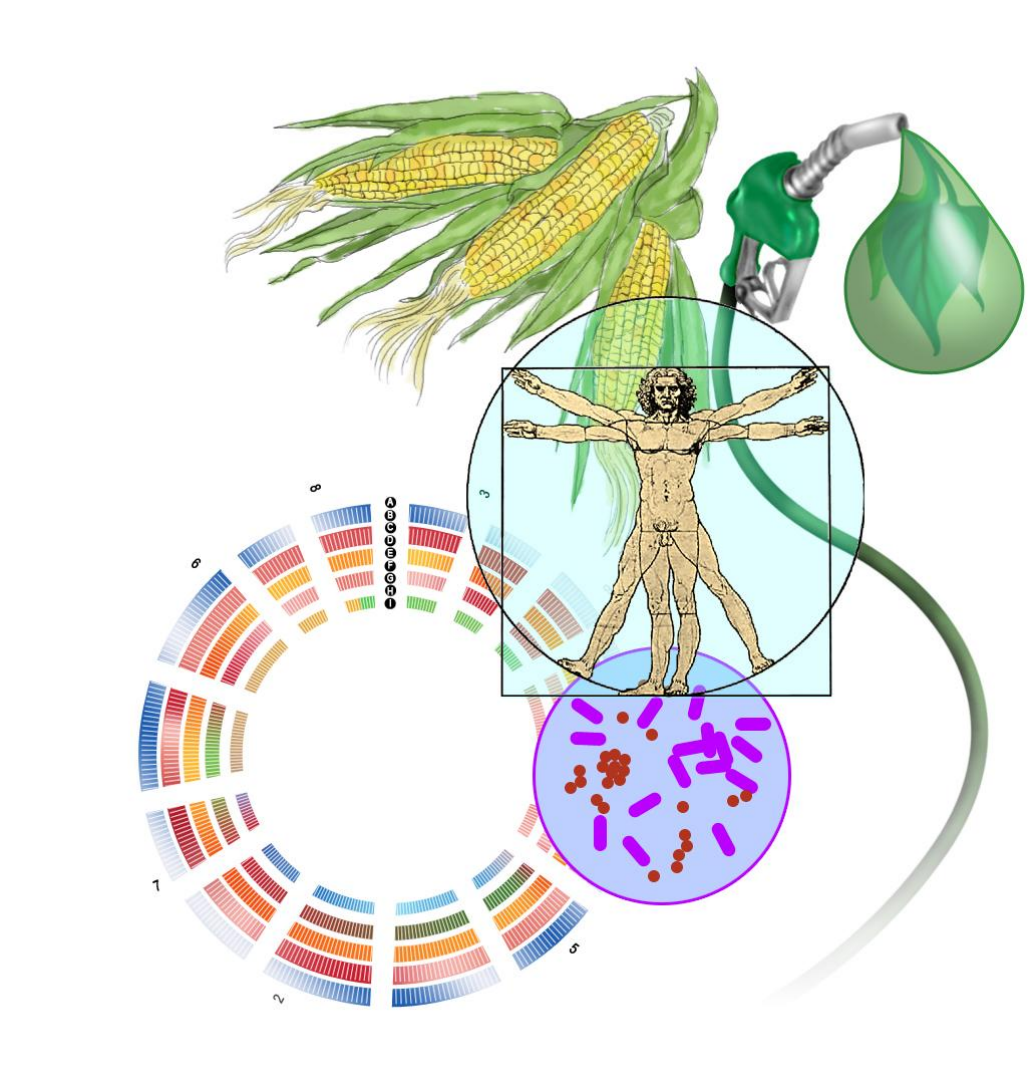
Phenomics: Genotype to Phenotype

A report of the Phenomics workshop sponsored by the USDA and NSF 2011



This report was prepared by the participants of the workshop. The workshop was sponsored by the National Science Foundation through Grant Number MCB - 1129780 to Michigan State University. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

Table of Contents

Cover art: Ramon F. Vinces

Phenomics: Genotype to Phenotype

Based on a NSF-USDA sponsored workshop held 31 March - 2 April 2011 in St. Louis, Missouri, USA[1]

Executive Summary

The question of how genetics and environment interact to influence phenotype has a long and important history. Recent advances in DNA sequencing and phenotyping technologies, in concert with analysis of large datasets have spawned '***phenomics', the use of large scale approaches to study how genetic instructions from a single gene or the whole genome translate into the full set of phenotypic traits of an organism.*** This workshop focused on analyzing phenotype, because it is frequently slower and more expensive than genomics due to the difficulties of measuring molecular, cellular, or organismal traits with sufficient throughput, resolution, and precision. Phenomics can be used across the full range of biological sciences - from studies of monocultures in well-defined and controlled laboratory environments through agricultural field conditions to populations of organisms under rapidly changing conditions. *Thus, phenomics has broad importance in applied and basic biology and is equally relevant to goals as disparate as yield improvement in food and energy crops, environmental remediation using microbes and plants and understanding complex networks that control fundamental life processes.*

Because it is inherently large scale and high throughput, phenomics can provide large amounts of phenotypic data at relatively low cost. The acquisition of high quality digital phenotypic data with explicit metadata (e.g., descriptions of protocols, growth conditions, etc. in a standardized format) provides opportunities for analysis and mathematical modeling of the molecular

[1] Participants: Timothy Close, UC Riverside, and Robert Last, Michigan State University (co-organizers), Mentewab Ayalew, Spelman College, Nitin Baliga, Institute for Systems Biology, C. Robin Buell, Michigan State University, David Clayton, University of Illinois, Katrien Devos, University of Georgia, Edison Fowlks, Hampton University, Caroline Harwood, University of Washington, Eva Huala, Carnegie Institution, H. Corby Kistler, USDA and University of Minnesota, Steven Knapp, Monsanto Corporation, Peter Langridge, Australian Centre for Functional Genomics, Jan Leach, Colorado State University, Mary Lipton, DOE Pacific Northwest National Laboratory, Joyce Loper, USDA and Oregon State University, Richard Michelmore, UC Davis, Jeff Ross-Ibarra, UC Davis, Ulrich Shurr, Forschungszentrum Jülich, Robert Sharp, University of Missouri, Mark Sorrells, Cornell University, Edgar Spalding, University of Wisconsin, Steven Strauss, Oregon State University, Matthew Vaughn, University of Texas.

networks controlling complex traits such as development, stress tolerance and metabolism or even the interactions of organisms in a community. For both microbes and plants, experimental design often can follow from well-defined agricultural, environmental or energy-related issues. A number of illustrative examples are provided in the appendix, "*Opportunities and Challenges for Microbial and Plant Phenomics".*

Recommendations

1. Balance of funding

For the US to maintain leadership in applied and fundamental microbial, plant, and non-medical animal research, balanced funding will be required for all stages of the research and development pipeline, from early knowledge discovery through solutions to societal problems (Figure 1). Some major recommendations that emerged from the Workshop are as follows:

A. Phenomics should be deployed to solve a variety of complex practical problems across simple and complex biological systems. Effective efforts will include a mix of projects ranging from single investigator through interdisciplinary multi-investigator teams. Examples include problems as diverse as studies of behavior of animals in communities and agricultural systems, complex microbial systems in the environment, plant-microbe interactions that improve yield in food or biomass crops and understanding the physiological basis for yield improvement in crop plants.

B. Strength should be maintained in foundational and application-oriented science in both reference[2] and non-reference organisms. While today's technologies can be applied to virtually any organism in response to a need for research to solve practical problems, progress in foundational 'basic' research fuels innovation that in turn benefits application-driven science. In fact many of the most ubiquitous and important technologies and techniques of biotechnology came from fundamental science, rather than targeted research. Maintaining the full research and development pipeline will allow the US research community to continue creating a strong base of fundamental knowledge that will serve as a foundation for work in economically important organisms and applications-oriented research. This approach is used by world-class research-

[2] In this document, the term 'reference organism' is used to describe a microbe, animal or plant species with a large collection of experimental tools and an active research community that make it excellent for studying diverse biological problems. In contrast, the term 'model' is used to describe a species, or group of related species, with excellent properties for studies of specific biological processes or to achieve specific goals such as crop yield improvement or bioremediation.

driven industrial organizations and should be supported by funding strategies set by NSF, USDA and other government and private sources.

C. Funding agencies should facilitate mechanisms encouraging communication and collaboration between groups working on fundamental research and those trying to solve 'real world' problems. These could range from relatively inexpensive Research Coordination Network type grant awards used at NSF (http://www.nsf.gov/pubs/2011/nsf11531/nsf11531.htm) through funding of research from early discovery to translation, or conversely by focusing a breadth of research activities on well-defined practical problems.

D. Funding priorities should be set based on science and technology goals, including both discovery-oriented and application-oriented research. Despite significant acceleration of progress on a range of economically and societally relevant issues by increasing the emphasis on organisms directly involved in these issues, recent trends have gone too far toward excluding 'basic' research on reference organisms from programs at NSF and USDA. In the longer term, this works against an end-to-end model of innovation and application. Consistent strong support for US research on reference organisms would maintain or increase the competiveness of US research and development.

2. Balance in scale of projects

Phenomics projects are by nature relatively large scale and infrastructure intensive and when done well generate large amounts of high quality data at relatively low cost. These projects usually require expertise from domain experts that do not traditionally collaborate, and are often in different institutions or units within an institution (such as Biology, Computer Science and Engineering). These and other factors can create barriers for collaboration and increase the amount of time, effort and cost required to achieve the desired goals. While large collaborative projects are often viewed favorably by funding agencies and university administrators, they can result in lack of recognition for individual participants and reduced training potential.

Phenomics research would benefit from funding of a portfolio of projects that range from creation of specific enabling technologies or proof-of-concept studies in single laboratories or small groups, through large-scale phenotyping projects and development of data and germplasm resources. Having a vision for how the smaller projects should impact phenomics

will be important prior to making requests for proposals so that the best science that fits into this larger vision can be funded. Innovations should be sought in experimental design and process, cyberinfrastructure and data analysis methods that can be applied to a broad range of organisms and growth conditions (ranging from controlled environment to field) and that increase data acquisition throughput, quality and utility.

3. Data Management

Biologists face the challenge of developing efficient and robust computational and bioinformatic methods to reduce large and diverse phenomics datasets into representations that can be interpreted in a biological context. Datasets of long-term value require data standards and metadata descriptions of the experiments in a format that enables computational approaches to data analysis.

Approaches should be sought for high-throughput data collection methods that promote high quality results and long-term utility of the data. Use of a laboratory information management system (LIMS) in large projects is essential to ensure collection of high quality data. Different project management models should be considered for large phenomics projects, depending on the desired outcomes. These could range from highly integrated projects at a single site to collaborative consortia of laboratories with world-class expertise in complementary areas of biology, phenotyping or data analysis (See Figure 2). As with all large projects, strong project management with milestones and quality metrics are essential. ***Funding is needed to meet data storage/archive needs and ensure the availability and utility of phenomic data for computational approaches, with a sustained long-term funding stream preferable to a high funding rate over a short duration.***

4. Considerations for workforce training

Large-scale phenomics projects should have integrated training plans that align the interests of the project with the needs of students and postdocs. Large-scale phenomics projects generally involve repetitive procedures, which are typically best done by IT, laboratory and field technical support staff, and in some cases by undergraduate students. Science training opportunities exist

for graduate students and postdoctorals in planning the project to create data useful for asking important questions, creating enabling technologies and mining the phenotypic data.

Initiatives in phenomics should include educational activities to enable biology trainees to think quantitatively and collaborate with experts in physical, computational and engineering sciences. Training also should be provided in management of, and participation in, large interdisciplinary and collaborative domestic and international projects. Training in computer science and experimental design, including statistics, is essential for biologists involved in phenomics research. A basic understanding of the logic and methods of programming, knowledge of command-line tools (e.g. UNIX shell), and a familiarity with the development of computational pipelines and workflows will be essential for scientists to acquire, analyze, and critically interpret genomic and phenomic data. ***Funding is needed for undergraduates, predoctorals and postdoctorals to be trained in computational thinking that will advance our ability to obtain, analyze and utilize large scale phenomics data.*** The recently announced National Plant Genome Initiative Postdoctoral Research Fellowship program (http://www.nsf.gov/pubs/2011/nsf11499/nsf11499.pdf) and USDA's National Institute of Food and Agriculture Fellowships (http://www.csrees.usda.gov/fo/afrinifafellowshipsgrantprogram.cfm) are examples of such programs, and there is need for others.

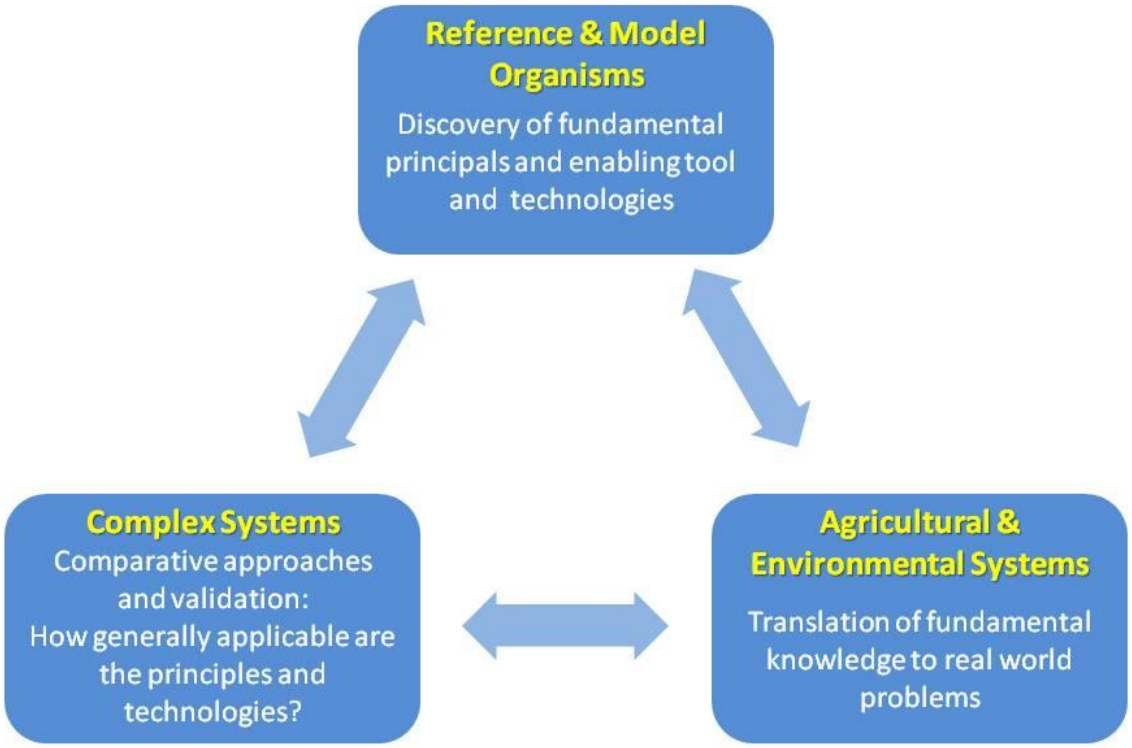


Figure 1. The context of phenomics research on reference and model organisms. All scales of basic and applied biology benefit from phenomic research on reference and model organisms.

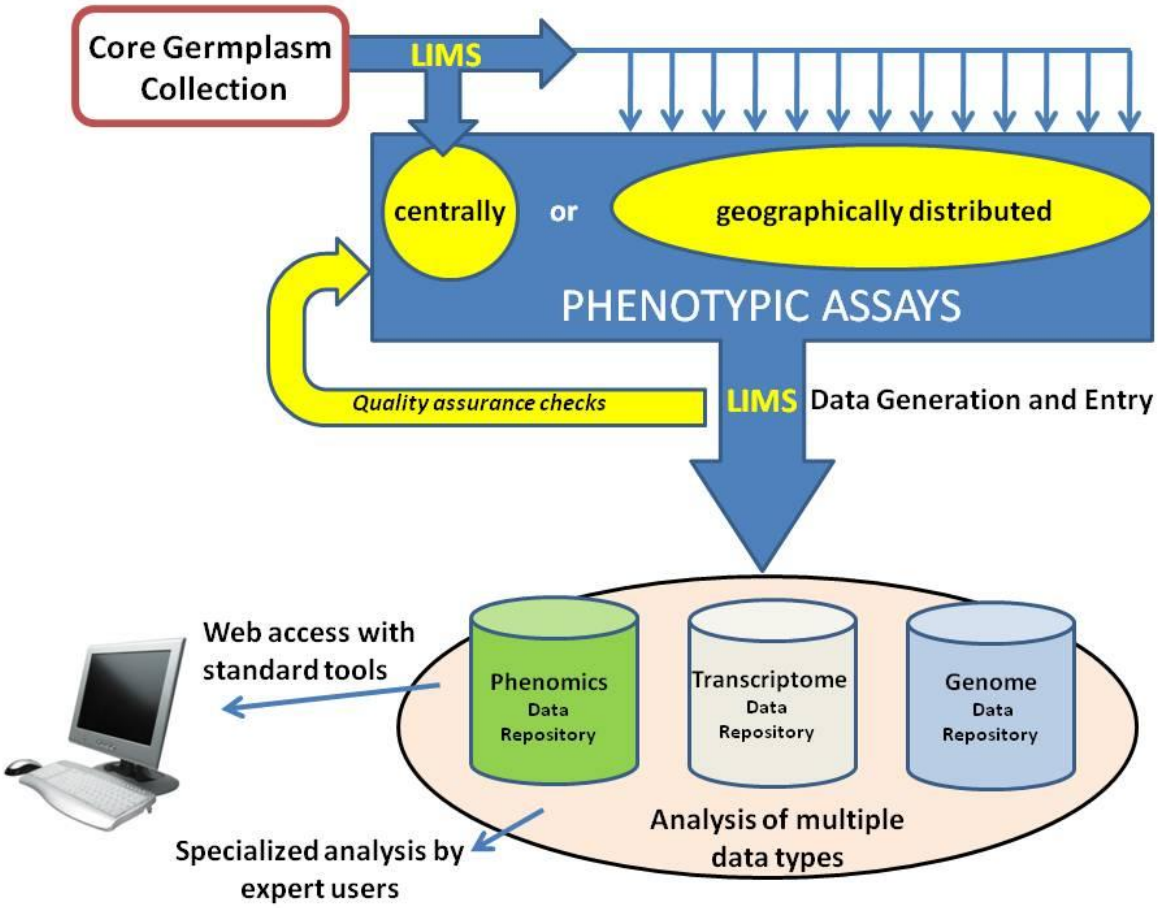


Figure 2. Schematic representation of phenomics approaches using a common germplasm. Phenomics is typically performed by running multiple phenotypic assays on a large set of genetic variants (preferably from a well annotated germplasm collection). Because enormous amounts of data are generated from large numbers of samples, laboratory information management systems (LIMS) for labeling and tracking samples should be in place before the project begins. For the highest quality results, growth of the organism and phenotypic assays can be performed in a single location or in a distributed manner, as long as standard operating procedures are in place (Massonnet et al., 2010) for analysis and metadata capture and the LIMS is used for all samples. As data are collected, quality assurance procedures should be in place to monitor quality and reproducibility of the data. For maximal impact, phenomics data can be made available to the community through analysis tools provided via web interface tools and to experts for specialized analysis. The phenomics data can also be analyzed with other data types.

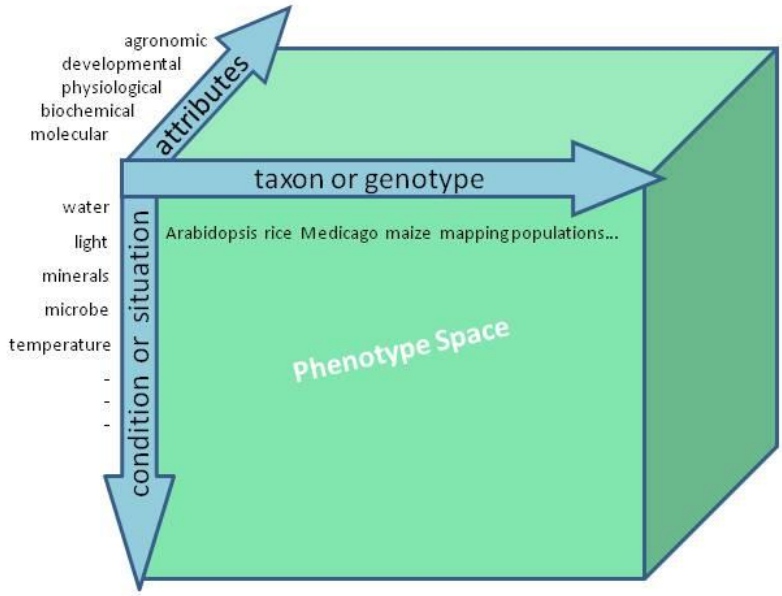


Figure 3. Phenotypes are complex. This simplified schematic diagram attempts to illustrate the multidimensionality of phenomics experiments and data. Genetic diversity can come in many forms such as induced mutants, natural variants, results of genetic crosses and populations of organisms. The environmental conditions used will impact the traits being measure and the variety of phenotypes ('attributes') measurable is nearly limitless. Not shown in this diagram are the dimensions of time (for example developmental progression of the organism) and differences in phenotypes in various cell, tissue or organ-types.

1. Introduction

A. Phenomics: From Genotype to Phenotype

One of the central principals of biology is the concept that a set of genetic instructions, or genotype, interacts with the environment to produce the characteristics, or phenotype, of an organism. Understanding how particular genotypes result in specific phenotypic properties is a core goal of modern biology, and enables development of organisms with commercially useful characteristics. However, prediction of phenotype from genotype is generally a difficult problem due to the large number of genes and gene products that contribute to most phenotypes in concert with complex and changeable environmental influences.

The last 20 years have created a revolution in our understanding of genotype: while genomes typically are quite large, with millions or billions of nucleotides, the relative chemical simplicity of DNA lends itself to large-scale analysis. We can now determine genotypes down to the level of individual nucleotides in whole genomes, and entire genomes are now rapidly sequenced at steadily declining costs and ever increasing speed. Next generation resequencing methods provide opportunities to get the complete genotype and epigenotype not only of a single representative of a genus or species, but of many representatives of a phylogenetic group or population. High-density single nucleotide polymorphism genotyping, first pioneered in the human HapMap project, has become tractable for any organism and now is routinely applied to plants and microbes for the characterization of natural genetic variation and to support trait-driven efforts to clone and understand specific genes. Thus genome science is moving beyond the era of reference and model organisms to study in depth any microbe, animal or plant that has characteristics of interest to science and society.

The study of phenotypes is quite different. Unlike a genotype, the phenotype of an organism can be described at many levels, from specific molecules to dynamic metabolic networks to complex cellular developmental and physiological systems, all the way to the aggregate or social behaviors of complex populations. Interactions with symbionts, pathogens or competing organisms create additional levels of phenotypic complexity. Moreover, phenotypes are dynamic

and the timescales in which they change vary tremendously. Consider for instance the rapid responses of a bacterium to nutrient changes (Segall et al., 1986) or the dynamic changes in photosynthesis of a leaf as a single cloud passes over the sun (Murchie and Niyogi, 2011), compared to the slow morphological changes in long lived plants or even the lifelong changes in the outward appearance of a human being. Phenotypes rarely have a single discrete description, and most phenotypic characters are better described as continuous functions as opposed to the discrete 'A,C,G,T' character codes of the genotype. Indeed a complete catalog of phenotypes (the phenome) can have essentially infinite complexity (See Figure 3).

Now that digital DNA data are available in abundance, we face an acute need to quantify individual phenotypes in a way that can be explicitly matched to individual genotypes. If this challenge can be mastered, we face the promise of gaining a deeper insight into the components of complex traits such as yield or stress resistance in economically important plants and animals or population dynamics for microbes that play key roles in global nutrient cycling. This can extend to a systems level description of the underlying processes, ultimately enabling predictions of emergent phenotypes such as fitness and survival in studies of ecology and evolution or yield and stress tolerance or other traits of economic value.

B. How is phenome different from phenotype?

Phenomics, the study of the phenome, is a rapidly emerging area of science, which seeks to characterize phenotypes in a rigorous and formal way, and link these traits to the associated genes and gene variants (alleles). Examples of phenotypic parameters include gross morphological measures such as cell size, tree height or wheat yield, dynamic measures such as rate of cell division of a unicellular organism, metabolism or nutrient uptake, and molecular measures such as mass spectrometry fingerprints and transcript profiles.

Formally, phenomics is the science of large-scale phenotypic data collection and analysis, whereas the phenome is the actual catalog of measurements. While it shares characteristics with classical mutant screening or quantitative trait analysis, it is distinguished from these traditional approaches in scale and scope (Winzeler et al., 1999; Lango Allen et al., 2010; Speliotes et al., 2010; Heffner et al., 2011; Lu et al., 2011; Nichols et al., 2011). ***First***, phenomic studies

typically employ large populations of genetic variants with the goal of sampling variation in many or all genes. ***Second***, each genotype is assayed for a large number of traits, typically using well-tested and high-throughput standard operating procedures with systems in place to maximize accuracy in sample tracking and data reproducibility. ***Third***, key features of the growth conditions are well defined and closely monitored. ***Finally***, the phenotypic data and metadata descriptions of the experimental conditions are captured in formats that allow detailed data analysis. These analyses would ideally identify relationships between genotype and phenotype as well as reveal correlations between seemingly unrelated phenotypes (Schauer et al., 2006; Lu et al., 2008) or genetic loci (Gerke et al., 2009).

Because most phenotypes are determined by the interactions of genes and environment, the ideal situation is to collect large numbers of measures across multiple environments, at different developmental stages, and for multiple cell/tissue/organ types. As it is unrealistic to sample all alleles for each measurable trait under every possible growth condition, the experimental design and methods of data analysis must be matched to the desired outcome, as described in (Shasha et al., 2001). For example, controlled growth conditions are often employed, whether 'optimal' conditions or to test interactions of the organism with specific controlled abiotic factors (e.g. temperature, photoperiod, nutrient availability) or in response to known mutualistic, parasitic or competitor organisms. In contrast, identification of crop plant genotypes that produce desirable traits such as yield or nutritional quality is best done under field conditions of soil, climate and biotic stress agents similar to where the final varieties will be grown commercially. Similarly, microorganisms can be phenotypically assayed under a variety of conditions, depending upon the relevant question being asked. These can range from a simple measure such as growth rate in a monoculture in the laboratory to analysis of their properties in the context of complex communities of microbes or in association with animals or plants.

2. What is needed to advance phenomics?

The phenotype may be considered a multi-scale description of an organism’s attributes displayed in space and through time. Sets of multidimensional, high resolution data quantitatively describing the phenotype in many meaningful conditions would enable mapping genetic

elements to biological function at the desired level of detail. This section considers the resources necessary for effective research on phenotypes from the metabolome to populations of organisms.

A. The roles of 'reference' and 'model' systems

Prior to the age of molecular genetics, biologists typically chose a model organism for study because of a match between the question being asked and the properties of the organism. This paradigm shifted during the last half of the 20th century when scientists gravitated towards reference organisms with facile genetics. Hallmarks of these reference organisms now include the availability of broadly useful genetics resources such as diverse germplasm, detailed knowledge of developmental, biochemical and physiological networks, advanced genomics tools and large, collaborative communities. Examples include microbes such as *Escherichia coli* and *Saccharomyces cerevisiae* (baker's yeast), the plants *Arabidopsis thaliana*, *Zea mays* (corn) and *Oryza sativa* (rice), and animals such as *Caenorhabditis elegans (*a soil roundworm), *Drosophila melanogaster* (fruit fly) and *Mus musculus* (mouse). These reference organisms have traditionally been a strong focus of research and funding because the available assets make them such efficient systems for primary investigations into a wide variety of biological phenomena. The large amount of information available about these systems - for examples, high quality gene annotation, deep understanding of developmental, physiological and metabolic networks - provides a valuable point of reference for researchers studying even distantly related organisms.

Although reference organisms have lasting value as experimental systems, they represent only part of the remarkable phenotypic diversity of the biological world. For example, while our sophisticated understanding of *Arabidopsis* root development serves as a representative or working hypothesis for many other plants (Benfey et al., 2010), this organism cannot be used for studies of symbiosis of plants and mycorrhizal fungi known to be important in nutrient assimilation in numerous plants (Smith and Smith, 2011). Similarly, *E. coli* does not have the genetic wherewithal to degrade pollutants such as benzene, whereas other bacteria have this capability (Irie et al., 1987). Even traits that are superficially shared (e.g. flowering time and rate of biomass accumulation) may have dramatically different structural and regulatory gene networks (Buckler et al., 2009; Salome et al., 2011). Finally, certain agronomically important

traits such as seed yield are likely to be controlled by such a complex array of factors that they are best studied in the target organism in their usual production environments.

The use of non-reference organisms for the study of specific biological processes is regaining popularity in part as a result of the revolution in sequencing and phenotyping technologies. Once an organism, organ or cell type is identified as being useful for study of a specific problem, transcriptome and/or genome sequence provides a starting place for characterizing the gene space and relative abundances of mRNAs. These sequences allow discovery of candidate genes (Cocuron et al., 2007), enable large scale proteomics analysis (Schilmiller et al., 2010) and a first pass construction of metabolic or regulatory networks (Keseler et al., 2009). Re-sequencing of genetic variants provides enormous numbers of molecular markers that can be used to find genes contributing to target phenotypes by traditional genetic mapping or more modern approaches such as genome-wide association studies (Rounsley and Last, 2010). Phenomics approaches such as digital descriptions of growth and development, physiological parameters, and protein and metabolite abundance may also increase the accessibility of non-reference model organisms.

B. High throughput data types and workflows

Some components currently considered key elements of a phenotype description such as gene expression profiles can be acquired by relatively accessible, high throughput technologies. In fact, large-scale transcript profiling studies in reference organisms have already proven of great value (EcoCyc for *E. coli*: http://ecocyc.org/; AtGenExpress for Arabidopsis: http://www.weigelworld.org/resources/microarray/AtGenExpress/ (Kilian et al., 2007). In contrast, technologies for characterizing downstream phenomics attributes are becoming increasingly powerful but are currently far from routine.

i. Proteome and Metabolome

Techniques for identification and quantification of large numbers of proteins and metabolites (Last et al., 2007) are increasingly sophisticated. However, many challenges exist for making these techniques accessible to the broad scientific community. These include expense of equipment, large range of concentrations and diverse chemical properties of these molecules. For example, analysis of stable, abundant and soluble metabolites and proteins is far easier than rare,

labile and insoluble molecules. Although most protein and metabolite analysis methods require tissue extraction, well established methods exist for non-invasive measurements such as near infrared transmittance for seed metabolites (Velasco et al., 1999) and more experimental methods such as MALDI-ToF (matrix assisted laser desorption/ionization time of flight) mass spectrometry for spatial resolution of metabolites (Shroff et al., 2008).

ii. Physiological attributes

Physiological measurements of processes such as photosynthesis, nutrient uptake and transport can be reproducible and sophisticated (for example, see (Baxter et al., 2008), but achieving the necessary throughput is challenging. Spatial variation of physiological parameters can increasingly be approached through imaging technologies. This opens a new window of analysis, since in many cases the heterogeneity of the response in space and time is a key feature of the phenotype, which contains significant information about the underlying biological principles (Jansen et al., 2009; Walter et al., 2009).

iii. Plant growth and development

Growth and development of multicellular organisms can be measured by quantitative parameters like biomass or by spatially resolving technologies based on cameras and image analysis if experimental systems and computer algorithms capable of measuring the feature or process of interest can be developed. High throughput phenotyping platforms based on image analysis are available for laboratory and greenhouse settings (http://www.lemnatec.com/; http://www.plantaccelerator.org.au/) but their use is far from widespread. Quantification of below-ground structure and behavior is a major hurdle, though culture in or on gelled media make image-based techniques practicable (Fang et al., 2009; Brooks et al., 2010; Clark et al., 2011). Tomographic systems can provide insight in the dynamics of structure and function of root systems (Jahnke et al., 2009) but currently are not able to handle high throughput due to technical limitations and the large amount of data generated.

iv. Phenomics *in situ*: measurements in the field

Ecologists, breeders and systematists have been practicing phenotyping in the field for decades. High throughput phenotyping for certain target and correlated traits is routine,

particularly in plant breeding where thousands of unique genotypes are evaluated seasonally. Remote sensing is increasingly powerful; for example canopy spectral reflectance is employed in plant breeding programs for measuring nitrogen- or water-use efficiency (Gutierrez et al., 2010). However, there are still many important traits that are difficult or costly to evaluate and phenomics technologies could bring new approaches that would enhance the identification of superior genotypes and effectively train prediction models.

C. Maximizing the value of phenomics resources

Because of their large scale, phenomics projects are resource intensive and generate large amounts of data. These large projects can be highly cost effective if the resulting data are of high quality and lasting utility to a large number of investigators. Several major factors contribute to the long term success of phenomics projects:

- the source of genetic diversity employed and whether it is preserved for future use;
- the quality of the growth conditions;
- the phenotypic assays performed;
- collection, storage and interpretation of data.

Genetic diversity

Germplasm collections are the starting point of many phenotypic investigations and are increasingly important as new phenotyping technologies emerge. Phenotyping tools are most useful for the task of understanding genetic function when they can be applied to the study of well-curated germplasm or genetic stocks appropriate to the problem being addressed. These collections can include lines generated by transgenesis or transposon mutagenesis, chemical or radiation induced mutagenesis, accessions of variants derived from natural populations and lines produced by crossing including breeding material, introgression lines and recombinant inbred lines (Eshed and Zamir, 1995; Alonso et al., 2003; Yu et al., 2008; Buckler et al., 2009).

Production and maintenance of large sets of germplasm is time and labor intensive and careful thought should be given to the balance of cost and utility of a population. For example, sets of germplasm that can be used for large numbers of studies or to measure broad sets of phenotypes may be of higher value than those custom designed for specific projects or narrow phenotypic

Germplasm

In addition to the costs of production and testing, existing germplasm collections face a number of difficulties, including lack of resources for storage and distribution and lack of quality control. For example, plant experimental germplasm collected by investigators is not typically accepted into the National Plant Germplasm System (NPGS; http://www.ars-grin.gov/), especially when these collections are large. Also, some germplasm collections do not fit in any of the currently available stock centers, for example mapping populations and transgenics. Models for sustaining the collections could include user fees or institutional subscriptions. Crop curators and the researchers need to define mutual responsibilities for quality assurance, replenishing depleting stock, and the projected duration for the NPGS's commitment to curate these materials. Communities of stakeholders with a common interest in securing germplasm collections and making them maximally effective should be brought together to address these problems.

studies. Having a system in place for capturing diverse phenotypic data for a germplasm collection in a central data repository is especially powerful (see 'geographically distributed phenotypic assays' in Figure 2; (Baxter et al., 2007; McMullen et al., 2009; Lu et al., 2011). This is in part because it allows members of the community to query multiple phenotypes and relate these traits to genotype. Finally, for a collection of germplasm to be of widest utility it should come without intellectual property restrictions or with a material transfer agreement that is simple and not onerous.

D. Making data and software Available

Storage and retrieval costs of genomic and phenomic data are rising due to the increasing quantity of data being collected. Thus, community standards are needed to address which datasets should be retained. Several criteria will be needed to determine long-term value of phenomic data including: 1) *Data quality* - do the data have sufficient value and broad importance to warrant archiving in a public resource? Are the experimental and data analysis methods and metadata sufficiently well described to make the data useful to other researchers? 2) *Data regeneration costs* - is it fiscally prudent to archive, rather than regenerate, the data when needed? 3) *Data collection technology* - is the data collection method up to date? Are higher quality datasets available using more recent technologies that supersede the older data? 4)

Availability of genetic stocks – Data are often more valuable if the corresponding genetic stocks are available for re-examination or for follow-up studies.

If a dataset warrants long-term storage, additional mechanisms will be required to determine:

- *how the data should be stored* to maximize utility both for expert and early career users;
- *where results should be stored* to ensure sufficiently long-term access at low cost to funding agencies, universities and scientific publishers;
- *the level of data and metadata curation required* and mechanisms for ensuring compliance, including oversight by funding agencies and publishers;
- *criteria to determine when a dataset no longer needs to be retained*.

The community must take an active role in maintaining phenomics data collections deemed worthy of long-term storage. Several models were considered by workshop participants. In any of these models, the iPlant Collaborative (http://www.iplantcollaborative.org/) may provide a gateway for community-driven collection, integration and curation of phenomics data.

(1) Data-type specific repositories. A comprehensive cross-species resource would be created for each major class of phenomics data, i.e. transcriptome, proteome, metabolome, physiological/morphological measurements. This model is highly suitable for large scale quantitative data and has already proven successful for transcriptomics data (ArrayExpress / GEO / Plexdb). There are a variety of advantages to this model including that data deposition could be more easily enforced by journals and funding sources, and compliance with data and metadata standards set by the community could be monitored by professional curators, facilitating computational reuse of data. This approach also provides an economy of scale by avoiding duplication of effort across multiple resources handling the same data types.

(2) A Wikipedia-like model. In this scenario research projects deposit their data to a central repository and data can be updated or curated by others. A funded curator would be required to initially set up the wiki database, but it is anticipated that the system would be self-regulating within a few years. Challenges of this model include how to motivate scientists to participate, and how to maintain consistent data standards and formats across many individual contributions

iPlant Collaborative

The iPlant Cyberinfrastructure Collaborative (http://www.iplantcollaborative.org/) offers both a compelling platform for sharing of source code for biological applications and an environment for building reproducible scientific workflows from published software. The phenomics community could be well served to encourage adoption of that platform by developers of bioinformatics applications and domain experts who can build best-practice workflows comprised of these software components. Encouraging participation of phenomics researchers and tool developers in iPlant-sponsored 'hack-a-thons' for collaborative code development and 'bring your own data' training workshops could ensure that this infrastructure contains tools appropriate to topics in phenomics. In addition, the iPlant platform encourages integration of visualization resources into its 'software ecosystem' as well as development of new information visualization applications using the Stanford Protovis toolkit, R, and the Javascript InfoViz Toolkit.

to enable extraction and computational analysis of datasets spanning multiple experiments. This model may be appropriate for specialized datasets with lower potential for widespread reuse and small, highly cohesive communities focused around specific research questions.

(3) Individual project or community databases. Generic Model Organism Database tools could be used by individual communities or projects. Interoperability between different databases that capture the results, metadata, and provenance is essential and could be facilitated by the establishment of common controlled vocabularies for phenotypic measurements. This model has a variety of challenges including how to motivate researchers to contribute data, mechanisms to ensure that interoperability is maintained and approaches to fund long-term data curation and storage for many such resources.

An increased reliance on computational approaches has resulted in development of a large number of software packages for a range of biological problems. However, documentation and widespread adaptability of the software is a major obstacle to re-use of code outside the developer's group. As with the reuse of data, software reuse is not always the optimal solution, but depends on the quality of available software, the effort required to adapt existing software vs. the effort of writing new

software, and the potential for continued refinement of pre-existing vs. newly-developed software. Nevertheless, availability of software and analysis pipelines for use and reuse will have a positive impact on phenomics research, allowing groups without in-house software expertise to carry out analyses, facilitating direct comparison of different datasets and allowing verification of published experimental results. Software tools and analysis pipelines could be made accessible for reuse by other groups within a community resource such as iPlant and documented as persistent publishable objects, referred to via a DOI or other identifier. This will serve dual purposes: First, experimental reproducibility can be enhanced because bioinformatics methods can be described not just in narrative terms, but in the context of a replayable series of events in an analytical infrastructure. Second, authors of tools and pipelines could receive publication credit as their DOIs are referenced in the literature. This should create incentives for such development activity.

E. Project scale: the effects of size

Until recently, most biological research involved single laboratories or small numbers of investigators who collaborated because of shared interests. The post-genome era dramatically changed this model. Now, research projects can vary in size from a small individual investigator group directed toward a defined focused project to large consortia of investigators working together toward a broad set of goals. Large group projects allow collaboration of domain experts in plant biology, microbiology, sequencing, high throughput omics, informatics and mathematical modeling**.** Bringing in expertise from disciplines such as statistics, engineering and computational sciences allows design of more efficient processes for obtaining high quality data and novel approaches to analyzing the large datasets.

Each scale of project in the continuum has advantages and disadvantages to participants and to the broader science community. For example, traditional single investigator science allows a tremendous amount of freedom to the investigator, strong training in problem solving and hypothesis testing to early career participants, and can generate deep understanding of specific areas of science. However, the small size may limit both the breadth of approaches and opportunities for multi-disciplinary training and creation of biological resources and data of longer-term value. In contrast, consortia can tackle larger questions through applying diverse

expertise and create multi-disciplinary training environments. However, successful management of these projects requires skills different from 'smaller' science: competent and trusted leadership, formal project management, and resources devoted to data management. Communication between laboratories can divert time and effort from data gathering, analysis and dissemination. Large projects with set goals and repetitive operations can stifle creativity and training. In addition, the academic system tends to emphasize individual achievement, especially for early and mid-career scientists, where first author publications and grant funding record are of paramount importance.

A proper balance of funding between the two has clear advantages. The large consortia can facilitate the larger scale experiments that would be beyond the means of the individual investigator. Funding of smaller projects enables the data to be mined extensively for hypothesis generation and testing.

F. Success depends on a trained workforce

21st Century Biology, including phenomics, critically depends upon a workforce trained differently from the traditional US model focused on graduate student and postdoctoral training in deep, and sometimes narrow, hypothesis-driven research. In addition to strong expertise in biology, the characterization of phenotypes is increasingly dependent on tools and activities at the interface of biological, computational and physical sciences. Any initiative in phenomics should include educational activities to enable biology students to think quantitatively and collaborate with scientists in domain areas such as chemistry, computer science and engineering.

Most current plant biology curricula are sorely lacking in explicit training in computational methods. A basic understanding of the logic and methods of programming, knowledge of command-line tools (e.g. Unix shell), and a familiarity with the development of computational pipelines and workflows will be essential for scientists to acquire, analyze, and critically interpret genomic and phenomic data. Such training should include data management and curation, fundamentals of information visualization, an understanding of basic data types, and best practices in terms of methodology documentation. Making such training a requisite or strongly recommended part of coursework (similar to funding agency requirements for ethics

training) could help to ensure broad adoption of these training standards. Increasing the numbers of students trained in quantitative genetics and modern plant breeding methods is another major workforce issue that will impact phenomics and should be vigorously addressed.

In addition to expanding capabilities in the public sector, creating a pool of students trained in large-scale phenotyping, plant breeding, large interdisciplinary projects and management of large datasets will greatly benefit US industrial research competitiveness. Training also should be provided in management of, and participation in, large interdisciplinary and collaborative projects.

3. Computation and modeling

The next decade will lead to major discoveries in biology, driven in large part by acquisition of massive amounts of genomic and phenomic data. While this creates unparalleled opportunities, computational methods in data analysis are lagging. Furthermore, the volume of data that can be generated exceeds both fiscal and logistical resources available for data storage and practical data mining. Because phenomics information encompasses a wide variety of disparate data types there is unlikely to be a 'one size fits all' solution to these computational challenges. Instead, a set of tools will be needed, with each tailored to one or more phenomics subdomains. An emphasis on interoperability among these tools will allow researchers to perform cross-cutting research that combines disparate types of phenomics data.

A. Data Capture

Acquisition and dissemination of phenomics data will require the availability and cross-compatibility of simple and cost-effective LIMS (laboratory information management systems) software (Baxter et al., 2007; Lu et al., 2011). Open-source LIMS for phenomics data would enable more rapid adoption of data collection standards and broader use of comprehensive data and metadata tracking. This would entail development of standardized schemas, barcoding software, and field-based data entry tools with mobile devices.

Detailed and well-structured information about environmental conditions and experimental treatments is an essential part of phenotype data collection. Environmental effects play a large

Ontologies

Ontologies provide a set of controlled vocabulary terms (e.g. 'seed', 'endosperm') and relationships between those terms (e.g. 'endosperm' is a part of a 'seed') that make biological statements accessible to computational approaches. The use of standardized, well-defined terms ensures that data generators and end users will use terms consistently, and defined relationships between terms ensure that computational reasoning can be applied to sets of annotations made using ontology terms (for example finding phenotypes affecting the endosperm and other parts of the seed in response to a query for all seed-related phenotypes). A partial list of ontologies relevant to annotation of plant phenotypes is included below as examples. Terms from different ontologies can be combined to generate a phenotype annotation including information about the plant part, developmental stage, attribute affected, and conditions under which the phenotype was observed, including growth conditions, chemical treatment, etc. Both qualitative and quantitative phenotypes can be captured in this way.

role in shaping phenotype and metadata standards should be developed that are analogous to MIAME: Minimal Information About a Microarray Experiment (Brazma et al., 2001). The metadata should describe the experimental design including environmental conditions and experimental treatments and data analysis methods, thereby providing sufficient information for the experiment to be replicated. Metadata collection should be simplified and automated whenever possible to minimize data entry errors and encourage collection of broadly useful quantitative data. Where user interaction is required, graphical user interfaces ('GUIs') for collecting metadata need to be designed to enable accurate data and metadata collection.

B. Data Integration

Integration of both similar and disparate data types is essential to developing a complete picture of the complex phenomics domain. While integration of similar datasets (e.g. two proteomics datasets generated with the same methods) presents certain challenges, integration of disparate datasets is far more daunting, requiring the same objects (e.g. genes, proteins, metabolites) or similar objects with a known relationship (e.g. orthologous proteins) to be found in two or more of the disparate datasets and

necessitating the community-wide adoption of unique identifiers for such objects. The use of ontologies to describe phenomics data and metadata will be essential to make datasets searchable and reusable and facilitate data integration and computational analyses. Several efforts are already underway in this area (see Ontologies sidebar); thus the challenge will be to encourage or enforce community adoption and input into these ongoing efforts.

Plant phenotype ontologies.

Acronym	Full Name	Scope	Link
GO	Gene Ontology	Biological process, molecular function , subcellular localization	www.geneontology.org
PO	Plant Ontology	Plant anatomical parts and developmental stages	www.plantontology.org
TO	Trait Ontology	Cereal plant traits	www.gramene.org/db/ontology/search?id=TO:0000387
CO	Crop Ontology	Crop plants (anatomy, developmental stage, trait)	www.generationcp.org/ontology
PATO	Phenotypic Attribute and Trait Ontology	Phenotypic qualities	obofoundry.org/wiki/index.php/PATO: Main_Page
CHeBI	Chemical Entities of Biological Interest	Chemical compounds	www.ebi.ac.uk/chebi/
EnvO	Environmental Ontology	Habitat	environmentontology.org
EO	Environment	Plant growth conditions incl. temperature, growth media, light regime, etc.	www.ebi.ac.uk/ontology-lookup/browse.do?ontName=EO
UO	Unit Ontology	Units (describing length, volume, density, irradiance, temperature, etc)	http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=UO

C. Data Analysis and Visualization

i. Visualization

There is active development of visualization software within the greater scientific community for "omic" phenotypic data. For example, Cytoscape modules are widely used for visualization networks that incorporate co-expression, protein-protein interaction, and biochemical pathway data. The Generic Model Organism Database (GMOD) group (http://gmod.org/) also develops tools focused on genome and pathway data analysis and visualization. Gaggle (http://gaggle.systemsbiology.net/docs/) provides a way to share datasets across different analysis and visualization tools. However, the large number of underlying components that must be visualized in multicellular organisms or communities and the large scale of some phenomics datasets presents a computational challenge. Development of scalable, interactive methods is needed to view and interpret genome-scale phenotype data. Data presented via such tools should be seamlessly linked to other datasets and resources to allow more efficient data exploration and mining.

ii. Image processing

Image-based phenotyping offers a way to capture and extract not just morphological and gross developmental phenotype data, but also to interrogate physiological status through non-destructive close-range or remote-sensing technologies. A major roadblock is the lack of extensible algorithms for performing quantitation, feature extraction, and summarization. Assembly of the necessary data storage, transmission, and computational pipelines required is also difficult due to logistical impediments and computational requirements for advanced image analysis. These issues could be addressed by encouraging collaboration between image processing experts in the computer science and engineering domains and plant biologists through image classification contests, use-case marketplaces, and other networking opportunities. Adoption of extensible, high-throughput image analysis platforms such as the BISQUE system from The Center for Bioimage Informatics,UC Santa Barbara (http://www.bioimage.ucsb.edu/) may help address the logistical and scalability issues.

D. Predictive Modeling

Accurate prediction of an organism's phenotype from its genotype and environment is a stringent test of our understanding of a biological system. The ability to make such predictions has both fundamental scientific applications and practical benefits, providing a way to generate and test hypotheses about biological mechanisms as well as facilitating plant breeding and microbe engineering.

i. Genomic selection

Advancements in high-throughput genotyping are rapidly decreasing the cost of whole-genome genotyping while phenotyping costs are stable or increasing. This is driving the use of marker-assisted selection for major genes in plant and animal breeding. Commonly employed marker-assisted selection strategies, however, are not well suited for complex traits controlled by many loci of small effect (Meuwissen et al., 2001; McMullen et al., 2009). Genomic selection is an emerging technology complementary to marker-assisted selection, which uses phenotypes and thousands of genetic markers covering the entire genome to develop complex prediction models that are used to calculate genomic estimated breeding values for complex traits. These models can then be used to predict phenotypes based only on genotype in related populations. Because selections for multiple traits are based solely on whole-genome genotypes, multiple cycles of selection can be made without phenotyping, resulting in increased annual genetic gain (Heffner et al., 2011). Due to the complexities of such modeling, however, more research is needed to assess model accuracies for populations differing in linkage disequilibrium, distribution of QTL, size, marker density, and especially subpopulation structure. Further, these methods emphasize the importance of accurate, high throughput phenotyping for complex traits that drive the gains in efficiency. Because plants and animals differ in several aspects affecting GS strategies, the statistical approaches may be similar but the outcomes and applications will differ substantially.

ii. Explicit mechanistic modeling

While a predictive model lacking explicit biological mechanisms connecting genotype and environment with phenotype can serve a practical purpose in designing new crops, microbes and

better performing animal food systems, a model based on known biological mechanisms provides a powerful tool to test our understanding of those mechanisms. Knowledge of mechanistic information facilitates construction of a model that is not only predictive but also aids in designing specific experiments to test our understanding and alter biological behavior. For instance, while modeling transcriptional regulatory networks, knowledge of the mechanism by which a transcriptional factor exerts control can be incorporated as a strategy for discovery of transcriptionally co-regulated genes that have similar expression patterns *and* share conserved *cis*-regulatory sequence patterns in their promoters.

The complexity of biological networks necessitates a multi-scale approach that incorporates abstractions depending on the scale at which the modeling is conducted. A systems model that incorporates thousands of genes of an organism has to sacrifice detail on variations in responses at a single cell level. On the other hand, a model for dynamic interactions among few genes could incorporate such detail. Efforts to model biological phenomena across a range of scales should be supported to maximize the payoff from this approach.

4. Summary

Technologies developed in the 21st century will enable major discoveries in biology. With the revolution in genomic technology, we begin to move past the constraints of studying reference organisms selected for their genetic tractability. Deep genotypic information can be collected on essentially any organism, allowing biologists to leverage the unique phenotypic characteristics of diverse organisms to create both fundamental knowledge and useful solutions to social challenges that improve the human condition.

The emerging science of phenomics will be central to realizing this vision once several challenges are overcome. Improved methods are needed for high-throughput collection of diverse phenotypic measures, in both natural and laboratory environments. Phenomic datasets can be large and complex, likely dwarfing the size of genomic datasets. Tools are needed not only for generating these datasets, but also for storing, analyzing and interrogating them, efficiently and affordably.

The power of phenomics will be multiplied when datasets can be combined and correlated across different studies, allowing increases in statistical power and the scope of analyses. For this to bear fruit, it will be critical for the field to develop formalized methods for data quality control, and for describing phenotypic measures and the circumstances under which they were collected. Many of these challenges can be met through creative application of computational and engineering technologies, further advancing the mission of research at the intersection of the physical and life sciences.

References Cited

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., Gadrinab, C., Heller, C., Jeske, A., Koesema, E., Meyers, C.C., Parker, H., Prednis, L., Ansari, Y., Choy, N., Deen, H., Geralt, M., Hazari, N., Hom, E., Karnes, M., Mulholland, C., Ndubaku, R., Schmidt, I., Guzman, P., Aguilar-Henonin, L., Schmid, M., Weigel, D., Carter, D.E., Marchand, T., Risseeuw, E., Brogden, D., Zeko, A., Crosby, W.L., Berry, C.C., and Ecker, J.R. (2003). Genome-wide insertional mutagenesis of Arabidopsis thaliana. Science **301,** 653-657.

Baxter, I., Ouzzani, M., Orcun, S., Kennedy, B., Jandhyala, S.S., and Salt, D.E. (2007). Purdue ionomics information management system. An integrated functional genomics platform. Plant Physiol **143,** 600-611.

Baxter, I.R., Vitek, O., Lahner, B., Muthukumar, B., Borghi, M., Morrissey, J., Guerinot, M.L., and Salt, D.E. (2008). The leaf ionome as a multivariable system to detect a plant's physiological status. Proc Natl Acad Sci U S A **105,** 12081-12086.

Benfey, P.N., Bennett, M., and Schiefelbein, J. (2010). Getting to the root of plant biology: impact of the Arabidopsis genome sequence on root research. Plant J **61,** 992-1000.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet **29,** 365-371.

Brooks, T.L., Miller, N.D., and Spalding, E.P. (2010). Plasticity of Arabidopsis root gravitropism throughout a multidimensional condition space quantified by automated image analysis. Plant Physiol **152,** 206-216.

Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C., Goodman, M.M., Harjes, C., Guill, K., Kroon, D.E., Larsson, S., Lepak, N.K., Li, H., Mitchell, S.E., Pressoir, G., Peiffer, J.A., Rosas, M.O., Rocheford, T.R., Romay, M.C., Romero, S., Salvo, S., Sanchez Villeda, H., da Silva, H.S., Sun, Q., Tian, F.,

Upadyayula, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S., and McMullen, M.D. (2009). The genetic architecture of maize flowering time. Science **325,** 714-718.

Clark, R.T., MacCurdy, R.B., Jung, J.K., Shaff, J.E., McCouch, S.R., Aneshansley, D.J., and Kochian, L.V. (2011). Three-dimensional root phenotyping with a novel imaging and software platform. Plant Physiol **156,** 455-465.

Cocuron, J.C., Lerouxel, O., Drakakaki, G., Alonso, A.P., Liepman, A.H., Keegstra, K., Raikhel, N., and Wilkerson, C.G. (2007). A gene from the cellulose synthase-like C family encodes a beta-1,4 glucan synthase. Proc Natl Acad Sci U S A **104,** 8550-8555.

Eshed, Y., and Zamir, D. (1995). An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics **141,** 1147-1162.

Fang, S., Yan, X., and Liao, H. (2009). 3D reconstruction and dynamic modeling of root architecture in situ and its application to crop phosphorus research. Plant J **60,** 1096-1108.

Gerke, J., Lorenz, K., and Cohen, B. (2009). Genetic interactions between transcription factors cause natural variation in yeast. Science **323,** 498-501.

Gutierrez, M., Reynolds, M.P., and Klatt, A.R. (2010). Association of water spectral indices with plant and soil water relations in contrasting wheat genotypes. J. Exp. Bot. **61,** 3291-3303.

Heffner, E.L., Jannink, J.-L., and Sorrells, M.E. (2011). Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. Plant Genome **4,** 65-75.

Irie, S., Doi, S., Yorifuji, T., Takagi, M., and Yano, K. (1987). Nucleotide sequencing and characterization of the genes encoding benzene oxidation enzymes of Pseudomonas putida. J Bacteriol **169,** 5174-5179.

Jahnke, S., Menzel, M.I., van Dusschoten, D., Roeb, G.W., Buhler, J., Minwuyelet, S., Blumler, P., Temperton, V.M., Hombach, T., Streun, M., Beer, S., Khodaverdi, M., Ziemons, K., Coenen, H.H., and Schurr, U. (2009). Combined MRI-PET dissects dynamic changes in plant structures and functions. Plant J **59,** 634-644.

Jansen, M., Gilmer, F., Biskup, B., Nagel, K.A., Rascher, U., Fischbach, A., Briem, S., Dreissen, G., Tittmann, S., Braun, S., De Jaeger, I., Metzlaff, M., Schurr, U., Scharr, H., and Walter, A. (2009). Simultaneous phenotyping of leaf growth and chlorophyll fluorescence via GROWSCREEN FLUORO allows detection of stress tolerance in Arabidopsis thaliana and other rosette plants. Funct. Plant Biol. **36,** 902-914.

Keseler, I.M., Bonavides-Martinez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R.P., Johnson, D.A., Krummenacker, M., Nolan, L.M., Paley, S., Paulsen, I.T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A.G., and Karp, P.D. (2009). EcoCyc: a comprehensive view of Escherichia coli biology. Nucleic Acids Res **37,** D464-470.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J., and Harter, K. (2007). The AtGenExpress

global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J **50,** 347-363.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., Ferreira, T., Wood, A.R., Weyant, R.J., Segre, A.V., Speliotes, E.K., Wheeler, E., Soranzo, N., Park, J.H., Yang, J., Gudbjartsson, D., Heard-Costa, N.L., Randall, J.C., Qi, L., Vernon Smith, A., Magi, R., Pastinen, T., Liang, L., Heid, I.M., Luan, J., Thorleifsson, G., Winkler, T.W., Goddard, M.E., Sin Lo, K., Palmer, C., Workalemahu, T., Aulchenko, Y.S., Johansson, A., Zillikens, M.C., Feitosa, M.F., Esko, T., Johnson, T., Ketkar, S., Kraft, P., Mangino, M., Prokopenko, I., Absher, D., Albrecht, E., Ernst, F., Glazer, N.L., Hayward, C., Hottenga, J.J., Jacobs, K.B., Knowles, J.W., Kutalik, Z., Monda, K.L., Polasek, O., Preuss, M., Rayner, N.W., Robertson, N.R., Steinthorsdottir, V., Tyrer, J.P., Voight, B.F., Wiklund, F., Xu, J., Zhao, J.H., Nyholt, D.R., Pellikka, N., Perola, M., Perry, J.R., Surakka, I., Tammesoo, M.L., Altmaier, E.L., Amin, N., Aspelund, T., Bhangale, T., Boucher, G., Chasman, D.I., Chen, C., Coin, L., Cooper, M.N., Dixon, A.L., Gibson, Q., Grundberg, E., Hao, K., Juhani Junttila, M., Kaplan, L.M., Kettunen, J., Konig, I.R., Kwan, T., Lawrence, R.W., Levinson, D.F., Lorentzon, M., McKnight, B., Morris, A.P., Muller, M., Suh Ngwa, J., Purcell, S., Rafelt, S., Salem, R.M., Salvi, E., Sanna, S., Shi, J., Sovio, U., Thompson, J.R., Turchin, M.C., Vandenput, L., Verlaan, D.J., Vitart, V., White, C.C., Ziegler, A., Almgren, P., Balmforth, A.J., Campbell, H., Citterio, L., De Grandi, A., Dominiczak, A., Duan, J., Elliott, P., Elosua, R., Eriksson, J.G., Freimer, N.B., Geus, E.J., Glorioso, N., Haiqing, S., Hartikainen, A.L., Havulinna, A.S., Hicks, A.A., Hui, J., Igl, W., Illig, T., Jula, A., Kajantie, E., Kilpelainen, T.O., Koiranen, M., Kolcic, I., Koskinen, S., Kovacs, P., Laitinen, J., Liu, J., Lokki, M.L., Marusic, A., Maschio, A., Meitinger, T., Mulas, A., Pare, G., Parker, A.N., Peden, J.F., Petersmann, A., Pichler, I., Pietilainen, K.H., Pouta, A., Ridderstrale, M., Rotter, J.I., Sambrook, J.G., Sanders, A.R., Schmidt, C.O., Sinisalo, J., Smit, J.H., Stringham, H.M., Bragi Walters, G., Widen, E., Wild, S.H., Willemsen, G., Zagato, L., Zgaga, L., Zitting, P., Alavere, H., Farrall, M., McArdle, W.L., Nelis, M., Peters, M.J., Ripatti, S., van Meurs, J.B., Aben, K.K., Ardlie, K.G., Beckmann, J.S., Beilby, J.P., Bergman, R.N., Bergmann, S., Collins, F.S., Cusi, D., den Heijer, M., Eiriksdottir, G., Gejman, P.V., Hall, A.S., Hamsten, A., Huikuri, H.V., Iribarren, C., Kahonen, M., Kaprio, J., Kathiresan, S., Kiemeney, L., Kocher, T., Launer, L.J., Lehtimaki, T., Melander, O., Mosley, T.H., Jr., Musk, A.W., Nieminen, M.S., O'Donnell, C.J., Ohlsson, C., Oostra, B., Palmer, L.J., Raitakari, O., Ridker, P.M., Rioux, J.D., Rissanen, A., Rivolta, C., Schunkert, H., Shuldiner, A.R., Siscovick, D.S., Stumvoll, M., Tonjes, A., Tuomilehto, J., van Ommen, G.J., Viikari, J., Heath, A.C., Martin, N.G., Montgomery, G.W., Province, M.A., Kayser, M., Arnold, A.M., Atwood, L.D., Boerwinkle, E., Chanock, S.J., Deloukas, P., Gieger, C., Gronberg, H., Hall, P., Hattersley, A.T., Hengstenberg, C., Hoffman, W., Lathrop, G.M., Salomaa, V., Schreiber, S., Uda, M., Waterworth, D., Wright, A.F., Assimes, T.L., Barroso, I., Hofman, A., Mohlke, K.L., Boomsma, D.I., Caulfield, M.J.,

Cupples, L.A., Erdmann, J., Fox, C.S., Gudnason, V., Gyllensten, U., Harris, T.B., Hayes, R.B., Jarvelin, M.R., Mooser, V., Munroe, P.B., Ouwehand, W.H., Penninx, B.W., Pramstaller, P.P., Quertermous, T., Rudan, I., Samani, N.J., Spector, T.D., Volzke, H., Watkins, H., Wilson, J.F., Groop, L.C., Haritunians, T., Hu, F.B., Kaplan, R.C., Metspalu, A., North, K.E., Schlessinger, D., Wareham, N.J., Hunter, D.J., O'Connell, J.R., Strachan, D.P., Wichmann, H.E., Borecki, I.B., van Duijn, C.M., Schadt, E.E., Thorsteinsdottir, U., Peltonen, L., Uitterlinden, A.G., Visscher, P.M., Chatterjee, N., Loos, R.J., Boehnke, M., McCarthy, M.I., Ingelsson, E., Lindgren, C.M., Abecasis, G.R., Stefansson, K., Frayling, T.M., and Hirschhorn, J.N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature **467,** 832-838.

Last, R.L., Jones, A.D., and Shachar-Hill, Y. (2007). Towards the plant metabolome and beyond. Nat Rev Mol Cell Biol **8,** 167-174.

Lu, Y., Savage, L.J., Larson, M.D., Wilkerson, C.G., and Last, R.L. (2011). Chloroplast 2010: a database for large-scale phenotypic screening of Arabidopsis mutants. Plant Physiol **155,** 1589-1600.

Lu, Y., Savage, L.J., Ajjawi, I., Imre, K.M., Yoder, D.W., Benning, C., Dellapenna, D., Ohlrogge, J.B., Osteryoung, K.W., Weber, A.P., Wilkerson, C.G., and Last, R.L. (2008). New connections across pathways and cellular processes: industrialized mutant screening reveals novel associations between diverse phenotypes in Arabidopsis. Plant Physiol **146,** 1482-1500.

Massonnet, C., Vile, D., Fabre, J., Hannah, M.A., Caldana, C., Lisec, J., Beemster, G.T., Meyer, R.C., Messerli, G., Gronlund, J.T., Perkovic, J., Wigmore, E., May, S., Bevan, M.W., Meyer, C., Rubio-Diaz, S., Weigel, D., Micol, J.L., Buchanan-Wollaston, V., Fiorani, F., Walsh, S., Rinn, B., Gruissem, W., Hilson, P., Hennig, L., Willmitzer, L., and Granier, C. (2010). Probing the reproducibility of leaf growth and molecular phenotypes: a comparison of three Arabidopsis accessions cultivated in ten laboratories. Plant Physiol **152,** 2142-2157.

McMullen, M.D., Kresovich, S., Villeda, H.S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S.E., Peterson, B., Pressoir, G., Romero, S., Oropeza Rosas, M., Salvo, S., Yates, H., Hanson, M., Jones, E., Smith, S., Glaubitz, J.C., Goodman, M., Ware, D., Holland, J.B., and Buckler, E.S. (2009). Genetic properties of the maize nested association mapping population. Science **325,** 737-740.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics **157,** 1819-1829.

Murchie, E.H., and Niyogi, K.K. (2011). Manipulation of photoprotection to improve plant photosynthesis. Plant Physiol **155,** 86-92.

Nichols, R.J., Sen, S., Choo, Y.J., Beltrao, P., Zietek, M., Chaba, R., Lee, S., Kazmierczak, K.M., Lee, K.J., Wong, A., Shales, M., Lovett, S., Winkler, M.E., Krogan, N.J., Typas, A., and Gross, C.A. (2011). Phenotypic landscape of a bacterial cell. Cell **144,** 143-156.

Rounsley, S.D., and Last, R.L. (2010). Shotguns and SNPs: how fast and cheap sequencing is revolutionizing plant biology. Plant J **61,** 922-927.

Salome, P.A., Bomblies, K., Laitinen, R.A., Yant, L., Mott, R., and Weigel, D. (2011). Genetic architecture of flowering-time variation in Arabidopsis thaliana. Genetics **188,** 421-433.

Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., Pleban, T., Perez-Melis, A., Bruedigam, C., Kopka, J., Willmitzer, L., Zamir, D., and Fernie, A.R. (2006). Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol **24,** 447-454.

Schilmiller, A.L., Miner, D.P., Larson, M., McDowell, E., Gang, D.R., Wilkerson, C., and Last, R.L. (2010). Studies of a biochemical factory: tomato trichome deep expressed sequence tag sequencing and proteomics. Plant Physiol **153,** 1212-1223.

Segall, J.E., Block, S.M., and Berg, H.C. (1986). Temporal comparisons in bacterial chemotaxis. Proc Natl Acad Sci U S A **83,** 8987-8991.

Shasha, D.E., Kouranov, A.Y., Lejay, L.V., Chou, M.F., and Coruzzi, G.M. (2001). Using combinatorial design to study regulation by multiple input signals. A tool for parsimony in the post-genomics era. Plant Physiol **127,** 1590-1594.

Shroff, R., Vergara, F., Muck, A., Svatos, A., and Gershenzon, J. (2008). Nonuniform distribution of glucosinolates in Arabidopsis thaliana leaves has important consequences for plant defense. Proc Natl Acad Sci U S A **105,** 6196-6201.

Smith, S.E., and Smith, F.A. (2011). Roles of arbuscular mycorrhizas in plant nutrition and growth: new paradigms from cellular to ecosystem scales. Annu Rev Plant Biol **62,** 227-250.

Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Allen, H.L., Lindgren, C.M., Luan, J., Magi, R., Randall, J.C., Vedantam, S., Winkler, T.W., Qi, L., Workalemahu, T., Heid, I.M., Steinthorsdottir, V., Stringham, H.M., Weedon, M.N., Wheeler, E., Wood, A.R., Ferreira, T., Weyant, R.J., Segre, A.V., Estrada, K., Liang, L., Nemesh, J., Park, J.H., Gustafsson, S., Kilpelainen, T.O., Yang, J., Bouatia-Naji, N., Esko, T., Feitosa, M.F., Kutalik, Z., Mangino, M., Raychaudhuri, S., Scherag, A., Smith, A.V., Welch, R., Zhao, J.H., Aben, K.K., Absher, D.M., Amin, N., Dixon, A.L., Fisher, E., Glazer, N.L., Goddard, M.E., Heard-Costa, N.L., Hoesel, V., Hottenga, J.J., Johansson, A., Johnson, T., Ketkar, S., Lamina, C., Li, S., Moffatt, M.F., Myers, R.H., Narisu, N., Perry, J.R., Peters, M.J., Preuss, M., Ripatti, S., Rivadeneira, F., Sandholt, C., Scott, L.J., Timpson, N.J., Tyrer, J.P., van Wingerden, S., Watanabe, R.M., White, C.C., Wiklund, F., Barlassina, C., Chasman, D.I., Cooper, M.N., Jansson, J.O., Lawrence, R.W., Pellikka, N., Prokopenko, I., Shi, J., Thiering, E., Alavere, H., Alibrandi, M.T., Almgren, P., Arnold, A.M., Aspelund, T., Atwood, L.D., Balkau, B., Balmforth, A.J., Bennett, A.J., Ben-Shlomo, Y., Bergman, R.N., Bergmann, S., Biebermann, H., Blakemore, A.I., Boes, T., Bonnycastle, L.L., Bornstein, S.R., Brown, M.J., Buchanan, T.A., Busonero, F., Campbell, H., Cappuccio, F.P., Cavalcanti-Proenca, C., Chen, Y.D., Chen, C.M., Chines, P.S., Clarke, R., Coin, L., Connell, J., Day, I.N., Heijer, M., Duan, J., Ebrahim, S., Elliott, P.,

Elosua, R., Eiriksdottir, G., Erdos, M.R., Eriksson, J.G., Facheris, M.F., Felix, S.B., Fischer-Posovszky, P., Folsom, A.R., Friedrich, N., Freimer, N.B., Fu, M., Gaget, S., Gejman, P.V., Geus, E.J., Gieger, C., Gjesing, A.P., Goel, A., Goyette, P., Grallert, H., Grassler, J., Greenawalt, D.M., Groves, C.J., Gudnason, V., Guiducci, C., Hartikainen, A.L., Hassanali, N., Hall, A.S., Havulinna, A.S., Hayward, C., Heath, A.C., Hengstenberg, C., Hicks, A.A., Hinney, A., Hofman, A., Homuth, G., Hui, J., Igl, W., Iribarren, C., Isomaa, B., Jacobs, K.B., Jarick, I., Jewell, E., John, U., Jorgensen, T., Jousilahti, P., Jula, A., Kaakinen, M., Kajantie, E., Kaplan, L.M., Kathiresan, S., Kettunen, J., Kinnunen, L., Knowles, J.W., Kolcic, I., Konig, I.R., Koskinen, S., Kovacs, P., Kuusisto, J., Kraft, P., Kvaloy, K., Laitinen, J., Lantieri, O., Lanzani, C., Launer, L.J., Lecoeur, C., Lehtimaki, T., Lettre, G., Liu, J., Lokki, M.L., Lorentzon, M., Luben, R.N., Ludwig, B., Manunta, P., Marek, D., Marre, M., Martin, N.G., McArdle, W.L., McCarthy, A., McKnight, B., Meitinger, T., Melander, O., Meyre, D., Midthjell, K., Montgomery, G.W., Morken, M.A., Morris, A.P., Mulic, R., Ngwa, J.S., Nelis, M., Neville, M.J., Nyholt, D.R., O'Donnell, C.J., O'Rahilly, S., Ong, K.K., Oostra, B., Pare, G., Parker, A.N., Perola, M., Pichler, I., Pietilainen, K.H., Platou, C.G., Polasek, O., Pouta, A., Rafelt, S., Raitakari, O., Rayner, N.W., Ridderstrale, M., Rief, W., Ruokonen, A., Robertson, N.R., Rzehak, P., Salomaa, V., Sanders, A.R., Sandhu, M.S., Sanna, S., Saramies, J., Savolainen, M.J., Scherag, S., Schipf, S., Schreiber, S., Schunkert, H., Silander, K., Sinisalo, J., Siscovick, D.S., Smit, J.H., Soranzo, N., Sovio, U., Stephens, J., Surakka, I., Swift, A.J., Tammesoo, M.L., Tardif, J.C., Teder-Laving, M., Teslovich, T.M., Thompson, J.R., Thomson, B., Tonjes, A., Tuomi, T., van Meurs, J.B., van Ommen, G.J., Vatin, V., Viikari, J., Visvikis-Siest, S., Vitart, V., Vogel, C.I., Voight, B.F., Waite, L.L., Wallaschofski, H., Walters, G.B., Widen, E., Wiegand, S., Wild, S.H., Willemsen, G., Witte, D.R., Witteman, J.C., Xu, J., Zhang, Q., Zgaga, L., Ziegler, A., Zitting, P., Beilby, J.P., Farooqi, I.S., Hebebrand, J., Huikuri, H.V., James, A.L., Kahonen, M., Levinson, D.F., Macciardi, F., Nieminen, M.S., Ohlsson, C., Palmer, L.J., Ridker, P.M., Stumvoll, M., Beckmann, J.S., Boeing, H., Boerwinkle, E., Boomsma, D.I., Caulfield, M.J., Chanock, S.J., Collins, F.S., Cupples, L.A., Smith, G.D., Erdmann, J., Froguel, P., Gronberg, H., Gyllensten, U., Hall, P., Hansen, T., Harris, T.B., Hattersley, A.T., Hayes, R.B., Heinrich, J., Hu, F.B., Hveem, K., Illig, T., Jarvelin, M.R., Kaprio, J., Karpe, F., Khaw, K.T., Kiemeney, L.A., Krude, H., Laakso, M., Lawlor, D.A., Metspalu, A., Munroe, P.B., Ouwehand, W.H., Pedersen, O., Penninx, B.W., Peters, A., Pramstaller, P.P., Quertermous, T., Reinehr, T., Rissanen, A., Rudan, I., Samani, N.J., Schwarz, P.E., Shuldiner, A.R., Spector, T.D., Tuomilehto, J., Uda, M., Uitterlinden, A., Valle, T.T., Wabitsch, M., Waeber, G., Wareham, N.J., Watkins, H., Wilson, J.F., Wright, A.F., Zillikens, M.C., Chatterjee, N., McCarroll, S.A., Purcell, S., Schadt, E.E., Visscher, P.M., Assimes, T.L., Borecki, I.B., Deloukas, P., Fox, C.S., Groop, L.C., Haritunians, T., Hunter, D.J., Kaplan, R.C., Mohlke, K.L., O'Connell, J.R., Peltonen, L., Schlessinger, D., Strachan, D.P., van Duijn, C.M., Wichmann, H.E., Frayling, T.M., Thorsteinsdottir, U., Abecasis, G.R.,

Barroso, I., Boehnke, M., Stefansson, K., North, K.E., McCarthy, M.I., Hirschhorn, J.N., Ingelsson, E., and Loos, R.J. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet **42,** 937-948.

Velasco, L., Pérez-Vich, B., and Fernández-Martínez, J.M. (1999). Estimation of seed weight, oil content and fatty acid composition in intact single seeds of rapeseed (*Brassica napus* L.) by near-infrared reflectance spectroscopy. Euphytica **106,** 79-85.

Walter, A., Silk, W.K., and Schurr, U. (2009). Environmental effects on spatial and temporal patterns of leaf and root growth. Annu Rev Plant Biol **60,** 279-304.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., El Bakkoury, M., Foury, F., Friend, S.H., Gentalen, E., Giaever, G., Hegemann, J.H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D.J., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J.L., Riles, L., Roberts, C.J., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R.K., Veronneau, S., Voet, M., Volckaert, G., Ward, T.R., Wysocki, R., Yen, G.S., Yu, K., Zimmermann, K., Philippsen, P., Johnston, M., and Davis, R.W. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science **285,** 901-906.

Yu, J., Holland, J.B., McMullen, M.D., and Buckler, E.S. (2008). Genetic design and statistical power of nested association mapping in maize. Genetics **178,** 539-551.

Appendix: Opportunities and Challenges for Microbial and Plant Phenomics

For both microbes and plants, critical decisions regarding experimental design and measurements usually can follow from a well defined starting point, which often derives from a practical agricultural, environmental or energy-related issue. A number of illustrative examples of goal-driven topics for phenomic research were mentioned in the main body of the White Paper and are provided here in more detail, as follows:

The Microbiome

- The plant microbiome
- Microbial quorum sensing
- Plant fungal interactions in mycorrhizae
- Understanding microbial communities related to bioenergy
- Understanding microbial communities for environmental remediation
- Linking plant health to the rhizosphere microbiome

The Plant Phenome

- General context
- Increasing yield
- Spatial and temporal understanding - disease resistance
- Root system developmental plasticity

I. The Microbiome

The Plant Microbiome. It has long been appreciated that plants and microbes form intimate associations. Studies of *Agrobacterium tumefaciens* and the crown gall tumors they cause spawned the entire field of plant genetic engineering (Valentine 2003). Symbiotic associations between rhizobia and legumes involve an intricate set of plant-bacterial communications resulting in root nodules that allow bacteria to receive carbon from the plant in exchange for fixed nitrogen (Downie 2010). Insights into how microbial pathogens interact with plants at the molecular level to induce innate immune responses or to cause disease are guiding the understanding of plant-pathogen co-evolution in natural and agricultural systems. Against this backdrop researchers understand that plants maintain complex communities of epiphytic, endophytic and rhizosphere microbes; plants have microbiomes. However, these communities are largely unstudied, and their influence on plant growth, yield and general health is essentially unknown. Filling this large knowledge gap will be a challenge. How do we define if a microbe is truly plant-associated? How do we identify microbial and plant phenotypes that are critical for maintaining stable interactions? How do we deal with extreme variation in plant-associated communities? Before these framework issues can be tackled we need to address the even more fundamental question of who is there. With recent technological advances in DNA sequencing and proteomics there is an unprecedented opportunity to inventory and functionally characterize plant microbiomes.

Microbial quorum sensing. Cell-cell communication between microbes and their plant hosts. Many bacteria can perceive and respond to one another by a process called quorum

sensing (Waters and Bassler 2005). This communication system influences colonization of plant and animal hosts by both symbionts and pathogens. Over 100 species of bacteria use small diffusible signaling molecules called *N*-acyl homoserinelactones (acyl-HSLs) to control gene expression by quorum sensing. Quorum sensing signal activity has been described for many plant-associated bacteria, including members of the *Pseudomonas*, *Sinorhizobium*, *Mesorhizobium*, and *Bradyrhizobium* genera (Dulla and Lindow 2009; Mathesius *et al.* 2003). It controls a variety of processes including motility, exopolysaccharide synthesis, plasmid transfer, and efficiencies of root nodulation and nitrogen-fixation. Acyl-HSLs can elicit responses in plants as well. In the legume *Medicago truncatula*, over 150 root proteins were differentially synthesized upon addition of acyl-HSLs. The identified proteins had predicted roles in host defense response, flavonoid metabolism, and hormone response, among others. The profiles of seed exudates also changed in response to acyl-HSL addition. In recent studies, the model plant *Arabidopsis thaliana* responded to acyl-HSL addition in a variety of ways; changes were seen in: root architecture, root hair development, and gene expression (Ortiz-Castro *et al.* 2008; von Rad *et al.* 2008). These results suggest that acyl-HSLs serve not only as intraspecies bacterial signals, but also as interkingdom signals to hosts.

Plant fungal interactions in mycorrhizae - benefits to plants. Microbial interactions have the potential to greatly influence plant growth and phenotype. The earliest fossil records of fungi are in intimate association with plants (Taylor *et al.* 1995;Yuan *et al.* 2005) and phylogenetic reconstruction suggest that plant/fungal co-evolution has occurred over hundreds of millions of years (Taylor and Berbee 2006). Distinct plant-fungal associations, known as mycorrhizae, have evolved in roots to promote plant growth by mobilization of micronutrients in the soil and provide adaptive phenotypes such as drought tolerance (Gianinazzi *et al.* 2010). More recently, many other fungi that grow within plants (endophytic fungi) have been discovered that have the potential to influence plant phenotype (Saunders *et al.* 2010). For example, an endophytic fungus from a highly thermotolerant plant is necessary for plant thermotolerance; separation of the plant from the endophyte renders both organisms sensitive to high temperature (Redman *et al.* 2002). Another common endophyte of maize is able to ward of infection by pathogenic fungi (Lee *et al.* 2009). Comprehensive examination of plant endophytic fungi has only just begun, yet has great potential for the discovery of microbes which positively or negatively impact plant productivity and fitness.

Understanding microbial communities related to bioenergy. Prior to NextGen sequencing methodologies, the challenge of characterizing microbial communities associated with plants was daunting. However, currently opportunities exist for determining how plants may affect microbial community structure of soils or plant surfaces and in turn how these microbes feed back to impact plant growth and development. Opportunities also exist for determining a comprehensive microbiome of degraded plant biomass in nature that will inform efforts aimed at use of these substrates for biofuels. The future of industrial microbiology lies in harnessing microbial communities to perform complex bioprocesses patterned on natural microbial communities (Sabra *et al.* 2010).

Understanding microbial communities for environmental remediation. The phenotypic characterization of a plant microbiome is a daunting task. Is it possible to learn from others who have applied genomic and proteomic techniques to simpler natural systems? Jill Banfield at the

University of California, Berkeley and her colleagues are conducting a comprehensive metagenomic project at the Iron Mountain Mine Superfund Site in Redding, California (Denef *et al.* 2010). This extreme environment is characterized by low pH, high metal concentrations, and a very limited number of species, making the microbial community very attractive for quantitative, genomic- and proteomic-based analyses of function. The Banfield group has reconstructed near-complete genomes of essentially all the bacterial and archaeal natural populations consistently detected in acid mine drainage biofilms, as well as one fungal genome. In addition, they have mapped changes in population structure and protein abundance over space, time, and biofilm developmental stages.

Linking plant health to the rhizosphere microbiome. Plant health and productivity are influenced by microbial communities resident at the soil-root interface, and this influence has long been attributed to microbial interactions that result in the suppression of soilborne plant pathogens. Nevertheless, the microorganisms and mechanisms involved in disease suppression are largely unknown. Recently, Mendes *et al*. (2011) used a PhyloChip-based approach to characterize the microbiome of the root surface (rhizosphere), detecting more than 33,000 bacterial and archaeal species in this environment. They compared the rhizosphere microbiomes of plants grown in a disease-conducive soil and a suppressive soil, where disease does not occur despite the presence of soilborne plant pathogens. Certain taxa of bacteria (Proteobacteria, Firmicutes, and Actinobacteria) were associated consistently with the rhizosphere of plants grown in the disease suppressive soil, providing direct evidence for the microbial basis of plant disease suppression. Furthermore, disease-suppressive soils were associated with specific genes involved in the biosynthesis of secondary metabolites. Many rhizosphere bacteria that suppress plant disease produce secondary metabolites that are toxic to plant pathogenic fungi and Oomycetes (Haas and Defago 2005) and these metabolites are key contributors to disease control. The study by Mendes *et al*. (2011) provides an exciting example of the power of phenomics for linking an ecosystem to a microbiome and microbial genes, thereby providing new insight into an important ecological process that has evaded scientific inquiry in the past.

II. The Plant Phenome.

General context. Identifying genetic variation in a phenotype is key to developing strategies to understanding the processes involved in that phenotype and for improvement of a phenotype (whether it be plant disease resistance, altering plant cell walls for biomass conversion properties, or identifying plants with tolerance to abiotic stresses). High throughput phenomic approaches will expedite identification of variation in natural (diverse germplasm collections, association genetic panels) and derived (e.g. recombinant inbred, mutant, wide introgression lines, etc.) genetic populations. **The challenge** for successful application of phenomic approaches will be in the design and adaptation of robust screens to allow high throughput, reliable, and meaningful comparisons. For efficient phenotyping, it is important not only to be able to make fast, accurate and reproducible measurements, but to know **what** phenotypes to measure. In this regard, there is often inadequate underlying understanding of responses and mechanisms of plant adaptation to environmental variation/stress. Greater understanding of responses/mechanisms is needed to increase the precision of phenotyping---there is a need to know **what type of data to collect and where and when to collect it.** This is true regardless of the level of phenotyping, whether molecular/biochemical or morphological. This is why

physiology-based breeding, or "physiological breeding", has been largely unsuccessful. To at least some extent the issue has been insufficient understanding of key responses/mechanisms to inform the phenotyping. To illustrate the problem, a visit to the doctor's office can result in a battery of measurements that accurately inform the condition of the patient. For example, hormone balance is very informative of growth and other characteristics. But, in plants, we do not have enough fundamental knowledge to be able to make equally informative screening assessments. For example, how does plant hormone balance respond to different stress conditions in different tissues and cell types, and what is the role of the different plant hormones in responses to these conditions? There is some such knowledge in the literature, but certainly not of sufficient comprehensiveness to allow equivalent metabolic fingerprinting/prediction of particular stress conditions. The lack of sufficient knowledge of this type is being addressed in part by the formation of information networks such as the International Plant Phenomics Network (http://www.plantphenomics.com/sites/), which is a science-based concept and technology resources addressing major challenges of plant performance including:

- Breeding plants for a changing environment
- Prognosis of plant performance in global change
- Innovative plant production for present and future crops based on the understanding of the complex interaction of plants with their environment and its dynamics
- Monitoring of plant performance in natural systems

Increasing yield. For crop plants, the key trait or phenotype of interest is yield. Yield can be a direct reflection of biomass or the proportion of biomass that is converted to the harvestable commodity; for grain this is the harvest index. Yield is the integration of many physiological processes over time and in environments that may be fluctuating or highly variable. The interactions between the genotype of the organisms and the environment can lead to variable or unstable phenotypes. There are various strategies that that be used to try and minimize these variability. The most obvious is to minimize the environmental variation by growing the organism under a tightly controlled or managed environment. This could involve growth in a glasshouse, growth room, or incubator but it can be difficult to accurately replicate the target or field environment; for example, physical constraints make it impractical to evaluate mature forest trees or a whole plant community, such as a crop, under a controlled environment. The alternative is to use the field environment but closely monitor the environmental variables or control specific aspects of the environment; for example, rainout shelters or irrigation can be used to simulate various levels of water availability. Field or *in situ* evaluation of single genotypes or populations has been the basis for phenotyping by plant breeders and ecologists. Many plant breeding program run hundreds of thousands of field plots every year to generate data on yield and yield components. Careful measurement of the environment has been critical and new measurement techniques have greatly improved the accuracy of environmental monitoring. This includes sophisticated weather stations that record very small changes in many parameters simultaneously and transmit these directly to researchers. Imaging and analysis of growth from the single plant level to the landscape level are now feasible using cameras on microscopes or satellites. For crop and ecosystem monitoring, the ability to measure reflected light can also provide useful data on plant health. Spectral reflectance measurements of crop plants can indicate plant health and response to climatic stresses or pathogen attack. Many of these techniques can be applied at high throughput, thus all fall under the moniker of phenomics.

Spatial and temporal understanding - disease resistance. A particularly appealing attribute of phenomic approaches is that automation allows for design of screens that better encompass responses over time. For example, rather than relying on 'yes/no' phenotypes (typical of qualitative traits), novel approaches for improving plant disease resistance could take into account a continuum from disease to resistance (typical of multigenic or quantitative traits), and therefore, would rely on assessing large plant populations for the amount of disease development over time in response to multiple pathogens or types of pathogens. A specific example would the slow stem rusting genes for *Puccinia graminis* race Ug99 resistance. Both race-specific "gene-for-gene" and polygenic resistance seem to exist in wild wheat and barley germplasm. The task is to locate the genomic positions of resistance determinants and introgress those regions into elite backgrounds without dragging agronomically undesirable neighboring genes (Hiebert *et al*. 2011). A well coordinated effort combining large scale field tests of breeding materials with automated observation of carefully chosen small subsets of these materials may address critical aspects of plant architecture, reproductive development and disease progression. Automated conditions should simulate different environments (solar irradiation, humidity, soil types, etc) to maximize the relevance of large datasets to the selection of traits for specific geographic areas and climatic conditions. Variations may address problems such as flood, salinity or drought tolerance. Simultaneous measurement of physiological and morphological status over time (photosynthesis/respiration rates, growth rates, composition of root exudates, etc.) may provide holistic insights into fundamental mechanisms of plant response or adaptation.

Root system developmental plasticity. An example of the complexity of morphological phenotyping is provided by root system responses to soil drying. Root system developmental responses (plasticity) in response to drought are complex (O'Toole and Bland 1987); different types of roots respond differently, and responses will be determined not only by water status but other interacting variables. As an example, lateral (secondary) root proliferation (number and length) can be stimulated in response to mild water deficit (Read and Bartlett 1972; Jupp and Newman 1987), but inhibited as the soil dries further. Thus, the phenotype of lateral root proliferation can be evaluated only under specific mild water deficit conditions. In addition, the response will be spatially and temporally variable as the soil profile dries. Root hair proliferation can also be stimulated in response to water deficits (Vasellati *et al*. 2001), but many phenotyping systems may not have the precision to evaluate this response. The complexity of phenotyping root system developmental responses to soil drying is further compounded by other interacting variables such as soil type (Sponchiado *et al*. 1989), rate of drying, interaction with other stress conditions (soil strength, temperature, etc), interaction with nutritional status, and interaction with rhizosphere microorganisms. For example, phosphorous (P) deficiency also causes lateral root and root hair proliferation (Zhu and Lynch 2004; Zhu *et al*. 2010). Since soil drying causes decreased P mobility in the soil, soil drying and soil P status may have interacting effects that will also be altered by interacting mycorrhiza, which also influence P uptake (e.g., Zhu *et al*. 2005). The complexity of the question is further compounded by consideration of varying microbial populations, disease pressures, climate, soil types, etc in different regions.

References Cited (Appendix)

Denef VJ, Mueller RS, Banfield JF (2010) AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *ISME Journal* **4**: 599-610.

Downie JA (2010) The roles of extracellular proteins, polysaccharides and signals in the interactions of rhizobia with legume roots. *FEMS Microbiol Rev* **34**: 150-170.

Dulla GF, Lindow SE (2009) Acyl-homoserine lactone-mediated cross talk among epiphytic bacteria modulates behavior of Pseudomonas syringae on leaves. *ISME Journal* **3**: 825-834.

Gianinazzi S, Gollotte A, Binet M-N, van Tuinen D, Redecker D, Wipf D (2010) Agroecology: the key role of arbuscular mycorrhizas in ecosystem services. *Mycorrhiza* **20**: 519-530.

Haas D, Defago G (2005). Biological control of soil-borne pathogens by fluorescent pseudomonads. *Nature Reviews Microbiology* **3**: 307-319.

Hiebert CW, Fetch TG, Zegeye T, Thomas JB, Somers DJ, Humphreys DG, McCallum BD, Cloutier S, Singh D, Knott DR (2011) Genetics and mapping of seedling resistance to Ug99 stem rust in Canadian wheat cultivars 'Peace' and 'AC Cadillac'. *Theor Appl Genetics* **122:** 143-149.

Jupp AP, Newman EI (1987) Morphological and anatomical effects of severe drought on the roots of *Lolium perenne* L. *Annals of Botany* **105**: 393-402.

Lee K, Pan JJ, May G (2009) Endophytic *Fusarium verticilliodes* reduces disease severity caused by *Ustilago maydis* on maize. *FEMS Microbiol Lett* **299**: 31-37.

Mathesius U, Mulders S, Gao M, Teplitski M, Caetano-Anolles G, Rolfe BG, Bauer WD (2003) Extensive and specific responses of a eukaryote to bacterial quorum-sensing signals. *Proc Natl Acad Sci USA* **100**: 1444-1449.

Mendes R, Kruijt M, de Bruijn I, Dekkers E, van der Voort M, Schneider JH, Piceno YM, DeSantis TZ, Andersen GL, Bakker PA, Raaijmakers JM (2011) Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332**: 1097-1100.

O'Toole JC, Bland WL (1987) Genotypic variation in crop plant root systems. *Advances in Agronomy* **41**: 91-145.

Ortiz-Castro R, Martinez-Trujillo M, Lopez-Bucio J (2008) N-acyl-L-homoserine lactones: a class of bacterial quorum-sensing signals alter post-embryonic root development in Arabidopsis thaliana. *Plant Cell Environ* **31**: 1497-1509.

Read DJ, Bartlett EM (1972) The physiology of drought resistance in the soy-bean plant (*Glycine max*). I. The relationship between drought resistance and growth. *Journal of Applied Ecology* **9**: 487-499.

Redman R, Sheehan KB, Stout RG, Rodriguez TJ, Henson JM (2002) Thermotolerance generated by plant/fungal symbiosis. *Science* **298**: 1581.

Sabra W, Dietz D, Tjahjasari D, Zeng A-P (2010) Biosystems analysis and engineering of microbial consortia for industrial biotechnology. *Eng Life Sci* **10**: 407-421.

Saunders M, Glenn AE, Kohn LM (2010) Exploring the evolutionary ecology of fungal endophytes in agricultural systems: using functional traits to reveal mechanisms in community processes. *Evolutionary Applications* **3**: 525–537.

Sponchiado BN, White JW, Castillo JA, Jones PG (1989) Root growth of four common bean cultivars in relation to drought tolerance in environments with contrasting soil types. *Experimental Agriculture* **25**: 249-257.

Taylor JW, Berbee ML (2006) Dating divergences in the fungal tree of life: review and new analyses. *Mycologia* **98**: 838-849.

Taylor TN, Remy W, Hass H, Kerp H (1995) Fossil arbuscular mycorrhizae from the Early Devonian. *Mycologia* **87**: 560-573.

Valentine L (2003) *Agrobacterium tumefaciens* and the plant: the David and Goliath of modern genetics. *Plant Physiol* **133**: 948-955.

Vasellati V, Oesterheld M, Medan D, Loreti J (2001) Effects of flooding and drought on the anatomy of *Paspalum dilatum*. *Annals of Botany* **88**: 355-360.

von Rad U, Klein I, Dobrev PI, Kottova J, Zazimalova E, Fekete A, Hartmann A, Schmitt-Kopplin P, Durner J (2008) Response of Arabidopsis thaliana to N-hexanoyl-DL-homoserine-lactone, a bacterial quorum sensing molecule produced in the rhizosphere. *Planta* **229**: 73-85.

Waters CM, Bassler BL (2005) Quorum sensing: cell-to-cell communication in bacteria. *Annu Rev Cell Dev Biol* **21**: 319-346.

Yuan X, Xiao S, Taylor TN (2005) Lichen-like symbiosis 600 million years ago. *Science* **308**: 1017-1020.

Zhu J, Kaeppler SM, Lynch JP (2005) Topsoil foraging and phosphorus acquisition efficiency in maize (*Zea mays*). *Functional Plant Biology* **32**: 749-762.

Zhu J, Lynch JP (2004) The contribution of lateral rooting to phosphorus acquisition efficiency in maize (*Zea mays*) seedlings. *Functional Plant Biology* **31**: 949-958.

Zhu J, Zhang C, Lynch JP (2010) The utility of phenotypic plasticity of root hair length for phosphorus acquisition. *Functional Plant Biology* **37**: 313-322.