*Application of Semantic Technology in Biodiversity Science*
*Anne E. Thessen (Ed.)*
*ISBN 978-3-89838-733-0*
*© 2018 AKA Verlag Berlin*

# Chapter 10

# 'Computable' Phenotypes Enable Comparative and Predictive Phenomics Among Plant Species and Across Domains of Life

Ian BRAUN[a], James P. BALHOFF[b**], Tanya Z. BERARDINI[c**], Laurel COOPER[d**], Georgios GKOUTOS[e**], Lisa HARPER[f,g**], Eva HUALA[c**], Pankaj JAISWAL[d**], Toni KAZIC[h**], Hilmar LAPP[i**], James A. MACKLIN[j**], Chelsea D. SPECHT[k**], Todd VISION[l**], Ramona L. WALLS[m**], and Carolyn J. LAWRENCE-DILL[a,n*]

[a]Department of Genetics, Development and Cell Biology and Interdepartmental Bioinformatics and Computational Biology, Iowa State University, Ames, Iowa, USA
[b]Renaissance Computing Institute, University of North Carolina, Chapel Hill, North Carolina, USA
[c]Phoenix Bioinformatics, Redwood City, California, USA
[d]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA
[e]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK
[f]USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, Iowa, USA and [g]USDA-ARS Plant Gene Expression Center, Albany, California, USA
[h]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, USA

[i]Center for Genomic and Computational Biology, Duke University, Durham, North Carolina, USA
[j]Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada
[k]School of Integrative Plant Sciences, Cornell University, Ithaca, New York, USA
[l]Biology Department, University of North Carolina, Chapel Hill, North Carolina, USA
[m]CyVerse, Bio5 Institute, University of Arizona, Tucson, Arizona, USA
[n]Department of Agronomy, Iowa State University, Ames, Iowa, USA
[*]communication triffid@iastate.edu
[**]alphabetical

Scientists are adept at comparing genomic sequences. The collection of more such data promises to increase our ability to determine gene function, discover and describe biological processes, and prioritize causative variants of interest that underlie disease response. Yet the question remains: Can we compare phenotypes or traits of interest across disciplines in a manner similar to how we compare genomic sequences? Here we present examples of 'semantic reasoning' - computational methodologies that enable computation across organized formal phenotypic representations. These methods facilitate the analysis of phenotype information across species, domains of knowledge, people, and computers. We review representative examples of successful semantic reasoning to recover known biological phenomena in medical and agricultural applications. Necessary changes in how we collect, analyze, and share data to enable such computations are presented, and database and analytic tool suites for these sorts of analyses are described.

## 1. Background

Phenotypic variation is the raw material acted on by both natural and artificial selection; it provides the diversity required for species to adapt and respond to changing environments. Linking phenotypes to genotypes across evolutionarily distant lineages provides researchers with the ability to predict phenotypic outcomes of genotypic changes, and to compare genetic strategies that give rise to like phenotypes. This information in turn enables researchers to develop medical and agricultural innovations and to assess and manage the organismal and community-level variation critical to maintaining ecosystem processes and adaptive responses to climate change.

Phenotypic data are extremely diverse, ranging in scope from expression profiles, to quantitative information, to summarizing textual descriptions of development. Data can be associated to individuals, populations, or species and can be described in comparative terms (e.g., mutant versus wild type) or absolute measurements (e.g., days to flowering). The documentation of these data can be at a summary level (e.g., average height of plants studied) or measured (a particular plant is measured to be 62 cm tall). For these reasons and myriad others, the documentation, integration, representation, and accessibility of phenotype data is notoriously challenging (reviewed in[1]). Adding to the complexity, new high-throughput measurements of phenotype can involve remote sensing, high-density imaging, and integration with geo-location data. Because phenotypic data are so diverse, and the rates, volumes, and complexities of data collection are only increasing, it is difficult to aggregate these complex datasets for downstream analyses (reviewed in[2]).

In an effort to leverage the wealth of phenotypic data available for making important biological inferences, McGary et al.[3] developed a method of candidate gene discovery involving phenologs, or orthologous phenotypes. As defined by the authors, phenotype *A* from species *A* and phenotype *B* from species *B* are phenologs if their two sets of known causal genes have a significant overlap in the form of orthologous genes. Once phenotypes *A* and *B* have been identified as phenologs, the authors' methodology identifies candidate genes as those genes which are known to be causal in one species, but are not currently associated with that phenotype in the other species. If Gene 1 is causal to phenotype *A*, then its ortholog in species *B* is a candidate gene for phenotype *B*. The authors demonstrated the utility of this methodology by discovering non-obvious model systems for human disease phenotypes, predicting specific novel candidate genes associated with those phenotypes, and verifying selected predictions. For example, a significant overlap in orthologous genes revealed a phenolog relationship between mammalian abnormal angiogenesis and reduced rates of growth in lovastatin-treated yeast. A predicted candidate gene for angiogenesis, *SOX13* (known to be causal to the yeast reduced growth rate phenotype), was experimentally confirmed through knockdown studies in both mouse and human cells.

The methodology of finding phenologs first proposed by McGary et al.[3] relies on previously known genotype to phenotype associations and the use of orthology relationships between genes to reveal related phenotypes. Phenologs, however, may also be proposed based solely on the characteristics of the phenotypes themselves, represented in the form of textual or other data. Relying on textual descriptions rather than genotype to phenotype associations to identify phenologs is advantageous in the case of phenotypes for which associated genotypes (causal genes) are not known, or causal genes are involved in similar pathways between the species but are not necessarily orthologous. Example phenolog sets that generate similar phenotypes are kinesin motor proteins that, when mutated, cause trichome branching defects in *Arabidopsis* and neuronal branching defects in mice (Fig. 1), and some genes involved in lesion formation in both humans and maize (Fig. 2).

The lesion phenotypes shown in panels A and B of Figure 2 are caused by reduced activity of uroporphyrinogen decarboxylase that leads to the accumulation of uroporphyrin and related metabolites[4–6]. The enzymes encoded by the *UROD* and *Les22* genes in humans and maize, respectively are 35% identical and so would be readily discovered by the method of McGary et al.[3,7] (GenBank: Accession No. NP_000365.3 and Kazic, unpublished). But approximately 54 other mutations producing lesion phenotypes in maize have been confirmed so far, and for most neither the gene nor the biochemical functions it encodes have been identified (Neuffer and Kazic, unpublished). All of these mutations produce discontiguous patches of chlorotic or necrotic tissue on leaves, often in response to light or developmental cues, and their morphology, behavior, spatial distribution, and time of onset are sensitive to genetic background and environmental perturbations. For example, Figure 2 panel C shows a classic oscillatory lesion phenotype produced by *Les1*[8]. Mapping has placed the locus on the short arm of chromosome 2, and no biochemical function has been described[9]. Only searching over phenologs would discover *Les1* and its phenotypes. Many other lesion mimic mutants of maize display these types of oscillations under appropriate conditions, and these provide important clues to the underlying causal mechanisms (Kazic, unpublished). Such searches would benefit considerably from also annotating other important dimensions of the phenotype, such as its spatiotemporal oscillation and sensitivity to ambient temperature and genetic background.

Regardless of the genotypic data associated with these phenotypes, they share characteristic morphologies across the species that are readily summarized by images. It is likely that morphological phenologs will eventually be directly discovered by sophisticated combinations of image analysis and pattern recognition techniques that can be used on distributed image databases, though these techniques are only just emerging and will require significant research. However, many dimensions of these phenotypes, and many other non-morphological phenotypes, are not neatly captured by an image. For example, the time of onset of a phenotype and the environmental perturbations that trigger or modify it are far more accurately and compactly expressed as text. Further, the genetic component that produces phenocopies is sharply reduced by definition, but these phenotypes can be particularly revealing: Type I *porphyria cutanea tarda* and lesion formation in response to pathogen infection both reveal important causal mechanisms. For these non-morphological phenotypes and phenotypic dimensions, computational analysis of phenotypic descriptions will be key to automating discovery of such associations, now and for the foreseeable future.

To enable computational discovery methods, free text descriptions of phenotypes need to be associated with standardized ontology terms which are placed in hierarchical directed acyclic graphs. The analysis of ontology terms and the relationship graph in which they are placed is called semantic reasoning; it is what allows machines to "understand" (reason over) domain knowledge. The graph nature of the terms and relationships in an ontology allows metrics utilizing graph theory to be applied to quantifying similarity between terms, and thus the phenotype descriptions linked to them. Following the identification of phenologs through semantic reasoning, the same method of candidate gene discovery suggested by McGary et al.[3] can be applied. Multiple research groups are currently using semantic reasoning for novel candidate gene discovery. Hoehndorf et al.[13] predicted the association between the genes *Adam19* and *Fgf15* with the Tetralogy of Fallot disease phenotype in humans, one result of the construction of a network of phenotype similarity scores among 86,203 phenotypes across five different species (yeast, fly, worm, mouse, zebrafish) called

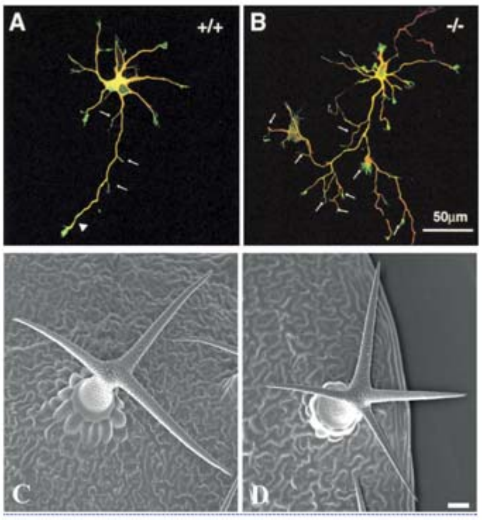


**Figure 1.** Branching defects shared between mouse and Arabidopsis cells. Homma et al.[10] reported increased branching in cultured neurons of mouse Kinesin-13 mutant KIF2A (A) wild type and (B) mutant. A mutation in the *Arabidopsis* ortholog KIN-13A (At3g16630) also shows increased trichome branching (C) wild type and (D) mutant[11]

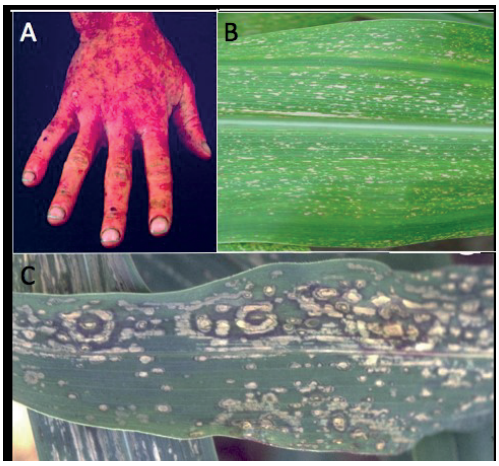


**Figure 2.** Lesion formation in humans and maize. A: *Porphyria cutanea tarda* in humans[12]. B: Lesion formation in a *Les22* mutant plant due to a mutation in the UROD enzyme (image courtesy John Gray). C: The classic oscillatory lesions displayed by *Les1* heterozygotes[8]

PhenomeNET (http://phenomebrowser.net/).

In seeking to construct a similar network to facilitate candidate gene discovery in plants, Oellrich et al.[14] developed and tested a workflow to curate and standardize existing plant phenotype datasets. This approach employed curated data to demonstrate the feasibility of semantic comparison across plant species. Data for six plant species, encompassing both model species and crop plants with established genetic resources, were integrated and analyzed using a common set of ontologies, annotation standards, formats, and best practices. The study focused on mutant phenotypes associated with genes of known sequence in *Arabidopsis thaliana* (L.) Heynh. (Arabidopsis), *Zea mays* L. subsp. *mays* (maize), *Medicago truncatula* Gaertn. (barrel medic or Medicago), *Oryza sativa* L. (rice), *Glycine max* (L.) Merr. (soybean), and *Solanum lycopersicum* L. (tomato). Curated phenotypes from taxon-specific databases were converted into a common format using the Ontologies for Plant Biology (described in[15]) that include Plant Ontology (PO[16]), Gene Ontology (GO[17]), Plant Experimental Conditions Ontology (PECO[15]), Chemical Entities of Biological Interest (ChEBI[18]) ontology, and Phenotype and Trait Ontology (PATO[19]). The ontology annotations were used to construct a matrix of semantic similarity scores for all possible pairs of inter- and intra-specific genotypes. The constructed ontology annotations representing each phenotype are referred to as EQ (Entity-Quality) statements[20], as they are composed of ontology term(s) representing a biological structure or process (entity), and term(s) representing an aspect or modification of that entity (its quality[21]).

From 2,866 genotypes yielding over 8 million possible combinations, 548,888 had non-zero semantic similarity scores. A similarity score of 0 indicates no semantic overlap with respect to the phenotype, while a similarity score of 1 indicates an identical semantic phenotype description (and therefore equivalent sets of EQs). Of these, 44% of the non-zero semantic similarity scores were below 0.1, indicating that many of the phenotypes show only a small overlap in their description, while 13% of the genotype pairs with non-zero scores fell into the 0.9 - 1 range. This indicates that for most of the genes the semantic similarity of their mutant phenotype descriptions with other genes is low. Some of the very high scores (scores near 1) are likely artifacts due to limited data curation. For example, if only some characteristics of genotypes have been annotated in the form of EQ statements, two genotypes may appear artificially much more similar or dissimilar than they would be had their phenotypes been annotated in full. Furthermore, not all phenotype changes may be reported in the literature for a given genotype in the first place. It is important to note that semantic similarity algorithms cannot compensate for such gaps in reporting or in annotation. Results of the semantic similarity analysis are provided through the Plant PhenomeNET[22] web interface, which was adapted from PhenomeNET. For each genotype, a detailed page provides information about similarity scores to any of the other genotypes as well as a link to an additional page providing the phenotype assigned by the curator and those translated to use terms represented in the ontologies.

The semantic similarity dataset was evaluated for its ability to enhance predictions of gene families, protein functions, and shared metabolic pathways that underlie informative plant phenotypes. In one example, Oellrich et al. were able to use Plant PhenomeNET to identify a set of maize gene models that participate in the initial reactions of flavonoid biosynthesis as part of the phenylpropanoid biosynthesis pathway. This result indicates that reasoning across curated phenotypes in plants is capable of recapitulating well-characterized biological phenomena and hints that, for plant species that are not genetically well-characterized, the ontological reasoning approach to predicting phenotypic associations can help with characterizing understudied species and assist in forward genetics approaches.

In a second example, Oellrich et al.[14] were able to place 2,741 EQ-annotated genes from all six species into 1,895 gene families, of which 42 contain between 5 and 12 genes with EQ statements. These families were assessed for how often homologous genes have similar functions. There were also 147 families containing EQ statements from two or more species, which allowed the authors to assess how often functions are conserved between orthologs. For most families in this sample, gene function was conserved or similar, but there were some cases in which annotated phenotypes were quite different across orthologs.

## 2. Current efforts

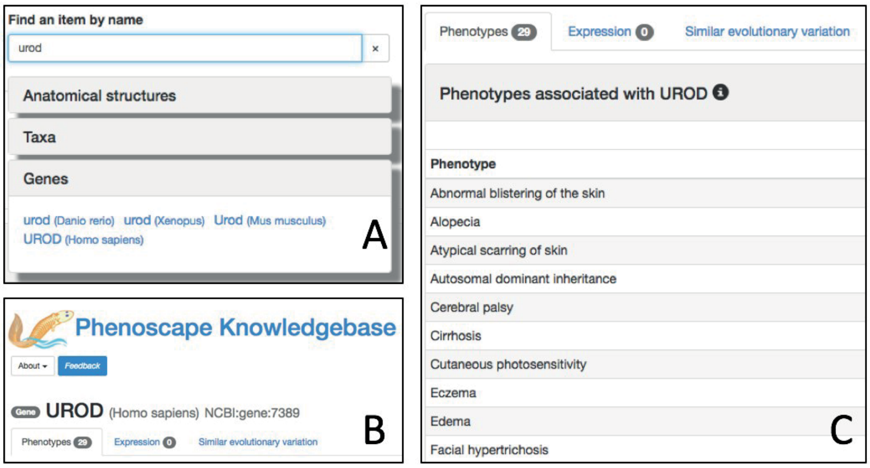
This method of comparing semantic similarity of mutant phenotypes has high potential for semantic prediction, but requires consistent, coherent, and complete phenotype annotations that computationally replicate the underlying biology of organisms, which in turn will require a much larger, more complete dataset. Within the plant kingdom, the Planteome group has begun working toward this goal.

The Planteome project[15,23] provides a suite of interconnected reference and species-specific ontologies associated with a database of plant gene expression and function, traits, phenotypes, QTLs, and germplasm annotations spanning 95 plant taxa. The reference ontologies include the Plant Ontology, Plant Trait Ontology, and Plant Experimental Conditions Ontology, developed by the Planteome project, as well as those developed by collaborating groups, such as the GO, PATO, and ChEBI. An important feature of the Planteome database is an integration of species-specific Crop Ontologies describing traits and phenotype scoring standards being utilized by international plant breeding projects. In the Planteome 2.0 Release (February 2017), the Planteome database includes trait ontologies for eight crop species: maize (*Zea mays*), sweet potato (*Ipomoea batatas*), soybean (*Glycine max*), pigeon pea (*Cajanus cajan*), rice (*Oryza sativa*), cassava (*Manihot esculenta*), lentil (*Lens culinaris*) and wheat (*Triticum aestivum*). Planteome database users can access the ontologies and annotated data from the project website and ontology browser, perform faceted searches for ontology terms, annotations and bioentities, and download custom datasets for further analysis. Other tools offered by the Planteome include web services[24] for ontology terms and annotated data, and the Planteome Noctua platform[25] for collaborative building of gene annotation models using ontology terms.

The current methods of annotating phenotypes are largely manual, which limits high volume data curation. Therefore, semi-automated methods dependent on data mining using reference ontologies and natural language processing methods have started to become available[26,27]. The Planteome project, among others, is working towards the goal of using ontologies and neural network-based methods to identify phenotypes and plant characters (phenotypes and traits) from both high-throughput phenotyping project data and plant taxonomic sample collections.

Within the context of vertebrates, the Phenoscape comparative framework and Knowledgebase ([28,29] and see Chapter 11) is another platform that, if
adapted to include plant data, may enable plant-centric analyses, as

well as cross-domain analyses including ones such as those shown in Figures 1 and 2. The Phenoscape Knowledgebase[30] currently combines computable morphological phenotype descriptions from phylogenetic systematics publications on comparative fish morphology and the vertebrate fin-to-limb transition, on the one hand, and from mutant screens and other genetic perturbation experiments in vertebrate model organisms, on the other. By way of example, Figure 3 shows the results of querying the Phenoscape Knowledgebase with the search term "urod" (a gene associated with human *porphyria cutanea tarda*, with phenotype shown in Figure 2A). Not only are orthologs in zebrafish, *Xenopus*, mouse, and human returned by the search, in each species the phenotypes associated with UROD are listed.



Find an item by name

urod

Anatomical structures

Taxa

Genes

urod (Danio rerio) urod (Xenopus) Urod (Mus musculus) UROD (Homo sapiens)

Gene UROD (Homo sapiens) NCBI:gene:7389

Phenotypes associated with UROD

| Phenotype |
|---|
| Abnormal blistering of the skin |
| Alopecia |
| Atypical scarring of skin |
| Autosomal dominant inheritance |
| Cerebral palsy |
| Cirrhosis |
| Cutaneous photosensitivity |
| Eczema |
| Edema |
| Facial hypertrichosis |

**Figure 3.** Results of a query on the Phenoscape Knowledgebase for UROD. A: The search term "urod" returns genes in zebrafish, *Xenopus*, mouse, and human. B: Clicking on the human UROD, the knowledgebase returns 29 phenotypes. C: A sample of the list of phenotypes associated with human UROD is shown.

Evolutionary phenotype profiles for taxa and clades are linked to the phenotypes of mutated genes in model organisms with the highest semantic similarity. This enables researchers to explore conservation of phenotype in distantly related organisms and leverage knowledge from model organisms to identify candidate genes for related phenotypes in non-model organisms.

An additional goal of the Phenoscape project is the development of natural language processing tools to increase the speed with which existing and new phenotypic data can be rendered computable. Such tools include CharaParser[31] and Phenex[32,33] that enable semi-autonomous encoding of morphological descriptions and facilitate mapping to ontology terms, respectively. Scaling the computable phenotype dataset of the Plant PhenomeNET (6 taxa) to the size of the Phenoscape Knowledgebase (5,211 taxa) through incorporation of existing and future natural language processing tools would drive the prediction of novel candidate genes for crop plants, helping to extract as much information as possible from the wealth of phenotypic data.

## 3. Future work

What would it take to create a PlantPhenoscape? To build such a resource would require phenotype data in the form of computable ontologies that are universal to all land plants, functional information in the form of Gene Ontology (GO) annotations and gene expression data, and evolutionary data in the form of phylogenetic relatedness among genes and species. Combining phenomic with genomic data in a single query-based database would enable researchers to advance understanding of the basic genomic mechanisms underlying plant development and evolution. Connection and integration of these resources with existing and emerging community data-centric projects (e.g., TAIR, MaizeGDB, Genomes to Fields, CyVerse, DivSeek, Planteome etc.) would ensure broad access and longevity for developed resources.

Once assembled, a PlantPhenoscape Knowledgebase could be used to combine existing phenotype data, including both natural variation and genetic mutant phenotypes, for agricultural and model plants. These data could be assembled based on the ontology-building workflow established by Oellrich et al.[14]. To be most useful, a Plant-Phenoscape Knowledgebase should include annotations for genes and gene orthologs - including molecular function, biological processes, and localization of gene products from GO and link these with the phenotype database using mutant EQ phenotypes extracted from PhenomeNet, Planteome, and PATO.

Once combined into a single database, the PlantPhenoscape Knowledgebase would allow researchers to ask questions about the evolutionary and phenotypic relatedness of particular structures, and to develop hypotheses concerning the genetic and developmental mechanisms underlying these particular changes. Such a resource could lead to the development of a fast semantic similarity engine for searching in real time across taxa or genotypes for shared phenotypic profiles. A phenotypic profile could include particular morphological or developmental descriptors or shared aspects of gene expression that result in phenotypic differences. While Phenoscape's existing architecture for semantic similarity tests and trait/character matrix capabilities could be used for comparative phylogenetic analyses, the breadth of Phenoscape's computable data types could also be expanded to include species interactions, developmental data, phenotype-genotype connections, gene network interactions, and genomic data.

By combining plant kingdom-wide ontologies for structural and anatomical characters with linked phenotypic and genotypic data from across plant genetic systems, one could describe plant diversity across many lineages and species and predict which genes and gene expression patterns may be responsible not only for ecologically driven and evolutionarily significant changes in plant form and function, but also for traits of interest in the world's major crops.

As noted earlier, the construction of the existing Plant PhenomeNET similarity matrix created by Oellrich et al.[14] required manual creation of EQ statements from free-text phenotype descriptions sourced from phenotypic databases and literature papers. This conversion from human-readable phenotypic data to their computable representations demanded extensive time and effort from domain experts for each of the plant species included in the project. In a similar fashion, the entity-quality relations of the Phenoscape Knowledgebase are manually generated from phenotypic and morphological data reported in literature and character state matrices, albeit assisted through purpose-built auto-completion tools such as Phenex, mentioned previously. The automation of this process - converting human-readable phenotype descriptions into computable EQ statements - has the potential to reduce the number of human hours spent on this task, allowing curators to focus primarily on ensuring the quality of the EQ representations rather than generating them. In addition, automating this process would expand the total amount of phenotypic data that can be processed within a given time frame, proportionally expanding the scope of the analyses that can be performed.

Information extraction, one of the problems at the core of parsing out EQ statements from phenotype descriptions, is an established problem in the field of natural language processing (NLP). Within the scope of information extraction, one of the most notable challenges is adapting algorithms and techniques for specific biological domains. Whereas a general case NLP algorithm may identify people or places, biological applications require recognition of items such as gene and protein names[34], and more complex ideas such as disease-phenotype relations[35] and descriptions of mutations within genes[36]. These specific information extraction algorithms have been applied and evaluated in the domain of biomedical texts, but differences in taxa notation and vocabulary and the wide variety of phenotypic information available (from anatomical phenotypes, to cellular concentrations, to biological processes, etc.), makes generalization difficult.

CharaParser, the computational tool reported in Cui et al.[31], addresses this problem of information extraction with respect to morphological phenotypes. From an input in a variety of natural language formats, CharaParser produces character-state formulated phenotype descriptions encoded in an XML file format. Words and phrases corresponding to characters and character-states are identified through an unsupervised learning algorithm[37], preventing the need for the creation of domain-specific annotated training data. Instead, a small number of seed characters and character-states are fed to the algorithm and used to identify patterns leading to the identification of more character and character states in the input text itself, in an iterative process. Following the unsupervised learning algorithm, the proposed characters and character states extracted from the input text are verified, altered, or removed by a human reviewer. The widely used NLP parsing tool Stanford Parser[38] is then used in combination with sets of heuristic rules to identify the relationship between characters and character states in the input text and produce the final XML representation of the phenotypes. CharaParser has been shown to perform well on plant data sets, with 90% precision and recall at the sentence level on a North American Flora data set, with slightly lower performance on an invertebrate data set[31].

In the XML representation of phenotypes produced by tools such as CharaParser, characters and character-states are computationally identified and organized, but semantically they are represented with the exact vocabulary with which they were represented in the input text. This prevents comparison between multiple encoded descriptions, as is possible with EQ statements. The task of representing predicted characters and character states in terms of entities and qualities drawn from ontologies is referred to as concept coding, or concept mapping. More generally, concept coding refers to mapping between any word or set of words and a corresponding ontology term. As with other natural language processing problems, tools have been built to address the problem of concept coding in the biomedical text domain. Notably, Aronson et al.[39,40] developed a concept coding algorithm called MetaMap for mapping text to concepts in the UMLS (Unified Medical Language System) Metathesaurus. The MetaMap algorithm accounts for possible variants of the input text (synonyms, abbreviations, etc.), and scores their similarity to available ontology terms through custom metrics, such as how many words were used to find the match, and how central those words are to the meaning of the input text.

Combining the functionality of an information extraction tool like CharaParser, capable of identifying entity and quality-like terms in phenotypic data, with a concept coder capable of mapping those terms to ontological concepts, would allow for near-complete automation of EQ statement generation. Provided that the information extraction algorithms can be developed to perform on a diverse collection of datasets, the variety of EQ statements such a system could generate would only be limited by the availability of ontologies. Anatomical, chemical, and biological process ontologies already provide the basis for representing an extremely wide range of phenotypic information.

## 4. Outlook

Given recent innovations in gene editing and the availability of tools to design specific changes to gene regions of interest (reviewed in[41]), predictive phenomics can be used to target desired phenotypes and test correlations between phenomes and genomes in any species of interest, bringing functional genomics tools to bear on all phenotypes across many species and even domains of life. Together, ontology-based phenotypic prediction, coupled with simplified, broadly accessible gene editing capabilities, will not only advance our understanding of basic biological mechanisms and principles, but has the potential to improve disease models and agricultural innovation.

## 5. Acknowledgements

Thanks to Gerry Neuffer and John Gray for sharing images of the lesion mimic mutant phenotypes in maize. These authors were brought together largely due to the Phenotype Ontology Research Coordination Network supported by National Science Foundation award DBI-0956049. The Planteome Project is supported by National Science Foundation award IOS-1340112. The Phenoscape II project is funded by NSF collaborative grants DBI-1062404 and DBI-1062542 grant, and is supported by the National Evolutionary Synthesis Center (NESCent), NSF #EF-0905606. Its predecessor, the Phenoscape I project, was funded by NSF grant BDI-0641025, and supported by NESCent, NSF #EF-0423641. Contributions by IRB and CJLD were supported by the Predictive Plant Phenomics NSF Research Traineeship (NRT) DGE-1545453 and the Iowa State University Plant Sciences Institute Faculty Scholars program.

## 6. References

[1] Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP, Blackburn DC, Blake JA, Burleigh JG, Chanet B, Cooper LD, Courtot M, Csösz S, Cui H, Dahdul W, Das S, Dececchi TA, Dettai A, Diogo R, Druzinsky RE, Dumontier M, Franz NM, Friedrich F, Gkoutos GV, Haendel M, Harmon LJ, Hayamizu TF, He Y, Hines HM, Ibrahim N, Jackson LM, Jaiswal P, James-Zorn C, Köhler S, Lecointre G, Lapp H, Lawrence CJ, Le Novère N, Lundberg JG, Macklin J, Mast AR, Midford PE, Mikó I, Mungall CJ, Oellrich A, Osumi-Sutherland D, Parkinson H, Ramírez MJ, Richter S, Robinson PN, Ruttenberg A, Schulz KS, Segerdell E, Seltmann KC, Sharkey MJ, Smith AD, Smith B, Specht CD, Squires RB, Thacker RW, Thessen A, Fernandez-Triana J, Vihinen M, Vize PD, Vogt L, Wall CE, Walls RL, Westerfeld M, Wharton RA, Wirkner CS, Woolley JB, Yoder MJ, Zorn AM, Mabee P (2015) Finding our way through phenotypes. PLoS Biol 13(1):e1002033. doi:10.1371/journal.pbio.1002033

[2] Thessen AE, Bunker DE, Buttigieg PL, Cooper LD, Dahdul WM, Domisch S, Franz NM, Jaiswal P, Lawrence-Dill CJ, Midford PE, Mungall CJ, Ramírez MJ, Specht CD, Vogt L, Vos RA,

Walls RL, White JW, Zhang G, Deans AR, Huala E, Lewis SE, Mabee PM. (2015) Emerging semantics to link phenotype and environment. PeerJ 3:e1470. doi:10.7717/peerj.1470

[3] McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte E (2010) Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proc Natl Acad Sci USA 107(14):6544-9. doi:10.1073/pnas.0910200107

[4] Griffiths C, Barker J, Bleiker T, Chalmers R, Creamer D (2016) Rook's Textbook of Dermatology. 9th ed. John Wiley and Sons, Inc. Chichester, UK.

[5] Hu G, Yalpani N, Briggs SP, Johal GS (1988) A porphyrin pathway impairment is responsible for the phenotype of a dominant disease lesion mimic mutant of maize. Plant Cell 10(7):1095-1105.

[6] Johal GS (2007) Disease lesion mimics mutants of maize. Online. APSnet Features. doi:10.1094/APSnetFeatures-2007-0707.

[7] GenBank: Accession No. NP_000365.3. uroporphyrinogen decarboxylase [Homo sapiens]. GenBank. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. `https://www.ncbi.nlm.nih.gov/protein/71051616`

[8] Neuffer MG, Calvert OH (1975) Dominant disease lesion mimics in maize. J Hered 66:265-270.

[9] Neuffer MG, Pawar SE (1980) Dominant disease lesion mutants. Maize Genet Coop News 54:34-37.

[10] Homma N, Takei Y, Tanaka Y, Nakata T, Terada S, Kikkawa M, Noda Y, Hirokawa N (2003) Kinesin superfamily protein 2A (KIF2A) functions in suppression of collateral branch extension. Cell 114(2):229-39.

[11] Lu L, Lee YR, Pan R, Maloof JN, Liu B (2005) An internal motor kinesin is associated with the Golgi apparatus and plays a role in trichome morphogenesis in *Arabidopsis*. Mol Biol Cell 16(2):811-23.

[12] Sarkany RPE (2008) Making sense of the porphyrias. Photodermatol Photoimmunol Photomed 24:102-108.

[13] Hoehndorf R, Schofield PN, Gkoutos GV (2011) PhenomeNET: A whole-phenome approach to disease gene discovery. Nucleic Acids Res 39(18):e119. doi:10.1093/nar/gkr538

[14] Oellrich A, Walls RL, Cannon EK, Cannon SB, Cooper L, Gar-

diner J, Gkoutos GV, Harper L, He M, Hoehndorf R, Jaiswal P, Kalberer SR, Lloyd JP, Meinke D, Menda N, Moore L, Nelson RT, Pujar A, Lawrence CJ, Huala E (2015) An ontology approach to comparative phenomics in plants. Plant Methods 11:10. doi:10.1186/s13007-015-0053-y

[15] Cooper L, Meier A, Laporte MA, Elser JL, Mungall C, Sinn BT, Cavaliere D, Carbon S, Dunn NA, Smith B, Qu B, Preece J, Zhang E, Todorovic S, Gkoutos G, Doonan JH, Stevenson DW, Arnaud E, Jaiswal P (2017) The Planteome database: An integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res 46:D1168-D1180. doi:10.1093/nar/gkx1152

[16] Cooper L, Walls RL, Elser J, Gandolfo MA, Stevenson DW, Smith B, Preece J, Athreya B, Mungall CJ, Rensing S, Hiss M, Lang D, Reski R, Berardini TZ, Li D, Huala E, Schaeffer M, Menda N, Arnaud E, Shrestha R, Yamazaki Y, Jaiswal P (2013) The Plant Ontology as a tool for comparative plant anatomy and genomic analyses. Plant Cell Physiol 54:e1. doi:10.1093/pcp/pcs163

[17] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: Tool for the unification of biology. Nat Genet 25(1):25-29. doi:10.1038/75556

[18] Hastings J, De Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C (2013) The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. Nucleic Acids Res 41(D1):456-463. doi:10.1093/nar/gks1146

[19] Gkoutos GV, Green ECJ, Mallon A, Hancock JM, Davidson D (2005) Using ontologies to describe mouse phenotypes. Genome Biol 6(1):R8. doi:10.1186/gb-2004-6-1-r8

[20] Mungall CJ, Gkoutos GV, Smith CL, Haendel MA, Lewis SE, Ashburner M (2010) Integrating phenotype ontologies across multiple species. Genome Biol 11:R2.

[21] Gkoutos GV, Schofield PN, Hoehndorf R (2017) The anatomy of phenotype ontologies: Principles, properties and applications. Brief Bioinform doi:10.1093/bib/bbx035

[22] http://phenomebrowser.net/plant/

[23] http://www.planteome.org

[24] http://planteome.org/web_services

[25] http://noctua.planteome.org/

[26] Wei CH, Kao HY, Lu Z (2013) PubTator: A web-based text mining tool for assisting biocuration. Nucleic Acids Res 41(Web Server Issue):518-522. doi:10.1093/nar/gkt441

[27] Xu W, Gupta A, Jaiswal P, Taylor C, Lockhart P (2016) Enhancing information accessibility of scientific publications with text mining and ontology. CEUR Workshop Proc 1747:2-3.

[28] http://www.phenoscape.org

[29] Mabee P, Balhoff JP, Dahdul WM, Lapp H, Midford PE, Vision TJ, Westerfield M (2012) 500,000 fish phenotypes: The new informatics landscape for evolutionary and developmental biology of the vertebrate skeleton. J Appl Ichthyol 28:300-305.

[30] http://kb.phenoscape.org

[31] Cui H (2012) CharaParser for fine-grained semantic annotation of organism morphological descriptions. J Assoc Inf Sci Technol 63(4):738-754. doi:10.1002/asi.22618

[32] Balhoff JP, Dahdul WM, Dececchi TA, Lapp H, Mabee P, Vision TJ (2014) Annotation of phenotypic diversity: decoupling data curation and ontology curation using Phenex. J Biomed Semant 5(1):45. doi:10.1186/2041-1480-5-45

[33] Balhoff JP, Dahdul WM, Kothari CR, Lapp H, Lundberg JG, Mabee P, Midford PE, Westerfield M, Vision TJ (2010) Phenex: Ontological annotation of phenotypic diversity. PLoS ONE 5(5):1-10. doi:10.1371/journal.pone.0010500

[34] Settles B (2005) ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 21(14):3191-3192. doi:10.1093/bioinformatics/bti475

[35] Xu R, Li L, Wang Q (2013) Towards building a disease-phenotype knowledge base: Extracting disease-manifestation relationship from literature. Bioinformatics 29(17):2186-2194. doi:10.1093/bioinformatics/btt359

[36] Horn F, Lau AL, Cohen FE (2004) Automated extraction of mutation data from the literature: Application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. Bioinformatics 20(4):557-568. doi:10.1093/bioinformatics/btg449

[37] Cui H, Boufford D, Selden P (2010) Semantic annotation of biosystematics literature without training examples. J Am Soc Inf Sci 61:522-542. doi:10.1002/asi.21246

[38] Klein D, Manning CD (2003) Accurate unlexicalized parsing. In: Matsumoto Y [Ed.] Proceedings of the 41st Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics p. 423-430.

[39] Aronson AR (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp 17. doi:D010001275

[40] Aronson AR, Lang FM (2010) An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 17(3):229-236. doi:10.1136/jamia.2009.002733.

[41] Brazelton VA Jr, Zarecor S, Wright DA, Wang Y, Liu J, Chen K, Yang B, Lawrence-Dill CJ (2015) A quick guide to CRISPR sgRNA design tools. GM Crops Food 6(4):266-76. doi:10.1080/21645698.2015.1137690