arabidopsis can be used to more quickly and efficiently test functional hypotheses and prioritize experiments for the more labor-, cost-, and time-intensive studies in crops.

The outcome of an active community of Arabidopsis researchers is the detailed curation of genes and pathways in the Arabidopsis genome, perfect for mining by data scientists. This curation data has been leveraged for annotating orthologous genetic components in other species, and thus is an invaluable resource. It is likely that many fundamental biological processes are conserved across plant species (McGary et al., 2010; Oellrich et al., 2015). As an example, agricultural biotechnology industries make use of this information through large-scale text mining algorithms combined with comparative genomics approaches to project annotations and associations onto crop models (Holtan et al., 2011; Preuss et al., 2012). The depth and breadth of these resources in Arabidopsis also positions this organism at the

---

[27] http://www.bbsrc.ac.uk/funding/filter/professional-internships/

[28] http://www.bbsrc.ac.uk/funding/filter/flexible-interchange-programme/

[29] http://www.conicyt.cl/wp-content/uploads/2012/07/Brochure-Institucional-2011-Inglés.pdf

forefront of predictive modeling in plants through systems biology approaches. Moving forward, there is an immediate need to make better use of existing data from Arabidopsis studies by developing new data integration paradigms aimed at predictive modeling and subsequent discovery. Using Arabidopsis as a framework for how to integrate diverse datasets should facilitate similar analyses in species with less developed resources.

On the other hand, Arabidopsis may not be the most appropriate model to understand traits related to domestication, physiology such as C4 photosynthesis, or other aspects of plant biology such as secondary metabolism. To address such questions alternative model systems are being established; this includes *Setaria viridis* and *Brachypodium distachyon* as model grass species (Brutnell, 2015; Brutnell et al., 2015) and *Camelina sativa* for metabolic engineering of co-products (Bansal and Durrett, 2016; Zhu et al., 2016). We recommend that the communities developing these new systems leverage best practices from the Arabidopsis community, particularly with reference to genome annotation and data curation for these species. Fostering such interactions between scientists could occur through cross-species conferences in plant science; for example, a Keystone Meeting focused on "Translational Plant Biology". Inclusion of data scientists in these forums will be critical to ensure maximal usefulness of these emerging model systems.

Collaborations

Taking advantage of large-scale datasets and technologies in order to reveal novel biological conclusions will require groups of people with diverse expertise, skill sets, and at different career levels to work well together. Thus, in order to train the next generation of Arabidopsis biologists in quantitative and computational biology, we also need to train scientists on how to initiate, define, manage and maintain effective collaborations.

Identifying collaborators

It is often difficult for biologists to develop their research questions to include tangible opportunities for quantitative experts, or to effectively articulate their specific needs in a vocabulary that is accessible to experts in those fields. Face-to-face communication is particularly important and thus we attribute the highest priority to the identification of regional collaborators. Inclusive, regular, cross-faculty and cross-institute interactions at all career levels, with the clear objective to also empower early-career researchers to take active roles, are required to initiate local collaborations. In order to implement role models for such collaborative efforts, hiring or recruiting researchers who already work across biological science and statistical, computational, or mathematical departments can be beneficial due to their ability to expose biological problems to theoreticians who might not typically see such data as valuable to analyze. However, the infrastructure for promotion and merit within most academic institutions has generally not advanced sufficiently to effectively hire and maintain theoreticians at the tenure-track level in biology departments.

Collaborations between disciplinary experts can be accelerated through intensive trainings and activities that promote networking and knowledge sharing. In-depth, week-long immersion sessions have proven effective at providing both the biologist and the quantitative expert with the proper, shared, vocabulary, resulting in productive collaborations. For example, the “Maths in the Plant Sciences” Study Group in the United Kingdom[30] has been successful in generating in short timeframes both new collaborations and funded grant applications.

Co-supervision of graduate students by a biologist and theoretician is another effective strategy to develop a collaboration. Initiating cross-disciplinary cohorts of graduate students is another approach. Complementing collaborative interactions, or in the absence of local cross-disciplinary opportunities, the availability of more high-quality online video material outlining advances in current plant biology, for example, in a jargon-free format would be useful for quantitative experts. In the long term, graduate students and postdoctoral scholars who have been trained in an interdisciplinary environment will likely generate the best collaborations. By working together from an early career stage, a deep appreciation of diverse abilities will be engendered and the ability to communicate freely will enable new research avenues to be pursued.

Defining collaborations

An effective multi-disciplinary collaboration must go beyond the mere provision of a service by a collaborator. As such, before initiating a project, all partners should jointly articulate and agree on the scientifically interesting research questions and discuss experimental design and data analysis. A management plan should involve contributors at all career levels and consider the benefit for each contributing individual. It is important for collaborators to recognize differences in cost bases for biological versus theoretical research (e.g. experimental laboratory-associated costs are quite high whereas in the theoretical sciences, experts command higher salaries than experimental biologists). A realistic assessment of project timelines and deliverables is critical. Furthermore, a plan to include periodic assessment of progress with respect to the defined timelines and deliverables should be implemented to allow for adaptation, with the understanding that things do not always proceed expectedly. Contingency plans are also ideal to establish at the start, as are plans for publications, since biological and theoretical fields have fundamentally differing authorship rules and norms, publication strategies, and career recognition criteria. It is important to discuss and specify the timeframes that are

[30] https://www.cpib.ac.uk/outreach/mpssg/

likely for the publication of biological data and how the development of novel theory or analysis tools could be published prior to their use in biological data analysis. To ensure recognition, CRedit[31] (through ORCid) comprises structured vocabulary for assigning author credit. It is also critical to put in place an explicit plan for the possibility of managing disagreements that may arise as well as the conditions under which a collaborator might exit a project.

In practice, project meetings between collaborators should be held at more frequent intervals than may normally occur in within-discipline collaborations. This is especially true at the beginning of the project where the development of mutual understanding and the building of close working relationships among the researchers are essential. If the collaboration is between local groups, regular, e.g. monthly, joint meetings would be ideal. If the collaboration involves partners at a considerable geographic distance, then monthly web-based meetings are necessary and the collaboration would benefit from face-to-face meetings with all team members, ideally once every six months at a minimum. Budgeting for necessary travel should be considered at the time of project design. Furthermore, the physical movement of postdoctoral scholars or graduate students between groups for reciprocal training or joint work contributes highly to the effective integration of projects. Appreciation of differences in language or culture should be conveyed, as should reciprocal trust and respect, interest in the mutual fields, and the willingness to learn from the expertise of a partner.

Case Study: Training Arabidopsis Biologists for High-Throughput Phenotyping

As a concrete example for how scientists can be trained and educated in an interdisciplinary, collaborative fashion using experimental biology and quantitative approaches, we consider phenomics as a case study. Phenomics is an emerging field at the intersection of plant biology, engineering, computer science and mathematics which has led to a deeper understanding of mechanisms for acclimation to environmental variation (Miller et al., 2007; Slovak et al., 2014; Campbell et al., 2015; Fahlgren et al., 2015; Rellán-Álvarez et al., 2015). These studies evolved from the need to characterize phenotypic traits across large numbers of genotypes (Chen et al., 2014; Cruz et al., 2016; Ge et al., 2016).

A project using phenomics can be considered as a pipeline with three identifiable stages: data acquisition, data analysis, and data modeling (Figure 1), all to answer a clear biological question. Generally, this question is: what genes or genetic regulation underlie a trait of interest? Generally, a consortium of scientists is needed to carry out a phenomic-scale project. Consortium members should have diverse skills, be able to interact collaboratively, and each researcher's role should be well defined. Prior to data acquisition, consortium members should collectively discuss and agree upon experimental

[31] http://casrai.org/credit

design, biological replicates, statistical power, the type of data to be acquired and appropriate models used for data analysis. The data acquisition stage includes the use of sensors such as cameras, fluorescent measurement devices, or any tool that can make a measurement when connected to a computer to measure a phenotype associated with a trait. This stage often leverages expertise in the engineering disciplines and may involve robotics. Input from biologists is needed in order to ensure that a physiologically relevant aspect of plant growth or response to the environment is being captured. The output of this stage is the generation of raw data files. The analysis stage includes the computer code needed to extract features from the raw data files - image analysis is a good example - to produce "measurements". This stage also includes 'workflow' software, which brings the raw data from the sensors to the analysis algorithms. The analysis phase passes processed data, or results, to the next stage. Again, input from an "experimental" biologist is needed to ensure that these data are within the expected range of values. The modeling stage involves synthesizing results for the purpose of generating new biological conclusions. A typical example would be integrating phenotype results with genotype information to complete a statistical genetic analysis. However, the modeling stage can also be conceptually general enough to include any sort of analysis that converts phenotype measurements into a new biological understanding.

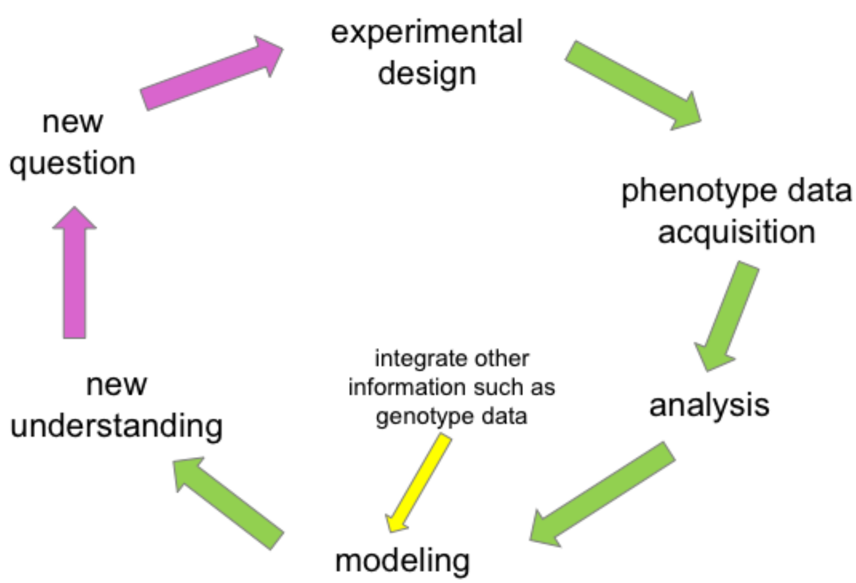

Figure 1. A computation-based phenotyping project requires a software continuum that takes raw data generated by acquisition activities, analyzes the raw data, integrates them with different data such as

genotype or environmental information, and then produces new understanding through modeling activities such as statistical associations. The new understanding leads to new questions.

Phenomic projects using Arabidopsis are ideal for training students in collaborative, innovative, and interdisciplinary approaches. Outreach and training modules on plant phenotyping naturally bridge multiple disciplines including plant biology, computer science, mathematics, and engineering, and provide alternative ways of attracting students to the plant sciences. Single-board computers like Raspberry Pi, Hummingboard, or Cubieboard are low-cost microcomputers originally built for educators, hobbyists and researchers, and are currently being incorporated into plant phenotyping research and teaching modules. Online resources provide tutorials to set up imaging systems (**Table 2**), however next-generation resources should be designed in collaboration with educational experts.

Table 2: Online resources providing tutorials to setup imaging systems.

Resource	Description	Website	Reference
Scikit-image examples and tutorials	Comprehensive list of imaging tasks with example code. Scikit-image is an imaging library for python	http://scikit-image.org/docs/dev/auto_examples/	(Van der Walt et al., 2014)
OpenCV tutorial	A collection of turtorial for openCV in C++. Open CV is a standard computer vision library available in C++, python and other languages	http://docs.opencv.org/2.4/doc/tutorials/tutorials.html	(Bradski and Kaehler, 2008)
Mahotas documentation	Mahotas is a python library written in C++. The documentation provides many examples for standard imaging tasks	http://mahotas.readthedocs.io/en/latest/	(Coelho, 2013)
DIRT tutorials and videos	DIRT (Digital Imaging of Root Traits) is an online root phenotyping platform that allows users to submit root images for phenotyping. The website contains tutorial and videos for non-technical users as well as documentation for developers. It's source code is freely available.	Online interface: http://dirt.iplantcollaborative.org/get-started Source Code: https://github.com/Computational-Plant-Science/DIRT	(Bucksch et al., 2014) (Das et al., 2015)

Phenotiki	Hardware (Raspberry Pi) and software for analyzing growth chamber collected phenotyping data	http://phenotiki.com/getting_started.html	(Minervini et al., 2014) (Giuffrida et al., 2015)

Executive Summary

Historically, the Arabidopsis research community has been able to effectively combine efforts internationally and to provide a collective voice regarding our needs to facilitate fundamental biological discoveries. We propose that such synergism be employed, using the specific recommendations in this commentary as a guide, in training this next generation of plant biologists to be able to understand and implement, in a rigorous manner, quantitative approaches in their research.

Specifically – for undergraduate and graduate training we recommend an overhaul in curriculum design for plant biology majors or plant biology graduate students that involves a seamless integration of concepts in math, physics, statistics and computation within courses that illustrate biological processes. This could be done according the recommendations of Bialek and Botsein (2004). We have adapted a set of core competencies and minimal skill sets, adapted from those of Tan et al. (2009), Rubinstein and Chor (2014), and Welch et al. (2014), and we strongly recommend that, when designing or revising curricula for this next generation of plant biologists, that these core competencies and skills are kept in mind. We have highlighted above a set of curricula based on these and which are publicly available either within the US or internationally; these may serve as a further resource. While there is no existing training standard for postdoctoral scholars in plant biology, we have identified a suite of fellowships for which postdocs may apply and which facilitate independent interdisciplinary training. We also advocate for programs which offer institutional support in identifying interdisciplinary and quantitative training for postdocs who wish to pursue such opportunities. Additional opportunities are outlined for faculty members who wish to undergo this training. Collaborations are often the cornerstone of successful quantitative projects and we provide concrete recommendations to promote effective and meaningful collaborations that we hope will guide institutional and cross-institutional interdisciplinary efforts. We collectively advocate for the continued use of Arabidopsis as an ideal organism for use in quantitative training efforts. For cases in which other organisms are more appropriate, we recommend leveraging best practices from the Arabidopsis community (e.g. efforts in genome annotation and data curation). Our case study in high-throughput Arabidopsis phenotyping provides an example of effective interdisciplinary and quantitative training and of the merging of quantitative and biological science integral for plant breeding in the future.

Acknowledgments

We would like to thank NSF-RCN 1518280 and NSF-RCN 1062348 for funding the workshop that led to generation of this commentary.

Literature Cited

Bansal S, Durrett TP (2016) Camelina sativa: An ideal platform for the metabolic engineering and field production of industrial lipids. Biochimie **120**: 9–16

Bialek W, Botstein D (2004) Introductory Science and Mathematics Education for 21st-Century Biologists. Science **303**: 788–790

Bradski G, Kaehler A (2008) Learning OpenCV. O'Reilly

Brutnell TP (2015) Model grasses hold key to crop improvement. Nat Plants **1**: 15062

Brutnell TP, Bennetzen JL, Vogel JP (2015) Brachypodium distachyon and Setaria viridis: Model Genetic Systems for the Grasses. Annu Rev Plant Biol **66**: 465–485

Bucksch A, Burridge J, York LM, Das A, Nord E, Weitz JS, Lynch JP (2014) Image-Based High-Throughput Field Phenotyping of Crop Roots. Plant Physiol **166**: 470–486

Campbell MT, Knecht AC, Berger B, Brien CJ, Wang D, Walia H (2015) Integrating Image-Based Phenomics and Association Analysis to Dissect the Genetic Architecture of Temporal Salinity Responses in Rice1[OPEN]. Plant Physiol **168**: 1476–1489

Chen D, Neumann K, Friedel S, Kilian B, Chen M, Altmann T, Klukas C (2014) Dissecting the Phenotypic Components of Crop Plant Growth and Drought Responses Based on High-Throughput Image Analysis. Plant Cell **26**: 4636–4655

Coelho LP (2013) Mahotas: Open source software for scriptable computer vision. J Open Res Softw. doi: 10.5334/jors.ac

Cruz JA, Savage LJ, Zegarac R, Hall CC, Satoh-Cruz M, Davis GA, Kovac WK, Chen J, Kramer DM (2016) Dynamic Environmental Photosynthetic Imaging Reveals Emergent Phenotypes. Cell Syst **2**: 365–377

Das A, Schneider H, Burridge J, Ascanio AKM, Wojciechowski T, Topp CN, Lynch JP, Weitz JS, Bucksch A (2015) Digital imaging of root traits (DIRT): a high-throughput computing and collaboration platform for field-based root phenomics. Plant Methods. doi: 10.1186/s13007-015-0093-3

Fahlgren N, Gehan MA, Baxter I (2015) Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. Curr Opin Plant Biol **24**: 93–99

Feldon DF, Jeong S, Peugh J, Roksa J, Maahs-Fladung C, Shenoy A, Oliva M (2017) Null effects of boot camps and short-format training for PhD students in life sciences. Proc Natl Acad Sci **114**: 9854–9858

Ge Y, Bai G, Stoerger V, Schnable JC (2016) Temporal dynamics of maize plant growth, water use, and leaf water content using automated high throughput RGB and hyperspectral imaging. Comput Electron Agric **127**: 625–632

Giuffrida MV, Minervini M, Tsaftaris S (2015) Learning to Count Leaves in Rosette Plants. British Machine Vision Association, p 1.1-1.13

Gutlerner JL, Van Vactor D (2013) Catalyzing Curriculum Evolution in Graduate Science Education. Cell **153**: 731–736

Holtan HE, Bandong S, Marion CM, Adam L, Tiwari S, Shen Y, Maloof JN, Maszle DR, Ohto M, Preuss S, et al (2011) BBX32, An Arabidopsis B-box Protein, Functions in Light Signaling by Suppressing HY5-Regulated Gene Expression And Interacting with STH2. Plant Physiol **156:** 2109-2123

McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM (2010) Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proc Natl Acad Sci U S A **107**: 6544–6549

Miller ND, Parks BM, Spalding EP (2007) Computer-vision analysis of seedling responses to light and gravity. Plant J **52**: 374–381

Minervini M, Abdelsamea MM, Tsaftaris SA (2014) Image-based plant phenotyping with incremental learning and active contours. Ecol Inform **23**: 35–48

Oellrich A, Walls RL, Cannon EK, Cannon SB, Cooper L, Gardiner J, Gkoutos GV, Harper L, He M, Hoehndorf R, et al (2015) An ontology approach to comparative phenomics in plants. Plant Methods **11**: 10

Preuss SB, Meister R, Xu Q, Urwin CP, Tripodi FA, Screen SE, Anil VS, Zhu S, Morrell JA, Liu G, et al (2012) Expression of the Arabidopsis thaliana BBX32 Gene in Soybean Increases Grain Yield. PLOS ONE **7**: e30717

Provart NJ, Alonso J, Assmann SM, Bergmann D, Brady SM, Brkljacic J, Browse J, Chapple C, Colot V, Cutler S, et al (2016) 50 years of Arabidopsis research: highlights and future directions. New Phytol **209**: 921–944

Rellán-Álvarez R, Lobet G, Lindner H, Pradier P-L, Sebastian J, Yee M-C, Geng Y, Trontin C, LaRue T, Schrager-Lavelle A, et al (2015) GLO-Roots: an imaging platform enabling multidimensional characterization of soil-grown root systems. eLife. doi: 10.7554/eLife.07597

Rosenwald AG, Pauley MA, Welch L, Elgin SCR, Wright R, Blum J (2016) The CourseSource Bioinformatics Learning Framework. CBE Life Sci Educ. doi: 10.1187/cbe.15-10-0217

Rubinstein A, Chor B (2014) Computational Thinking in Life Science Education. PLOS Comput Biol **10**: e1003897

Searls DB (2012) An Online Bioinformatics Curriculum. PLoS Comput Biol. doi: 10.1371/journal.pcbi.1002632

Slovak R, Göschl C, Su X, Shimotani K, Shiina T, Busch W (2014) A Scalable Open-Source Pipeline for Large-Scale Root Phenotyping of Arabidopsis. Plant Cell **26**: 2390–2403

Tan TW, Lim SJ, Khan AM, Ranganathan S (2009) A proposed minimum skill set for university graduates to meet the informatics needs and challenges of the “-omics” era. BMC Genomics **10**: S36

Walt S van der, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T (2014) scikit-image: image processing in Python. PeerJ **2**: e453

Welch L, Lewitter F, Schwartz R, Brooksbank C, Radivojac P, Gaeta B, Schneider MV (2014) Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. PLOS Comput Biol **10**: e1003496

Wingreen N, Botstein D (2006) Back to the future: education for systems-level biologists. Nat Rev Mol Cell Biol **7**: 829–832

Zhu L-H, Krens F, Smith MA, Li X, Qi W, Loo EN van, Iven T, Feussner I, Nazarenus TJ, Huai D, et al (2016) Dedicated Industrial Oilseed Crops as Metabolic Engineering Platforms for Sustainable Industrial Feedstock Production. Sci Rep **6**: 22181