

**Computational methods for association studies utilizing free-text and spoken plant
phenotype descriptions**

by

Colleen Frances Yanarella

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology (Predictive Plant Phenomics)

Program of Study Committee:
Carolyn J. Lawrence-Dill, Major Professor
Matthew B. Hufford
Baskar Ganapathysubramanian
Kris De Brabanter
Qi Li

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2023

Copyright © Colleen Frances Yanarella, 2023. All rights reserved.

DEDICATION

This dissertation is dedicated to my parents, Douglas F. and Christine A. Yanarella, who instilled my lifelong love of learning and wonderment; Travis L. Anderson, my darling and rock; and Cash P. Yanarella, for always bringing me joy.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vi
LIST OF FIGURES	vii
ACKNOWLEDGMENTS	viii
ABSTRACT	x
CHAPTER 1. GENERAL INTRODUCTION	1
1.1 Introduction	1
1.2 Research Goals	3
1.3 Dissertation Organization	4
1.4 References	5
CHAPTER 2. COMPUTING ON PHENOTYPIC DESCRIPTIONS FOR CANDIDATE GENE DISCOVERY AND CROP IMPROVEMENT	7
2.1 Abstract	7
2.2 Background	8
2.3 What Do Phenotype Networks Look Like and How Can They Be Used?	10
2.4 What Seems Unexpected (to Us) about the Use of Automated Methods for Com- puting on Phenotypic Descriptions?	11
2.5 Conflicts of Interest	12
2.6 Authors' Contributions	12
2.7 Acknowledgments	12
2.8 References	12
2.9 Figures and Tables	14
2.10 Appendix: Consent To Include Co-Authored Article in Dissertation	16
CHAPTER 3. THE CASE FOR RETAINING NATURAL LANGUAGE DESCRIPTIONS OF PHENOTYPES IN PLANT DATABASES AND A WEB APPLICATION AS PROOF OF CONCEPT	17
3.1 Abstract	17
3.1.1 Database URLs	18
3.2 Keywords	18
3.3 Introduction	18
3.4 Materials and Methods	21
3.4.1 Datasets	21
3.4.2 Measure of gene pair similarity	23

3.4.3	Formulating Biologically Relevant Questions	28
3.5	Results	29
3.5.1	Text-based approaches recover biological relationships	29
3.5.2	Enabling biologists to use these methods and dataset	33
3.6	Discussion	35
3.7	Data and Code Availability	37
3.8	Author Contribution	38
3.9	Acknowledgements	38
3.10	Funding	39
3.11	References	39
3.12	Figures and Tables	46
3.13	Appendix: Supplementary Tables	55
3.14	Appendix: Consent To Include Co-Authored Article in Dissertation	58
CHAPTER 4.	WISCONSIN DIVERSITY PANEL PHENOTYPES: SPOKEN DESCRIPTIONS OF PLANTS AND SUPPORTING DATA	59
4.1	Abstract	59
4.1.1	Objectives	59
4.1.2	Data description	59
4.2	Keywords	60
4.3	Objective	60
4.4	Data description	61
4.5	Limitations	63
4.6	Abbreviations	63
4.7	Availability of data and materials	63
4.8	Declarations	63
4.8.1	Ethics approval and consent to participate	63
4.8.2	Consent for publication	64
4.8.3	Competing interests	64
4.9	Funding	64
4.10	Authors' contributions	65
4.11	Acknowledgements	65
4.12	References	66
4.13	Figures and Tables	67
4.14	Appendix: Institutional Review Board Exemption Letter	70
CHAPTER 5.	GWAS FROM SPOKEN PHENOTYPIC DESCRIPTIONS: A PROOF OF CONCEPT FROM MAIZE FIELD STUDIES	72
5.1	Abstract	72
5.2	Keywords	73
5.3	Introduction	73
5.4	Materials and Methods	74
5.4.1	Spoken Phenotype Collection Summary	75
5.4.2	Phenotype Detection and Descriptions	75
5.4.3	Preprocessing Genotypic Dataset	76
5.4.4	Preprocessing Phenotypic Datasets	76

5.4.5	Genome-Wide Association Studies	78
5.4.6	Genome-Wide Association Study Analyses	78
5.5	Results and Discussion	79
5.5.1	Detecting Phenotypes from Spoken Descriptions	79
5.5.2	Extracting Phenotype Data for Plant Height from Spoken Descriptions	80
5.5.3	Association Studies using Phenotypes Derived from Speech	81
5.5.4	Investigating GO Terms from GWAS Results	83
5.6	Conclusion	84
5.7	Web Resources	85
5.8	Data Availability	86
5.9	Acknowledgments	86
5.10	Funding	87
5.11	Conflicts of Interest	87
5.12	Author Contributions	87
5.13	References	87
5.14	Figures and Tables	93
5.15	Appendix: Institutional Review Board Exemption Letter	103
CHAPTER 6. GENERAL CONCLUSION		105
6.1	Summary	105
6.2	Collaborative Project Outcomes	106
6.2.1	Predictive Plant Phenomics (P3) Research Trainee-ship Symposium Sessions	106
6.2.2	Gene Ontology Annotations for Plant Species	106
6.3	Future Work	107
6.4	References	108
6.5	Appendix: Manuscripts and Datasets From Collaborate Projects	114
6.5.1	Manuscripts	114
6.5.2	Datasets	114

LIST OF TABLES

	Page
Table 3.1	Scope and scale of the complete dataset. 52
Table 3.2	Biological relationships tested in each task. 53
Table 3.3	Number of genes and gene pairs used for each task. 53
Table 3.4	Similarities among datasets across biological tasks. 53
Table 3.5	Comparing F_1 scores and group significance rates for phenotype and pathway relationships. 54
Table 3.6	Comparing F_1 scores for associations and orthologous gene pair relationships. 55
Table 3.7	Comparing F_1 scores for pathways for intraspecies and interspecies gene pairs. 57
Table 4.1	Overview of data files/data sets. 67
Table 5.1	Proportion of word usage for describing positive control accessions. 99
Table 5.2	Observation retention by spoken phenotype method. 100
Table 5.3	Plant height gene models count identified from publications. 101
Table 5.4	Appearance of GO terms by plant hormone for each method. 102
Table 6.1	Contributions to manuscripts out of the scope of the described research aims. 114
Table 6.2	Contributions to datasets out of the scope of the described research aims. . 114

LIST OF FIGURES

		Page
Figure 2.1	Phenotypic similarity.	14
Figure 3.1	Phenotype description text length distributions.	46
Figure 3.2	Overlap among vocabularies.	47
Figure 3.3	Comparing the groups of gene pair similarity approaches.	47
Figure 3.4	Cohesiveness of phenotype and pathway gene groups.	48
Figure 3.5	Querying plant genes, annotations, and phenotype descriptions.	50
Figure 3.6	Querying with genes in QuOATS.	51
Figure 5.1	Comparison of intersections of WiDiv dataset taxa.	93
Figure 5.2	Spoken phenotype process overview.	94
Figure 5.3	Distributions for each student participants word count per observation.	94
Figure 5.4	Example of detecting phenotypes from positive control accession descriptions.	95
Figure 5.5	Manhattan plot of measured height phenotypic data.	96
Figure 5.6	Manhattan plot of tall query semantic similarity phenotypic data.	97
Figure 5.7	Manhattan plot of binned plant height phrases phenotypic data.	98

ACKNOWLEDGMENTS

The research described here required the input of many individuals, and I am humbled and grateful for their assistance and encouragement. Firstly, I thank my major professor, Dr. Carolyn Lawrence-Dill, for encouraging, supporting, believing in, and challenging me. Thank you for championing my research project. Thank you for assembling such a fun and diverse research group that has been a pleasure to work with and has had memorable adventures outside the lab.

I thank my committee members. I thank Dr. Matthew Hufford for suggesting classes that immeasurably helped me form the foundation for my research. I thank Dr. Baskar Ganapathysubramanian for asking difficult questions that challenge my thinking. I thank Dr. Kris De Brabanter for the helpful discussions about statistical analysis and for steering me to new methods. I thank Dr. Qi Li for guiding me through Natural Language Processing techniques that shaped my methods and for always being easily available; I cannot understate how much I appreciate your assistance. I thank Dr. Ramona Walls for serving on my committee as an arbiter honorarius medium committee member, for your input into my project, and for being welcoming during my time in Arizona.

I thank the former and current lab members for their insight, willingness to share their expertise, and uplifting attitudes. I thank Darwin Campbell for his willingness to answer questions, even when limits were applied. I thank Scott Zarecor, Kokulapalan Wimalanathan, Mingze He, Ian Braun, Carla Mann, Sweta Roy-Carson, Vincent Antonio Brazelton, and Jyothi Prasanth Durairaj Rajeswari. I would like to recognize Leila Fattel for being an incredible lab mate, scientist, cheerleader, and friend. Leila, I cannot understate how much I appreciate you and how much of a pleasure it has been working with and learning from you throughout graduate school.

I thank the amazing, patient, and helpful folks at the Iowa State University High-Performance Computing Facility. In particular, Marina Kraeva and Yasyasvy "Yash" Nanyam for responding to my numerous calls for help and many workarounds that prevented my research from stalling multiple times.

I appreciate everyone who has given me feedback on my research and writing. I am forever grateful.

ABSTRACT

Recording observations for plant traits is a time-consuming and costly process called phenotyping. These observations of the phenotypes of plant traits are valuable for improving crops by identifying genetic regions of interest identified through association studies. Genome-Wide Association Studies (GWAS) require genotypic marker information across all of the chromosomes for members of a population and phenotypic data that characterizes a trait. Improvements in genotyping technologies and the encouragement of data sharing have made genotypic datasets for maize populations accessible. Concurrently, high-throughput phenotyping methods are being developed to improve data collection and involve expensive automated machinery and sensors to collect measurement and scoring data. Natural language descriptions of plants contain a wealth of underutilized phenotypic information. Recent research efforts use structured descriptions of data that apply ontologies, which are structured data that represent relatedness to other terms and ease the computational burden of determining semantic similarity or word meaning similarity. Comparative analysis of gene interaction demonstrated the utility of structured language descriptions of plant phenotypes. Methods to automate the development of structured descriptions of plants indicate that humans who curate these data may use new computational methods to generate such terms. Additionally, computational pre-trained natural language models enable computing on free-text (i.e., unstructured data) to categorize and predict gene interactions. New methods for generating and analyzing plant descriptions are pertinent because of the success of extracting biologically meaningful information from free-text descriptions of plant phenotypes. This work culminates in developing processes to collect, process, and analyze spoken language descriptions of plants recorded in a field environment. Descriptions of the accessions in the Wisconsin Diversity panel are used to perform GWAS to identify regions of interest of the genome associated with the plant height trait.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Introduction

Phenotyping plants is a process in which the traits of plants are observed. These observations are a valuable source of information that can be obtained through scoring (e.g. scoring disease prevalence (Menkir and Ayodele (2005) and Beyene et al. (2017))), measuring with digital or ruled apparatus' (e.g. plant height or leaf width (Tilly et al. (2014) and Pearce et al. (1975))), or even using machinery that detect features that cannot be seen with the human eye (e.g. infrared sensors (Spielbauer et al. (2009) and Masuka et al. (2012))). Plant breeders and geneticists have traditionally relied on selecting plants to breed based on these physical features, and in the past thirty years, vast improvements in genotyping have enabled and dominated the acceleration of breeding and genetics (summarized in Durmaz et al. (2015)). These physical features of plants are a result of the elaborate interaction of genetics and the environment in which the plants grow (reviewed in Rutherford (2000)). Plant Biologists and Agronomists historically recorded descriptions of plant features along with their measuring and scoring data. An example includes the vivid descriptions reported by Gregor Mendel while studying pea seeds, pods, and flowers (Mendel (1865)). A great deal of research effort has been placed on making the process of phenotyping more automated and less labor intensive (reviewed in Gage et al. (2019)). These methods generally result in numeric measures or scores that make performing association studies, which identify regions of the genome that are of interest for the traits, plausible.

Efforts to investigate the biological meaning of text-based phenotypic descriptions of plants have effectively been investigated using word meaning similarity, known as semantic similarity methods. Research that computes on and compares structured text descriptions of plants (Oellrich et al. (2015)) and Braun and Lawrence-Dill (2020) use ontologies (i.e., Gene Ontology (GO; Ashburner et al. (2000)) and Plant Ontology (PO; Cooper et al. (2012))). Ontologies are

directed acyclic graphs (DAGs) where nodes contain descriptions of gene functions, and child nodes have more specific descriptions than parental nodes. Additional structures that ease computing on text descriptions are Entity-Quality (EQ) statements (reviewed in [Mungall et al. \(2010\)](#)), an example being the "whole plant" entity, where a quality of the "whole plant" entity is "dwarf-like". While computing on structured language is intrinsically less computationally intensive, the process of structuring data involves a large curation effort that is human labor intensive.

Research efforts have demonstrated that the formation of EQ statements can be automated, which may reduce the need for human curation ([Braun and Lawrence-Dill \(2020\)](#)). Additionally, the advent of pre-trained models (Bidirectional Encoder Representations from Transformers (BERT; [Devlin et al. \(2018\)](#)), BioBERT ([Lee et al. \(2019\)](#)), Wikipedia ([Lau and Baldwin \(2016\)](#)), etc.) enables using free-text descriptions of plants for tasks such as categorizing genes into pairs using semantic similarity ([Braun et al. \(2021\)](#)). Using free-text descriptions of plant phenotypes is advantageous to structured data because the researchers are not confined to a syntax or predefined structure in which they must make their observations. Despite advancements in natural language processing, few researchers have protocols for collecting whole plant descriptions in the field environment, which limits data availability for developing methods to apply these spoken descriptions to biological hypotheses.

Of the few researchers who collect descriptive data in the field, spoken descriptions of plant phenotypes are collected ([Kazic \(2020\)](#)). The collection of spoken data bottleneck has been a need for methods to process these data and to enhance investigating biological insights from these data. As tools for speech-to-text become more accurate, accessible, and inexpensive, collecting spoken data is becoming viable. Additionally, these data can be processed and used, in the same manner as measured or scored data, as input for Genome-Wide Association Study (GWAS) pipelines, such as Genome Association and Prediction Integrated Tool (GAPIT; [Lipka et al. \(2012\)](#)), to identify regions implicated for traits of interest. The work described in this

dissertation includes the development of datasets and pipelines to perform association studies on natural language descriptions of plant phenotypes.

1.2 Research Goals

The essence of the research described in this dissertation is to enable researchers to collect natural language descriptions of plants in a manner that is not mechanical. Structuring language for the ease of silicon burdens data collectors, and our objective is to develop tools that allow researchers to use natural language in a humanistic way.

The first goal of this research was to contribute to developing tools to predict biological relationships from descriptive text datasets. This work included catalyzing the reproducibility of a pipeline that processes text descriptions of plant phenotypes to predict shared phenotypes. The biological applicability of relationships of gene pairs generated from various computational approaches was compared to those determined from a benchmark curation-based approach.

The second goal of this research was to develop datasets to investigate whether biologically relevant results can be extracted from spoken descriptions of plants collected in a field setting. The robust dataset we designed, collected, and disseminated includes audio processing procedures and methods, camera images, drone images and footage, field layout details, human prompting procedures, measuring and scoring data, weather data, and a set of sample of audio collected by volunteers. Numeric scoring and measuring techniques accompanied spoken data as ground truth for subsequent genome-wide association studies.

The third goal of this research was to demonstrate the use of spoken descriptions of plants for association studies. These recordings of descriptions were processed using computational tools, and association studies were performed using Genome-Wide Association Study (GWAS) tools. Principally, we used spoken phenotype descriptions to recover known genes of interest for the plant height trait and discover new candidate genes of interest for plant height.

1.3 Dissertation Organization

This dissertation contains chapters including this chapter (Chapter 1), which contains the general introduction followed by publications (Chapters 2-5) and a general conclusion (Chapter 6). Chapter 2, a perspective published in *Plant Phenomics*, addresses investigating biological questions and applications of natural language descriptions of plant phenotypes. Chapter 3, a manuscript submitted to *Database*, describes investigating and comparing biological relationships that can be generated using various computational approaches on text descriptions of plant phenotypes. Chapter 4, a data note under review in *BMC Data Notes*, reports and describes the dataset we developed to investigate field-generated spoken phenotype descriptions of plants for association studies. Chapter 5, a manuscript to be submitted to *G3: Genes|Genomes|Genetics*, details using spoken descriptions of plants to perform association studies to recover known and new genes of interest for plant traits. Chapter 6 includes a general conclusion summarizing the work described in this dissertation, future directions, and a summary of projects that enhance my doctoral training but remain outside the breadth of my direct research.

My contributions to these works include, for Chapter 2, I conceptualized methods for using natural language descriptions of plants, and I wrote and edited the perspective article. For Chapter 3, I reviewed and edited the manuscript draft and accompanying code and directed the reproducibility review of the scripts to perform the analysis described in the manuscript. For Chapter 4, I assisted in obtaining Institutional Review Board (IRB) Exempt Status, designed and planted an experimental field, designed and organized in-filed data collection, managed student participants, wrangled data, obtained a Digital Object Identifier (DOI) for the data, and wrote the first draft of the data note. For Chapter 5, I performed preprocessing, performed Genome-Wide Association Studies (GWAS), performed the data analysis, submitted to obtain a DOI for the analysis data, and wrote the first draft of the manuscript.

1.4 References

- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.
- Beyene, Y., Gowda, M., Suresh, L. M., Mugo, S., Olsen, M., Oikeh, S. O., Juma, C., Tarekegne, A., and Prasanna, B. M. (2017). Genetic analysis of tropical maize inbred lines for resistance to maize lethal necrosis disease. *Euphytica*, 213(9).
- Braun, I. R. and Lawrence-Dill, C. J. (2020). Automated Methods Enable Direct Computation on Phenotypic Descriptions for Novel Candidate Gene Prediction. *Frontiers in Plant Science*, 10.
- Braun, I. R., Yanarella, C. F., Rajeswari, J. P. D., Bassham, D. C., and Lawrence-Dill, C. J. (2021). The Case for Retaining Natural Language Descriptions of Phenotypes in Plant Databases and a Web Application as Proof of Concept. *bioRxiv*.
- Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T. Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y., and Jaiswal, P. (2012). The Plant Ontology as a Tool for Comparative Plant Anatomy and Genomic Analyses. *Plant and Cell Physiology*, 54(2):e1–e1.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.
- Durmaz, A. A., Karaca, E., Demkow, U., Toruner, G., Schoumans, J., and Cogulu, O. (2015). Evolution of Genetic Techniques: Past, Present, and Beyond. *BioMed Research International*, 2015:1–7.
- Gage, J. L., Richards, E., Lepak, N., Kaczmar, N., Soman, C., Chowdhary, G., Gore, M. A., and Buckler, E. S. (2019). In-field whole-plant maize architecture characterized by subcanopy rovers and latent space phenotyping. *The Plant Phenome Journal*, 2(1):1–11.
- Kazic, T. (2020). Chloe: Flexible, efficient data provenance and management. *bioRxiv*.
- Lau, J. H. and Baldwin, T. (2016). An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. *arXiv preprint arXiv:1607.05368*.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

- Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., and Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, 28(18):2397–2399.
- Masuka, B., Araus, J. L., Das, B., Sonder, K., and Cairns, J. E. (2012). Phenotyping for abiotic stress tolerance in MaizeF. *Journal of Integrative Plant Biology*, 54(4):238–249.
- Mendel, G. (1865). Experiments in plant hybridization. *Verhandlungen des naturforschenden Vereins Brünn*. Available online: 1996, *Electronic Scholarly Publishing Project*, <http://old.esp.org/foundations/genetics/classical/gm-65-a.pdf>.
- Menkir, A. and Ayodele, M. (2005). Genetic analysis of resistance to gray leaf spot of midaltitude maize inbred lines. *Crop Science*, 45(1):163–170.
- Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2.
- Oellrich, A., Walls, R. L., Cannon, E. K., Cannon, S. B., Cooper, L., Gardiner, J., Gkoutos, G. V., Harper, L., He, M., Hoehndorf, R., Jaiswal, P., Kalberer, S. R., Lloyd, J. P., Meinke, D., Menda, N., Moore, L., Nelson, R. T., Pujar, A., Lawrence, C. J., and Huala, E. (2015). An ontology approach to comparative phenomics in plants. *Plant Methods*, 11(1).
- Pearce, R. B., Mock, J. J., and Bailey, T. B. (1975). Rapid method for estimating leaf area per plant in maize. *Crop Science*, 15(5):691–694.
- Rutherford, S. L. (2000). From genotype to phenotype: buffering mechanisms and the storage of genetic information. *BioEssays*, 22(12):1095–1105.
- Spielbauer, G., Armstrong, P., Baier, J. W., Allen, W. B., Richardson, K., Shen, B., and Settles, A. M. (2009). High-throughput near-infrared reflectance spectroscopy for predicting quantitative and qualitative composition phenotypes of individual maize kernels. *Cereal Chemistry Journal*, 86(5):556–564.
- Tilly, N., Hoffmeister, D., Schiedung, H., Hütt, C., Brands, J., and Bareth, G. (2014). Terrestrial laser scanning for plant height measurement and biomass estimation of maize. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-7:181–187.

CHAPTER 2. COMPUTING ON PHENOTYPIC DESCRIPTIONS FOR CANDIDATE GENE DISCOVERY AND CROP IMPROVEMENT

Ian R. Braun^{1,2*}, Colleen F. Yanarella^{1,2*}, and Carolyn J. Lawrence-Dill^{1,2,3}

¹Interdepartmental Bioinformatics and Computational Biology, Iowa State University, Ames, IA
50011, USA

²Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011,
USA

³Department of Agronomy, Iowa State University, Ames, IA 50011, USA

*Joint first authors

Modified from a manuscript published in *Plant Phenomics*

2.1 Abstract

Many newly observed phenotypes are first described, then experimentally manipulated. These language-based descriptions appear in both the literature and in community datastores. To standardize phenotypic descriptions and enable simple data aggregation and analysis, controlled vocabularies and specific data architectures have been developed. Such simplified descriptions have several advantages over natural language: they can be rigorously defined for a particular context or problem, they can be assigned and interpreted programmatically, and they can be organized in a way that allows for semantic reasoning (inference of implicit facts). Because researchers generally report phenotypes in the literature using natural language, curators have been translating phenotypic descriptions into controlled vocabularies for decades to make the information computable. Unfortunately, this methodology is highly dependent on human curation, which does not scale to the scope of all publications available across all of plant biology. Simultaneously, researchers in other domains have been working to enable computation on

natural language. This has resulted in new, automated methods for computing on language that are now available, with early analyses showing great promise. Natural language processing (NLP) coupled with machine learning (ML) allows for the use of unstructured language for direct analysis of phenotypic descriptions. Indeed, we have found that these automated methods can be used to create data structures that perform as well or better than those generated by human curators on tasks such as predicting gene function and biochemical pathway membership. Here, we describe current and ongoing efforts to provide tools for the plant phenomics community to explore novel predictions that can be generated using these techniques. We also describe how these methods could be used along with mobile speech-to-text tools to collect and analyze in-field spoken phenotypic descriptions for association genetics and breeding applications.

2.2 Background

The volume of data related to phenotyping of plants is enormous and growing consistently. While sensor-based high-throughput technologies (described elsewhere in this issue) are responsible for much of this growth in phenotype data, text-based phenotype descriptions also contribute significantly. The scientific literature serves as the primary source of phenotype descriptions, where an example might look something like “maize line X with specific mutation Y exhibits delayed flowering under stress condition Z .” Some phenotype descriptions find their way into model organism databases (e.g., TAIR, MaizeGDB, and SGN) through dedicated curation efforts ([Berardini et al. \(2015\)](#), [Portwood et al. \(2018\)](#), [Fernandez-Pozo et al. \(2014\)](#)).

Given the volume of phenotype descriptions available and the relevance of these descriptions to biological problems generally, interest in finding ways to compute on phenotypic descriptions is quite high. The most common method for making phenotypic descriptions computable involves representing the data using terms from large but finite and highly structured vocabularies such as the gene ontology (GO; [Ashburner et al. \(2000\)](#)), the plant ontology (PO; [Cooper et al. \(2012\)](#)), or the plant trait ontology (TO; [Cooper et al. \(2017\)](#)), among others (reviewed in [Braun and Lawrence-Dill \(2020\)](#)). The utility of using such vocabularies has been immense across the life

sciences generally, with over 27,000 citations to the first GO publication alone (see [Ashburner et al. \(2000\)](#)). Use of these controlled vocabularies allows for increased consistency in how phenotypes are described, and the architecture of these data structures makes querying over a large volume of phenotypes realistic. Their hierarchical nature also enhances the meaning of each phenotype collected as a data point by inheriting implicit knowledge. For example, the GO hierarchy (Figure 2.1(a)) specifies that fruit ripening is a type of aging, so the association of a phenotype related to fruit ripening with this term allows that phenotype to be recovered by a query for aging, without that association being explicitly stated.

Despite the computational and inferential advantages that this type of annotation confers, detailed manual curation comes at the cost of the time and effort required to construct high-quality annotations for the large number of phenotypes observed, and the simplification of phenotypic descriptions to match the architecture of a particular knowledge representation necessarily reduces the specificity of a phenotypic description, thus losing some shades of meaning that are conveyed using natural language directly. How can these shortcomings be addressed? There are several applications for which unannotated natural language is becoming directly computable, a fact which has been largely underexploited in the biological disciplines.

The field of natural language processing (NLP) has made great advancement in recent years. NLP methods are used to compute on language directly to gain insights from semantic (meaning-based) and syntactic (structural) patterns. In the field of human health, applications of NLP with machine learning (ML) have been used to discover hidden patterns which can aid in informing patient care decisions. Such applications include text mining of medical records to predict probabilities of disease, machine translation of physician notes, and automated identification of articles relevant to disease phenotypes, to name just a few (reviewed in [Ohno-Machado \(2011\)](#)). These types of text analyses typically involve representing natural language using numerical vectors, which can then be used as inputs for ML models or to derive similarity scores (Figure 2.1(b)).

In a recent publication, we used NLP and ML to encode descriptions of plant phenotypes and measured pairwise similarity to construct similarity networks (Braun and Lawrence-Dill (2020)). These computationally generated networks were shown to recover underlying gene functions and to predict membership in biochemical pathways, even on datasets distributed across multiple species. Most importantly, these computationally generated networks outperformed networks constructed using high-quality, ontology-based manual annotations in many cases, demonstrating that for these types of predictive tasks involving large datasets, applying computational methods over natural language descriptions yields comparable results to what can be achieved using a slower, labor-intensive, manual curation-based approach. Although high-quality curation plays an invaluable role in organizing phenotypic data, our findings suggest that there is much to be gained by applying purely computational approaches to phenotypic descriptions in plants.

2.3 What Do Phenotype Networks Look Like and How Can They Be Used?

Figures 2.1(c) and 2.1(d) illustrate what two types of similarity networks inferred from natural language descriptions of phenotypes look like. The first is useful for novel candidate gene prediction, and the second could become useful for genome-wide association studies (GWAS) through specification of a concept we call “synthetic traits” where clustered phenotypes are treated as a single trait.

For the novel candidate gene prediction application (Figure 2.1(c)), each node in the network refers to a particular gene and its corresponding phenotype. The similarity between two nodes implies an increased probability that the pair of genes is involved in a common regulatory network, biochemical pathway, or similar shared process. For example, two genes associated with phenotype descriptions that mention leaf size and shape are predicted to be involved in the same pathway or process. This sort of data structure enables researchers to generate new hypotheses about which genes may be involved in processes that generate a given phenotype.

For gene discovery, computationally generated phenotype similarity networks would be generated with no associations to genes asserted within the network (Figure 2.1(d)). In such a

network, highly related phenotypes would create clusters, which we are defining as “synthetic traits.” Sequence data from plants with and without each synthetic trait could then be analyzed with well-understood GWAS approaches (Visser et al. (2017)) to correlate specific genetic loci with the synthetic traits. This methodology could lead to the discovery of genes related to some phenotype properties that a researcher was not specifically looking to discover but that may be well represented in a specific growing environment by the germplasm under observation. For example, the graph may contain a cluster with words or phrases related to aerial root mucilage (Figure 2.1(d)) enabling this property to be used as a trait in downstream analyses like GWAS, even if this phenotype was not previously well understood (Van Deynze et al. (2018)). For collecting these data in a field environment, we envision phenotypic descriptions of plants being spoken and recorded, translated to text, then parsed computationally into specific statements. As such, this methodology is applicable to qualitative descriptions, rather than continuous numerical measurements. From there, the networks are created, highly interconnected clusters are identified as synthetic traits, and those traits are associated with genomic variants.

2.4 What Seems Unexpected (to Us) about the Use of Automated Methods for Computing on Phenotypic Descriptions?

The diversity of phenotype descriptions is beneficial to (rather than a hindrance to) this method of computing on the data. It is not necessary to standardize the words used to describe phenotypes for computational analysis, and the diversity of descriptions actually improves the quality of the result if enough phenotypic observations are recorded. By using data-driven approaches to specify synthetic traits, the concept of a trait becomes objective. This objectivity in grouping observations means that scientists may discover phenotype and trait groups that have not yet been conceived of and described previously. We are at the beginning of a new era for computing on phenotypic descriptions. In the past, researchers had to create simplified and structured descriptions to make phenotypes computable. Put another way, researchers were asked to think and behave like computers. Now, computational methods can accommodate the rich

language that experts use to describe phenotypes. With NLP and ML, computers are able to reason like humans.

2.5 Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the contents of this manuscript or its publication.

2.6 Authors' Contributions

IRB, CFY, and CJLD contributed to the conception of the ideas presented here and to the writing and revision of the manuscript. Ian R. Braun and Colleen F. Yanarella contributed equally to this work.

2.7 Acknowledgments

We thank the Iowa State University Plant Science Institute Faculty Scholars and the Iowa State University Predictive Plant Phenomics graduate students for helpful discussions on the topics discussed in this manuscript. The authors were supported by an Iowa State University Presidential Interdisciplinary Research Seed Grant (CJLD), an Iowa State University Plant Sciences Institute Faculty Scholar Award (CJLD), the Predictive Plant Phenomics NSF Research Traineeship (DGE-1545453; CJLD is a coprincipal investigator; IRB and CFY are trainees), and a seed grant from National Science Foundation (OAC-1636865) that partially supported some work by IRB.

2.8 References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

- Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2015). The Arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. *Genesis*, 53(8):474–485.
- Braun, I. R. and Lawrence-Dill, C. J. (2020). Automated Methods Enable Direct Computation on Phenotypic Descriptions for Novel Candidate Gene Prediction. *Frontiers in Plant Science*, 10.
- Cooper, L., Meier, A., Laporte, M.-A., Elser, J. L., Mungall, C., Sinn, B. T., Cavaliere, D., Carbon, S., Dunn, N. A., Smith, B., Qu, B., Preece, J., Zhang, E., Todorovic, S., Gkoutos, G., Doonan, J. H., Stevenson, D. W., Arnaud, E., and Jaiswal, P. (2017). The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research*, 46(D1):D1168–D1180.
- Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T. Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y., and Jaiswal, P. (2012). The Plant Ontology as a Tool for Comparative Plant Anatomy and Genomic Analyses. *Plant and Cell Physiology*, 54(2):e1–e1.
- Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., Yan, A., and Mueller, L. A. (2014). The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. *Nucleic Acids Research*, 43(D1):D1036–D1041.
- Ohno-Machado, L. (2011). Realizing the full potential of electronic health records: the role of natural language processing. *Journal of the American Medical Informatics Association*, 18(5):539–539.
- Portwood, J. L., Woodhouse, M. R., Cannon, E. K., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., Walsh, J. R., Sen, T. Z., Cho, K. T., Schott, D. A., Braun, B. L., Dietze, M., Dunfee, B., Elsik, C. G., Manchanda, N., Coe, E., Sachs, M., Stinard, P., Tolbert, J., Zimmerman, S., and Andorf, C. M. (2018). MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Research*, 47(D1):D1146–D1154.
- Van Deynze, A., Zamora, P., Delaux, P.-M., Heitmann, C., Jayaraman, D., Rajasekar, S., Graham, D., Maeda, J., Gibson, D., Schwartz, K. D., Berry, A. M., Bhatnagar, S., Jospin, G., Darling, A., Jeannotte, R., Lopez, J., Weimer, B. C., Eisen, J. A., Shapiro, H.-Y., Ané, J.-M., and Bennett, A. B. (2018). Nitrogen fixation in a landrace of maize is supported by a mucilage-associated diazotrophic microbiota. *PLOS Biology*, 16(8):e2006352.
- Visser, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1):5–22.

2.9 Figures and Tables

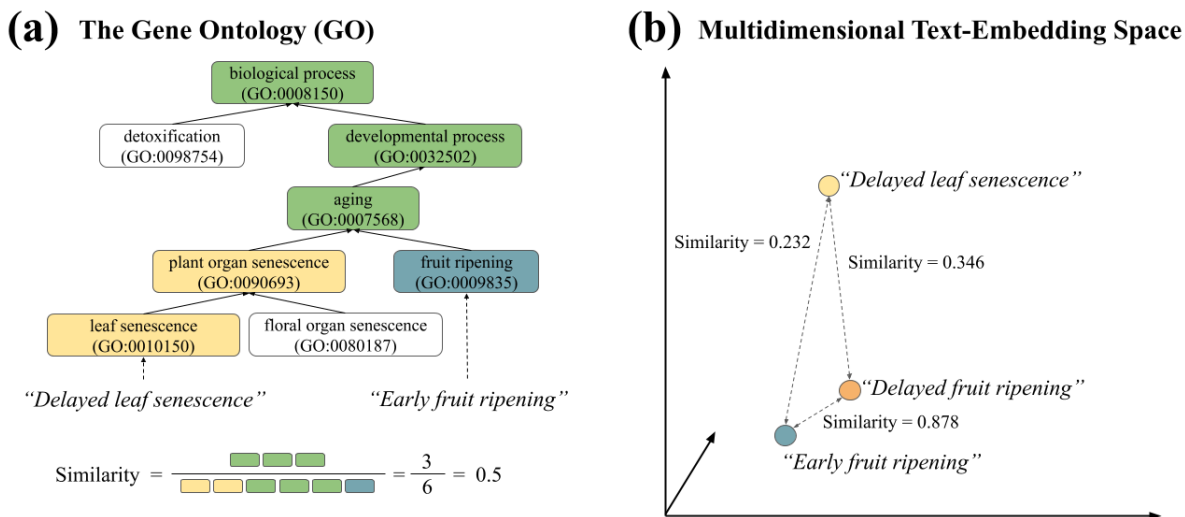
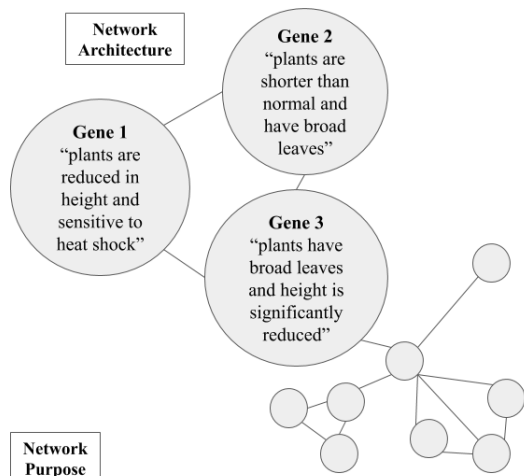
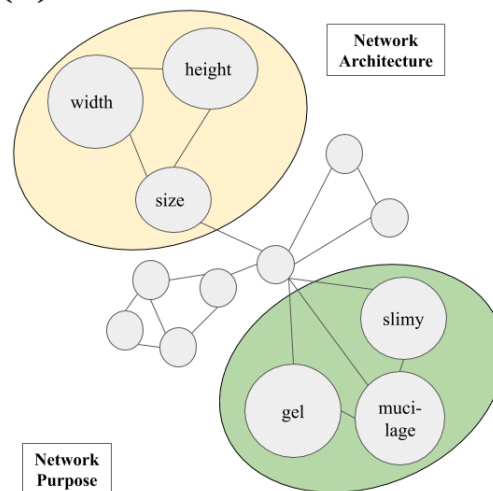


Figure 2.1 Phenotypic similarity. (a) For the GO, the similarity between two concepts can be evaluated based on the relationship between the sets of terms from the ontology that represent those concepts. This relationship can be quantified using metrics such as Jaccard similarity (shown). (b) Natural language processing technique such as sentence embedding using machine learning models or presence and absence of individual words can be used to produce high-dimensional vector representations of concepts, where their position within the vector space allows for quantification of similarity. The example shown plots concepts within three dimensions.

(c) Functional Phenomics

Genes that are associated with phenotypes that have similar text (connected by an edge) are predicted to be biologically associated, such as being involved in a shared pathway. We can use these networks to predict associations between genes and pathways or functional groups.

(d) Gene and Trait Discovery

Words or phrases from a vocabulary of phenotypes descriptions that cluster in the network can define *synthetic traits*, which may be broader or different than a predefined phenotype. We can use these networks to discover represented traits and/or define traits that are mapped to sets of plants in order to perform GWAS to look for genes associated with these synthetic traits.

Figure 2.1 Phenotypic similarity *continued*. (c) Example phenotypic similarity network where nodes represent genes and any associated phenotypic text descriptions. (d) Example phenotypic similarity networks where nodes represent words or phrases drawn from a set of descriptions about some population of plants.

2.10 Appendix: Consent To Include Co-Authored Article in Dissertation

Consent to include co-authored article in dissertation.**THE PARTIES**

Student Author (Full Name, Major, and Institution)	Colleen Frances Yanarella, Bioinformatics and Computational Biology, Iowa State University
List other student co-authors and their institutions.	Ian Robert Braun, Bioinformatics and Computational Biology, Iowa State University
Title(s) of the co- authored section (Chapter, etc.)	Computing on Phenotypic Descriptions for Candidate Gene Discovery and Crop Improvement
Journal Name, Book Title, etc. (if applicable)	Plant Phenomics

DISTRIBUTION OF TASKS AND RESPONSIBILITIES

In this research publication, I, Colleen F. Yanarella, was responsible for the following roles:
(Select all roles that apply.)

- Conceptualization
- Data curation
- Formal analysis
- Funding acquisition
- Investigation
- Methodology
- Resources
- Software
- Supervision
- Validation
- Visualization
- Writing – original draft
- Writing – review & editing
- Other: Please describe briefly: [Click or tap here to enter text.](#)

The CRediT taxonomy is taken from <https://credit.niso.org/>. Go to the link to see the descriptions of contributor roles.

CHAPTER 3. THE CASE FOR RETAINING NATURAL LANGUAGE DESCRIPTIONS OF PHENOTYPES IN PLANT DATABASES AND A WEB APPLICATION AS PROOF OF CONCEPT

Ian R. Braun^{1,2,*}, Colleen F. Yanarella^{1,2,*}, Jyothi Prasanth Durairaj Rajeswari^{3,*}, Diane C. Bassham⁴, and Carolyn J. Lawrence-Dill^{1,2,3}

¹Interdepartmental Bioinformatics and Computational Biology, Iowa State University

²Department of Agronomy, Iowa State University

³Department of Electrical and Computer Engineering, Iowa State University

⁴Department of Genetics, Development and Cell Biology, Iowa State University

*Joint first authors

Modified from a manuscript submitted to *Database*

3.1 Abstract

Similarities in phenotypic descriptions can be indicative of shared genetics, metabolism, and stress responses, to name a few. Finding and measuring similarity across descriptions of phenotype is not straightforward, with previous successes in computation requiring a great deal of expert data curation. Natural language processing of free text descriptions of phenotype is often less resource intensive than applying expert curation. It is therefore critical to understand the performance of natural language processing techniques for organizing and analyzing biological datasets and for enabling biological discovery. For predicting similar phenotypes, a wide variety of approaches from the natural language processing domain perform as well as curation-based methods. These computational approaches also show promise both for helping curators organize and work with large datasets and for enabling researchers to explore relationships among available phenotype descriptions. Here we generate networks of phenotype similarity and share a web

application for querying a dataset of associated plant genes using these text mining approaches. Example situations and species for which application of these techniques is most useful are discussed.

3.1.1 Database URLs

The database and analytical tool called QuOATS are available at <https://quoats.dill-picl.org/>. Code for the web application is available at <https://git.io/Jtv9J>. Datasets are available for direct access via <https://zenodo.org/record/7947342#.ZGwAK0zMK3I> (Braun et al. (2023)). The code for the analyses performed for the publication is available at <https://github.com/Dill-PICL/Plant-data> and <https://github.com/Dill-PICL/NLP-Plant-Phenotypes>.

3.2 Keywords

Phenotype, Ontology, NLP, Machine Learning, Plant Biology

3.3 Introduction

Phenotypes, defined as measurable characteristics or properties of an organism that result from interactions between genetics and the environment, comprise an enormous portion of the biological data considered important across a wealth of domains in the life sciences and beyond. Phenotypes are everything we see or measure in biology. On a more practical note, phenotypes encompass critical information related to human health and medicine, and important agronomic traits such as plant height and biomass of crop species. The scope of phenotypic information also ranges widely, from cellular phenotypes such as membrane composition or chemical concentrations, to community-level phenotypes like total leaf surface area in a field of crops. The extreme diversity in how phenotypes can be observed and represented makes handling this information on a computational level fundamentally different than genomic data, which lends

itself to computational means of representation and analysis based on the existing natural codes of bases and amino acids (reviewed in [Braun et al. \(2018\)](#)). This is especially true for phenotypes that are comparative, such as two different leaf morphologies, rather than phenotypes that are easily translated into a quantitative value, such as height (reviewed in [Braun et al. \(2020\)](#)).

Despite these challenges, bio-ontologies have greatly helped to enable computation on phenotypic information by providing standardized, hierarchical sets of descriptors (terms) that can be used to annotate phenotypic information. Doing so enables comparison between data points, including comparisons across multiple species, studies, and sources in a meaningful way, which has contributed to the use of these data structures in recent years. Using terms from the Gene Ontology (GO; [Ashburner et al. \(2000\)](#)) to describe cellular components, functions, and processes allows researchers to quickly find genes related to a biological concept of interest, and to understand which biological processes are potentially carried out or influenced by a group of genes of interest ([Huang et al. \(2008\)](#)). Using this same ontology as the format for predictions about gene functions allows datasets of predicted gene functions to be seamlessly incorporated with and compared to known information ([Zhou et al. \(2019\)](#)). Biomedical vocabulary graphs such as the Human Phenotype Ontology (HPO; [Robinson et al. \(2008\)](#)) and Disease Ontology (DO; [Schriml et al. \(2011\)](#)) allow for organization and interoperability of the vast and growing body of knowledge surrounding human medicine. Efforts such as Phenoscope ([Edmunds et al. \(2015\)](#)), the Monarch Initiative ([Mungall et al. \(2016\)](#)), and Planteome ([Cooper et al. \(2017\)](#)), use ontologies to provide common data representations and allow for comparisons across diverse species or across evolutionary history.

At the same time, both the performance and availability of natural language processing (NLP) and machine learning (ML) methods for working with natural language and text data have continually improved. Large language models now include artificial intelligence (AI) tools such as ChatGPT, developed by OpenAI ([OpenAI \(2023\)](#)). The release of ChatGPT is popular, in part, because of its public availability and conversational nature. ChatGPT uses a generative pre-trained transformer (GPT) model ([Radford et al. \(2018\)](#)), a deep-learning language model,

and has elicited an excited response within the language processing community (Dwivedi et al. (2023)). Improvements in language processing are due to recent and continued innovations in how neural networks are designed to handle this type of information (Mikolov et al. (2013), Le and Mikolov (2014), Vaswani et al. (2017)), and how they can be trained on massive volumes of unlabeled data (such as Wikipedia or PubMed) to provide systems for accurately modeling text in computable formats, and allowing for transfer to other domains and fine-tuning for more specific problems (Devlin et al. (2018), Wolf et al. (2020)). One result of this progress is that such techniques now represent a complementary approach to computationally handle the diversity of phenotypic information, at least for cases where phenotypes are represented as text descriptions. Given that phenotypes have been described in academic articles for more than a century, sources for phenotypic descriptions abound. Although the vast majority of phenotypes described in the literature have not been extracted and represented in computationally accessible community databases, some databases do exist that contain phenotype descriptions in free text fields.

Previously, we demonstrated that for some organizational tasks (like grouping functionally similar genes together), computational approaches that process text descriptions of phenotypes can work as well as, or better than, curated ontology term annotations for the creation of meaningful similarity measurements (Braun and Lawrence-Dill (2020)). Here, we demonstrate that this finding holds true for a larger dataset of the available phenotype text descriptions from across six different plant species. This means that, where available, text descriptions of phenotypes have the potential to provide useful biological insight when combined with a variety of methods from the field of NLP. We therefore make a case for expanded inclusion of free text descriptions as a valuable component of biological databases going forward, whether as a supplemental data type to more standardized ontology term annotations, or as a potential short-term alternative for species currently lacking the curatorial resources to produce large scale datasets of high-confidence, curated annotations.

In demonstrating the utility of analyzing text descriptions of phenotypes with NLP approaches, we focus on what can be learned from evaluating similarity between descriptions as a

measure of gene pair similarity. This is closely comparable to the ongoing problem in NLP of measuring sentence similarity, which has applications for text querying, text classification, and other tasks (De Boom et al. (2016)). An enormous variety of solutions have been put forward for this problem, including both general solutions as well as more narrowly focused solutions for working in particular domains, such as biomedical literature (Soğancıoğlu et al. (2017), Chen et al. (2019)). The number of solutions to this task is related to the fact that virtually all approaches for dealing with text computationally involve representing words or sentences as numerical vectors, on top of which similarity or distance metrics can then be applied to quantify relatedness between the two texts. In other words, all approaches for vectorizing text, which is typically the first step in handling any problem with NLP, can subsequently be used to find similarity between two texts by applying similarity metrics to their vector representations. This enables the generation of networks for organizing data across large datasets. In this work, we assess the performance of a variety of both simple and state-of-the-art methods for translating plant phenotype descriptions into numerical vectors and build networks that can be used to make inferences from pairwise similarities.

We also discuss and demonstrate how these same techniques can be applied for organizing and analyzing large phenotype description datasets, accounting for phenotypic characteristics that have not yet been explicitly defined by the input data. Finally, we provide a web application that enables others to explore and make use of phenotypic similarities identified. The application, called QuOATS (Querying with Ontology Annotations and Text Similarity), can be used to search for plant genes with similar phenotypic descriptions using gene identifiers, ontology terms, keywords, or similarity to searched phenotype descriptions as input.

3.4 Materials and Methods

3.4.1 Datasets

Species included for our analyses included *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). We collected a dataset of available phenotype descriptions that have been mapped to specific plant genes, primarily through mutation studies, from the model species databases for Arabidopsis - TAIR (Berardini et al. (2015)), maize - MaizeGDB (Portwood et al. (2018)), and solanaceous plants - SGN (Fernandez-Pozo et al. (2014)), and combined these data with as a dataset of phenotype descriptions created by Oellrich et al. (2015) that includes all six species. After merging data from multiple sources and preprocessing the texts, the combined dataset consisted of 7,907 genes from the 6 plant species, with the quantity of genes and the text describing their associated phenotypes varying across species (Table 3.1). The distributions of sentences and words quantities present per gene also vary broadly across species (Figure 3.1). Portions of the vocabulary used to describe phenotypes in each of the species are unique to that particular species, but in all cases more than 80% of the vocabulary was shared with at least one other species (Figure 3.2).

For the genes in this dataset, we also collected three types of ontology term annotations: Gene Ontology (GO; Ashburner et al. (2000)) annotations, Plant Ontology (PO; Jaiswal et al. (2005), Cooper et al. (2012)) annotations, and entity-quality (EQ) statements composed of multiple ontology terms. For in-depth discussion on how EQ statements are composed and compared to one another, see (Hoehndorf et al. (2011), Oellrich et al. (2015), Braun and Lawrence-Dill (2020)). GO and PO annotations were additionally sourced from the model species databases (Berardini et al. (2015), Portwood et al. (2018), Wimalanathan et al. (2018)) and Planteome (Cooper et al. (2017), <http://www.planteome.org>), and were limited to those with evidence codes indicating they were either experimentally determined or created through author or curator statements (Consortium et al. (2012), Giglio et al. (2018)). The EQ statements were sourced from the dataset of curator-defined EQ statements created by Oellrich et al. (2015). Not all genes in the dataset had at least one annotation of each type, and these quantities are given in Table 3.1. The preprocessed, merged, and cleaned dataset described here is available and further described through a dedicated repository (see Section 3.7 Data and Code Availability).

We also mapped the genes in this dataset to objects from additional bioinformatics resources, namely biochemical pathways in KEGG (Kanehisa (2002)) and PlantCyc (Schläpfer et al. (2017)), protein-protein associations in STRING (Szklarczyk et al. (2016)), ortholog relationships in PANTHER (Thomas (2003)), and a hierarchical Arabidopsis gene classification based on phenotypes (Lloyd and Meinke (2012)). A subset of the genes in the complete dataset are found in each of these resources (Table 3.1).

3.4.2 Measure of gene pair similarity

We used a set of approaches for generating n by n pairwise similarity matrices, where n is the number of genes in the dataset, and the values in the matrix are some measure of the similarity between a given pair of genes. Each approach yields one matrix. The approaches belong to two main groups: text-based approaches that translate the text descriptions of phenotype(s) associated with each gene into numerical vectors, so that gene pair similarity can then be found using cosine similarity, and curator-based approaches, that rely on similarities between existing annotations for each gene (GO terms, PO terms, or EQ statements) to quantify gene pair similarity. Each of the text-based approaches used is described in overview here, as well as how the curator-based approaches determine gene pair similarity from annotations.

3.4.2.1 Tokenizing sentences

For each of the text-based approaches, we determined the effects of treating the entirety of the phenotype descriptions associated with a gene as one concatenated text, and comparing between those texts for pairs of genes to measure gene pair similarity, or by first tokenizing (separating) the phenotype descriptions into individual sentences, and treating those sentences as individual text instances. Then the maximum similarity scores obtained by any pair of sentences was taken as the gene pair relatedness score. This measure is intended to alleviate the effects of genes with longer phenotype descriptions seeming to appear unrelated to ones with shorter ones, and is analogous to looking for local alignments in the text, rather than global ones. In the subsequent

Methods sections, we use the word ‘text’, to mean either the concatenation of all phenotype descriptions associated with a gene, or a single sentence from those descriptions, depending on which of these two methods is being described. Sentence tokenization was done with the NLTK package (Bird et al. (2009)).

3.4.2.2 Baseline approach

Some genes in the collected dataset have identical phenotype descriptions. As a baseline approach against which to compare the subsequently described approaches, we include an approach that simply yields a similarity value of 1 for gene pairs that have identical texts, and 0 for gene pairs with texts that differ in any way, after preprocessing.

3.4.2.3 TF-IDF

Constructing tf-idf (term frequency-inverse document frequency) vectors is one of simplest ways of representing text in a computable format. With this approach, phenotype descriptions are treated as a bag-of-words, and translated to a vector which is the same length as the total number of unique words in the dataset vocabulary, where each position in the vector corresponds to a particular word. The value at the position in the vector for a particular word is the number of times that word appears in the phenotype description (term frequency) weighted by the inverse of the fraction of phenotype descriptions in which that word appears (inverse document frequency). Weighting by the inverse document frequency emphasizes the importance of rarer words (e.g., ‘gametophyte’) and de-emphasizes the importance of more common words (e.g., ‘plant’) in the vector encoding. In addition to this straightforward implementation of the tf-idf approach, we also used as a bigram approach where positions in the vector represent a sequence of two consecutive words (as opposed to the unigram approach, where positions are a single word, as described above). We also used a tf-idf monogram approach where the phenotype descriptions in the datasets are first subset to only include words that are over-represented in journal articles abstracts related to plant phenotypes. The criteria for inclusion was that a word appeared at

least twice as frequently in the dataset of plant phenotype related abstracts compared to a general domain corpus. In all cases, cosine similarity was used to calculate gene pair similarity after phenotype descriptions were translated into vectors.

3.4.2.4 Computational annotation (NOBLE Coder)

NOBLE Coder ([Tseytlin et al. \(2016\)](#)) is a computational tool for annotating text with ontology terms. We used NOBLE Coder to annotate phenotype descriptions with terms from a set of bio-ontologies (GO, PO, and PATO), inheriting additional terms using the hierarchical structure of the ontologies. We used NOBLE Coder with both the exact and partial match parameters, which alters how strictly an ontology term must match a text string for an annotation to be assigned. After assigning terms to phenotype descriptions for genes by this method, each gene is represented by a set of terms rather than a set words, and the process of translating these representations into numerical vectors and calculating gene pair relatedness using cosine similarity is the same as with the tf-idf approach, with positions in the resulting vectors referencing terms instead of words. Again, cosine similarity was applied to yield similarity matrices from these resulting vectors.

3.4.2.5 Topic modeling (LDA and NMF)

We used Latent Dirichlet-Allocation (LDA; [Blei et al. \(2003\)](#)) and Non-negative Matrix Factorization (NMF; [Lee and Seung \(1999\)](#)) to perform topic modelling on the dataset of phenotype descriptions. These are decomposition algorithms that are widely used in NLP applications (reviewed in [Jelodar et al. \(2018\)](#)), and result in translating a document-term matrix into a document-topic matrix (in our case, documents are phenotype descriptions). If the algorithm is run to learn 10 topics, then the outcome is that each phenotype is represented by a vector of length 10 where each position indicates the probability that the phenotype is derived from that particular topic. Determining the appropriate number of topics to use for a particular dataset is often a matter of trying a range of values, and looking at which value produces the

most coherent or logical topics given the subject matter. Based on the word probability distributions created using a range of topic quantities, we used our best judgement to elect to use 50 topics and 100 topics for our embedding approaches using each of these algorithms.

3.4.2.6 Neural network-based embeddings (Word2Vec, Doc2Vec, BERT, BioBERT)

We also used machine learning approaches designed to find vector embeddings that represent the semantics of input text in a compressed space, with positions in the embedding representing abstract semantic features. Word2Vec ([Mikolov et al. \(2013\)](#)) is an approach for generating word embeddings based on the contexts in which words appear in a corpus. We used a skip-gram model, where a shallow network is trained to take one word at a time from our corpus as input and predict surrounding context words. The result of this self-supervised training step is a vector embedding for each word that occurs in the dataset of descriptions that reflects the context those words appear in, in a compressed feature space (200 dimensions). To supplement our dataset of phenotype descriptions to build a larger corpus for self-supervised training, we shuffled in sentences accessed from PubMed that were present in abstracts retrieved with queries for the word ‘phenotype’ and any of the names of the species present in our dataset. Hyperparameters for model construction were selected through a validation task of predicting whether ontology term names and synonyms from PATO and PO were parent-child or sibling pairs, or more distantly related. This validation task led to the selection of a skip-gram model using a window size of 8, and a hidden layer size of 200 (see `gensim` package ([Rehurek and Sojka \(2010\)](#)) for parameter details). In addition, as a point of comparison, we also used pre-trained published models trained on PubMed ([Pyysalo et al. \(2013\)](#)) and Wikipedia ([Lau and Baldwin \(2016\)](#)).

Doc2Vec is an extension of Word2Vec that either exclusively learns embeddings for documents (texts with multiple words) or learns embeddings for documents simultaneously with word embeddings. We used a distributed bag of words architecture where the arbitrary document tags are used as an input in a self-supervised process to predict randomly selected words from the

input documents, resulting in network architecture that can be used to infer document-specific embeddings (Le and Mikolov (2014)). We utilized the same training approach as for word embeddings, using only concept pairs with multiple words as validation data. In addition, we used a pre-trained Doc2Vec model trained on Wikipedia (Lau and Baldwin (2016)).

BERT (Bidirectional Encoder Representations from Transformers) is a large-scale neural network architecture trained on large unlabeled text datasets to predict masked words in sentences and predict whether one sentence follows another in a corpus (Devlin et al. (2018)). This results in a network where the encoder can be used to generate context-specific vector embeddings for words in an input sentence. We used both the BERT base model (Devlin et al. (2018)) and BioBERT models fine-tuned on abstracts from PubMed and articles from PubMed Central (Lee et al. (2019)).

The Doc2Vec models were used to directly infer vector embeddings for phenotype descriptions. The Word2Vec and BERT models generate vector embeddings for each word in phenotype descriptions, so these individual word-embeddings were combined to produce a single vector embedding for each phenotype description. Whether the vectors are summed or averaged is a hyperparameter choice, along with how many encoder layers are used to build the BERT word vectors, and whether those layers should be summed or concatenated. These hyperparameter choices were made using performance on the validation task described previously for the networks trained on phenotype descriptions, and for the pre-trained models we selected hyperparameters based on their performance on a related biomedical sentence similarity problem with the BIOSSES dataset (Soğancıoğlu et al. (2017)), and went forward with the hyperparameters that provided the best results on that separate dataset. As with the other approaches, cosine similarity was applied to the resulting vectors to yield similarity matrices.

3.4.2.7 Using embeddings to generate meaningful vectors with word replacement

Producing the most informative vector representations of phenotype descriptions requires combining the tf-idf approach of explicitly representing the quantity of each particular word from

the vocabulary that is present in each phenotype description, and also accounting for semantics through learning vector embeddings of particular words relative to their own meanings in this vocabulary or their meaning relative to the words around them in these phenotypes. We used an approach where pairwise word-similarity matrices for each word in the vocabulary as represented by our Word2Vec models were used to replace each word in all descriptions with the most common word in the vocabulary out of the word itself and the three other most similar words predicted by that model (algorithm detailed in [Pontes et al. 2016](#)). This results in substitutions such as ‘susceptible’ to ‘resistance’ that may allow comparisons to be made between phenotypes that simpler bag-of-words approaches would consider as distinct. The resulting vector representations are tf-idf vectors, but the semantic relationships between words as informed by the neural network models is already accounted for prior to encoding.

3.4.2.8 Curated annotations (GO, PO, EQ statements)

For a point of comparison to the text-based approaches described above, we also used the curator-based annotations to quantify gene pair relatedness. For GO and PO annotations, we calculated similarities as the maximum information content of any single term shared between the annotation sets for a given pair of genes. The more similar two sets of annotations are, the more specific (with higher information content) the terms shared between the two sets will be with respect to the ontology graph structure, leading to greater similarity. In this case, information content is transformed to be in the range of 0 to 1, so that it can be used as a similarity metric compatible with the other approaches used. To quantify similarity between genes using EQ statements, we used the pairwise similarities provided in [Oellrich et al. \(2015\)](#).

3.4.3 Formulating Biologically Relevant Questions

We used additional bioinformatic resources (KEGG, PlantCyc, STRING, PANTHER, etc.) to assess representation of biologically relevant relationships between gene pairs in the dataset, that each approach described above can attempt to recover by quantifying the similarity for that pair

of genes, allowing for direct comparison among the approaches (Table 3.2). Because not all genes in the dataset map to each resource (Table 3.1), the number of gene pairs that are applicable to each question are not consistent (Table 3.3). Although these questions are likely related to one another in terms of true biology (e.g., if a pair of genes are related to the same observable phenotype, they are probably more likely to act in a shared pathway), these questions are neither identical nor redundant in the context of this work, because different questions apply to different portions of the dataset, and even within the overlaps of gene pairs that apply to multiple questions, the set of positives (gene pairs for which the correct answer is ‘true’) are not the same (Table 3.4). For example, the two most similar tasks are ‘Associations’ and ‘Pathways’, where 1,271,297 of the same gene pairs are considered in both tasks, and the Jaccard similarity between the two sets of target values (‘true’, ‘false’) between those gene pairs is only 0.172 (Table 3.4). For this reason, we looked at the results of each of these questions individually rather than combining them.

3.5 Results

3.5.1 Text-based approaches recover biological relationships

Using each of the text-based approaches as well as using similarity metrics over the existing curated annotations, we calculated gene pair similarity values for all pairs of genes in our dataset. We measured the success of each approach for (1) predicting whether two genes were orthologs (as specified in PANTHER), (2) predicting known protein associations specified in STRING, (3) predicting whether two genes functioned in at least one of the same biochemical pathways (as specified in PlantCyc and KEGG), and (4) at predicting whether two Arabidopsis genes belonged to one of the phenotype categories specified by [Lloyd and Meinke \(2012\)](#). For each of these biological questions, a given approach for measuring gene similarity is considered useful if the distribution of values for gene pairs for which the answer to the question is true is distinct from the distribution of values for gene pairs for which the answer to question is false. The success of each approach for each biological question was calculated in terms of the maximum F_1 statistic.

We also recalculated the maximum F_1 statistic for just the genes for which we have GO annotations, PO annotations, and EQ statements, to directly compare performance of each approach on each question with approaches that are based on curation (Table 3.5, Supplemental Table 3.6).

3.5.1.1 Text-based approach performance is dependent on biological query type

Of the four biological questions assessed for this analysis, predicting whether two genes were orthologous, whether two genes shared an association, or whether two genes belonged to a shared biochemical pathway were infeasible for any of the text-based or curation-based approaches, in terms of broad performance measured with maximum F_1 statistics (Table 3.5, Supplemental Table 3.6). The largest F_1 statistic obtained across all three of these tasks for any approach was 0.140 using the curated GO annotations, with all other approaches yielding F_1 values less than 0.12 (Table 3.5, Supplemental Table 3.6). However, F_1 statistics were much higher for the task of predicting whether two genes belonged to the same phenotypic category, an expected result given that this prediction follows directly from the explicit contents of the phenotypic descriptions (Table 3.5). This was true for both the text-based and curation-based approaches, but the best performance was achieved using text-based approaches (Table 3.5). Performance on this task of predicting whether two genes share a phenotypic category can be broken down by general classes of approach (Figure 3.3).

As previously stated, all approaches were unsuccessful in predicting ortholog relationships (Supplemental Table 3.6). In addition, all approaches were completely unsuccessful in predicting whether two genes from different species were involved in a common biochemical pathway (Supplemental Table 3.7). Even though the maximum F_1 statistics for predicting whether two genes share a pathway were already low, these values were even lower when filtering the dataset to only look at interspecies gene pairs, and marginally greater when filtering the dataset to only look at intraspecies gene pairs (Supplemental Table 3.7). Therefore, even the very small amount of biological information recovered only applies to looking at genes from within the same species.

This indicates that comparing the text of phenotype descriptions across different species is not biologically informative in this case. This might not be true for all species or all phenotypes, but it does not generalize across the current dataset of available plant phenotype descriptions.

3.5.1.2 Significant description similarity within individual phenotype and pathway gene groups

We evaluated similarities of phenotypic descriptions as a percentile based on F_1 score, and to visualize the results, plotted the average gene-to-gene similarity for phenotypic categories (Figure 3.4 (a)). Next, we imposed the same evaluation and visualization for genes mapped to each individual pathway (Figure 3.4 (b) and (c)). Although predicting whether two genes shared a biochemical pathway was generally unsuccessful (low maximum F_1), this is in part a consequence of the fact that pathways vary greatly in how related the phenotype descriptions for their component genes are. In Table 3.5 we report calculated p-values. This was accomplished by randomly sampling groups of genes at each value of n then calculating p-values for each phenotype category and pathway based on the probability of each approach generating a mean similarity value between genes in that group that is that large or larger, controlling for false discovery rate for each approach with the Benjamini–Hochberg procedure. For text-based approaches using sentence tokenization, 81% to 100% of the phenotypic categories had a significantly large average similarity value (with respect to the Benjamini–Hochberg procedure), while between 6% and 39% of the pathways obtained significant average similarity values, for these same approaches, with an average of 23% (Table 3.5). Taken together, these results indicate that while text-based similarity values are not broadly indicative of whether or not two genes share a pathway, there is a significant subset of known pathways for which this is the case. In the case of groups of genes belonging to the same pathway that do have similar phenotype descriptions, these are generally either due to mentions of downstream phenotypic effects of pathway disruption, or more direct mentions of the pathway function or role. For example, the descriptions associated with genes in the chlorophyll degradation pathway include mentions of

necrotic lesions, and the descriptions associated with genes in the phospholipid desaturation pathway include mentions of fatty acid levels or composition.

3.5.1.3 Combining syntactic and semantic approaches improves recovery of phenotypic categories

The purely syntactic text-based approaches (tf-idf) were among the most successful in terms of maximum F_1 statistic for predicting whether gene pairs belonged to the same phenotypic category (Table 3.5, Figure 3.3). In general, semantic approaches that use ML techniques to drastically reduce the dimensionality of the vector encoding for each text instance were comparably successful (Table 3.5, Figure 3.3). However, the combined approaches where semantic techniques were used to augment the information in the tf-idf vectors by replacing words with similar words prior to encoding provided a boost in performance over other approaches (Table 3.5, Figure 3.3). Taken together, this indicates that this dataset contains phenotype descriptions for genes in the same phenotypic category that are similar both in terms of explicitly shared words (where syntactic approaches are most helpful), as well as genes that are similar only in terms of shared meaning but not specific words (where semantic approaches provide an advantage). Using word embedding models trained on plant phenotype specific data provided marginal improvement over models trained on PubMed generally or the Wikipedia corpus, but all three models provided the same boost over other approaches when applied to word replacement, indicating that useful associations between words for recovering common phenotypic categories from descriptions are not limited to relationships only represented in a narrow corpus of text related to plant phenotypes. Given that using bio-ontologies for this same task did not perform as well as text-based approaches, and one of the main functions of such ontologies in this case is to inject domain-specific inferences into the similarity metrics, this result is not surprising.

3.5.1.4 Sentence tokenization is important for comparing phenotypes

For all the text-based approaches on all the biological questions posed, the preprocessing step of tokenizing phenotype descriptions into sentences and evaluating gene pair relatedness as the maximum pairwise sentence similarity resulted in greater F_1 statistics (Table 3.5, Supplemental Table 3.6). Unexpectedly, this held true even for approaches that are generally intended for use with larger input texts, such as Doc2Vec, and topic modeling algorithms LDA and NMF. This indicates that when predicting whether two genes share a common role, it is important to account for ‘local alignments’ in their associated phenotype descriptions, as the similarity might exist between single sentences associated with those genes while other sentences act as noise obscuring this relationship.

3.5.2 Enabling biologists to use these methods and dataset

3.5.2.1 Web application (QuOATS)

We have developed a web application called QuOATS (Querying with Ontology Annotations and Text Similarity) for querying the dataset described here through leveraging the computational methods described here (Figure 3.5 (a)). The underlying dataset of plant genes is the same as is described previously (Table 3.1), and can be filtered to include particular species (Figure 3.5 (b)). The application supports four different query types (Figure 3.5 (d)), with the primary purpose being to obtain lists of genes that are related to phenotypes described similarly to some phenotypic characteristic(s) of interest. Firstly, a free text query can be used to search the dataset for any genes related to phenotypes that are described similarly to text strings separated by periods in the query (Figure 3.5 (e)). Secondly, a keyword query can be used to input any number of strings of any length, and genes whose phenotype descriptions contain those strings (after preprocessing including stemming and case-normalization) are returned (Figure 3.5 (f)). Thirdly, an ontology term query can be used to search for any genes annotated by curators with one or more ontology terms, either directly or inherited through the ontology hierarchy (Figure 3.5 (g)). Lastly, a gene identifier query can be carried out to search for any gene name, protein name, gene

model, or any other gene identifier potentially represented in the dataset. Selecting a gene from the returned list of candidates that match the query will auto-complete a second query that returns genes related to phenotypes that are described similarly to the selected genes (Figure 3.5 (h)). The similarity scores used to rank genes in the returned list are calculated using approaches described here, selected from a drop-down menu in the web application (Figure 3.5 (c)).

3.5.2.2 Proof of concept applications of the web tool

In our previous findings illustrated in [Braun and Lawrence-Dill \(2020\)](#), we discussed how a set of genes related to anthocyanin biosynthesis could be used to demonstrate recovering gene groups by querying specifically with phenotype descriptions or computationally generated annotations from those descriptions. Specifically, we looked at a dataset of 16 maize genes ([Li et al. \(2019\)](#)) and 21 genes from Arabidopsis ([Appelhagen et al. \(2014\)](#)) but only 10 of the maize genes and 16 of the Arabidopsis genes were present in the dataset. Our expanded dataset in this work includes 14 of those maize genes and 18 of the Arabidopsis genes. We now evaluate the results of querying with each of these genes in the web application QuOATS, to recover both genes in the same species from these sets and genes in the alternate species. Over the 64 total queries (32 within the same species and 32 between species), we quantified the average and standard deviation of the number of target genes contained in bins of ranks in the query results, in bin sizes of 10 up to 50, and a final bin for genes that obtain ranks higher than 50 (Figure 3.6). Additionally, we also repeated this analysis for a set of 9 core autophagy genes in Arabidopsis (Figure 3.6). These queries illustrate a proof-of-concept whereby the web application can be used to query with phenotypic descriptions associated with one gene to recover other related genes. This application demonstrates the utility of applying text-based algorithms in cases where ontology annotations are either not present, are insufficient, or could simply be augmented by allowing additional, less rigidly-defined phenotype descriptions to be searchable as well.

3.6 Discussion

The difficulty in computing on phenotypic data is largely a consequence of extreme variability with which these data are represented, and the diversity of ways that phenotypes are measured, quantified, and described. This is in contrast with sequence data; biology as a field has enormously benefited from the ways in which sequence data are intuitively computed on, given the naturally occurring nucleic acid and amino acid coding systems. Sequencing technology provided the datasets to compute on, and algorithms and applications like BLAST provided the means to make use of these data (Altschul et al. (1990)). Ontologies have begun to provide a similar means for making phenotypic data computable, and processing of natural language provides an additional avenue by which we can make biological inferences if we have the datasets on which to apply them. The combination of biological ontologies, machine learning approaches, and NLP provide strategies for handling phenotypic descriptions and learning from it where it exists.

Plant phenotypes are frequently described as text within academic papers or research notes. However, these text descriptions are rarely incorporated into relevant research community databases, associated with a specific gene or genotype, and made readily available as part of the growing data resources for that species. This could be the case for a variety of reasons, including the difficulties involved with extracting phenotype descriptions from larger texts, the curatorial effort necessary to produce high quality datasets of phenotypes descriptions associated with genes, or because these text representations of phenotypes are considered a non-valuable data type, and are instead represented by annotations using structured vocabularies of hierarchical terms such as biological ontologies. Notable exceptions to this situation exist, including The Arabidopsis Information Resource (TAIR), which contains thousands of text descriptions of phenotypes mapped to specific Arabidopsis genes (Berardini et al. (2015)).

In this work, we have shown that a variety of NLP approaches for vectorizing phenotype descriptions in order to generate gene pair similarity matrices are equally or more predictive in general of known phenotype categorizations compared to using existing curated annotations for this task. Based on these results, we argue that it is worthwhile for databases that contain

gene-to-phenotype information to include natural language descriptions of phenotypes. In addition, these descriptions should be made accessible to emergent AI tools, like ChatGPT, to enable the creation of additional resources that can make text descriptions of plant phenotypes more findable and accessible for biological analytics.

The natural language descriptions of phenotypes are useful, and when combined with NLP approaches for computationally representing text can be leveraged to provide a way for researchers to quickly identify genes associated with phenotypes similar to the ones that they are observing or studying. Natural language descriptions can also be used to organize genes computationally on a large scale and discover which categorizations of phenotypes are present in a dataset, with techniques like clustering and topic modeling. In some cases, this natural language data may be easier to generate than ontology annotations. In situations where curators are not available (or have limited time) to generate the high-confidence ontology term annotation datasets, it may be faster or still possible for authors or someone else to at least identify the free-text portions of the manuscript that include phenotype descriptions, and the genes associated with them. In the near future, NLP techniques for parsing full-texts may also progress to the point where this phenotype identification could be done automatically as well. In these instances, we argue it is worthwhile to generate and make accessible this free text phenotype information. In other cases, these text data might already be generated, but are potentially discarded. In situations where curators are actively involved in generating ontology annotations from papers, this process often involves the tasks of highlighting text from the paper, or possibly writing down the phenotype descriptions first then producing the ontology term representation of those associations. Given that the free text itself is useful, we argue it should be retained in the final mapping in the resulting database or dataset rather than being discarded as an intermediate data form. It is possible that for some applications the ontology annotations will be more useful than the natural language descriptions, for example when making comparisons between species, but we have shown that this is not always the case, and if it is being generated regardless, it makes sense to retain the natural language and make it available.

The area where the application of these methods would likely make the most difference is for species where phenotypes are still largely described in general biological terms rather than in cases where phenotype descriptions are limited to special vocabularies and/or where phenotype data have been carefully curated in a pervasive, large-scale way (as is the case with human phenotypes and diseases). In addition, these approaches would make the most sense to use when high quality, curated datasets of ontologized phenotypes are either not available or curation into those data forms are not financially feasible. In these cases, if at a minimum phenotype descriptions are extracted from literature and associated with specific genes in an accessible community database, NLP methods can be applied to organize these data and to group by genes into sets that impact similar phenotypes, therefore allowing researchers to search based on linguistic similarity.

Not only do plant scientists understand their phenotypes and use rich language to describe them, there is a diversity of algorithms available to enable computation on phenotypic descriptions so that the scope of data any single researcher can access becomes quite expansive. In 2011, Mike Freeling made an impression by saying, “Ontologies are for people who don’t understand their phenotype,” to CJLD at the Annual Maize Meeting Genetics Conference in response to a request to review the completeness of the MaizeGDB Phenotypic Controlled Vocabulary (Michael Freeling personal communication). While ontologies have proved invaluable for managing and analyzing the massive quantity of data that biologists deal with, we think that this quote emphasizes the key finding for the efforts here: that we should not undervalue the utility of free text as a datatype, and that it should be made available through bioinformatic resources that provide phenotypic data to the research community, given that we have the computational tools to leverage it in useful ways.

3.7 Data and Code Availability

The dataset of plant genes collected from other sources for this work is available at [Braun et al. \(2023\)](#), along with all the code for preprocessing, reshaping, and merging this data <https://github.com/Dill-PICL/Plant-data>. The code for carrying out the analysis shown

here has its own repository at <https://github.com/Dill-PICL/NLP-Plant-Phenotypes>. The results given here can be reproduced using code and datasets at those locations. In addition, a Python package called OATS (Ontology Annotation and Text Similarity) for working with gene-phenotype datasets, ontology annotations, and free-text was developed in parallel with this work. This package was used extensively for this analysis, and can be found at <https://git.io/JTuqv>, with documentation available at <https://irbraun-oats.readthedocs.io>. We have combined the dataset and some of the techniques for identifying similar texts into a streamlit web application named QuOATS available at <https://quoats.dill-picl.org/>. Use this tool for looking up genes by phenotype keywords or phrases, or finding genes with similar descriptions to a searched phenotype description. The code for this web application is available at <https://git.io/Jtv9J>.

3.8 Author Contribution

IRB and CJLD conceived the idea for this work. IRB created the first demonstration of all results and drafted the initial version of the manuscript. DCB conceived and reviewed autophagy analyses included as a proof of concept. CFY directed the reproducibility review and edited the manuscript, accordingly. JPDR reproduced previous results and updated the manuscript accordingly. All authors read, edited, and approved the manuscript.

3.9 Acknowledgements

The authors thank Carson Andorf, John Portwood, and Naama Menda for their assistance in obtaining the dataset of plant phenotype descriptions and annotations. The authors thank Darwin Campbell and Scott Zarecor for their assistance with dataset creation and hosting and maintenance of the web application discussed here. The authors thank Marna Yandea-Nelson, Iddo Friedberg, Baskar Ganapathysubramanian, Annette O'Connor, and Qi Li for useful discussions on the topics discussed here, and particularly thank Marna Yandea-Nelson for assistance in editing and Iddo Friedberg for assistance in reviewing the design of the web

application. The authors thank Leila Fattel for helpful discussions and support. The authors also thank the High Performance Computing support team at Iowa State University for their help throughout the project.

3.10 Funding

This work has been supported by an Iowa State University Presidential Interdisciplinary Research Seed Grant (PI Diane Bassham; CJLD is a coPI), the Iowa State University Plant Sciences Institute Faculty Scholars Program to CJLD, the Predictive Plant Phenomics NSF Research Traineeship (#DGE-1545453) to CJLD (IRB and CFY are trainees), a USDA NIFA Agriculture and Food Research Initiative Predoctoral Research Grant (#2020-67034-31745) to IRB, and the NSF and USDA-NIFA AI Research Institutes program for AI Institute: for Resilient Agriculture (#2021-67021-35329) to CJLD and supporting CFY.

3.11 References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.
- Appelhagen, I., Thiedig, K., Nordholt, N., Schmidt, N., Huep, G., Sagasser, M., and Weisshaar, B. (2014). Update on transparent testa mutants from *Arabidopsis thaliana*: characterisation of new alleles from an isogenic collection. *Planta*, 240(5):955–970.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.
- Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2015). The *Arabidopsis* information resource: Making and mining the “gold standard” annotated reference plant genome. *Genesis*, 53(8):474–485.
- Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly Media, Inc.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

- Braun, I., Balhoff, J. P., Berardini, T. Z., Cooper, L., Gkoutos, G., Harper, L., Huala, E., Jaiswal, P., Kazic, T., Lapp, H., Macklin, J. A., Specht, C. D., Vision, T., Walls, R. L., and Lawrence-Dill, C. J. (2018). ‘*Computable*’ Phenotypes Enable Comparative and Predictive Phenomics Among Plant Species and Across Domains of Life, volume 33, page 187–205. IOS Press.
- Braun, I. R. and Lawrence-Dill, C. J. (2020). Automated Methods Enable Direct Computation on Phenotypic Descriptions for Novel Candidate Gene Prediction. *Frontiers in Plant Science*, 10.
- Braun, I. R., Yanarella, C. F., and Lawrence-Dill, C. J. (2020). Computing on Phenotypic Descriptions for Candidate Gene Discovery and Crop Improvement. *Plant Phenomics*, 2020.
- Braun, I. R., Yanarella, C. F., Rajeswari, J. P. D., Bassham, D. C., and Lawrence-Dill, C. J. (2023). Datasets for “The Case for Retaining Natural Language Descriptions of Phenotypes in Plant Databases and a Web Application as Proof of Concept”.
- Chen, Q., Peng, Y., and Lu, Z. (2019). BioSentVec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.
- Consortium, G. O., Blake, J. A., Dolan, M., Drabkin, H., Hill, D. P., Li, N., Sitnikov, D., Bridges, S., Burgess, S., Buza, T., McCarthy, F., Peddinti, D., Pillai, L., Carbon, S., Dietze, H., Ireland, A., Lewis, S. E., Mungall, C. J., Gaudet, P., Chrisholm, R. L., Fey, P., Kibbe, W. A., Basu, S., Siegele, D. A., McIntosh, B. K., Renfro, D. P., Zweifel, A. E., Hu, J. C., Brown, N. H., Tweedie, S., Alam-Faruque, Y., Apweiler, R., Auchinchloss, A., Axelsen, K., Bely, B., Blatter, M. C., Bonilla, C., Bouguerleret, L., Boutet, E., Breuza, L., Bridge, A., Chan, W. M., Chavali, G., Coudert, E., Dimmer, E., Estreicher, A., Famiglietti, L., Feuermann, M., Gos, A., Gruaz-Gumowski, N., Hieta, R., Hinz, C., Hulo, C., Huntley, R., James, J., Jungo, F., Keller, G., Laiho, K., Legge, D., Lemercier, P., Lieberherr, D., Magrane, M., Martin, M. J., Masson, P., Mutowo-Muellenet, P., O’Donovan, C., Pedruzzi, I., Pichler, K., Poggioli, D., Millán, P. P., Poux, S., Rivoire, C., Roechert, B., Sawford, T., Schneider, M., Stutz, A., Sundaram, S., Tognolli, M., Xenarios, I., Foulgar, R., Lomax, J., Roncaglia, P., Khodiyar, V. K., Lovering, R. C., Talmud, P. J., Chibucos, M., Giglio, M. G., Chang, H. Y., Hunter, S., McAnulla, C., Mitchell, A., Sangrador, A., Stephan, R., Harris, M. A., Oliver, S. G., Rutherford, K., Wood, V., Bahler, J., Lock, A., Kersey, P. J., McDowall, D. M., Staines, D. M., Dwinell, M., Shimoyama, M., Laulederkind, S., Hayman, T., Wang, S.-J., Petri, V., Lowry, T., D’Eustachio, P., Matthews, L., Balakrishnan, R., Binkley, G., Cherry, J. M., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hitz, B. C., Hong, E. L., Karra, K., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Weng, S., Wong, E. D., Berardini, T. Z., Huala, E., Mi, H., Thomas, P. D., Chan, J., Kishore, R., Sternberg, P., Auken, K. V., Howe, D., and Westerfield, M. (2012). Gene Ontology Annotations and Resources. *Nucleic Acids Research*, 41(D1):D530–D535.

- Cooper, L., Meier, A., Laporte, M.-A., Elser, J. L., Mungall, C., Sinn, B. T., Cavaliere, D., Carbon, S., Dunn, N. A., Smith, B., Qu, B., Preece, J., Zhang, E., Todorovic, S., Gkoutos, G., Doonan, J. H., Stevenson, D. W., Arnaud, E., and Jaiswal, P. (2017). The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research*, 46(D1):D1168–D1180.
- Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T. Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y., and Jaiswal, P. (2012). The Plant Ontology as a Tool for Comparative Plant Anatomy and Genomic Analyses. *Plant and Cell Physiology*, 54(2):e1–e1.
- De Boom, C., Van Canneyt, S., Demeester, T., and Dhoedt, B. (2016). Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koochang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., Chowdhury, S., Crick, T., Cunningham, S. W., Davies, G. H., Davison, R. M., Dé, R., Dennehy, D., Duan, Y., Dubey, R., Dwivedi, R., Edwards, J. S., Flavián, C., Gauld, R., Grover, V., Hu, M.-C., Janssen, M., Jones, P., Junglas, I., Khorana, S., Kraus, S., Larsen, K. R., Latreille, P., Laumer, S., Malik, F. T., Mardani, A., Mariani, M., Mithas, S., Mogaji, E., Nord, J. H., O’Connor, S., Okumus, F., Pagani, M., Pandey, N., Papagiannidis, S., Pappas, I. O., Pathak, N., Pries-Heje, J., Raman, R., Rana, N. P., Rehm, S.-V., Ribeiro-Navarrete, S., Richter, A., Rowe, F., Sarker, S., Stahl, B. C., Tiwari, M. K., van der Aalst, W., Venkatesh, V., Viglia, G., Wade, M., Walton, P., Wirtz, J., and Wright, R. (2023). Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71:102642.
- Edmunds, R. C., Su, B., Balhoff, J. P., Eames, B. F., Dahdul, W. M., Lapp, H., Lundberg, J. G., Vision, T. J., Dunham, R. A., Mabee, P. M., and Westerfield, M. (2015). Phenoscope: Identifying Candidate Genes for Evolutionary Phenotypes. *Molecular Biology and Evolution*, 33(1):13–24.
- Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., Yan, A., and Mueller, L. A. (2014). The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. *Nucleic Acids Research*, 43(D1):D1036–D1041.

- Giglio, M., Tauber, R., Nadendla, S., Munro, J., Olley, D., Ball, S., Mitraka, E., Schriml, L. M., Gaudet, P., Hobbs, E. T., Erill, I., Siegele, D. A., Hu, J. C., Mungall, C., and Chibucos, M. C. (2018). ECO, the Evidence & Conclusion Ontology: community standard for evidence information. *Nucleic Acids Research*, 47(D1):D1186–D1194.
- Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2011). PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 39(18):e119–e119.
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13.
- Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., Reiser, L., Rhee, S. Y., Sachs, M. M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Ware, D., and Zapata, F. (2005). Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages. *Comparative and Functional Genomics*, 6(7-8):388–397.
- Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2018). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.
- Kanehisa, M. (2002). The KEGG Database. In *‘In Silico’ Simulation of Biological Processes: Novartis Foundation Symposium 247, Volume 247*, pages 91–103. Wiley Online Library.
- Lau, J. H. and Baldwin, T. (2016). An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. *arXiv preprint arXiv:1607.05368*.
- Le, Q. V. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *arXiv preprint arXiv:1405.4053*.
- Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Li, T., Zhang, W., Yang, H., Dong, Q., Ren, J., Fan, H., Zhang, X., and Zhou, Y. (2019). Comparative transcriptome analysis reveals differentially expressed genes related to the tissue-specific accumulation of anthocyanins in pericarp and aleurone layer for maize. *Scientific Reports*, 9(1).
- Lloyd, J. and Meinke, D. (2012). A Comprehensive Dataset of Genes with a Loss-of-Function Mutant Phenotype in Arabidopsis. *Plant Physiology*, 158(3):1115–1129.

- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.
- Mungall, C. J., McMurry, J. A., Köhler, S., Balhoff, J. P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., Gourdine, J., Jacobsen, J. O., Keith, D., Laraway, B., Lewis, S. E., NguyenXuan, J., Shefchek, K., Vasilevsky, N., Yuan, Z., Washington, N., Hochheiser, H., Groza, T., Smedley, D., Robinson, P. N., and Haendel, M. A. (2016). The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45(D1):D712–D722.
- Oellrich, A., Walls, R. L., Cannon, E. K., Cannon, S. B., Cooper, L., Gardiner, J., Gkoutos, G. V., Harper, L., He, M., Hoehndorf, R., Jaiswal, P., Kalberer, S. R., Lloyd, J. P., Meinke, D., Menda, N., Moore, L., Nelson, R. T., Pujar, A., Lawrence, C. J., and Huala, E. (2015). An ontology approach to comparative phenomics in plants. *Plant Methods*, 11(1).
- OpenAI (2023). Introducing ChatGPT.
- Pontes, E. L., Huet, S., Torres-Moreno, J.-M., and Linhares, A. C. (2016). Automatic Text Summarization with a Reduced Vocabulary Using Continuous Space Vectors. In *Natural Language Processing and Information Systems*, pages 440–446. Springer International Publishing.
- Portwood, J. L., Woodhouse, M. R., Cannon, E. K., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., Walsh, J. R., Sen, T. Z., Cho, K. T., Schott, D. A., Braun, B. L., Dietze, M., Dunfee, B., Elsik, C. G., Manchanda, N., Coe, E., Sachs, M., Stinard, P., Tolbert, J., Zimmerman, S., and Andorf, C. M. (2018). MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Research*, 47(D1):D1146–D1154.
- Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., and Ananiadou, S. (2013). Distributional semantics resources for biomedical text processing. *Proceedings of LBM*, pages 39–44.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.
- Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.
- Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *The American Journal of Human Genetics*, 83(5):610–615.

- Schläpfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., Dreher, K., Chavali, A. K., Nilo-Poyanco, R., Bernard, T., Kahn, D., and Rhee, S. Y. (2017). Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. *Plant Physiology*, 173(4):2041–2059.
- Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W. A. (2011). Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):D940–D946.
- Soğancıoğlu, G., Öztürk, H., and Özgür, A. (2017). BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.
- Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., and von Mering, C. (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362–D368.
- Thomas, P. D. (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Research*, 31(1):334–341.
- Tseytlin, E., Mitchell, K., Legowski, E., Corrigan, J., Chavan, G., and Jacobson, R. S. (2016). NOBLE – Flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics*, 17(1).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.
- Wimalanathan, K., Friedberg, I., Andorf, C. M., and Lawrence-Dill, C. J. (2018). Maize GO Annotation—Methods, Evaluation, and Review (maize-GAMER). *Plant Direct*, 2(4).
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsóh, B. Z., Crocker, A. W., Lewis, K. A., Georghiou, G., Nguyen, H. N., Hamid, M. N., Davis, L., Dogan, T., Atalay, V., Rifaioğlu, A. S., Dalkıran, A., Atalay, R. C., Zhang, C., Hurto, R. L., Freddolino, P. L., Zhang, Y., Bhat, P., Supek, F., Fernández, J. M., Gemovic, B., Perovic, V. R., Davidović, R. S., Sumonja, N., Veljkovic, N., Asgari, E., Mofrad, M. R., Profiti, G., Savojardo, C., Martelli, P. L., Casadio, R., Boecker, F., Schoof, H., Kahanda, I., Thurlby, N., McHardy, A. C., Renaux, A., Saidi, R., Gough, J., Freitas, A. A., Antczak, M., Fabris, F., Wass, M. N., Hou, J., Cheng, J., Wang, Z., Romero, A. E., Paccanaro, A., Yang, H., Goldberg, T., Zhao, C., Holm, L., Törönen, P., Medlar, A. J., Zosa, E., Borukhov, I., Novikov, I., Wilkins, A., Lichtarge, O., Chi, P.-H., Tseng, W.-C., Linial, M., Rose, P. W., Dessimoz, C., Vidulin, V., Dzeroski, S., Sillitoe, I., Das, S., Lees, J. G., Jones, D. T., Wan, C., Cozzetto, D., Fa, R., Torres, M., Vesztröcy, A. W., Rodriguez, J. M., Tress, M. L., Frasca, M., Notaro, M., Grossi, G., Petrini, A., Re, M., Valentini, G., Mesiti, M., Roche, D. B., Reeb, J., Ritchie, D. W., Aridhi, S., Alborzi, S. Z., Devignes, M.-D., Koo, D. C. E., Bonneau, R., Gligorijević, V., Barot, M., Fang, H., Toppo, S., Lavezzo, E., Falda, M., Berselli, M., Tosatto, S. C., Carraro, M., Piovesan, D., Rehman, H. U., Mao, Q., Zhang, S., Vucetic, S., Black, G. S., Jo, D., Suh, E., Dayton, J. B., Larsen, D. J., Omdahl, A. R., McGuffin, L. J., Brackenridge, D. A., Babbitt, P. C., Yunes, J. M., Fontana, P., Zhang, F., Zhu, S., You, R., Zhang, Z., Dai, S., Yao, S., Tian, W., Cao, R., Chandler, C., Amezola, M., Johnson, D., Chang, J.-M., Liao, W.-H., Liu, Y.-W., Pascarelli, S., Frank, Y., Hoehndorf, R., Kulmanov, M., Boudelloua, I., Politano, G., Carlo, S. D., Benso, A., Hakala, K., Ginter, F., Mehryary, F., Kaewphan, S., Björne, J., Moen, H., Tolvanen, M. E., Salakoski, T., Kihara, D., Jain, A., Šmuc, T., Altenhoff, A., Ben-Hur, A., Rost, B., Brenner, S. E., Orengo, C. A., Jeffery, C. J., Bosco, G., Hogan, D. A., Martin, M. J., O'Donovan, C., Mooney, S. D., Greene, C. S., Radivojac, P., and Friedberg, I. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1).

3.12 Figures and Tables

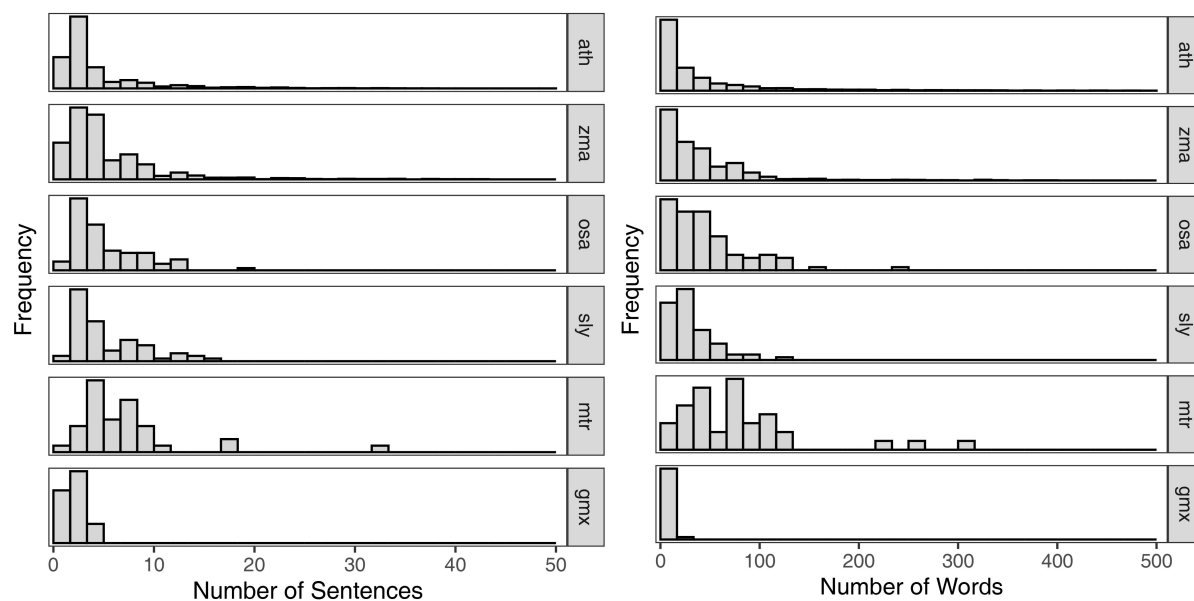


Figure 3.1 Phenotype description text length distributions across six plant species. The distributions for quantities of text in terms of both sentences (Left) and words (Right) describing phenotypes for genes in each of the plant species. Outliers with very long descriptions are not shown, which includes $<1\%$ of the genes belonging to Arabidopsis and $<0.1\%$ of the genes belonging to maize. The y-axis is scaled to be proportional to the quantity of genes for each individual species.

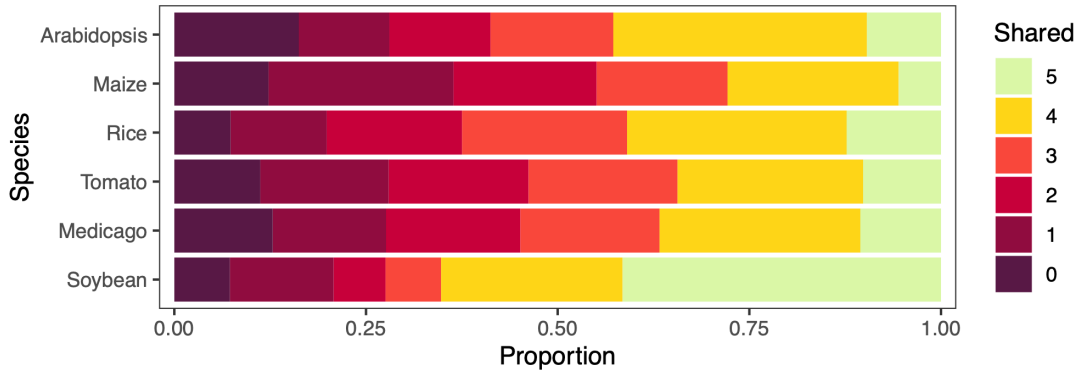


Figure 3.2 Overlap among vocabularies used to describe phenotypes in each species. For each of the six species in the dataset, listed on the left, the proportion of words in the total vocabulary used in all phenotype descriptions of that species that are shared with the vocabularies of a given additional number of species are shown, with colors indicated on the right. For example, plum/purple indicates the proportion of words used only in that species, and light green indicates the proportion of that species vocabulary that is shared with the vocabulary of all five other species.

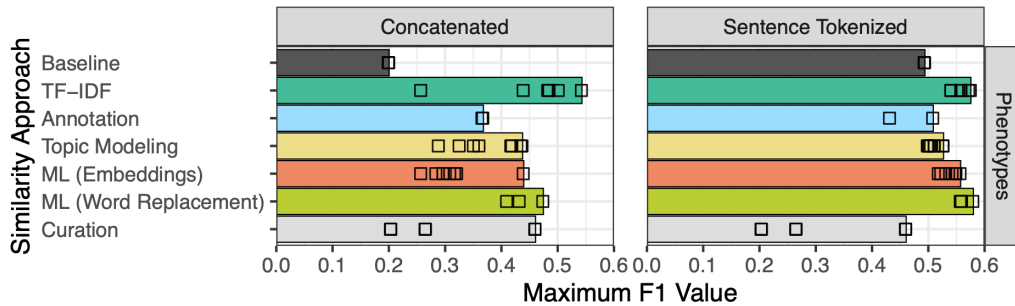


Figure 3.3 Comparing the groups of gene pair similarity approaches. The maximum F_1 statistics for each approach in each broad category for measuring gene similarity is shown, with the bar indicating the best F_1 statistics among all the approaches in that general group. Bars on the left indicate performance when phenotype descriptions are treated as one concatenated piece of text, and bars on the right indicate performance when the descriptions are sentence tokenized first.

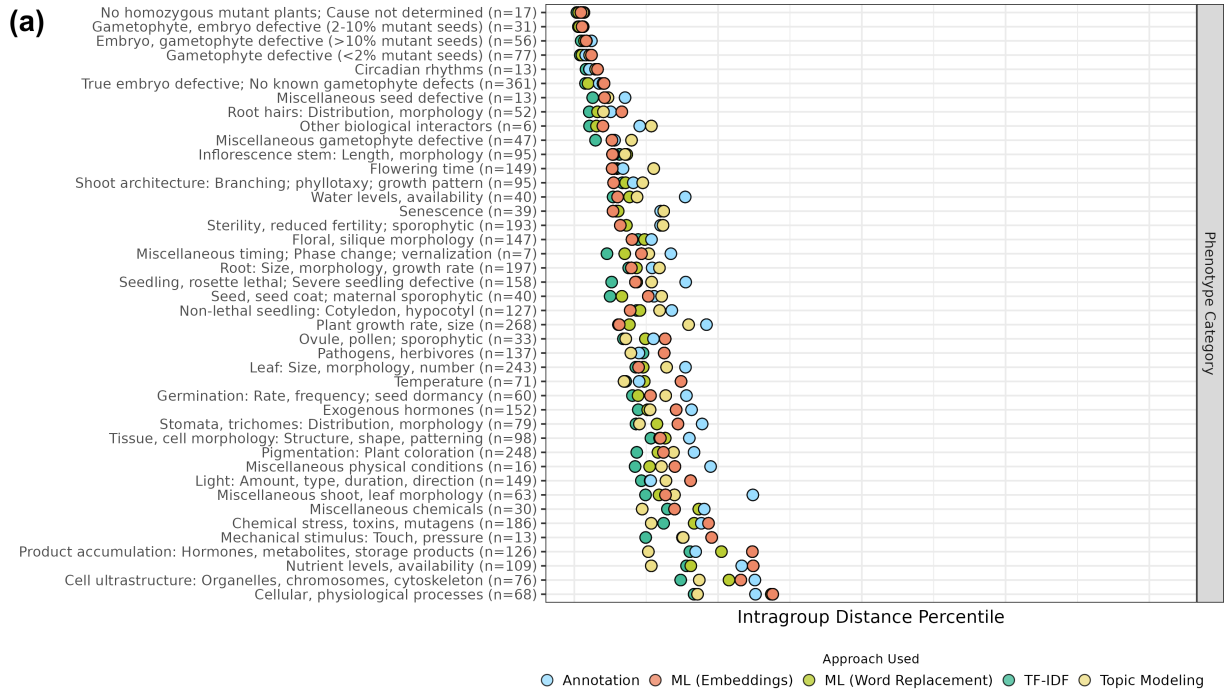


Figure 3.4 Cohesiveness of phenotype and pathway gene groups. Phenotype categories (a) are listed, with the number of genes in these datasets belonging to each group listed to the right of the group’s name. The x-axis indicates group cohesiveness, given as the percentile against all pairwise gene distances that the average distance between any two genes in that group falls in. The minimum value of this metric achieved by any approach that is in the listed category is shown. For example, the location of the yellow dot in a particular row indicates the smallest intragroup distance percentile obtained by any approach in the topic modeling category of text-based approaches for that particular group of genes.

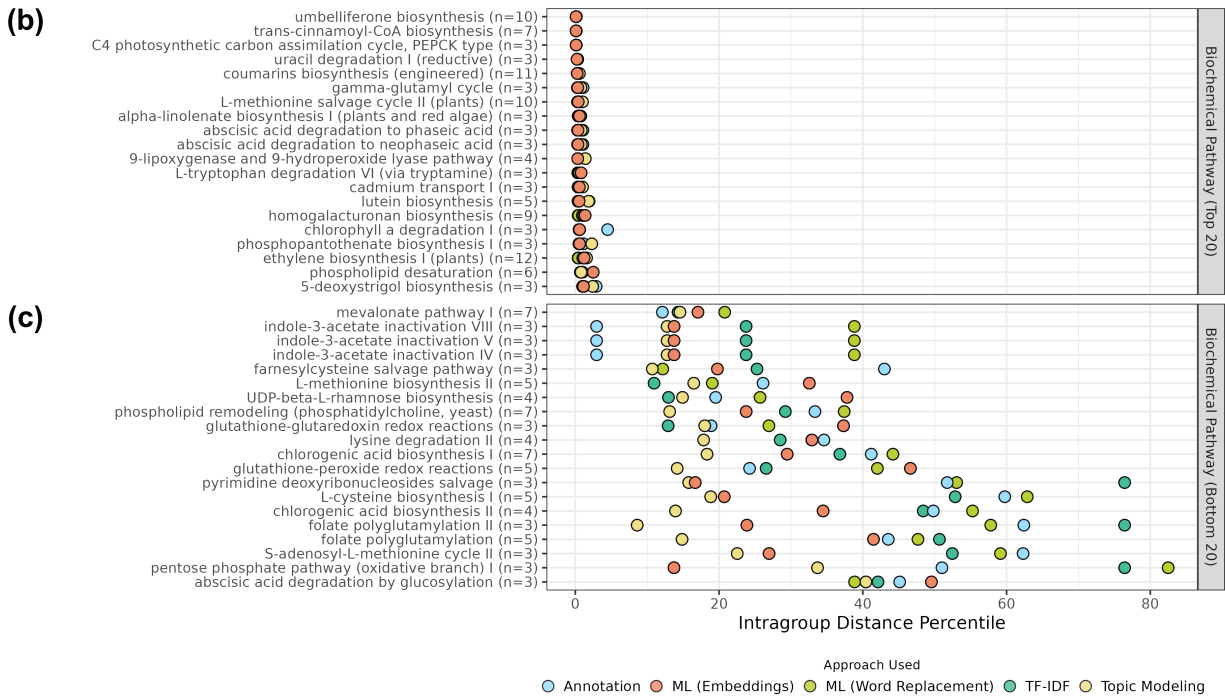


Figure 3.4 Cohesiveness of phenotype and pathway gene groups *continued*. Cohesiveness of phenotype and pathway gene groups continued. Biochemical Pathways (b) and (c) are listed, with the number of genes in these datasets belonging to each group listed to the right of the group’s name. The x-axis indicates group cohesiveness, given as the percentile against all pairwise gene distances that the average distance between any two genes in that group falls in. The minimum value of this metric achieved by any approach that is in the listed category is shown. For example, the location of the yellow dot in a particular row indicates the smallest intragroup distance percentile obtained by any approach in the topic modeling category of text-based approaches for that particular group of genes.

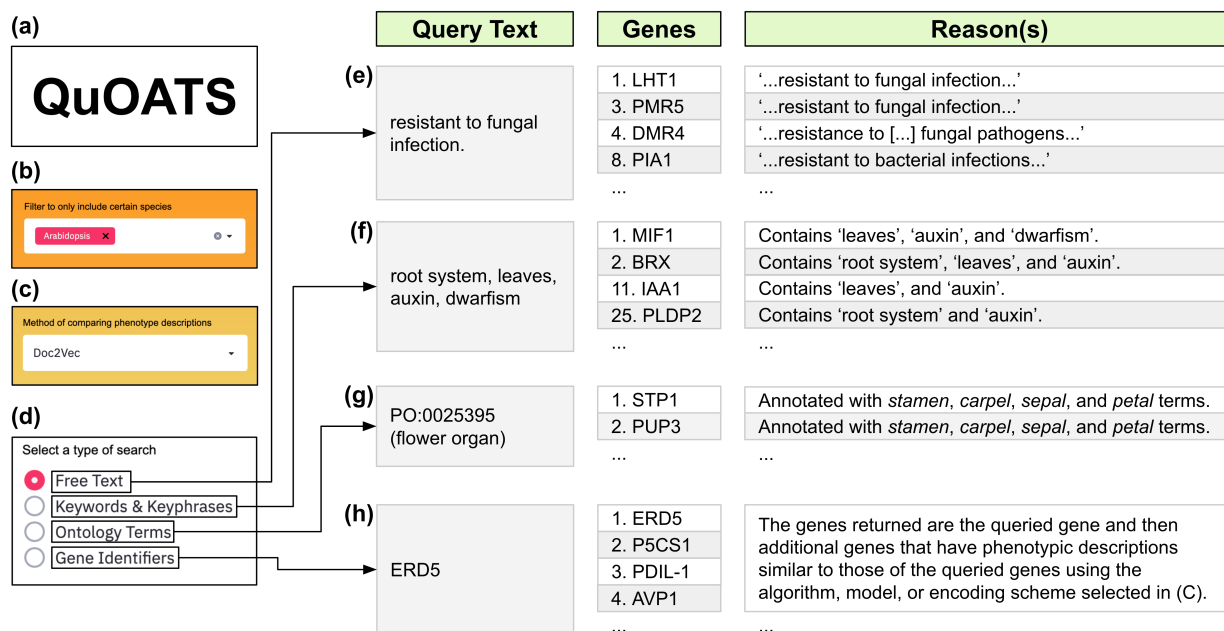


Figure 3.5 Querying plant genes, annotations, and phenotype descriptions. (a) The name of the web application we have developed. (b) Option to subset the available dataset to only include certain species. (c) Option to select the algorithm or method used to compare phenotype descriptions. (d) Four different types of querying are supported. (e), (f), (g), (h). The information given here for each query type is presented when using the webtool, but has been re-organized and truncated for the sake of illustration. The queries listed are the text strings that are entered into the search bar to generate the results shown. The returned genes appear in the results in the row indicated by the number to the left of the gene names. The reasons that these genes appear in this order given these particular queries are described to the right of the gene names.

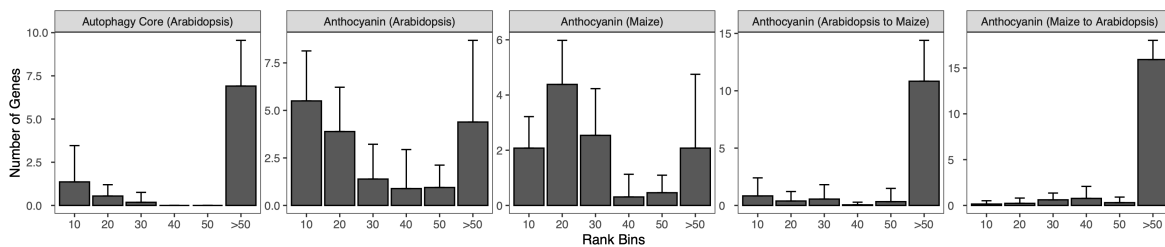


Figure 3.6 Querying with autophagy core genes and anthocyanin biosynthesis genes in QuOATS. The labels above each plot indicate the set of genes, the species of the genes used as the queries, and then the species for which the resulting ranked genes were filtered (in the case of the left three plots the species is the same for queries and targets). Bars represent bins of rank values for returned genes. Their height indicates the average number of genes with those ranks returned in each query. The error bar indicates the standard deviation in each case. Bars in each plot are labeled with the rank that falls in the right-most edge of that bin. For example, the bar labelled 20 represents genes that were ranked between 11 and 20 in the query results.

Table 3.1 Scope and scale of the complete dataset.

Species	Number of Genes		Phenotype Descriptions			Annotations			Other Databases			
	Total	Sentences	Total Words	Unique Words	Unique to Species	Mapped to EQ Stmt(s)	Mapped to GO Term(s)	Mapped to PO Term(s)	Mapped in PlantCyc	Mapped in KEGG	Mapped in STRING	Mapped in PANTHER
Arabidopsis	6,274	30,123	261,422	6,792	5,084	2,393	5,984	4,395	843	1,435	4,317	377
Maize	1,405	7,512	47,139	1,722	498	114	190	761	133	157	229	443
Rice	92	478	3,689	760	97	92	46	92	3	0	45	86
Tomato	69	359	1,678	552	99	72	25	64	2	18	11	10
Medicago	37	263	2,447	671	123	40	30	32	2	0	13	0
Soybean	30	62	222	78	12	30	28	27	0	1	5	0
Total	7,907	38,797	316,597	7,663	5,913	2,741	6,303	5,371	983	1,611	4,620	916

Table 3.2 Biological relationships tested in each task.

Task	Description (Are genes A and B...)	Knowledge Source
Phenotypes	...impacting the same phenotype?	Lloyd and Meinke, 2012
Pathways	...functioning in the same pathway?	PlantCyc, KEGG
Associations	...known to share a function or process?	STRING
Orthologs	...orthologous to one another?	PANTHER

Table 3.3 Number of genes and gene pairs used for each task.

Question	All Text Data				With Annotations			
	Genes	Pairs	Positive Pairs		Genes	Pairs	Positive Pairs	
Phenotypes	2,356	2,774,190	303,009	10.92%	2,284	2,607,186	293,221	11.25%
Pathways	1,838	1,688,203	45,847	2.72%	1,045	545,490	14,853	2.72%
Associations	4,620	9,343,325	147,271	1.58%	2,377	2,530,556	52,541	2.08%
Orthologs	921	248,913	65	0.03%	368	43,187	23	0.05%

Table 3.4 Similarities among datasets across biological tasks.

Task 1	Task 2	Overlap Size	Jaccard (Pairs)	Jaccard (Values)
Associations	Pathways	1,271,297	0.130	0.172
Phenotypes	Pathways	511,566	0.129	0.050
Phenotypes	Associations	2,687,721	0.285	0.032
Pathways	Orthologs	29,654	0.016	0.012
Phenotypes	Orthologs	0	0.000	
Associations	Orthologs	0	0.000	

Table 3.5 Comparing F_1 scores and group significance rates for phenotype and pathway relationships.

Approach	Category	Concat	Phenotypes (F_1 , % Significant)				Pathways (F_1 , % Significant)			
			All Genes	Curated	All Genes	Curated				
Baseline	Baseline	Yes	0.197	17%	0.202	19%	0.052	7%	0.053	3%
TF-IDF (Unigrams)	TF-IDF	Yes	0.465	100%	0.473	100%	0.105	61%	0.114	57%
TF-IDF (Unigrams & Bigrams)	TF-IDF	Yes	0.473	100%	0.482	100%	0.109	64%	0.120	58%
TF-IDF (Plant Article Unigrams)	TF-IDF	Yes	0.462	100%	0.471	100%	0.100	59%	0.110	59%
NOBLE Coder (Precise)	Annotation	Yes	0.364	91%	0.370	91%	0.075	34%	0.079	23%
NOBLE Coder (Partial)	Annotation	Yes	0.372	100%	0.380	100%	0.087	42%	0.094	33%
LDA (50 Topics)	Topic Modeling	Yes	0.327	88%	0.336	88%	0.075	12%	0.088	24%
LDA (100 Topics)	Topic Modeling	Yes	0.329	88%	0.337	91%	0.072	25%	0.083	26%
NMF (50 Topics)	Topic Modeling	Yes	0.436	100%	0.448	100%	0.090	44%	0.103	46%
NMF (100 Topics)	Topic Modeling	Yes	0.418	100%	0.423	100%	0.093	49%	0.102	48%
Doc2Vec (Wikipedia)	ML (Embeddings)	Yes	0.313	86%	0.322	86%	0.068	22%	0.068	15%
Doc2Vec (Plants)	ML (Embeddings)	Yes	0.237	45%	0.240	50%	0.059	20%	0.060	15%
Word2Vec (Wikipedia)	ML (Embeddings)	Yes	0.276	98%	0.284	98%	0.087	12%	0.096	21%
Word2Vec (PubMed)	ML (Embeddings)	Yes	0.320	93%	0.327	98%	0.093	14%	0.108	21%
Word2Vec (Plants)	ML (Embeddings)	Yes	0.443	98%	0.452	98%	0.102	26%	0.111	36%
BERT	ML (Embeddings)	Yes	0.289	93%	0.296	95%	0.084	4%	0.096	15%
BioBERT	ML (Embeddings)	Yes	0.309	81%	0.316	83%	0.087	5%	0.098	17%
Word2Vec (Wikipedia)	ML (Word Replacement)	Yes	0.387	98%	0.396	100%	0.099	46%	0.107	43%
Word2Vec (PubMed)	ML (Word Replacement)	Yes	0.417	100%	0.426	100%	0.104	50%	0.111	45%
Word2Vec (Plant Phenotypes)	ML (Word Replacement)	Yes	0.456	98%	0.464	100%	0.103	59%	0.114	53%
Baseline	Baseline	No	0.465	74%	0.476	76%	0.089	19%	0.097	39%
TF-IDF (Unigrams)	TF-IDF	No	0.544	95%	0.555	100%	0.100	51%	0.108	61%
TF-IDF (Unigrams & Bigrams)	TF-IDF	No	0.540	95%	0.552	95%	0.100	58%	0.107	63%
TF-IDF (Plant Article Unigrams)	TF-IDF	No	0.555	98%	0.562	100%	0.098	33%	0.106	52%
NOBLE Coder (Precise)	Annotation	No	0.457	83%	0.466	83%	0.089	5%	0.103	33%
NOBLE Coder (Partial)	Annotation	No	0.501	95%	0.510	98%	0.095	24%	0.102	47%
NMF (50 Topics)	Topic Modeling	No	0.509	88%	0.520	88%	0.092	13%	0.099	30%
NMF (100 Topics)	Topic Modeling	No	0.517	93%	0.527	95%	0.090	18%	0.100	37%
LDA (50 Topics)	Topic Modeling	No	0.500	98%	0.511	100%	0.089	11%	0.099	35%
LDA (100 Topics)	Topic Modeling	No	0.500	98%	0.510	100%	0.097	19%	0.104	41%
Doc2Vec (Wikipedia)	ML (Embeddings)	No	0.518	100%	0.529	100%	0.100	23%	0.107	51%
Doc2Vec (Plants)	ML (Embeddings)	No	0.561	95%	0.571	95%	0.098	26%	0.104	52%
Word2Vec (Wikipedia)	ML (Embeddings)	No	0.521	98%	0.532	98%	0.098	12%	0.106	38%
Word2Vec (PubMed)	ML (Embeddings)	No	0.529	95%	0.540	100%	0.100	19%	0.108	42%
Word2Vec (Plants)	ML (Embeddings)	No	0.558	98%	0.569	98%	0.103	30%	0.111	51%
BERT	ML (Embeddings)	No	0.500	100%	0.511	100%	0.102	25%	0.111	43%
BioBERT	ML (Embeddings)	No	0.515	100%	0.526	100%	0.104	24%	0.113	48%
Word2Vec (Wikipedia)	ML (Word Replacement)	No	0.558	100%	0.568	100%	0.102	33%	0.110	53%
Word2Vec (PubMed)	ML (Word Replacement)	No	0.554	100%	0.566	100%	0.101	34%	0.109	51%
Word2Vec (Plant Phenotypes)	ML (Word Replacement)	No	0.566	95%	0.577	93%	0.099	37%	0.107	51%
GO	Curation				0.249	64%			0.140	38%
PO	Curation				0.215	17%			0.056	9%
EQs	Curation				0.475	74%			0.093	51%

3.13 Appendix: Supplementary Tables

Table 3.6 Comparing F_1 scores for associations and orthologous gene pair relationships.

Approach	Category	Concat	Associations (F_1)		Orthologs (F_1)	
			All Genes	Curated	All Genes	Curated
Baseline	Baseline	Yes	0.031	0.041	0.001	0.001
TF-IDF (Unigrams)	TF-IDF	Yes	0.049	0.068	0.010	0.061
TF-IDF (Unigrams & Bigrams)	TF-IDF	Yes	0.051	0.072	0.016	0.054
TF-IDF (Plant Article Unigrams)	TF-IDF	Yes	0.048	0.067	0.008	0.057
NOBLE Coder (Precise)	Annotation	Yes	0.037	0.049	0.012	0.022
NOBLE Coder (Partial)	Annotation	Yes	0.042	0.060	0.003	0.016
LDA (50 Topics)	Topic Modeling	Yes	0.039	0.053	0.002	0.005
LDA (100 Topics)	Topic Modeling	Yes	0.038	0.053	0.005	0.004
NMF (50 Topics)	Topic Modeling	Yes	0.042	0.060	0.007	0.013
NMF (100 Topics)	Topic Modeling	Yes	0.043	0.061	0.006	0.020
Doc2Vec (Wikipedia)	ML (Embeddings)	Yes	0.033	0.047	0.015	0.029
Doc2Vec (Plants)	ML (Embeddings)	Yes	0.031	0.041	0.001	0.007
Word2Vec (Wikipedia)	ML (Embeddings)	Yes	0.042	0.059	0.003	0.003
Word2Vec (PubMed)	ML (Embeddings)	Yes	0.042	0.060	0.006	0.007
Word2Vec (Plants)	ML (Embeddings)	Yes	0.052	0.070	0.012	0.065
BERT	ML (Embeddings)	Yes	0.045	0.059	0.003	0.003
BioBERT	ML (Embeddings)	Yes	0.046	0.062	0.009	0.020
Word2Vec (Wikipedia)	ML (Word Replacement)	Yes	0.046	0.065	0.006	0.029
Word2Vec (PubMed)	ML (Word Replacement)	Yes	0.048	0.066	0.023	0.133
Word2Vec (Plant Phenotypes)	ML (Word Replacement)	Yes	0.048	0.069	0.018	0.080

Table 3.6 Comparing F_1 scores for associations and orthologous gene pair relationships
continued.

Approach	Category	Concat	Associations (F_1)		Orthologs (F_1)	
			All Genes	Curated	All Genes	Curated
Baseline	Baseline	No	0.044	0.067	0.004	0.008
TF-IDF (Unigrams)	TF-IDF	No	0.051	0.072	0.006	0.008
TF-IDF (Unigrams & Bigrams)	TF-IDF	No	0.051	0.072	0.005	0.007
TF-IDF (Plant Article Unigrams)	TF-IDF	No	0.051	0.073	0.004	0.007
NOBLE Coder (Precise)	Annotation	No	0.051	0.068	0.004	0.003
NOBLE Coder (Partial)	Annotation	No	0.048	0.069	0.004	0.005
NMF (50 Topics)	Topic Modeling	No	0.049	0.070	0.004	0.007
NMF (100 Topics)	Topic Modeling	No	0.049	0.071	0.005	0.006
LDA (50 Topics)	Topic Modeling	No	0.047	0.069	0.005	0.006
LDA (100 Topics)	Topic Modeling	No	0.048	0.069	0.006	0.008
Doc2Vec (Wikipedia)	ML (Embeddings)	No	0.051	0.071	0.007	0.008
Doc2Vec (Plants)	ML (Embeddings)	No	0.051	0.071	0.006	0.010
Word2Vec (Wikipedia)	ML (Embeddings)	No	0.049	0.070	0.006	0.005
Word2Vec (PubMed)	ML (Embeddings)	No	0.048	0.071	0.007	0.009
Word2Vec (Plants)	ML (Embeddings)	No	0.053	0.074	0.006	0.008
BERT	ML (Embeddings)	No	0.048	0.070	0.005	0.008
BioBERT	ML (Embeddings)	No	0.048	0.071	0.005	0.008
Word2Vec (Wikipedia)	ML (Word Replacement)	No	0.052	0.073	0.005	0.007
Word2Vec (PubMed)	ML (Word Replacement)	No	0.052	0.073	0.005	0.006
Word2Vec (Plant Phenotypes)	ML (Word Replacement)	No	0.050	0.072	0.006	0.007
GO	Curation			0.094		0.059
PO	Curation			0.048		0.001
EQs	Curation			0.063		0.014

Table 3.7 Comparing F_1 scores for pathways for intraspecies and interspecies gene pairs.

Approach	Category	Concat	Pathways, All Genes (F_1)		Pathways, Curated (F_1)	
			Intraspecies	Interspecies	Intraspecies	Interspecies
Baseline	Baseline	Yes	0.053	0.051	0.054	0.049
TF-IDF (Unigrams)	TF-IDF	Yes	0.107	0.067	0.116	0.094
TF-IDF (Unigrams & Bigrams)	TF-IDF	Yes	0.111	0.069	0.123	0.097
TF-IDF (Plant Article Unigrams)	TF-IDF	Yes	0.102	0.067	0.112	0.092
NOBLE Coder (Precise)	Annotation	Yes	0.078	0.055	0.082	0.073
NOBLE Coder (Partial)	Annotation	Yes	0.088	0.058	0.097	0.072
LDA (50 Topics)	Topic Modeling	Yes	0.084	0.060	0.089	0.076
LDA (100 Topics)	Topic Modeling	Yes	0.078	0.060	0.086	0.065
NMF (50 Topics)	Topic Modeling	Yes	0.092	0.073	0.104	0.097
NMF (100 Topics)	Topic Modeling	Yes	0.091	0.068	0.104	0.081
Doc2Vec (Wikipedia)	ML (Embeddings)	Yes	0.069	0.051	0.070	0.062
Doc2Vec (Plants)	ML (Embeddings)	Yes	0.060	0.055	0.062	0.049
Word2Vec (Wikipedia)	ML (Embeddings)	Yes	0.089	0.053	0.102	0.067
Word2Vec (PubMed)	ML (Embeddings)	Yes	0.095	0.056	0.114	0.074
Word2Vec (Plants)	ML (Embeddings)	Yes	0.105	0.071	0.115	0.107
BERT	ML (Embeddings)	Yes	0.087	0.052	0.102	0.059
BioBERT	ML (Embeddings)	Yes	0.089	0.051	0.104	0.060
Word2Vec (Wikipedia)	ML (Word Replacement)	Yes	0.100	0.063	0.110	0.088
Word2Vec (PubMed)	ML (Word Replacement)	Yes	0.105	0.062	0.114	0.088
Word2Vec (Plant Phenotypes)	ML (Word Replacement)	Yes	0.106	0.071	0.115	0.108
Baseline	Baseline	No	0.091	0.051	0.101	0.049
TF-IDF (Unigrams)	TF-IDF	No	0.102	0.067	0.109	0.099
TF-IDF (Unigrams & Bigrams)	TF-IDF	No	0.102	0.069	0.109	0.093
TF-IDF (Plant Article Unigrams)	TF-IDF	No	0.100	0.066	0.107	0.098
NOBLE Coder (Precise)	Annotation	No	0.093	0.057	0.106	0.081
NOBLE Coder (Partial)	Annotation	No	0.096	0.058	0.105	0.069
NMF (50 Topics)	Topic Modeling	No	0.094	0.054	0.102	0.070
NMF (100 Topics)	Topic Modeling	No	0.093	0.058	0.103	0.070
LDA (50 Topics)	Topic Modeling	No	0.091	0.056	0.102	0.069
LDA (100 Topics)	Topic Modeling	No	0.098	0.070	0.107	0.077
Doc2Vec (Wikipedia)	ML (Embeddings)	No	0.103	0.056	0.110	0.070
Doc2Vec (Plants)	ML (Embeddings)	No	0.101	0.063	0.106	0.077
Word2Vec (Wikipedia)	ML (Embeddings)	No	0.100	0.055	0.108	0.069
Word2Vec (PubMed)	ML (Embeddings)	No	0.103	0.060	0.112	0.082
Word2Vec (Plants)	ML (Embeddings)	No	0.106	0.072	0.113	0.104
BERT	ML (Embeddings)	No	0.104	0.057	0.115	0.069
BioBERT	ML (Embeddings)	No	0.106	0.057	0.116	0.079
Word2Vec (Wikipedia)	ML (Word Replacement)	No	0.104	0.070	0.112	0.102
Word2Vec (PubMed)	ML (Word Replacement)	No	0.104	0.064	0.111	0.090
Word2Vec (Plant Phenotypes)	ML (Word Replacement)	No	0.103	0.073	0.108	0.108
GO	Curation				0.137	0.191
PO	Curation				0.057	0.107
EQs	Curation				0.097	0.049

3.14 Appendix: Consent To Include Co-Authored Article in Dissertation

Consent to include co-authored article in dissertation.**THE PARTIES**

Student Author (Full Name, Major, and Institution)	Colleen Frances Yanarella, Bioinformatics and Computational Biology, Iowa State University
List other student co-authors and their institutions.	Ian Robert Braun, Bioinformatics and Computational Biology, Iowa State University
Title(s) of the co- authored section (Chapter, etc.)	The Case for Retaining Natural Language Descriptions of Phenotypes in Plant Databases and a Web Application as Proof of Concept
Journal Name, Book Title, etc. (if applicable)	Database

DISTRIBUTION OF TASKS AND RESPONSIBILITIES

In this research publication, I, Colleen F. Yanarella, was responsible for the following roles:
(Select all roles that apply.)

- Conceptualization
- Data curation
- Formal analysis
- Funding acquisition
- Investigation
- Methodology
- Resources
- Software
- Supervision
- Validation
- Visualization
- Writing – original draft
- Writing – review & editing
- Other: Please describe briefly: Directed reproducibility review.

The CRediT taxonomy is taken from <https://credit.niso.org/>. Go to the link to see the descriptions of contributor roles.

CHAPTER 4. WISCONSIN DIVERSITY PANEL PHENOTYPES: SPOKEN DESCRIPTIONS OF PLANTS AND SUPPORTING DATA

Colleen F. Yanarella¹, Leila Fattel¹, Ásrún Ý. Kristmundsdóttir¹, Miriam D. Lopez², Jode W. Edwards², Darwin A. Campbell¹, Craig A. Abel², and Carolyn J. Lawrence-Dill¹

¹ Iowa State University, Ames, IA, 50011, USA

² USDA ARS, Ames, IA, 50011, USA

Modified from a manuscript under review in *BMC Research Notes*

4.1 Abstract

4.1.1 Objectives

Phenotyping plants in a field environment can involve a variety of methods including the use of automated instruments and labor-intensive manual measurement and scoring. Researchers also collect language-based phenotypic descriptions and use controlled vocabularies and structures such as ontologies to enable computation on descriptive phenotype data, including methods to determine phenotypic similarities. In this study, spoken descriptions of plants were collected and observers were instructed to use their own vocabulary to describe plant features that were present and visible. Further, these plants were measured and scored manually as part of a larger study to investigate whether spoken plant descriptions can be used to recover known biological phenomena.

4.1.2 Data description

Data comprise phenotypic observations of 686 accessions of the maize Wisconsin Diversity panel, and 25 positive control accessions that carry visible, dramatic phenotypes. The data include the list of accessions planted, field layout, data collection procedures, student participants' and volunteers' observation transcripts, volunteers' audio data files, terrestrial and aerial images

of the plants, Amazon Web Services method selection experimental data, and manually collected (measurements and scores) phenotypes (e.g., plant height, ear and tassel features, etc.). Data were collected during the summer of 2021 at Iowa State University’s Agricultural Engineering and Agronomy Research Farms.

4.2 Keywords

Phenotyping, Maize, Association Studies, Audio Recordings, Text Transcripts, Images, Wisconsin Diversity panel

4.3 Objective

Formative research using free text descriptions of plant phenotypes along with Natural Language Processing (NLP) methods has demonstrated that computing on plant phenotypes alone can recover known genotype-phenotype associations (Oellrich et al. (2015), Braun and Lawrence-Dill (2020)). Building on these successes, continued efforts to generate plant phenotype descriptions that are both structured (e.g., ontologies) and unstructured (i.e., free text) hold great promise for enabling researchers to advance analytics for phenotypes and traits, especially when these data are made publicly accessible (Braun et al. (2021)).

We developed this dataset as a foundation for analyzing large volumes of spoken phenotype descriptions in a field environment. These phenotype observations were drawn from the Wisconsin Diversity panel, which contains sufficient phenotypic diversity in a field environment for various genotype-to-phenotype analyses (Hansey et al. (2011), Hirsch et al. (2014), Mazaheri et al. (2019)). Observers generating the datasets were not confined to rigid vocabularies and were not strictly limited to a list of traits to comment on.

Supplemental to spoken descriptions of plant phenotypes and the text derived from these observations, measurements and scores for traits of interest were also collected as ground truth. Field layout and weather data are reported, along with images of the rows in the field and aerial images from a drone. Consequently, this dataset may be useful to investigators interested in data

collected from diversity panels and to those interested in processing natural language and its use in describing scientific phenomena.

Additional computational work on this dataset is unpublished, but a manuscript is in preparation that demonstrates the use of the dataset for investigating biological relevance and utility, including developed tools to assist in the use of spoken descriptions for field-based plant phenotype analytics.

4.4 Data description

This dataset ([Yanarella et al. \(2023\)](#)) was collected and derived from observations of an experimental field at Iowa State University’s Agricultural Engineering and Agronomy Research Farms in Boone, Iowa. The Wisconsin Diversity panel (686 accessions), an environmental control line (B73, the maize reference line used for genetics and genomics), and 25 positive control accessions were planted in two replicates, and observations were generated over the summer of 2021. This dataset includes the following elements ([Table 4.1](#)).

- Audio text processing data contains the spoken data collected by the volunteers (WAV files) and descriptions of the recordings generated by student participants using Sony ICD-UX570 recorders. Additionally included are metadata (summary statistics) derived from the recordings and code to generate these statistics. Further, all intermediate files (JSON, TXT, and EXCEL files) and code to generate the final cleaned transcripts for all student participant recordings and a subset of the volunteer’s recordings are included. These files provide a resource to investigators to utilize field-collected spoken natural language descriptions of maize plants.
- Methods selection data includes data and code for generating transcriptions using various Amazon Web Services (AWS) Transcribe methods. These methods include using an individualized custom vocabulary for each student participant and an example of the process using volunteer Whiskey’s data, a generalized custom vocabulary for each student

participant and volunteer Whiskey's data, and no custom vocabulary. A subset of data was selected to process and compare to a gold standard transcription manually generated to calculate a similarity score to determine the method for transcribing all spoken descriptions collected during the summer of 2021.

- A Canon EOS Rebel T7 camera and Cannon EF-S 18-55mm Image Stabilizer Macro 0.25m/0.8ft set to 18mm with AF Stabilizer ON were used to capture images of each row in the field.
- A Mavic 2 Pro drone by DJI was used to capture aerial still images and footage of the experimental field.
- The field data information layout demonstrates the randomizations of the accessions planted and positive controls used. Additionally, seeds planted per row and the Iowa Phytosanitary Corn Field Inspection report conducted on the experimental field are included.
- Field prompting data includes the cards provided to student participants to prompt their behavior while collecting spoken observations in the field, information about the assigned card for each day, and logs for worker data collection. Volunteer field guide cards and logs for data collection are present. Each student participant was instructed to make three complete passes of the field, and volunteers were instructed to make one complete pass of the field. Volunteer India completed observations for replicate one only.
- Measurement and scoring data were collected through manual measuring and scoring by student participants and a volunteer. Plastic measuring sticks with hash marks every 10 cm were used to measure plant height.
- Weather information was collected and reported by The Iowa Environmental Mesonet through Iowa State University ([Herzmann \(2023\)](#)). Data for the weather stations nearest the Agricultural Engineering and Agronomy Research Farms for March 2021 to September 2021 and September 2020 to September 2021 are provided.

4.5 Limitations

Some audio observations were incomplete due to technical difficulties, including microphone disengagement from the recording devices or observers recording observations for the incorrect row. Also, speech-to-text pipelines and post-process cleaning steps are fallible, leading to transcription inaccuracies. These data were taken over approximately seven weeks, and there were apparent growth and developmental changes throughout the duration of the study. Additionally, the observations within this dataset are for two replicates in the same environment, and additional years, plots, and environments could supplement these available speech data for a more robust dataset.

4.6 Abbreviations

AWS: Amazon Web Services

NLP: Natural Language Processing

4.7 Availability of data and materials

The data described in this Data Note can be freely and openly accessed on CyVerse under Digital Object Identifiers (DOI) <https://doi.org/10.25739/pvx4-5j31>. Please see Table 4.1 and the references list for details and links to the data.

4.8 Declarations

4.8.1 Ethics approval and consent to participate

The ethics approval for this study was waived by Iowa State University (ISU), Ames, Iowa Internal Review Board (IRB) (Study Number 21-179-00). Informed consent was obtained from all participants involved in the behavioral component of this study. All participants' data remains de-identified, and their audio recordings cannot be released as ISU IRB requires for Study 21-179-00. All volunteer observers willingly consented to participate in spoken data collection. All

experiments were performed in accordance with relevant guidelines and regulations (such as the Declaration of Helsinki).

4.8.2 Consent for publication

Not applicable.

4.8.3 Competing interests

The authors declare no competing interests.

4.9 Funding

This work was supported by an Iowa State University Plant Sciences Institute Faculty Scholars Award (CJLD) and the Iowa State Predictive Plant Phenomics NSF Research Traineeship (DGE-1545453; CJLD is a co-principal investigator, and CFY is a trainee). The Predictive Plant Phenomics Program provided the Program Trainee Research Materials and Supplies Grant and the Program Small Research Grant (CFY). This article was also supported by the NSF and USDA-NIFA AI Institute for Resilient Agriculture (#2021-67021-35329) to CJLD and supporting CFY and LF. This article is also a product of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. IOW04714 (CJLD) which is supported by USDA/NIFA and State of Iowa funds. This research was supported in part by the US Department of Agriculture, Agricultural Research Service (USDA ARS) Project No. 5030-21000-066-000D and 5030-21000-019-000D. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

4.10 Authors' contributions

CFY and CJLD conceived this work and wrote the manuscript. All authors read, offered suggestions, and approved the final version of the manuscript. CFY, LF, ÁÝK, and CJLD are volunteers who recorded spoken observations of plants for dissemination. CJLD, DAC, and CFY managed IRB compliance, obtained IRB exemption, and coordinated student participants and individuals in the behavioral research component of this work. CFY, MDL, and JWE performed randomizations and approved the field layout design. CJLD and CFY selected the maize panel germplasm. CAA, MDL, and JWE planted the seed and approved field management practices. CFY performed image acquisition, data processing, and data organization. CFY and LF managed DOI acquisition.

4.11 Acknowledgements

We acknowledge our summer 2021 student participants, known throughout this project by their NATO code names for generating descriptive recordings and collecting scoring and measurement data. Delta, Golf, Kilo, Lima, Mike, Quebec, Victor, Yankee, and Zulu, thank you for lending your voices to this dataset and research. We thank the research groups of Nick Lauter, Marna Yandean-Nelson, Dior Kelley, and Justin Walley for bulking the seed, making the seed available for this research, and planting. We thank Nick Lauter for his guidance on germplasm selection and his passion for working with the Wisconsin Diversity panel. We thank John Golden for assisting with seed treatment and planting. We thank Marty Sachs and the Maize Genetics Cooperation Stock Center for discussing germplasm for positive controls and making stock available for this research. We acknowledge the Predictive Plant Phenomics NSF NRT for a Program Trainee Research Materials and Supplies Grant to CFY for recording supplies, camera supplies, and lending us a drone. We thank Ashlyn Rairdin for instructing us on the use of drones for aerial images. We appreciate the Agricultural Engineering Agronomy and Central Iowa Research Farms field crew for preparing and managing the field. We appreciate the ISU Office of

Research Ethics for guidance on the safety of human research subjects and for implementing their COVID-19 mitigation plan.

4.12 References

- Braun, I. R. and Lawrence-Dill, C. J. (2020). Automated Methods Enable Direct Computation on Phenotypic Descriptions for Novel Candidate Gene Prediction. *Frontiers in Plant Science*, 10.
- Braun, I. R., Yanarella, C. F., Rajeswari, J. P. D., Bassham, D. C., and Lawrence-Dill, C. J. (2021). The Case for Retaining Natural Language Descriptions of Phenotypes in Plant Databases and a Web Application as Proof of Concept. *bioRxiv*.
- Hansey, C. N., Johnson, J. M., Sekhon, R. S., Kaeppler, S. M., and de Leon, N. (2011). Genetic Diversity of a Maize Association Population with Restricted Phenology. *Crop Science*, 51(2):704–715.
- Herzmann, D. (2023). Iowa Environmental Mesonet.
- Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., de Leon, N., Kaeppler, S. M., and Buell, C. R. (2014). Insights into the Maize Pan-Genome and Pan-Transcriptome. *The Plant Cell*, 26(1):121–135.
- Mazaheri, M., Heckwolf, M., Vaillancourt, B., Gage, J. L., Burdo, B., Heckwolf, S., Barry, K., Lipzen, A., Ribeiro, C. B., Kono, T. J. Y., Kaeppler, H. F., Spalding, E. P., Hirsch, C. N., Buell, C. R., de Leon, N., and Kaeppler, S. M. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. *BMC Plant Biology*, 19(1).
- Oellrich, A., Walls, R. L., Cannon, E. K., Cannon, S. B., Cooper, L., Gardiner, J., Gkoutos, G. V., Harper, L., He, M., Hoehndorf, R., Jaiswal, P., Kalberer, S. R., Lloyd, J. P., Meinke, D., Menda, N., Moore, L., Nelson, R. T., Pujar, A., Lawrence, C. J., and Huala, E. (2015). An ontology approach to comparative phenomics in plants. *Plant Methods*, 11(1).
- Yanarella, C. F., Fattel, L., Ásrún Ý. Kristmundsdóttir, Lopez, M. D., Edwards, J. W., Campbell, D. A., Abel, C. A., and Lawrence-Dill, C. J. (2023). Carolyn_Lawrence_Dill_Maize_WiDiv_Summer_2021_Dataset_June_2023.

4.13 Figures and Tables

Table 4.1 Overview of data files/data sets.

Label	Name of data file /data set	File types (file extension)	Data repository and identifier (DOI or accession number)
2021 Wisconsin Diversity Panel Dataset	Carolyn_Lawrence_Dill _Maize_WiDiv_Summer _2021_Dataset_June_2023	Directory	CyVerse (Yanarella et al. (2023)) (https://doi.org/10.25739/pvx4-5j31)
	/README.txt	A file with file type: .txt Provides details regarding the subdirectories.	CyVerse (Yanarella et al. (2023)) (https://doi.org/10.25739/pvx4-5j31)
	/audio_text _processing_data	A subdirectory containing file types: .csv, .json, .py, .tar.gz, .txt, .xlsx, and .yaml Demonstrates processing of spoken observations to text files.	CyVerse (Yanarella et al. (2023)) (https://doi.org/10.25739/pvx4-5j31)
	/aws_method _selection	A subdirectory containing file types: .csv, .json, .out, .py, .R, .txt, .wav, .xlsx, and .yaml Demonstrates methods for transcribing speech-to-text using AWS.	CyVerse (Yanarella et al. (2023)) (https://doi.org/10.25739/pvx4-5j31)

Table 4.1 Overview of data files/data sets *continued (1)*.

Label	Name of data file /data set	File types (file extension)	Data repository and identifier (DOI or accession number)
	/camera_images	A subdirectory containing file types: .pptx, .tar.gz, .txt, and .xlsx Contains images of the experimental field rows.	CyVerse (Yanarella et al. (2023)) (https://doi.org/10.25739/pvx4-5j31)
	/drone_images	A subdirectory containing file types: .jpg, .mp4, and .txt Consists of drone-captured still images and videos of the experimental field.	CyVerse (Yanarella et al. (2023)) (https://doi.org/10.25739/pvx4-5j31)
	/field_layout_and_seed _information	A subdirectory containing file types: .pdf, .txt, and .xlsx Includes information about taxa planted and location of taxa in the experimental field.	CyVerse (Yanarella et al. (2023)) (https://doi.org/10.25739/pvx4-5j31)
	/field_prompting _cards_information	A subdirectory containing file types: .png, .psd, .txt, and .xlsx Demonstrates field prompting cards and directions provided to observers.	CyVerse (Yanarella et al. (2023)) (https://doi.org/10.25739/pvx4-5j31)

Table 4.1 Overview of data files/data sets *continued (2)*.

Label	Name of data file /data set	File types (file extension)	Data repository and identifier (DOI or accession number)
	/measurement_and _scoring_data	A subdirectory containing file types: .txt and .xlsx Records of measurement and scoring data of the experimental field.	CyVerse (Yanarella et al. (2023)) (https://doi.org /10.25739/pvx4-5j31)
	/weather_data	A subdirectory containing file types: .txt and .xlsx Reports weather data collected from stations in close proximity to the experimental field.	CyVerse (Yanarella et al. (2023)) (https://doi.org /10.25739/pvx4-5j31)

4.14 Appendix: Institutional Review Board Exemption Letter

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office of Research Ethics
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

Date: 06/04/2021
To: Carolyn Lawrence-Dill
From: Office of Research Ethics
Title: CFY Field Project 2021
IRB ID: 21-179
Submission Type: Initial Submission
Exemption Date: 05/18/2021

The project referenced above has been declared exempt from most requirements of the human subject protections regulations as described in 45 CFR 46.104 or 21 CFR 56.104 because it meets the following federal requirements for exemption:

2018 - 3 (i.B): Research involving benign behavioral interventions in conjunction with the collection of information from an adult subject through verbal or written responses or audiovisual recording when the subject prospectively agrees to the intervention and information collection and any disclosure of the human subjects' responses outside the research would not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation. - 3 (ii) If research involves deception, it is prospectively authorized by the subject.

The determination of exemption means that:

- **You do not need to submit an application for continuing review. Instead, you will receive a request for a brief status update every three years. The status update is intended to verify that the study is still ongoing.**
- **You must carry out the research as described in the IRB application.** Review by IRB staff is required prior to implementing modifications that may change the exempt status of the research. In general, review is required for any *modifications to the research procedures* (e.g., method of data collection, nature or scope of information to be collected, nature or duration of behavioral interventions, use of deception, etc.), any change in *privacy or confidentiality protections*, modifications that result in the *inclusion of participants from vulnerable populations*, removing plans for informing participants about the study, any *change that may increase the risk or discomfort to participants*, and/or any change such that the revised procedures do not fall into one or more of the [regulatory exemption categories](#). The purpose of review is to determine if the project still meets the federal criteria for exemption.
- **All changes to key personnel** must receive prior approval.
- **Promptly inform the IRB of any addition of or change in federal funding for this study.** Approval of the protocol referenced above applies only to funding sources that are specifically identified in the corresponding IRB application.

IRB 07/2020

Detailed information about requirements for submitting modifications for exempt research can be found on our [website](#). For modifications that require prior approval, an amendment to the most recent IRB application must be submitted in IRBManager. A determination of exemption or approval from the IRB must be granted before implementing the proposed changes.

Non-exempt research is subject to many regulatory requirements that must be addressed prior to implementation of the study. Conducting non-exempt research without IRB review and approval may constitute non-compliance with federal regulations and/or academic misconduct according to ISU policy.

Additionally:

- All research involving human participants must be submitted for IRB review. **Only the IRB or its designees may make the determination of exemption**, even if you conduct a study in the future that is exactly like this study.
- **Please inform the IRB if the Principal Investigator and/or Supervising Investigator end their role or involvement with the project** with sufficient time to allow an alternate PI/Supervising Investigator to assume oversight responsibility. Projects must have an [eligible PI](#) to remain open.
- **Immediately inform the IRB of (1) all serious and/or unexpected [adverse experiences](#) involving risks to subjects or others; and (2) any other [unanticipated problems involving risks](#) to subjects or others.**
- **Approval from other entities may also be needed.** For example, access to data from private records (e.g., student, medical, or employment records, etc.) that are protected by FERPA, HIPAA or other confidentiality policies requires permission from the holders of those records. Similarly, for research conducted in institutions other than ISU (e.g., schools, other colleges or universities, medical facilities, companies, etc.), investigators must obtain permission from the institution(s) as required by their policies. **An IRB determination of exemption in no way implies or guarantees that permission from these other entities will be granted.**
- Your research study may be subject to [post-approval monitoring](#) by Iowa State University's Office for Responsible Research. In some cases, it may also be subject to formal audit or inspection by federal agencies and study sponsors.
- Upon completion of the project, transfer of IRB oversight to another IRB, or departure of the PI and/or Supervising Investigator, please initiate a Project Closure in IRBManager to officially close the project. For information on instances when a study may be closed, please refer to the [IRB Study Closure Policy](#).

Please don't hesitate to contact us if you have questions or concerns at 515-294-4566 or IRB@iastate.edu.

CHAPTER 5. GWAS FROM SPOKEN PHENOTYPIC DESCRIPTIONS: A PROOF OF CONCEPT FROM MAIZE FIELD STUDIES

Colleen F. Yanarella^{1,2}, Leila Fattel^{1,3}, and Carolyn J. Lawrence-Dill^{1,2,3,4,5}

¹ Department of Agronomy, Iowa State University, Ames, IA 50011

² Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011

³ Interdepartmental Genetics and Genomics Program, Iowa State University, Ames, IA 50011

⁴ Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, IA 50011

⁵ College of Agriculture and Life Sciences, Iowa State University, Ames, IA 50011

Modified from a manuscript to be submitted to *G3: Genes|Genomes|Genetics*

5.1 Abstract

Speech-derived phenotypic descriptions analyzed using existing Genome-Wide Association Study (GWAS) methods recover genomic regions involved in the maize plant height trait, demonstrating that non-structured, spoken descriptions of phenotypes can be used for association genetics. We collected phenotypes of *Zea mays* by recording spoken descriptions of plant traits such as height, color, leaf width, and feel of the texture of the leaves. To examine the relevance of spoken phenotypic descriptions for association genetics, we phenotyped the Wisconsin Diversity panel and developed two methods to process these spoken descriptions. To measure semantic similarity, we generated a score that indicates how alike each observation is in meaning to the query "tall". For the second method, we binned manually scored phrases related to plant height, then assigned scores to each observation. These were compared to published genomic locations associated with plant height (and with data we manually collected). Both methods recover known plant height associations.

5.2 Keywords

Spoken Descriptions, Association Studies, Phenotyping, Maize, Genes, Plant Height

5.3 Introduction

Collecting phenotype data can be slow, which limits the speed of association genetics and genomic studies for trait improvement. High-throughput phenotyping methods are an area of development that concentrates on engineering sensors and unmanned vehicles to collect (mainly visual) data about traits of various crop species (reviewed in [Yang et al. \(2020\)](#)). These methods are beneficial for collecting large amounts of data in an automated fashion, but there are difficulties in deploying these tools in a field environment, and some traits are not detectable by images alone. Additionally, manually collecting phenotypes with pen-and-paper or tablet-and-stylus is time-consuming and generally requires predefined traits of interest. Sensors, imaging, and barcodes make data organization easier for large quantities of data ([Yao et al. \(2021\)](#), [Sarić et al. \(2022\)](#), [Kazic \(2020\)](#)). An underdeveloped area of in-field phenotyping ripe for exploration is using natural language descriptions of plants. Platforms exist where audio descriptions are recorded ([Kazic \(2020\)](#)). However, the biologically relevant data in spoken phenotypes thus far has remained inaccessible for association studies and other applications.

Natural language datasets for plant species are beneficial tools for investigating plant phenotypes; the development of these datasets was demonstrated by [Oellrich et al. \(2015\)](#). Using structured language data, such as ontologies or entity quality (EQ) statements ([Mungall et al. \(2010\)](#)) (where an entity is a feature, e.g., whole plant, and quality is a describer, e.g., dwarf-like) results in less intensive computations. Semantic or word-meaning similarity methods have also shown promise in ascertaining biologically meaningful genetic associations ([Braun et al. \(2020\)](#), [Braun and Lawrence-Dill \(2020\)](#)). Additionally, pre-trained models have enabled free-text descriptions of plant phenotypes for association studies based on semantics ([Braun et al. \(2021\)](#)). These developments in the computational processing of natural language plant phenotypes in an

unstructured text format contribute to conceptualizing methods for recording spoken descriptions of phenotypes.

We reasoned that a well-characterized diversity panel would be required to perform Genome-Wide Association Studies (GWAS) with field-collected natural language phenotype data, so we chose the Wisconsin Diversity (WiDiv) panel. WiDiv was developed to grow in the upper midwestern states of the United States, have a restricted phenology for flowering, and have genetic and phenotypic diversity (Hansey et al. (2011)). Research using the WiDiv panel have increased the included accessions, investigated flowering time and biomass yield traits, and generated genetic marker data for the panel (Hansey et al. (2011), Hirsch et al. (2014), Mazaheri et al. (2019), Mural et al. (2022b)).

Because the WiDiv panel data has expansive genotypic and phenotypic trait data available, we collected spoken descriptions of phenotypes for numerous traits (height, leaf width, color, etc.) during the summer of 2021 (Yanarella et al. (2023)). The objectives of this research were to (1) detect phenotype descriptions from spoken descriptions, (2) demonstrate techniques for extracting phenotype data for a proof of concept analysis involving the plant height trait for GWAS, (3) perform GWAS with phenotypes derived from speech, and (4) use available gene function data to review and assess known and novel gene trait associations.

5.4 Materials and Methods

We used a genotypic dataset that includes WiDiv panel taxa (lines). A dataset of 18 million SNP markers (Mural et al. (2022a)) obtained from RNA-Seq and resequencing techniques for 1,051 taxa (described in (Mural et al. (2022b))). Phenotypic datasets, described in a Data Note currently under review at *BMC Data Notes*, which contain 686 unique WiDiv panel taxa (Yanarella et al. (2023)), hereafter referred to as the Yanarella *et al.* dataset. This dataset contains an additional 25 taxa (Supplementary Table 1) that were positive controls for the analysis of spoken descriptions of phenotypes, as these plants were expected to have noticeable and describable phenotypes, though these taxa are not members of the WiDiv panel. Informed

consent for the spoken data from participants was collected per Iowa State University's Institutional Review Board's (IRB) Exempt Project status, and volunteers provided informed consent for using their spoken observations. Phenotypic data obtained from these data include measurements for plant height and spoken descriptions of plants grown in a field environment. The Mural *et al.* and Yanarella *et al.* datasets have an intersection of 653 taxa (Figure 5.1), which were used for further analyses.

5.4.1 Spoken Phenotype Collection Summary

Phenotype descriptions recorded by de-identified student workers in the Yanarella *et al.* dataset were analyzed. The field in which the recordings were taken included two replicates planted in a randomized incomplete block design. The first block consisted of 31 WiDiv panel taxa for a seed increase, and the second block consisted of 8 B73 experimental control rows, 25 positive control taxa, and 655 unique WiDiv panel taxa. The second block contained two rows of the WiDiv panel line MEF156-55-2. Therefore, the recordings were taken over 720 rows in each replicate.

Each of the de-identified student workers selected NATO code names. The students who were undergraduate Agronomy, Biology, and Genetics Majors at Iowa State University are known as "Delta," "Golf," "Kilo," "Lima," "Mike," "Quebec," "Victor," "Yankee," and "Zulu." Each participant was instructed to state their NATO code name and the row tag number before observing the plants in each row (Figure 5.2 (a)). This procedure ensured the participant's de-identified connection to the row number and spoken observation while enabling the parsing of each observation so that multiple row observations could be recorded in the same file.

5.4.2 Phenotype Detection and Descriptions

A subset of spoken observation transcripts containing the positive control accessions were parsed. 4-6 terms from the description and phenotype records for each accession drawn from MaizeGDB (Woodhouse *et al.* (2021)). One term from these lists was used to collect synonyms

from Merriam-Webster ([Merriam-Webster \(2023\)](#)) and WordHippo ([Kat IP Pty Ltd \(2008\)](#)) thesaurus services (Table 5.1). The number of rows containing at least one synonym related to each accession’s descriptions and phenotype records was calculated as a proportion to the number of observations for that accession.

5.4.3 Preprocessing Genotypic Dataset

Trait Analysis by aSSociation, Evolution and Linkage (TASSEL) Version 5.0 Standalone ([Bradbury et al. \(2007\)](#)) was used to convert the Mural *et al.* genotypic data from a variant call format (vcf) formatted file to a HapMap (hmp) formatted file. The data was then processed to contain marker information for the 653 taxa shared with the Yanarella *et al.* dataset, these data were grouped by chromosome, and HapMap files were generated for each chromosome. The chromosome files were sorted by maker position from lowest position to highest position.

The sorted chromosome files were reformatted to vcf files using TASSEL Version 5.0 Standalone, then vcftools v.0.1.14 ([Danecek et al. \(2011\)](#)) concatenated these files, and the resulting file was zipped. PopLDdecay v3.42 ([Zhang et al. \(2018\)](#)) was used to analyze and visualize Linkage Disequilibrium (LD) Decay of the Mural *et al.* genotypic data.

5.4.4 Preprocessing Phenotypic Datasets

5.4.4.1 Plant Height Measurement Data

R Scripts (v.4.2.2 and v.4.3.1) ([R Core Team \(2023\)](#)) were developed to process the measuring and scoring data from Yanarella *et al.* dataset such that only plant height observations were retained for each of the three observation groups. Replicate number was programmatically added to these data, and positive control were removed. The 653 taxa shared between the datasets were retained.

Best Linear Unbiased Estimators (BLUE) values were calculated for each taxa using R’s built-in lm function to perform linear regressions and the emmeans v.1.8.7 package ([Lenth \(2023\)](#)), where taxa and replicate were fit as fixed effects. Best Linear Unbiased Prediction

(BLUP) values were calculated for each taxa with the `lmer` function of the `lme4` v.1.1-34 package (Bates et al. (2015)) where taxa, replicate, and row number were fit as random effects.

Visualization of diagnostics plots of the models (Supplementary Figure 1, Supplementary Figure 2) were generated using the `ggResidpanel` v.0.3.0 package (Goode and Rey (2022)).

5.4.4.2 Semantic Similarity for Plant Height Spoken Data

Text transcripts of spoken data from the Yanarella *et al.* dataset were processed using Python v.3.8.2 (Van Rossum and Drake (2009)). The `spaCy` v.3.5.1 package (Honnibal and Montani (2023)) and the TensorFlow v.2.12.0 (Abadi et al. (2016)) `spaCy` Universal Sentence Encoder v.0.4.6 (Mensio (2023)) were used to process the transcripts to obtain semantic similarity scores. Three phrases, "tall," "tall plant," and "tall height," were compared to each row observation through `spaCy`'s similarity function using the pre-trained large English universal sentence encoder (`en_use_lg`) from TensorFlow. A dataset of similarity scores in the form of values from 0 to 1 was generated (Figure 5.2 (b)).

Similarity scores for the 653 taxa shared by both datasets were retained, encompassing 35,709 rows or 91.92% of the original rows observed (Table 5.2). The similarity scores for the "tall" query were used as input to calculate BLUEs and BLUPs in the same manner as described in the *Plant Height Measurement Data* section, and visualizations of diagnostics plots of the models (Supplementary Figure 3, Supplementary Figure 4) were generated.

5.4.4.3 Binning for Plant Height Spoken Data

The transcripts of spoken plant descriptions were reviewed for phrases directly related to narrations about plant height. A set of 797 plant height phrases were manually curated and binned from 0 to 7, where 0: no growth, 1: very short plants, 2: short plants, 3: short-medium height plants, 4: medium height plants, 5: medium-tall height plants, 6: tall plants, 7: very tall plants. Bin values were assigned to observations for the 653 taxa shared by both datasets (Figure

5.2 (b)). Of the total text transcripts, there were 34,209 or 88.06% row observations that were retained and binned (Table 5.2).

BLUEs and BLUPs were calculated as described in the *Plant Height Measurement Data* Section, and visualizations of diagnostics plots of the models (Supplementary Figure 5, Supplementary Figure 6) were generated. Additionally, the R `nnet` v.7.3-19 package (Venables and Ripley (2002)) was used to perform multinomial logistic regression to predict height phrase bins, where taxa and replicate were fit as fixed effects.

5.4.5 Genome-Wide Association Studies

Genome Association and Prediction Integrated Tool (GAPIT) 3 v.3.1, 2022.4.16 (Lipka et al. (2012), Tang et al. (2016), Wang and Zhang (2021)) was used to perform Fixed and random model Circulating Probability Unification (FarmCPU) (Liu et al. (2016)) and Mixed Linear Model (MLM) (Yu et al. (2005)) on each of the phenotypic datasets using the Mural *et al.* marker dataset for genotypic input. Each chromosome was run individually, and PCA.total parameter was set to 3 for all analyses (Figure 5.2 (c)). This manuscript focuses on the FarmCPU analyses, the MLM processing and results are available as described in *Web Resources* section.

5.4.6 Genome-Wide Association Study Analyses

Manhattan plots from the resulting GAPIT analyses were generated using the `ggplot2` v.3.4.3 package (Wickham (2016)). We used the RAINBOWR v.0.1.29 package's (Hamazaki and Iwata (2020)) `CalcThreshold` to determine the Bonferroni threshold for each analysis with a `sig.level` of 0.05. SNPs that were identified as above the Bonferroni threshold for each analysis were viewed on MaizeGDB's implementation of GBrowse2 (Generic Genome Browser v.2.55) (Stein (2013)) using Maize B73 RefGen_v4; gene IDs were collected within +/- 300 kilobases (kb), based on the LD decay curve generated by `PopLDdecay`, of the identified SNPs (Supplementary Figure 7).

We collected a list of genes shown to influence plant height (Table 5.3, Supplementary Table 2) and compared the gene IDs from the GWAS analyses using a web-based intersection and Venn

diagram tool (Sterck (2021)) to determine if these previously published plant height genes were identified within the +/- 300 kb region indicated by the LD decay curve for each of our analyses. Additionally, Gene Ontology (GO) terms for the gene IDs within +/- 300 kb of SNPs identified as significant were obtained using the B73 RefGen_V4 Zm00001d.2 annotations generated by Maize Go Annotation - Methods, Evaluation, and Review (maize-GAMER) tool (Wimalanathan and Lawrence-Dill (2017), Wimalanathan et al. (2018)) and the R package GO.db v.3.17.0 (Carlson (2023)) was used to collect terms associated to the GO IDs (Supplementary Table 3).

5.5 Results and Discussion

5.5.1 Detecting Phenotypes from Spoken Descriptions

Student participants made three complete passes of the field (~4,320 observations per student participant (Table 5.2, *Total Rows Observed Count*) and used their individual wording and phraseology to describe the phenotypes in the field. Student participants recorded observations between 2 and 241 words in length (Figure 5.3). To explore the ability of the student participants to identify and describe phenotypes for traits of interest, 25 positive control accessions that, if grown in the appropriate environmental conditions, would show visually "dramatic" phenotypes.

The 25 positive control accessions were observed 53-55 times over all nine student participants (Table 5.1). We utilized Merriam-Webster (Merriam-Webster (2023)) and WordHippo (Kat IP Pty Ltd (2008)) thesaurus services to determine the participant's ability to identify words synonymous with descriptors that describe the positive control phenotypes as demonstrated in Figure 5.4 (a-c). For example, accessions M241C *A1 A2 B1 C1 C2 Pl1 Pr1 R1-r* and 219L *B1-S; R1-r pl1-McClintock* (gene name *colored1* and *colored plant1*, Supplementary Table 1) had at least one synonym in each of the observations made by the student participants as indicated in Table 5.1 by the proportion of 1.000 for both Merriam-Webster and WordHippo synonyms. While accessions U740G *Fbr1-N1602* (gene name, *few branched1*), 703J *Rs1-O 1*, and 703K *Rs1-Z (rough sheath*, Supplementary Table 1) had low proportions of observations having at least one synonym for each observation as indicated in Table 5.1.

Our findings indicate that participants, unaware of expected phenotypes, can identify and describe them in their own words. These results are limited in the number of synonyms identified for phenotype descriptions and the environment in which the plants were grown. The thesaurus services could deflate the proportion of observations with at least one synonym if the participants used informal descriptors. Additionally, the student participants would only have described our intended positive control phenotypes if the field environment and weather conditions were conducive to displaying the expected phenotypes.

5.5.2 Extracting Phenotype Data for Plant Height from Spoken Descriptions

Two methods were employed to preprocess text transcriptions of spoken descriptions of plants. The semantic similarity method of comparing the term "tall" to each row observation retained 91.92% of the full set of row recordings captured (including the 25 positive controls and 33 accessions unique to the Yanarella *et al.* dataset) by the student participants, demonstrating that 35,709 row observations were made with taxa in both datasets (Table 5.2). The manual bin method of identifying phrases related to plant height and binning them based on apparent semantic similarity retained 86.06% of the full set of row recordings captured by the student participants and 95.80% of the observations with plant height phrases made with taxa in both datasets, which results in 34,209 row observations for manually binned data (Table 5.2).

Both methods parse information about the plant height traits and process the data into a format appropriate as input into available GWAS tools and models. Using a query term for plant height and semantic similarity requires less manual curation and was implemented on a larger subset of data. The benefit of the binning method is that it reduces the noise; only observations with plant height-related terms were considered.

A limitation of the query term and semantic similarity method is retaining noisy data because this method compares the "tall" query to each observation and relies on pre-trained models. An example of noise comes from the participant whose NATO code name "Victor's" recording for row 1,456 on 07/16/2021, *tall and height green all the way to the bottom ... super short hairs on*

top that are quite prickly ... in general there are the brace roots are short fat and light green in color, in which the similarity function of spaCy University Sentence Encoder (Mensio (2023)) when compared to the "tall" query string determine the semantic similarity score as 0.0848. The shortcomings of the binning method are the time-consuming nature of curating lists of phrases relevant to the trait of interest and the loss of data where phrases directly related to plant height were not specified.

5.5.3 Association Studies using Phenotypes Derived from Speech

We performed association studies using FarmCPU (Liu et al. (2016)) on three categories of phenotype data. The first phenotype category was ground truth (measured) plant height data using BLUEs (Figure 5.5 (a)) and BLUPs (Figure 5.5 (b)). For the BLUE analysis, 21 significant SNPs above the Bonferroni threshold of 8.55 were identified, and of those SNPs, we discovered 10 (Supplementary Table 3) in which at least one plant height gene was detected in the literature within +/- 300 kb (Supplementary Table 2). The BLUP analysis identified 29 significant SNPs, 9 (Supplementary Table 3) where at least one plant height gene was discovered in the literature within +/- 300 kb (Table 5.3, Supplementary Table 2).

The second phenotype category used BLUEs (Figure 5.6 (a)) and BLUPs (Figure 5.6 (b)) for tall query and semantic similarity of spoken phenotype descriptions. These analyses identified 27 and 23, respectively, significant SNPs (Supplementary Table 3), respectively, above the Bonferroni threshold of 8.55. Of these, 9 and 8 genes, respectively, were formerly detected for plant height within +/- 300 kb of the SNP (Table 5.3, Supplementary Table 2).

The third phenotype category used BLUEs (Figure 5.7 (a)) and BLUPs (Figure 5.7 (b)) for manual binning of spoken phenotype descriptions with plant height terms. These analyses identified 32 and 33, respectively, significant SNPs (Supplementary Table 3), above the Bonferroni threshold of 8.55 respectively. Of these, 13 and 12 have genes formerly reported within +/- 300 kb of the SNP (Table 5.3, Supplementary Table 2). An additional analysis was completed using predicted values from a multinomial regression (Supplemental Figure 8, Supplementary Table 2)

in which 21 SNPs were significant, and 3 had genes detected within +/- 300 kb of the SNP in the literature (Table 5.3, Supplementary Table 2).

Having demonstrated that transcripts of spoken descriptions of plants can be computationally processed and applied as phenotypic input for existing GWAS tools, we further sought to determine whether the data collected in this experiment are practical for association studies. To do this, we utilized ground truth measurements of plant height and identified SNPs associated with plant height that have been reported in other publications.

The semantic similarity method with the tall query to generate phenotype values for each spoken observation using spaCy's similarity function (Honnibal and Montani (2023)) and the Universal Sentence Encoder (Mensio (2023)) was successful. We were able to perform GWAS, and there were significant SNPs from regions associated with plant height. As this is a proof of concept study, we acknowledge that other pre-trained models exist capable of calculating semantic similarity or models that can be adapted to generate similarity scores related to plant height such as BioBERT (Lee et al. (2019)), those implemented by the Python gensim package (Řehůřek and Sojka (2010)), or others reviewed in Koroleva et al. (2019). Further, additional queries could be employed for relating the text observations to a height value.

The binning method for plant height phrases appears to be a promising method for association studies with phenotype data extracted from spoken descriptions. This method reduces the noisiness of the transcription data and scores observations on only phrases detailing features of plant height. Additionally, the GWAS performed with BLUE and BLUP values generated from the binning method detected more known regions associated with plant height formerly reported than the manually measured and semantic similarity query methods.

We demonstrate our use of a multinomial regression to generate phenotypic input with binned data, although we recognize that FarmCPU is not optimized for multinomial input. GWAS tools that utilize an ordered multinomial regression model to predict multinomial values for association studies were developed in the medical research field (German et al. (2019)). Regardless,

FarmCPU with multinomial binned input for plant height detected regions of the genome associated with plant height.

While participant language was not constrained, input that is less noisy and with lower data loss could be attainable if an emphasis were placed on stating specific aspects of the plant accompanied by a descriptor and reducing literary descriptive comments. An example of describing a specific aspect of a plant is "tall, green, and long" compared to "this row has tall plants." The former statement is unclear whether the whole plant is described or a specific aspect of the plant, while the latter makes it clearer that the total plant height is described. Literary language descriptions are more difficult to compute because context is necessary to determine the meaning behind a phrase, an example, candy cane stripe. While candy cane stripe may induce a mental image of a candy cane, unless a computation model is trained to identify the literary description, the model would not be able to discern the spoken description of phenotype as a particular striped pattern.

5.5.4 Investigating GO Terms from GWAS Results

After identifying the gene IDs associated with the regions +/- 300 kb of the significant SNPs, we investigated the GO terms annotated to these genes. The full list of GO terms for each model discussed above is available in Supplementary Table 3 and Supplementary Table 4. To examine how these terms align with plant height terms, we queried the dataset for the words auxin, brassinosteroid, and gibberellin because of their known functions in plant height regulation ([Li et al. \(2020\)](#) reviews the importance of these hormones). The term "auxin" was more frequently present in these datasets when compared to brassinosteroid or gibberellin (Table 5.4).

Additionally, other GO annotations identified in our analyses have functions that may affect plant height. Examples of these GO terms include developmental growth (GO:0048589), anatomical structure formation involved in morphogenesis (GO:0048646), and shoot system development (GO:0048367). Further, using GWAS, we found genomic regions with functional

annotations related to plant hormone functions that were not identified by the literature in Table 5.3. These regions can be of potential interest for the plant height trait.

While examining GO terms, descriptions that do not relate to plant functions but were annotated to gene IDs occurred. Errant assignments of GO terms for plant-specific tasks has been described in [Fattel et al. \(2022\)](#). An example is the gene ID *Zm00001d008201* being assigned the term animal organ development (GO:0048513). Interestingly, this ID was also assigned the terms auxin-activated signaling pathway (GO:0009734) and post-embryonic development (GO:0009791). These results demonstrate a compelling argument for reviewing GO terms produced by plant-specific annotation tools to remove non-plant terms.

5.6 Conclusion

We developed methods to process recordings of spoken observations of plants in a field environment as phenotypic input for commonly used GWAS tools. Here, we report that two methods for generating phenotypic data from transcripts of spoken observations recovered known genomic regions of interest for plant height. Additionally, novel regions were identified and could be investigated for their role in the plant height trait. These methods provide a framework for further exploratory research, including expanding the methods employed to obtain biologically relevant information from spoken data and expanding the number of traits in which spoken descriptions are the basis of phenotypic data for GWAS. Additional traits were collected for this WiDiv dataset by speech and manual scoring to contribute to these prospective research undertakings and are publicly available ([Yanarella et al., 2023](#)).

Beyond these demonstrations that spoken, unstructured phenotypic descriptions can be used to recover known associations and to identify potentially new regions of the genome contributing to well understood traits, there are two other conceptual benefits that should be considered. Firstly, when people are describing what they see in the field rather than exclusively collecting predefined traits, the potential to uncover novel phenomena is perhaps increased. Secondly, it is the case that for many years we have used computers to analyzed structured data, so those

collecting the data have limited themselves to documenting data in a structured, computer-friendly format. This is, in effect, asking people to structure their thinking and documentation like and for a computer. With the methods described here, the people collecting the data are enabled to think and behave in a more naturally human way for data collection. This has implications for the rate of data collection and for cognitive burden as follows. Over three weeks of data collection, each participant made three complete passes of the field, recording spoken observations. However, when the participants collected manual scoring and measurement data, none were able to make a complete pass of the field. Our experimental design enabled the student participants to speak and describe plant traits using their unique vocabulary and speech patterns. Participant "Zulu" reported that recording spoken observations was simpler and easier than measuring and scoring because they could make more detailed observations about different parts of the plants because recording spoken observations was both less strenuous and less mentally taxing.

Some may wonder why having a person describe phenotypes matters at all given that image-based data collection is improving all the time, and seems to be a great way to collect data for image-based machine learning analytics of phenotypes and traits (reviewed in [Xiao et al. \(2022\)](#)). Clearly image analysis can be useful for many traits, but for those traits that involve tactile, odor, or other sensory observations, image-based data collection cannot collect relevant data. Coupled with image-based data collection and analysis, it is becoming clear that language-based annotations, both spoken and written, are poised to both allow and enable the human perception to fill nuanced understanding of phenotypes and traits.

5.7 Web Resources

Code to recreate the analysis in this manuscript will be available on CyVerse, and is currently pending a DOI.

5.8 Data Availability

The de-identified spoken data described in this manuscript is exempted by Iowa State University’s Institutional Review Board (IRB ID: 21-179-00). Phenotypic data was obtained from (Yanarella et al. (2023)) and is available from:

https://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence_Dill_Maize_WiDiv_Summer_2021_Dataset_June_2023. Genotypic data was obtained from (Mural et al. (2022b), Mural et al. (2022a)) and can be accessed from:

https://figshare.com/articles/dataset/Maize_WiDiv_SAM_1051Genotype_vcf_gz_genotype_file/19175888/1. Gene Ontology data was obtained from (Wimalanathan and Lawrence-Dill (2017)) and is available from

https://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Carolyn_Lawrence-Dill_maize-GAMER_maize.B73_RefGen_v4_Zm00001d.2_Oct_2017.r1.

Supplementary material will be available at G3 online.

5.9 Acknowledgments

We acknowledge the High-Performance Computing (HPC) facility at Iowa State University for assisting this research through computing resources and technical support. We appreciate discussions with Toni Kazic about the preliminary conceptualizations for this research project and making audio data available to assist with experimental design. We appreciate discussions with Qi Li regarding the binning methods and Kris De Brabanter for discussions about the statistics underlying GWAS. We thank Brian Dilkes, Rajdeep Khangura, and Amanpreet Kaur for providing genotypic data, guidance, and examples of data processing techniques for association studies. We thank Tyler Foster and Yu-Ru Chen for enlightening discussions about GWAS methods and tools.

5.10 Funding

This work was supported by the Iowa State University Plant Sciences Institute Faculty Scholars Award (CJLD) and the Iowa State Predictive Plant Phenomics NSF Research Traineeship (DGE-1545453); CJLD is a co-principal investigator, and CFY is a trainee. This reasearch was also supported by the NSF and USDA-NIFA AI Research Institutes program for AI Institute: for Resilient Agriculture (#2021-67021-35329) to CJLD and supporting CFY and LF.

5.11 Conflicts of Interest

The authors declare no conflicts of interest.

5.12 Author Contributions

CFY and CJLD conceived the idea for this project. CFY performed the analyses and drafted the initial version of the manuscript. CFY and LF evaluated the GO term analysis. All authors have read, edited, and approved the manuscript.

5.13 References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, pages 265–283.
- Austin, D. F. and Lee, M. (1996). Genetic resolution and verification of quantitative trait loci for flowering and plant height with recombinant inbred lines of maize. *Genome*, 39(5):957–968.
- Azodi, C. B., Pardo, J., VanBuren, R., de los Campos, G., and Shiu, S.-H. (2019). Transcriptome-Based Prediction of Complex Traits in Maize. *The Plant Cell*, 32(1):139–151.
- Bai, W., Zhang, H., Zhang, Z., Teng, F., Wang, L., Tao, Y., and Zheng, Y. (2009). The evidence for non-additive effect as the main genetic component of plant height and ear height in maize using introgression line populations. *Plant Breeding*.
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

- Blakeslee, J. J., Peer, W. A., and Murphy, A. S. (2005). Auxin transport. *Current Opinion in Plant Biology*, 8(5):494–500.
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19):2633–2635.
- Braun, I. R. and Lawrence-Dill, C. J. (2020). Automated Methods Enable Direct Computation on Phenotypic Descriptions for Novel Candidate Gene Prediction. *Frontiers in Plant Science*, 10.
- Braun, I. R., Yanarella, C. F., and Lawrence-Dill, C. J. (2020). Computing on Phenotypic Descriptions for Candidate Gene Discovery and Crop Improvement. *Plant Phenomics*, 2020.
- Braun, I. R., Yanarella, C. F., Rajeswari, J. P. D., Bassham, D. C., and Lawrence-Dill, C. J. (2021). The Case for Retaining Natural Language Descriptions of Phenotypes in Plant Databases and a Web Application as Proof of Concept. *bioRxiv*.
- Brooks, L., Strable, J., Zhang, X., Ohtsu, K., Zhou, R., Sarkar, A., Hargreaves, S., Elshire, R. J., Eudy, D., Pawlowska, T., Ware, D., Janick-Buckner, D., Buckner, B., Timmermans, M. C. P., Schnable, P. S., Nettleton, D., and Scanlon, M. J. (2009). Microdissection of Shoot Meristem Functional Domains. *PLoS Genetics*, 5(5):e1000476.
- Carlson, M. (2023). *GO.db: A set of annotation maps describing the entire Gene Ontology*. R package version 3.17.0.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and and, R. D. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.
- Fattel, L., Psaroudakis, D., Yanarella, C. F., Chiteri, K. O., Dostalík, H. A., Joshi, P., Starr, D. C., Vu, H., Wimalanathan, K., and Lawrence-Dill, C. J. (2022). Standardized genome-wide function prediction enables comparative functional genomics: a new application area for Gene Ontologies in plants. *GigaScience*, 11.
- Gallavotti, A. (2013). The role of auxin in shaping shoot architecture. *Journal of Experimental Botany*, 64(9):2593–2608.
- Geisler, M. and Murphy, A. S. (2005). The ABC of auxin transport: The role of p-glycoproteins in plant development. *FEBS Letters*, 580(4):1094–1102.
- German, C. A., Sinsheimer, J. S., Klimentidis, Y. C., Zhou, H., and Zhou, J. J. (2019). Ordered multinomial regression for genetic association analysis of ordinal phenotypes at Biobank scale. *Genetic Epidemiology*, 44(3):248–260.

- Goode, K. and Rey, K. (2022). *ggResidpanel: Panels and Interactive Versions of Diagnostic Plots using 'ggplot2'*. R package version 0.3.0.
- Hamazaki, K. and Iwata, H. (2020). RAINBOW: Haplotype-based genome-wide association study using a novel SNP-set method. *PLOS Computational Biology*, 16(2):e1007663.
- Hanse, C. N., Johnson, J. M., Sekhon, R. S., Kaeppler, S. M., and de Leon, N. (2011). Genetic Diversity of a Maize Association Population with Restricted Phenology. *Crop Science*, 51(2):704–715.
- Hartwig, T., Chuck, G. S., Fujioka, S., Klempien, A., Weizbauer, R., Potluri, D. P. V., Choe, S., Johal, G. S., and Schulz, B. (2011). Brassinosteroid control of sex determination in maize. *Proceedings of the National Academy of Sciences*, 108(49):19814–19819.
- Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., de Leon, N., Kaeppler, S. M., and Buell, C. R. (2014). Insights into the Maize Pan-Genome and Pan-Transcriptome. *The Plant Cell*, 26(1):121–135.
- Honnibal, M. and Montani, I. (2023). spaCy v3.5.1 spancat for multi-class labeling, fixes for textcat+transformers and more. To appear.
- Jansson, S. (1994). The light-harvesting chlorophyll ab-binding proteins. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1184(1):1–19.
- Kat IP Pty Ltd (2008). WordHippo.
- Kazic, T. (2020). Chloe: Flexible, efficient data provenance and management. *bioRxiv*.
- Koroleva, A., Kamath, S., and Paroubek, P. (2019). Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. *Journal of Biomedical Informatics*, 100:100058.
- Lawit, S. J., Wych, H. M., Xu, D., Kundu, S., and Tomes, D. T. (2010). Maize DELLA proteins dwarf plant8 and dwarf plant9 as modulators of plant development. *Plant and Cell Physiology*, 51(11):1854–1868.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Lenth, R. V. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.7.

- Li, H., Wang, L., Liu, M., Dong, Z., Li, Q., Fei, S., Xiang, H., Liu, B., and Jin, W. (2020). Maize Plant Architecture Is Regulated by the Ethylene Biosynthetic Gene *ZmACS7*. *Plant Physiology*, 183(3):1184–1199.
- Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., and Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, 28(18):2397–2399.
- Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLOS Genetics*, 12(2):e1005767.
- Mazaheri, M., Heckwolf, M., Vaillancourt, B., Gage, J. L., Burdo, B., Heckwolf, S., Barry, K., Lipzen, A., Ribeiro, C. B., Kono, T. J. Y., Kaepler, H. F., Spalding, E. P., Hirsch, C. N., Buell, C. R., de Leon, N., and Kaepler, S. M. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. *BMC Plant Biology*, 19(1).
- Mensio, M. (2023). Martinomensio/spacy-universal-sentence-encoder: Google use (universal sentence encoder) for spaCy.
- Merriam-Webster (2023). Merriam-Webster Online Thesaurus.
- Multani, D. S., Briggs, S. P., Chamberlin, M. A., Blakeslee, J. J., Murphy, A. S., and Johal, G. S. (2003). Loss of an MDR Transporter in Compact Stalks of Maize *br2* and Sorghum *dw3* Mutants. *Science*, 302(5642):81–84.
- Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2.
- Mural, R., Sun, G., Grzybowski, M., Tross, M. C., Jin, H., Smith, C., Newton, L., Thompson, A. M., Sigmon, B., and Schnable, J. C. (2022a). Maize_WiDiv_SAM_1051Genotype.vcf.gz genotype file.
- Mural, R. V., Sun, G., Grzybowski, M., Tross, M. C., Jin, H., Smith, C., Newton, L., Andorf, C. M., Woodhouse, M. R., Thompson, A. M., Sigmon, B., and Schnable, J. C. (2022b). Association mapping across a multitude of traits collected in diverse environments in maize. *GigaScience*, 11.
- Oellrich, A., Walls, R. L., Cannon, E. K., Cannon, S. B., Cooper, L., Gardiner, J., Gkoutos, G. V., Harper, L., He, M., Hoehndorf, R., Jaiswal, P., Kalberer, S. R., Lloyd, J. P., Meinke, D., Menda, N., Moore, L., Nelson, R. T., Pujar, A., Lawrence, C. J., and Huala, E. (2015). An ontology approach to comparative phenomics in plants. *Plant Methods*, 11(1).

- Peiffer, J. A., Romay, M. C., Gore, M. A., Flint-Garcia, S. A., Zhang, Z., Millard, M. J., Gardner, C. A. C., McMullen, M. D., Holland, J. B., Bradbury, P. J., and Buckler, E. S. (2014). The Genetic Architecture Of Maize Height. *Genetics*, 196(4):1337–1356.
- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.
- Salvi, S., Corneti, S., Bellotti, M., Carraro, N., Sanguineti, M. C., Castelletti, S., and Tuberosa, R. (2011). Genetic dissection of maize phenology using an intraspecific introgression library. *BMC Plant Biology*, 11(1):4.
- Sarić, R., Nguyen, V. D., Burge, T., Berkowitz, O., Trtílek, M., Whelan, J., Lewsey, M. G., and Čustović, E. (2022). Applications of hyperspectral imaging in plant phenotyping. *Trends in Plant Science*, 27(3):301–315.
- Stein, L. D. (2013). Using GBrowse 2.0 to visualize and share next-generation sequence data. *Briefings in Bioinformatics*, 14(2):162–171.
- Sterck, L. (2021). Calculate and draw custom Venn diagrams.
- Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z., Pan, Y., Liu, D., Lipka, A. E., Buckler, E. S., and Zhang, Z. (2016). GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *The Plant Genome*, 9(2).
- Teng, F., Zhai, L., Liu, R., Bai, W., Wang, L., Huo, D., Tao, Y., Zheng, Y., and Zhang, Z. (2012). *ZmGA3ox2*, a candidate gene for a major QTL, *qPH3.1*, for plant height in maize. *The Plant Journal*, 73(3):405–416.
- Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.
- Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.
- Wallace, J. G., Zhang, X., Beyene, Y., Semagn, K., Olsen, M., Prasanna, B. M., and Buckler, E. S. (2016). Genome-wide Association for Plant Height and Flowering Time across 15 Tropical Maize Populations under Managed Drought Stress and Well-Watered Conditions in Sub-Saharan Africa. *Crop Science*, 56(5):2365–2378.
- Wang, J. and Zhang, Z. (2021). GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genomics, Proteomics & Bioinformatics*, 19(4):629–640.

- Weng, J., Xie, C., Hao, Z., Wang, J., Liu, C., Li, M., Zhang, D., Bai, L., Zhang, S., and Li, X. (2011). Genome-Wide Association Study Identifies Candidate Genes That Affect Plant Height in Chinese Elite Maize (*Zea mays L.*) Inbred Lines. *PLoS ONE*, 6(12):e29229.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- Wimalanathan, K., Friedberg, I., Andorf, C. M., and Lawrence-Dill, C. J. (2018). Maize GO Annotation—Methods, Evaluation, and Review (maize-GAMER). *Plant Direct*, 2(4).
- Wimalanathan, K. and Lawrence-Dill, C. (2017). maize-GAMER Annotations for maize B73 RefGen_V4 Zm00001d.2.
- Winkler, R. G. and Helentjaris, T. (1995). The maize Dwarf3 gene encodes a cytochrome P450-mediated early step in Gibberellin biosynthesis. *The Plant Cell*, 7(8):1307–1317.
- Woodhouse, M. R., Cannon, E. K., Portwood, J. L., Harper, L. C., Gardiner, J. M., Schaeffer, M. L., and Andorf, C. M. (2021). A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biology*, 21(1).
- Wu, A.-M., Rihouey, C., Seveno, M., Hörnblad, E., Singh, S. K., Matsunaga, T., Ishii, T., Lerouge, P., and Marchant, A. (2009). The arabidopsis IRX10 and IRX10-LIKE glycosyltransferases are critical for glucuronoxylan biosynthesis during secondary cell wall formation. *The Plant Journal*, 57(4):718–731.
- Xiao, Q., Bai, X., Zhang, C., and He, Y. (2022). Advanced high-throughput plant phenotyping techniques for genome-wide association studies: A review. *Journal of Advanced Research*, 35:215–230.
- Yanarella, C. F., Fattel, L., Kristmundsdóttir, Á.Ý., Lopez, M. D., Edwards, J. W., Campbell, D. A., Abel, C. A., and Lawrence-Dill, C. J. (2023). Carolyn_Lawrence_Dill_Maize_WiDiv_Summer_2021_Dataset_June_2023.
- Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., Xiong, L., and Yan, J. (2020). Crop Phenomics and High-Throughput Phenotyping: Past Decades, Current Challenges, and Future Perspectives. *Molecular Plant*, 13(2):187–214.
- Yao, L., van de Zedde, R., and Kowalchuk, G. (2021). Recent developments and potential of robotics in plant eco-phenotyping. *Emerging Topics in Life Sciences*, 5(2):289–300.
- Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2005). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208.

Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M., and Yang, T.-L. (2018). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, 35(10):1786–1788.

5.14 Figures and Tables

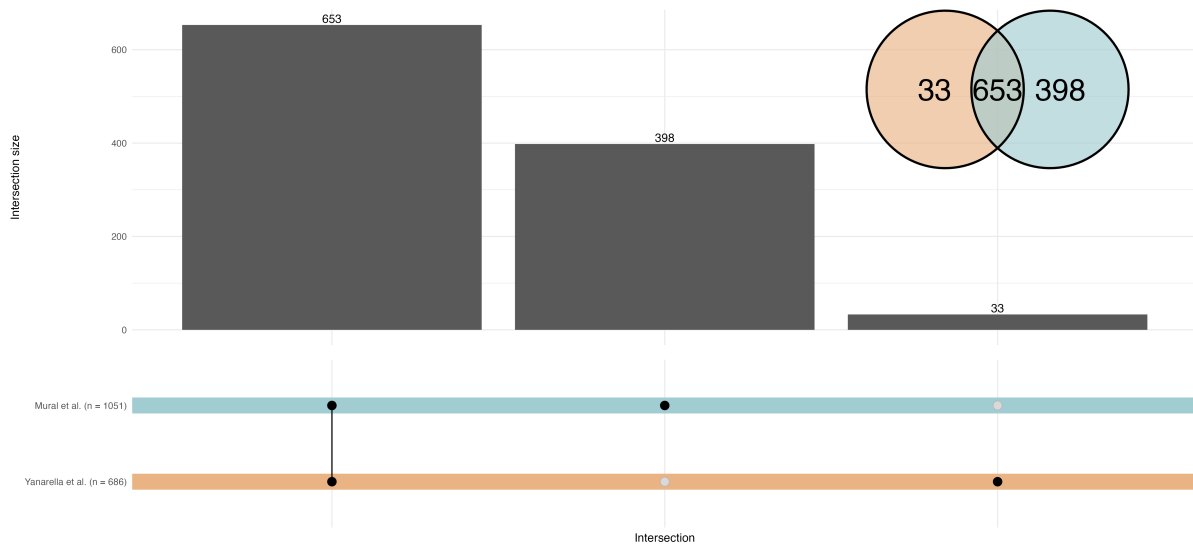


Figure 5.1 Comparison of intersections of Mural *et al.* and Yanarella *et al.* WiDiv dataset taxa (positive controls not included), where n is the number of unique taxa in each dataset. Mural *et al.* dataset (in blue), and Yanarella *et al.* dataset (in orange).

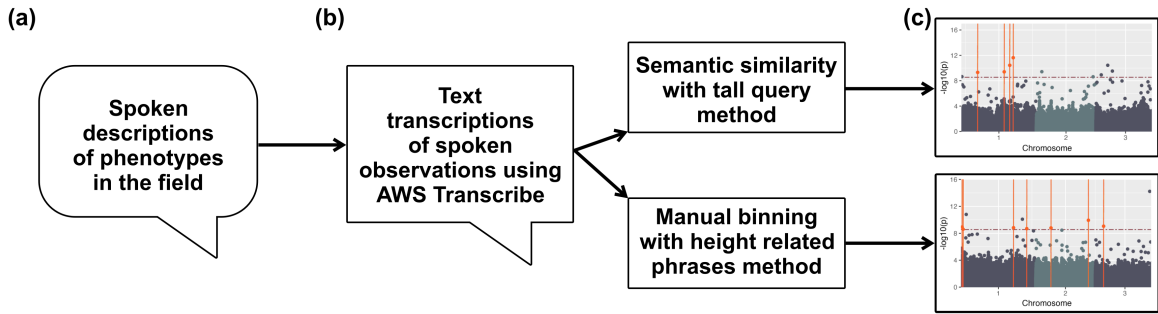


Figure 5.2 Spoken phenotype process overview. (a) In field spoken phenotype descriptions collection, (b) Spoken phenotype data processing, including transcript production and methods for generating numeric representations of phenotypes for traits, and (c) GWAS using data derived from spoken observations.

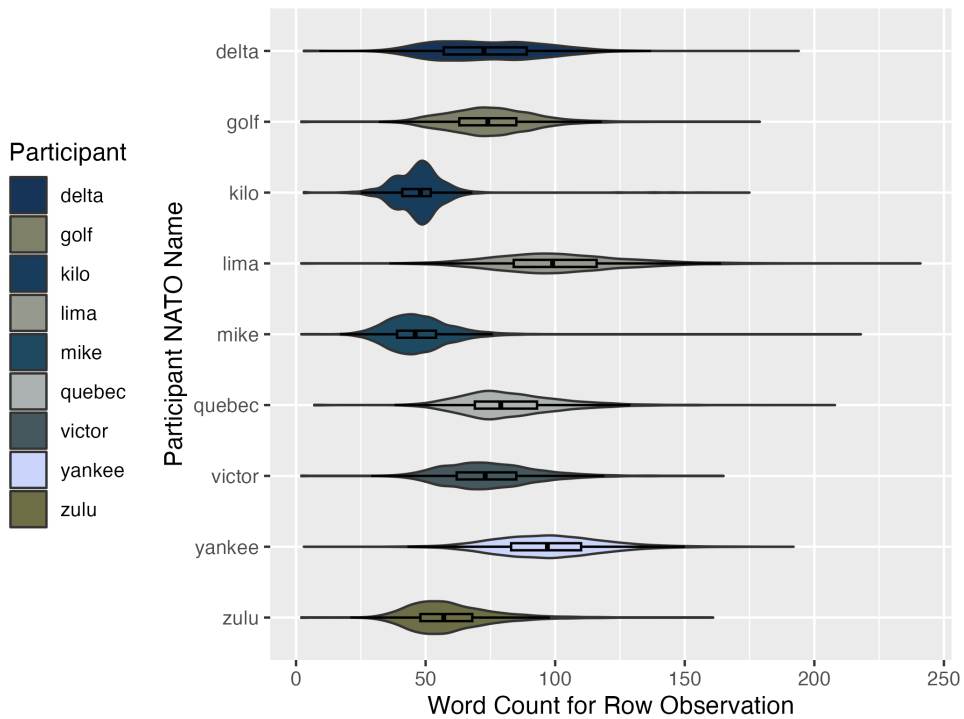


Figure 5.3 Distributions for each student participants word count per observation. A boxplot is included in each of the violin plots, individual outlier points not represented.

(a) 1248: "these plants are overall these plants are overall tall they are unique andy have purple stalks the stalks are skinny the midribs and the leaves are purple and the tassels they have are purple there's no ears on them the leaves are long and narrow there's some white spotting around the margins some of the leaves from the lower leaves are curling somewhat there's some wrinkles and bumps around the outer edges of some of the leaves the brace roots are purple with a little bit of white mixed in and these plants overall are are very unique due to their color patterns"


<p>(b)</p> <p style="text-align: center;">M241C</p> <p><i>A1 A2 B1 C1 C2 P11 Pr1 R1-r</i></p> 	<p>(c)</p> <p>Phenotype Descriptions: purple, red, marbled, pigment, striped, and spotted</p> <p>Merriam-Webster</p> <p>Synonyms Searched for: red</p> <p>Phenotype Descriptions and Synonyms: blooming, bloomy, blowsy, blowzy, blushing, bronzed, brown, cherubic, florid, flush, flushed, full-blooded, glowing, marbled, pigment, pink, pinkish, purple, red, rosy, rubicund, ruddy, sanguine, spotted, striped, suntanned, tanned, and warm</p>	<p>WordHippo</p> <p>Synonyms Searched for: purple</p> <p>Phenotype Descriptions and Synonyms: amaranthine, amethyst, blue, bluish, dark, heliotrope, lavender, lilac, magenta, mauve, mulberry, orchid, periwinkle, perse, plum, pomegranate, purple, red, reddish, violaceous, violet, wine, purple, red, marbled, pigment, striped, and spotted</p>
<p>At least one term from "Phenotype Description and Synonyms" present:</p>		
	<p>Yes</p>	<p>Yes</p>

Figure 5.4 Example of detecting phenotypes from positive control accession descriptions. (a) Transcript of a spoken description for positive control accession M241C *A1 A2 B1 C1 C2 P11 Pr1 R1-r*, (b) Image of positive control accession M241C *A1 A2 B1 C1 C2 P11 Pr1 R1-r*, (c) Example of phenotype description terms, and synonyms from Marriam-Webster and WordHippo, to demonstrate a description having at least one instance of a synonymous word for the phenotype of interest.

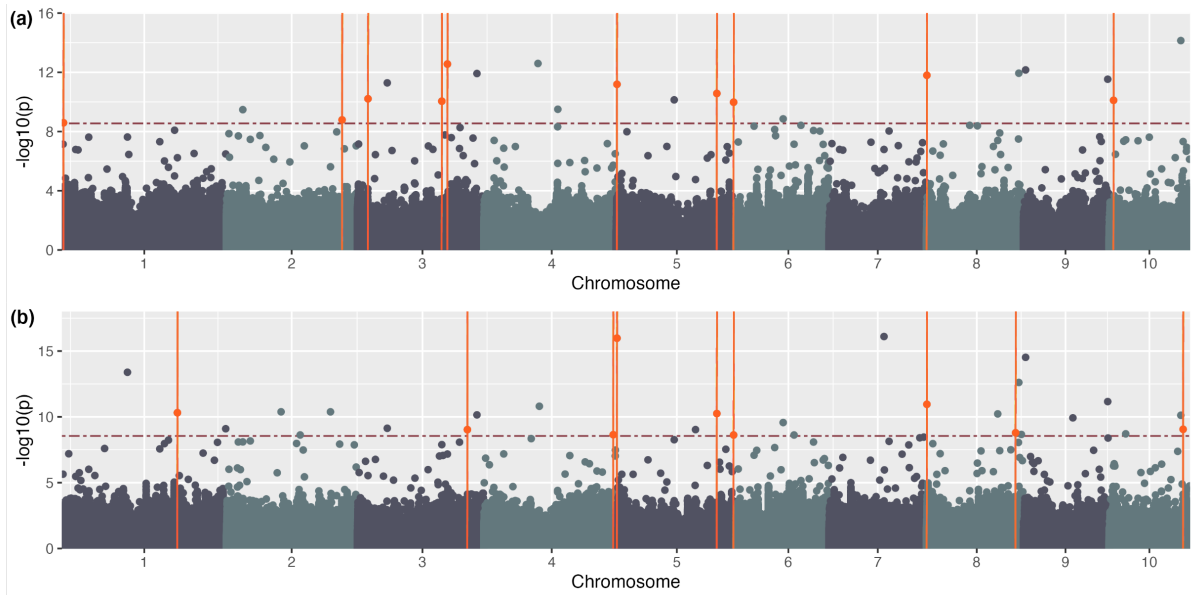


Figure 5.5 Manhattan plot of measured height phenotypic data. Plot generated using GAPIT and FarmCPU using measured height data (a) BLUEs and (b) BLUPs using Mural *et al.* genotypic data. The red dashed line indicates the Bonferroni threshold ((a) $-\log_{10}(p) = 8.55$; (b) $-\log_{10}(p) = 8.55$), orange points indicate identified SNPs with known plant height genes within ± 300 kb, and orange vertical lines indicate positions ± 300 kb identified SNPs with known plant height genes.

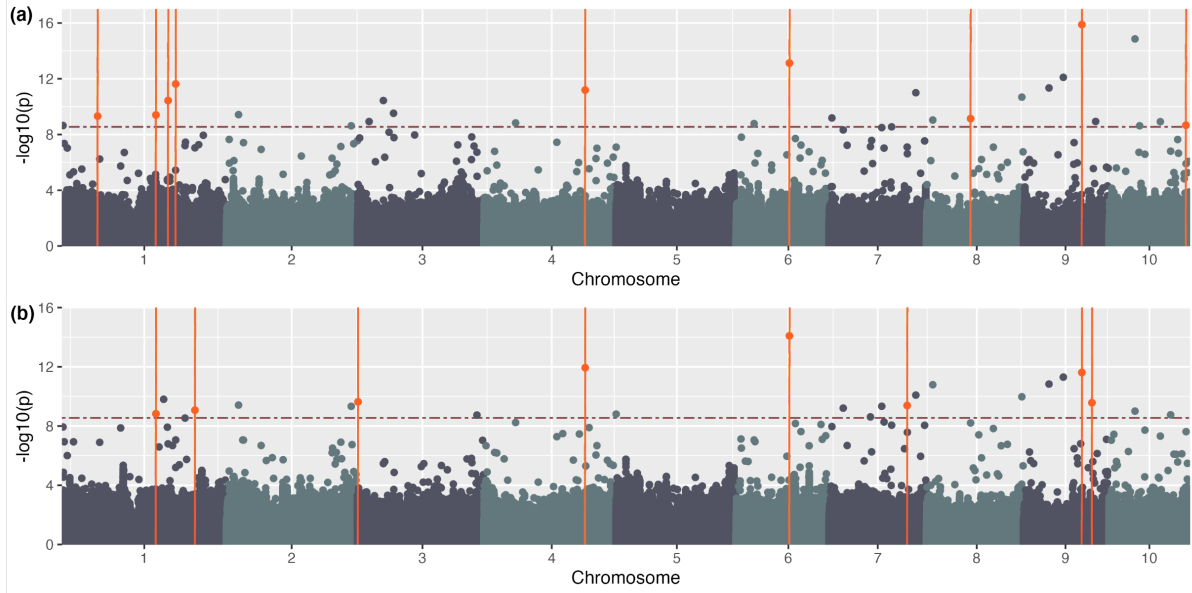


Figure 5.6 Manhattan plot of tall query semantic similarity phenotypic data. Plot generated using GAPIT and FarmCPU using semantic similarity score data (a) BLUEs and (b) BLUPs using Mural *et al.* genotypic data. The red dashed line indicates the Bonferroni threshold ((a) $-\log_{10}(p) = 8.55$; (b) $-\log_{10}(p) = 8.55$), orange points indicate identified SNPs with known plant height genes within +/- 300 kb, and orange vertical lines indicate positions +/- 300 kb identified SNPs with known plant height genes.

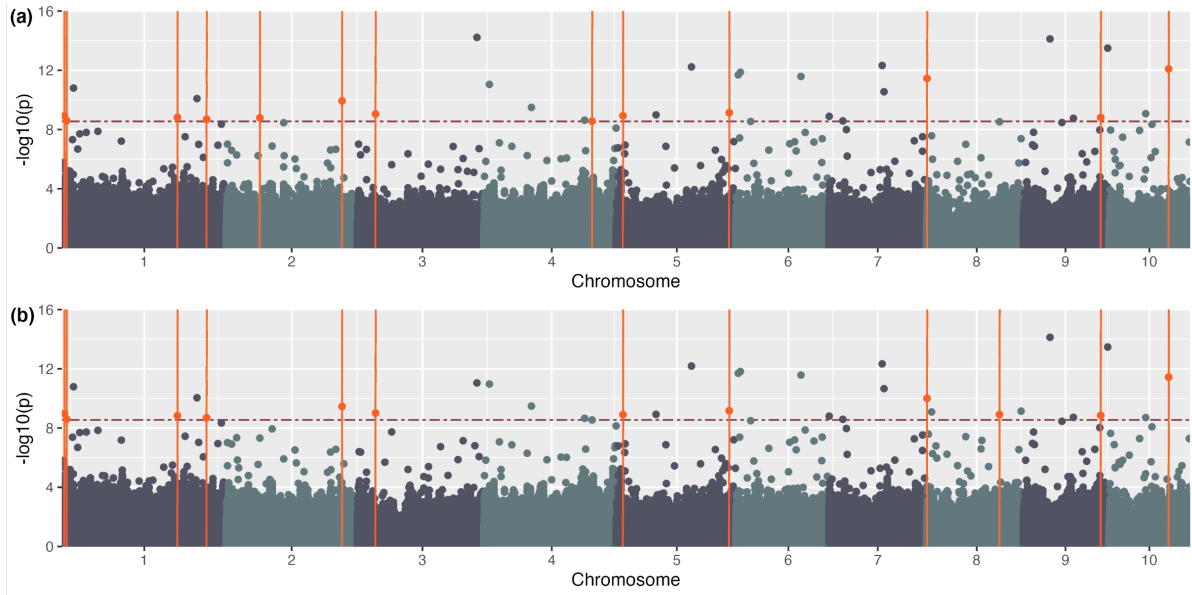


Figure 5.7 Manhattan plot of binned plant height phrases phenotypic data. Plot generated using GAPIT and FarmCPU using binned plant height data (a) BLUEs and (b) BLUPs using Mural *et al.* genotypic data. The red dashed line indicates the Bonferroni threshold ((a) $-\log_{10}(p) = 8.55$; (b) $-\log_{10}(p) = 8.55$), orange points indicate identified SNPs with known plant height genes within ± 300 kb, and orange vertical lines indicate positions ± 300 kb identified SNPs with known plant height genes.

Table 5.1 Proportion of word usage for describing positive control accessions.

Accession	Merriam-Webster			WordHippo		
	Rows Observed	Count	Number of Synonyms	Proportion	Number of Synonyms	Proportion
528A <i>Hsf1-N1595</i>	53		32	0.2264	85	0.7358
528AA <i>Hsf1-N2559</i>	55		32	0.2000	85	0.6727
528J <i>Hsf1-N1603</i>	54		32	0.2778	85	0.8148
X28F <i>cr4-6143</i>	54		34	0.4815	83	0.5000
X29C <i>cr4-N590C</i>	54		34	0.5370	83	0.4630
X29D <i>cr4-N647</i>	54		34	0.6296	83	0.5741
702B <i>o2 v5 ra1-Ref gl1</i>	53		34	0.3208	73	0.4528
703D <i>o2 ra1-Ref gl1</i>	54		34	0.4074	73	0.5185
711H <i>ra1-N408E</i>	54		34	0.5000	73	0.5370
117D <i>tb1</i>	53		32	0.2075	71	0.3208
117DA <i>tb1-8963</i>	54		32	0.3889	71	0.4630
X03A <i>sr3</i>	55		32	0.5455	22	0.4364
X03J <i>sr3-N504A</i>	54		32	0.5926	22	0.6296
104B <i>rli1-N2302A</i>	54		49	0.1667	77	0.1667
104C <i>rli1-N2276</i>	55		49	0.2182	77	0.2000
703J <i>Rs1-O</i>	54		49	0.1667	77	0.2037
703K <i>Rs1-Z</i>	54		49	0.1667	77	0.2593
219L <i>B1-S; R1-r pl1-McClintock</i>	54		28	1.0000	28	1.0000
617A <i>pl1-0 c1-EMS Sh1 B1-S R1-r</i>	55		28	0.2182	28	0.2727
M241C A1 A2 B1 C1 C2 P11 Pr1 R1-r¹	54		28	1.0000	28	1.0000
M341B <i>A1 A2 B1 C1 C2 pl1 Pr1 R1-r</i>	55		28	0.8364	28	0.8364
U740G <i>Fbr1-N1602</i>	54		31	0.0000	6	0.0370
916G <i>Trn1-N1597</i>	54		28	0.9815	57	0.9815
708C <i>o15-N1117</i>	54		71	0.3333	120	0.7037
119E <i>Ts6</i>	54		43	0.2778	47	0.8889

¹ Accession used for example illustrated in Figure 5.4.

Table 5.2 Observation retention by spoken phenotype method.

Participant	Total Rows Observed Count ¹	Spoken Phenotype Processing Method	
		Tall Query ²	Manual Bins ³
Delta	4,326	3,974 (0.9186)	3,911 (0.9041; 0.9841)
Golf	4,317	3,969 (0.9194)	3,711 (0.8596; 0.9350)
Kilo	4,321	3,973 (0.9195)	3,960 (0.9165; 0.9967)
Lima	4,322	3,975 (0.9197)	3,747 (0.8670; 0.9426)
Mike	4,302	3,953 (0.9189)	3,937 (0.9152; 0.9960)
Quebec	4,308	3,959 (0.9190)	3,241 (0.7523; 0.8186)
Victor	4,329	3,980 (0.9194)	3,946 (0.9115; 0.9915)
Yankee	4,308	3,960 (0.9192)	3,943 (0.9153; 0.9957)
Zulu	4,314	3,966 (0.9193)	3,813 (0.8839; 0.9614)
Sum	38,847	35,709 (0.9192)	34,209 (0.8806; 0.9580)

¹ Rows observed count includes data collected for all taxa within the Yanarella *et al.* dataset, including 25 positive control taxa and 33 taxa not found in the Mural *et al.* dataset.

² The count of row observations utilized for semantic similarity for spoken data methods represents the data 653 intersecting taxa between Yanarella *et al.* and Mural *et al.*, and the data in parentheses is the proportion of data retained from total rows observed.

³ The count of row observations utilized for binning for spoken data methods represents the data 653 intersecting taxa between Yanarella *et al.* and Mural *et al.*, and the data in parentheses is the proportion of data retained from total rows observed (left) and the proportion of data retained from the 653 intersecting taxa which have plant height terms (right).

Table 5.3 Plant height gene models count identified from publications.

Publication	Plant Height Related Gene Model Count ¹
Jansson (1994)	1
Winkler and Helentjaris (1995)	1
Austin and Lee (1996)	3
Multani et al. (2003)	2
Blakeslee et al. (2005)	1
Geisler and Murphy (2005)	1
Brooks et al. (2009)	1
Wu et al. (2009)	7
Bai et al. (2009)	6
Lawit et al. (2010)	2
Hartwig et al. (2011)	1
Salvi et al. (2011)	4
Weng et al. (2011)	1
Teng et al. (2012)	1
Gallavotti (2013)	1
Peiffer et al. (2014)	7
Wallace et al. (2016)	312
Mazaheri et al. (2019)	10
Azodi et al. (2019)	876
Mural et al. (2022b)	347

¹ Count of gene model IDs identified for gene model set Zm00001d.2 gene model set and Zm-B73-REFERENCE-GRAMENE-4.0 assembly version.

Table 5.4 Appearance of GO terms by plant hormone for each method.

Analysis	auxin		brassinosteroid		gibberellin	
	Literature ¹	This Study ²	Literature	This Study	Literature	This Study
Measured BLUEs GO Terms	2	15	0	3	0	4
Measured BLUPs GO Terms	3	22	0	8	0	6
Semantic Similarity Tall Query BLUEs GO Terms	12	5	0	5	0	7
Semantic Similarity Tall Query BLUPs GO Terms	2	15	0	3	0	6
Manual Term Bin BLUEs GO Terms	1	16	0	8	1	10
Manual Term Bin BLUPs GO Terms	1	15	0	6	1	9
Manual Term Bin Multinomial GO Terms	0	10	0	3	0	1

¹ Appearance count for GO terms of prior literature in Table 5.3.² Appearance count for GO terms unique to this study.

5.15 Appendix: Institutional Review Board Exemption Letter

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office of Research Ethics
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

Date: 06/04/2021
To: Carolyn Lawrence-Dill
From: Office of Research Ethics
Title: CFY Field Project 2021
IRB ID: 21-179
Submission Type: Initial Submission
Exemption Date: 05/18/2021

The project referenced above has been declared exempt from most requirements of the human subject protections regulations as described in 45 CFR 46.104 or 21 CFR 56.104 because it meets the following federal requirements for exemption:

2018 - 3 (i.B): Research involving benign behavioral interventions in conjunction with the collection of information from an adult subject through verbal or written responses or audiovisual recording when the subject prospectively agrees to the intervention and information collection and any disclosure of the human subjects' responses outside the research would not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation. - 3 (ii) If research involves deception, it is prospectively authorized by the subject.

The determination of exemption means that:

- **You do not need to submit an application for continuing review. Instead, you will receive a request for a brief status update every three years. The status update is intended to verify that the study is still ongoing.**
- **You must carry out the research as described in the IRB application.** Review by IRB staff is required prior to implementing modifications that may change the exempt status of the research. In general, review is required for any *modifications to the research procedures* (e.g., method of data collection, nature or scope of information to be collected, nature or duration of behavioral interventions, use of deception, etc.), any change in *privacy or confidentiality protections*, modifications that result in the *inclusion of participants from vulnerable populations*, removing plans for informing participants about the study, any *change that may increase the risk or discomfort to participants*, and/or any change such that the revised procedures do not fall into one or more of the [regulatory exemption categories](#). The purpose of review is to determine if the project still meets the federal criteria for exemption.
- **All changes to key personnel** must receive prior approval.
- **Promptly inform the IRB of any addition of or change in federal funding for this study.** Approval of the protocol referenced above applies only to funding sources that are specifically identified in the corresponding IRB application.

IRB 07/2020

Detailed information about requirements for submitting modifications for exempt research can be found on our [website](#). For modifications that require prior approval, an amendment to the most recent IRB application must be submitted in IRBManager. A determination of exemption or approval from the IRB must be granted before implementing the proposed changes.

Non-exempt research is subject to many regulatory requirements that must be addressed prior to implementation of the study. Conducting non-exempt research without IRB review and approval may constitute non-compliance with federal regulations and/or academic misconduct according to ISU policy.

Additionally:

- All research involving human participants must be submitted for IRB review. **Only the IRB or its designees may make the determination of exemption**, even if you conduct a study in the future that is exactly like this study.
- **Please inform the IRB if the Principal Investigator and/or Supervising Investigator end their role or involvement with the project** with sufficient time to allow an alternate PI/Supervising Investigator to assume oversight responsibility. Projects must have an [eligible PI](#) to remain open.
- **Immediately inform the IRB of (1) all serious and/or unexpected [adverse experiences](#) involving risks to subjects or others; and (2) any other [unanticipated problems involving risks](#) to subjects or others.**
- **Approval from other entities may also be needed.** For example, access to data from private records (e.g., student, medical, or employment records, etc.) that are protected by FERPA, HIPAA or other confidentiality policies requires permission from the holders of those records. Similarly, for research conducted in institutions other than ISU (e.g., schools, other colleges or universities, medical facilities, companies, etc.), investigators must obtain permission from the institution(s) as required by their policies. **An IRB determination of exemption in no way implies or guarantees that permission from these other entities will be granted.**
- Your research study may be subject to [post-approval monitoring](#) by Iowa State University's Office for Responsible Research. In some cases, it may also be subject to formal audit or inspection by federal agencies and study sponsors.
- Upon completion of the project, transfer of IRB oversight to another IRB, or departure of the PI and/or Supervising Investigator, please initiate a Project Closure in IRBManager to officially close the project. For information on instances when a study may be closed, please refer to the [IRB Study Closure Policy](#).

Please don't hesitate to contact us if you have questions or concerns at 515-294-4566 or IRB@iastate.edu.

CHAPTER 6. GENERAL CONCLUSION

6.1 Summary

The research described in the preceding chapters of this dissertation details the development of bioinformatic procedures for processing and utilizing natural language descriptions of plants for association studies. We proposed using natural language descriptions of plants for investigating gene pair similarity based on semantic similarity and association studies with natural language descriptions of plants. We utilized pre-trained language models to determine gene pair similarity for descriptions with structure (ontologies) and unstructured free text. These models vary in complexity from employing semantic similarity methods, syntactic similarity approaches, and using a combination of semantic and syntactic similarity procedures to predict gene pair similarity. Our results indicate we can recover known biological relationships from natural language descriptions.

Additionally, we designed, performed, and disseminated the data associated with collecting spoken descriptions of plants using the Wisconsin Diversity (WiDiv) panel in a field environment. These data are available to guide others who want to collect spoken phenotypes or investigate the descriptions recorded in our experiments. We performed association studies using the plant height data collected during the summer of 2021. We demonstrated two methods for processing speech for GWAS; these methods included using semantic similarity and a query term "tall" to determine a score for each observation, and the second method involved assigning a bin number for observations that include plant height descriptions. We identified regions of the genome previously reported as associated with plant height using data derived from spoken descriptions of plants. Also, we identified regions not found in previously reported literature with GO terms related to plant height. The code set to perform association studies using text transcriptions of spoken phenotypes is publicly available and reproducible.

6.2 Collaborative Project Outcomes

Throughout my graduate research and training, I participated in projects outside of the scope of the research described in this dissertation. Below, two projects I contributed to are described, and the publications associated with these are presented in Appendix 6.5 Table 6.1 and datasets found in Appendix 6.5 Table 6.2.

6.2.1 Predictive Plant Phenomics (P3) Research Trainee-ship Symposium Sessions

The P3 graduate student trainees organized and led a symposium session and workshop at the American Society of Plant Biologists's (ASPB) Phenome 2020 Conference with the support of Corteva Agriscience's Plant Sciences Symposia Series (PSSS). The sessions' activities included running a symposium session with four speakers and a workshop to present activities associated with interdisciplinary plant-related research.

My contribution to these events included co-leading the preparation for the events, including speaker selection, session schedule, and overseeing workshop development. Another contribution was co-authoring a perspective, published on The Open Science Framework (OSF), manuscript describing planning conference events as a graduate student (Yanarella et al. (2021)).

6.2.2 Gene Ontology Annotations for Plant Species

Gene Ontology Meta Annotator for Plants (GOMAP) is a pipeline for assigning functional annotations to genome assemblies (Wimalanathan and Lawrence-Dill (2021)). The pipeline produces high-coverage GO annotations for various plant species by aggregating sequence similarity, protein domain presence, and mixed methods approaches developed for the Critical Assessment of Functional Annotation (CAFA) Challenge (Zhou et al. (2019)). Annotations generated by GOMAP are resources for the research community to perform comparative analyses, to conduct enrichment analyses, and to locate functional annotations for genes identified in association studies for traits of interest. Research that has stemmed from GOMAP annotations includes generating parsimony and neighbor-joining trees using GO datasets for multiple plant

species to determine how closely these trees resemble known phylogenetic relationships based on an analysis of plant GO terms (Fattel et al. (2022)).

My contribution to the GOMAP project and work on comparative functional genomics was to test the pipeline and suggest modifications for usability while annotating *Hordeum vulgare* (Yanarella et al. (2019)) and *Pinus lambertiana* (Yanarella et al. (2020)). Additionally, I wrote the code to generate a preliminary diagram of the process flow of the GOMAP pipeline (DILL-PICL et al. (2022)) and assisted in training undergraduate and graduate students to generate annotations using GOMAP for a wide range of plant species (Appendix 6.5, Table 6.2).

6.3 Future Work

The work described in this dissertation focuses on generating and processing natural language descriptions of plant phenotypes. We utilized pre-trained models to determine gene pair similarity. We disseminated transcripts of spoken descriptions of plants so that the research community could continue to investigate the traits that we recorded spoken observations of phenotypes during the summer of 2021. Further, the methods we utilized to gather speech as a phenotype can be employed for various species. Additionally, we have demonstrated two methods for performing GWAS from spoken descriptions. As natural language processing methods improve, new methods can be developed for spoken descriptions of plants for association studies. Future research can be performed to identify genomic regions of interest for traits using spoken descriptions.

While evaluating the identified regions on the genome for our GWAS methods, we used GO annotations for genes within regions where linkage disequilibrium could occur. We uncovered animal function GO terms assigned to our plant-specific data, a concern described by Fattel et al. (2022). This research area of potential is the work of Leila Fattel (Lawrence-Dill Lab, Iowa State University), who is investigating the best practices for "trimming" the GO Directed Acyclic Graph (DAG) to the most specific term relating to plant functions.

In the early phases of developing the spoken phenotyping protocol, a group of naysayers adamantly defended their requirement of having only expert phenotypers collect data in the field.

While the role and need for experts must be recognized, using crowd-sourced data for phenotyping has been a valuable tool for performing science efficiently (see Zhou et al. (2018)). Throughout this research, I have acquired an appreciation for investing in developing, encouraging, and empowering undergraduate students to participate in research. Although our undergraduate participants are not considered experts at phenotyping, their ability to detect and vocalize plant phenotype descriptions has enabled us to perform association studies and identify genomic regions of interest for our intended trait. I encourage the facilitation of collecting spoken phenotypes so that we may work to uncover additional fascinating biological phenomena.

6.4 References

- Chiteri, K., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2020). Carolyn_Lawrence_Dill_GOMAP_Cannabis_NCBI-cs10_January_2020.r1.
- DILL-PICL, Yanarella, C., Fattel, L., Wimalanathan, K., Campbell, D., and Lawrence-Dill, C. (2022). GOMAP Process Flowchart.
- Dostalík, H., Fattel, L., Yanarella, C., and Lawrence-Dill, C. (2021). Carolyn_Lawrence_Dill_GOMAP_Grape_Genoscope_12x_January_2021.r1.
- Fattel, L., Psaroudakis, D., Yanarella, C. F., Chiteri, K. O., Dostalík, H. A., Joshi, P., Starr, D. C., Vu, H., Wimalanathan, K., and Lawrence-Dill, C. J. (2022). Standardized genome-wide function prediction enables comparative functional genomics: a new application area for Gene Ontologies in plants. *GigaScience*, 11.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2021a). Carolyn_Lawrence_Dill_GOMAP_Banana_NCBI_ASM31385v2_February_2021.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2021b). Carolyn_Lawrence_Dill_GOMAP_Blueberry_GigaDB_v1.0_June_2021.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2021c). Carolyn_Lawrence_Dill_GOMAP_Cacao_NCBI_CriolloV2_March_2021.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2021d). Carolyn_Lawrence_Dill_GOMAP_Coffee_CGH_v1.0_June_2021.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023a). Carolyn_Lawrence_Dill_GOMAP_Barley_IPK_cv_Morex_V3_June_2023_v1.r1.

- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023b).
Carolyn_Lawrence_Dill.GOMAP_Blueberry_GigaDB_Draper_v1.0_May_2023_v2.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023c).
Carolyn_Lawrence_Dill.GOMAP_Coffee_CGH_v1.0_May_2023_v2.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023d).
Carolyn_Lawrence_Dill.GOMAP_Cotton_DOE-JGI_v2.1_June_2023_v2.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023e).
Carolyn_Lawrence_Dill.GOMAP_Grape_Genoscope_12x_May_2023_v2.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023f).
Carolyn_Lawrence_Dill.GOMAP_Hop_HopBase_Cascade_July_2023_v2.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023g).
Carolyn_Lawrence_Dill.GOMAP_Maize_CyVerse_Mo17_CAU_2.0_July_2023.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023h).
Carolyn_Lawrence_Dill.GOMAP_Pepper_PGP_cvCM334_June_2023_v2.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023i).
Carolyn_Lawrence_Dill.GOMAP_Rapeseed_BnPIR_ZS11_June_2023_v2.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023j).
Carolyn_Lawrence_Dill.GOMAP_Rice_IRGSP_1.0_June_2023_v2.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023k).
Carolyn_Lawrence_Dill.GOMAP_Sorghum_DOE-JGI_v3.1.1_March_2023_v1.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023l).
Carolyn_Lawrence_Dill.GOMAP_Soybean_LIS_Wm82_IGA1008.gnm1.ann1.FGN6
_May_2023_v1.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023m).
Carolyn_Lawrence_Dill.GOMAP_Stiff_brome_DOE-JGI_Bd21_v3.2_May_2023_v1.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023n).
Carolyn_Lawrence_Dill.GOMAP_Sugar_Pine_TreeGenes_v1.5_June_2023_v2.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023o).
Carolyn_Lawrence_Dill.GOMAP_Tea_Teabase_CSS_ChrLev_20200506_June_2023_v1.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023p).
Carolyn_Lawrence_Dill.GOMAP_Tomato_SGN_SL4.0_July_2023_v2.r1.

- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023q).
Carolyn_Lawrence_Dill_GOMAP_Wheat_URGI_IWGSC_RefSeq_v2.1_May_2023_v1.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023r).
Carolyn_Lawrence_Dill_GOMAP_Wild_Tomato_SGN_LA716_June_2023_v2.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 B73).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_B73_NAM_5.0_October_2022_v2.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 B97).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_B97_NAM_1.0_October_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 CML103).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_CML103_NAM_1.0_October_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 CML228).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_CML228_NAM_1.0_October_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 CML247).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_CML247_NAM_1.0_October_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 CML277).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_CML277_NAM_1.0_October_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 CML322).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_CML322_NAM_1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 CML333).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_CML333_NAM_1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 CML52).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_CML52_NAM_1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 CML69).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_CML69_NAM_1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 HP301).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_HP301_NAM_1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 II14H).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_II14H_NAM_1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 Ki11).
Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_Ki11_NAM_1.0_November_2022.r1.

- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 Ki3).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_Ki3_NAM.1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 Ky21).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_Ky21_NAM.1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 M162W).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_M162W_NAM.1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 M37W).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_M37W_NAM.1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 Mo18W).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_Mo18W_NAM.1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 Ms71).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_Ms71_NAM.1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 NC350).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_NC350_NAM.1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 NC358).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_NC358_NAM.1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 Oh43).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_Oh43_NAM.1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 Oh7B).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_Oh7B_NAM.1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 P39).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_P39_NAM.1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 Tx303).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_Tx303_NAM.1.0_November_2022.r1.
- Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023 Tzi8).
 Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_Tzi8_NAM.1.0_November_2022.r1.
- Idris, N., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023a).
 Carolyn_Lawrence_Dill_GOMAP_Banana_NCBI_ASM31385v2_December_2022_v2.r1.
- Idris, N., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023b).
 Carolyn_Lawrence_Dill_GOMAP_Barrel_Clover_LIS_A17.gnm5.ann1.6.L2RX
 _November_2022_v1.r1.

- Idris, N., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023c). Carolyn_Lawrence_Dill_GOMAP_Barrel_Clover_LIS_R108.gnmHiC_1.ann1.Y8NH_November_2022_v1.r1.
- Idris, N., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023d). Carolyn_Lawrence_Dill_GOMAP_Cacao_NCBI_CriolloV2_December_2022_v2.r1.
- Idris, N., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023e). Carolyn_Lawrence_Dill_GOMAP_Cannabis_NCBI_cs10_2.0_October_2022_v1.r1.
- Idris, N., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023f). Carolyn_Lawrence_Dill_GOMAP_CommonBean_LIS_G19833_November_2022_v2.r1.
- Idris, N., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023g). Carolyn_Lawrence_Dill_GOMAP_Cowpea_JGI_IT97K-499-35_December_2022_v1.r1.
- Idris, N., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023h). Carolyn_Lawrence_Dill_GOMAP_Peanut_Tifrunner_IPGI.2.0_November_2022_v1.r1.
- Johnson, O., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2022). Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_B73_NAM_5.0_December_2021.r.
- Joshi, P., Yanarella, C., Psaroudakis, D., Wimalanathan, K., and Lawrence-Dill, C. (2020). Carolyn_Lawrence_Dill_GOMAP_Gossypium_raimondii_JGI_v2.1_January_2020.r1.
- Ngara, B., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023a). Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_Mo17_CAU_1.0_May_2023_v2.r1.
- Ngara, B., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023b). Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_PH207_NS-UIUC_UMN_1.0_May_2023_v2.r1.
- Ngara, B., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2023c). Carolyn_Lawrence_Dill_GOMAP_Maize_MaizeGDB_W22_NRGENE_2.0_May_2023_v2.r1.
- Starr, D., Fattel, L., Yanarella, C., Wimalanathan, K., and Lawrence-Dill, C. (2021). Carolyn_Lawrence_Dill_GOMAP_Canola_BnPIR_ZS11_March_2021.r1.
- Wimalanathan, K. and Lawrence-Dill, C. J. (2021). Gene ontology meta annotator for plants (GOMAP). *Plant Methods*, 17(1).
- Yanarella, C., Cook, T., Panelo, J., and Chiteri, K. (2021). Graduate Student Perspective on Organizing PSSS Events During Phenome 2020.

- Yanarella, C., Psaroudakis, D., Wimalanathan, K., and Lawrence-Dill, C. (2019). Carolyn_Lawrence_Dill.GOMAP_Barley_IBSC_PGSD-1.0_May_2019.r1.
- Yanarella, C., Psaroudakis, D., Wimalanathan, K., and Lawrence-Dill, C. (2020). Carolyn_Lawrence_Dill.GOMAP_SugarPine_TreeGenesDB-1.5_January_2020.r1.
- Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsóh, B. Z., Crocker, A. W., Lewis, K. A., Georghiou, G., Nguyen, H. N., Hamid, M. N., Davis, L., Dogan, T., Atalay, V., Rifaioglu, A. S., Dalkiran, A., Atalay, R. C., Zhang, C., Hurto, R. L., Freddolino, P. L., Zhang, Y., Bhat, P., Supek, F., Fernández, J. M., Gemovic, B., Perovic, V. R., Davidović, R. S., Sumonja, N., Veljkovic, N., Asgari, E., Mofrad, M. R., Profiti, G., Savojardo, C., Martelli, P. L., Casadio, R., Boecker, F., Schoof, H., Kahanda, I., Thurlby, N., McHardy, A. C., Renaux, A., Saidi, R., Gough, J., Freitas, A. A., Antczak, M., Fabris, F., Wass, M. N., Hou, J., Cheng, J., Wang, Z., Romero, A. E., Paccanaro, A., Yang, H., Goldberg, T., Zhao, C., Holm, L., Törönen, P., Medlar, A. J., Zosa, E., Borukhov, I., Novikov, I., Wilkins, A., Lichtarge, O., Chi, P.-H., Tseng, W.-C., Linial, M., Rose, P. W., Dessimoz, C., Vidulin, V., Dzeroski, S., Sillitoe, I., Das, S., Lees, J. G., Jones, D. T., Wan, C., Cozzetto, D., Fa, R., Torres, M., Vesztröcy, A. W., Rodriguez, J. M., Tress, M. L., Frasca, M., Notaro, M., Grossi, G., Petrini, A., Re, M., Valentini, G., Mesiti, M., Roche, D. B., Reeb, J., Ritchie, D. W., Aridhi, S., Alborzi, S. Z., Devignes, M.-D., Koo, D. C. E., Bonneau, R., Gligorijević, V., Barot, M., Fang, H., Toppo, S., Lavezzo, E., Falda, M., Berselli, M., Tosatto, S. C., Carraro, M., Piovesan, D., Rehman, H. U., Mao, Q., Zhang, S., Vucetic, S., Black, G. S., Jo, D., Suh, E., Dayton, J. B., Larsen, D. J., Omdahl, A. R., McGuffin, L. J., Brackenridge, D. A., Babbitt, P. C., Yunes, J. M., Fontana, P., Zhang, F., Zhu, S., You, R., Zhang, Z., Dai, S., Yao, S., Tian, W., Cao, R., Chandler, C., Amezola, M., Johnson, D., Chang, J.-M., Liao, W.-H., Liu, Y.-W., Pascarelli, S., Frank, Y., Hoehndorf, R., Kulmanov, M., Boudellioua, I., Politano, G., Carlo, S. D., Benso, A., Hakala, K., Ginter, F., Mehryary, F., Kaewphan, S., Björne, J., Moen, H., Tolvanen, M. E., Salakoski, T., Kihara, D., Jain, A., Šmuc, T., Altenhoff, A., Ben-Hur, A., Rost, B., Brenner, S. E., Orengo, C. A., Jeffery, C. J., Bosco, G., Hogan, D. A., Martin, M. J., O'Donovan, C., Mooney, S. D., Greene, C. S., Radivojac, P., and Friedberg, I. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1).
- Zhou, N., Siegel, Z. D., Zarecor, S., Lee, N., Campbell, D. A., Andorf, C. M., Nettleton, D., Lawrence-Dill, C. J., Ganapathysubramanian, B., Kelly, J. W., and Friedberg, I. (2018). Crowdsourcing image analysis for plant phenomics to generate ground truth data for machine learning. *PLOS Computational Biology*, 14(7):e1006337.

6.5 Appendix: Manuscripts and Datasets From Collaborate Projects

6.5.1 Manuscripts

Table 6.1 Contributions to manuscripts out of the scope of the described research aims.

Manuscript Title	Reference
Graduate Student Perspective on Organizing PSSS Events During Phenome 2020	Yanarella et al. (2021)
Standardized genome-wide function prediction enables comparative functional genomics: a new application area for Gene Ontologies in plants	Fattel et al. (2022)
Gene Function Annotations for the Maize NAM Founder Lines	Under Review: BMC Research Notes

6.5.2 Datasets

Table 6.2 Contributions to datasets out of the scope of the described research aims.

Dataset Title	Contribution Type	Reference
GOMAP Process Flowchart	Main Contributor	DILL-PICL et al. (2022)
Carolyn_Lawrence_Dill_GOMAP_Barley_IBSC_PGSB-1.0_May_2019.r1	Main Contributor	Yanarella et al. (2019)
Carolyn_Lawrence_Dill_GOMAP_SugarPine_TreeGenesDB-1.5_January_2020.r1	Main Contributor	Yanarella et al. (2020)
Carolyn_Lawrence_Dill_GOMAP_Banana_NCBI_ASM31385v2_December_2022_v2.r1	Contributor	Idris et al. (2023a)
Carolyn_Lawrence_Dill_GOMAP_Banana_NCBI_ASM31385v2_February_2021.r1	Contributor	Fattel et al. (2021a)
Carolyn_Lawrence_Dill_GOMAP_Barley_IPK_cv_Morex_V3_June_2023_v1.r1	Contributor	Fattel et al. (2023a)
Carolyn_Lawrence_Dill_GOMAP_Barrel_Clover_LIS_A17.gnm5.ann1.6.L2RX_November_2022_v1.r1	Contributor	Idris et al. (2023b)

Table 6.2 Contributions to datasets out of the scope of the described research aims *continued (1)*.

Dataset Title	Contribution Type	Reference
Carolyn_Lawrence_Dill_GOMAP_Barrel _Clover_LIS_R108.gnmHiC.1.ann1.Y8NH _November_2022_v1.r1	Contributor	Idris et al. (2023c)
Carolyn_Lawrence_Dill_GOMAP_Blueberry _GigaDB_Draper_v1.0_May_2023_v2.r1	Contributor	Fattel et al. (2023b)
Carolyn_Lawrence_Dill_GOMAP_Blueberry _GigaDB_v1.0_June_2021.r1	Contributor	Fattel et al. (2021b)
Carolyn_Lawrence_Dill_GOMAP_Cacao _NCBI_CriolloV2_December_2022_v2.r1	Contributor	Idris et al. (2023d)
Carolyn_Lawrence_Dill_GOMAP_Cacao _NCBI_CriolloV2_March_2021.r1	Contributor	Fattel et al. (2021c)
Carolyn_Lawrence_Dill_GOMAP_Cannabis _NCBI_cs10_2.0_October_2022_v1.r1	Contributor	Idris et al. (2023e)
Carolyn_Lawrence_Dill_GOMAP_Cannabis _NCBI_cs10_January_2020.r1	Contributor	Chiteri et al. (2020)
Carolyn_Lawrence_Dill_GOMAP_Canola _BnPIR_ZS11_March_2021.r1	Contributor	Starr et al. (2021)
Carolyn_Lawrence_Dill_GOMAP_Coffee _CGH_v1.0_June_2021.r1	Contributor	Fattel et al. (2021d)
Carolyn_Lawrence_Dill_GOMAP_Coffee _CGH_v1.0_May_2023_v2.r1	Contributor	Fattel et al. (2023c)
Carolyn_Lawrence_Dill_GOMAP _CommonBean_LIS_G19833_November _2022_v2.r1	Contributor	Idris et al. (2023f)
Carolyn_Lawrence_Dill_GOMAP _Cotton_DOE-JGI_v2.1_June_2023_v2.r1	Contributor	Fattel et al. (2023d)
Carolyn_Lawrence_Dill_GOMAP_Cowpea _JGI_IT97K-499-35_December_2022_v1.r1	Contributor	Idris et al. (2023g)
Carolyn_Lawrence_Dill_GOMAP_Gossypium _raimondii_JGI_v2.1_January_2020.r1	Contributor	Joshi et al. (2020)
Carolyn_Lawrence_Dill_GOMAP_Grape _Genoscope_12x_January_2021.r1	Contributor	Dostalík et al. (2021)
Carolyn_Lawrence_Dill_GOMAP_Grape _Genoscope_12x_May_2023_v2.r1	Contributor	Fattel et al. (2023e)
Carolyn_Lawrence_Dill_GOMAP_Hop _HopBase_Cascade_July_2023_v2.r1	Contributor	Fattel et al. (2023f)

Table 6.2 Contributions to datasets out of the scope of the described research aims *continued (2)*.

Dataset Title	Contribution Type	Reference
Carolyn_Lawrence_Dill_GOMAP_Maize _CyVerse_Mo17_CAU_2.0_July_2023.r1	Contributor	Fattel et al. (2023g)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_B73_NAM_5.0 _December_2021.r1	Contributor	Johnson et al. (2022)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_B73_NAM_5.0 _October_2022_v2.r1	Contributor	Fattel et al. (2023 B73)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_B97_NAM_1.0 _October_2022.r1	Contributor	Fattel et al. (2023 B97)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_CML103_NAM_1.0 _October_2022.r1	Contributor	Fattel et al. (2023 CML103)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_CML228_NAM_1.0 _October_2022.r1	Contributor	Fattel et al. (2023 CML228)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_CML247_NAM_1.0 _October_2022.r1	Contributor	Fattel et al. (2023 CML247)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_CML277_NAM_1.0 _October_2022.r1	Contributor	Fattel et al. (2023 CML277)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_CML322_NAM_1.0 _November_2022.r1	Contributor	Fattel et al. (2023 CML322)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_CML333_NAM_1.0 _November_2022.r1	Contributor	Fattel et al. (2023 CML333)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_CML52_NAM_1.0 _November_2022.r1	Contributor	Fattel et al. (2023 CML52)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_CML69_NAM_1.0 _November_2022.r1	Contributor	Fattel et al. (2023 CML69)

Table 6.2 Contributions to datasets out of the scope of the described research aims *continued (3)*.

Dataset Title	Contribution Type	Reference
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_HP301_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 HP301)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_II14H_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 II14H)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_Ki11_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 Ki11)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_Ki3_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 Ki3)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_Ky21_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 Ky21)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_M162W_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 M162W)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_M37W_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 M37W)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_Mo17_CAU_1.0_May_2023_v2.r1	Contributor	Ngara et al. (2023a)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_Mo18W_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 Mo18W)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_Ms71_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 Ms71)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_NC350_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 NC350)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_NC358_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 NC358)

Table 6.2 Contributions to datasets out of the scope of the described research aims *continued (4)*.

Dataset Title	Contribution Type	Reference
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_Oh43_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 Oh43)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_Oh7B_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 Oh7B)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_P39_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 P39)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_PH207_NS-UIUC_UMN.1.0 _May_2023_v2.r1	Contributor	Ngara et al. (2023b)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_Tx303_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 Tx303)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_Tzi8_NAM.1.0 _November_2022.r1	Contributor	Fattel et al. (2023 Tzi8)
Carolyn_Lawrence_Dill_GOMAP_Maize _MaizeGDB_W22_NRGENE.2.0 _May_2023_v2.r1	Contributor	Ngara et al. (2023c)
Carolyn_Lawrence_Dill_GOMAP_Peanut _Tifrunner.IPGI.2.0_November_2022_v1.r1	Contributor	Idris et al. (2023h)
Carolyn_Lawrence_Dill_GOMAP_Pepper _PGP_cvCM334_June_2023_v2.r1	Contributor	Fattel et al. (2023h)
Carolyn_Lawrence_Dill_GOMAP_Rapeseed _BnPIR_ZS11_June_2023_v2.r1	Contributor	Fattel et al. (2023i)
Carolyn_Lawrence_Dill_GOMAP_Rice _IRGSP_1.0_June_2023_v2.r1	Contributor	Fattel et al. (2023j)
Carolyn_Lawrence_Dill_GOMAP_Sorghum _DOE-JGI_v3.1.1_March_2023_v1.r1	Contributor	Fattel et al. (2023k)
Carolyn_Lawrence_Dill_GOMAP_Soybean _LIS_Wm82_IGA1008.gnm1.ann1.FGN6 _May_2023_v1.r1	Contributor	Fattel et al. (2023l)
Carolyn_Lawrence_Dill_GOMAP_Stiff_brome _DOE-JGI_Bd21_v3.2_May_2023_v1.r1	Contributor	Fattel et al. (2023m)

Table 6.2 Contributions to datasets out of the scope of the described research aims *continued (5)*.

Dataset Title	Contribution Type	Reference
Carolyn_Lawrence_Dill_GOMAP_Sugar_Pine _TreeGenes_v1.5_June_2023_v2.r1	Contributor	Fattel et al. (2023n)
Carolyn_Lawrence_Dill_GOMAP_Tea _Teabase_CSS_ChrLev_20200506 _June_2023_v1.r1	Contributor	Fattel et al. (2023o)
Carolyn_Lawrence_Dill_GOMAP_Tomato _SGN_SL4.0_July_2023_v2.r1	Contributor	Fattel et al. (2023p)
Carolyn_Lawrence_Dill_GOMAP_Wheat _URGI_IWGSC_RefSeq_v2.1_May_2023_v1.r1	Contributor	Fattel et al. (2023q)
Carolyn_Lawrence_Dill_GOMAP_Wild_Tomato _SGN_LA716_June_2023_v2.r1	Contributor	Fattel et al. (2023r)