SPEECH-BASED PHENOTYPING METHODS FOR FIELD STUDIES

Abstract

Field-based phenotyping of maize is time-consuming, cumbersome, and generally requires the evaluation of predefined traits of interest. We aim to expand researchers' field phenotyping "toolbox" by developing methods for collecting computationally tractable speech-based phenotyping useful for applications including, but not limited to, association genetics research. As a proof of concept, we designed an experiment that compares speech-derived trait concepts with traditional quantitative trait data collected by hand using the Wisconsin Diversity panel (grown in Ames, Iowa summer 2021). Details on methods, expectations, and current status for the project will be described.

Background

- Association panels, such as the Wisconsin Diversity panel, contain diverse genotypes that enable examining genetic marker associations with traits of agronomic interest [1,2,3].
- Genetic markers datasets for association studies, including Single Nucleotide Polymorphism (SNP) datasets, are available for the Wisconsin Diversity panel [1,2,3].
- Researchers can measure and score agronomic traits manually or through automated platforms that employ sensors [4,5,6,7,8].
- New field-based phenotyping methods are being developed to make observing plant phenotypes faster and less labor-intensive [7,8].
- Through computational methods, descriptions of plants are used to generate biologically meaningful connections between plant phenotypes and genes [9,10,11].
- Free-form speech descriptions of plants are useful for generating networks of semantic similarity or similarity of word meaning [11].

We are developing methods to do association genetics using recorded speech-based observations of plant phenotypes.

Methods

- In July of 2021, nine student workers recorded speech descriptions of accessions from the Wisconsin Diversity panel (Image 1).
- Speech-to-text tools, for example, Amazon Web Services (AWS) Transcribe [12], automate the transcription process to transform the raw speech data into computable text data (Figure 1: Audio Processing Component).
- Semantic similarity, or word meaning similarity, we can generate networks through Natural Language Processing (NLP) techniques from the text descriptions [9,10,11] (Figure 1: Association Study Component).
- We hypothesize that thresholding methods will result in clusters of descriptions with high semantic similarity, which we call synthetic phenotypes [10] (Figure 1: Association Study Component; Figure 2).
- A set of ~20 million SNPs were derived from whole genome resequencing and imputation and are stored as files compatible with association study tools bigsnpr [13], GAPIT [14], and FarmCPU [15].

Significance

- Describing visible phenotypes of plants with **speech tends to be faster and less restrictive** than traditional pen and paper scoring.
- We are **developing methods for in-field speech-based phenotyping**.
- **We aim to use synthetic phenotypes (Figure 2) as input for association studies** of the diverse genotypes of the Wisconsin Diversity panel **for marker trait associations.**

Next Steps

- Recreate previous association studies completed using the Wisconsin Diversity panel [2].
- Perform Association Study Component (Figure 1: Association Study Component; Figure 2) using tools described in [13, 14, 15].

Citations

[1] C. N. Hansey, J. M. Johnson, R. S. Sekhon, S. M. Kaeppler, and N. de Leon, "Genetic Diversity of a Maize Association Population with Restricted Phenology," *Crop Science*, vol. 51, no. 2, pp. 704–715, Mar. 2011.

[2] M. Mazaheri *et al.*, "Genome-wide association analysis of stalk biomass and anatomical traits in maize," *BMC Plant Biol*, vol. 19, no. 1, pp. 45–45, Jan. 2019.

[3] R. Bukowski *et al.*, "Construction of the third-generation Zea mays haplotype map," *GigaScience*, vol. 7, no. 4, p. gix134, Apr. 2018.

[4] D. Pauli *et al.*, "The Quest for Understanding Phenotypic Variation via Integrated Approaches in the Field Environment," *Plant Physiol*, vol. 172, no. 2, pp. 622–634, Oct. 2016.

[5] "Genomes to Fields Phenotyping Handbook." https://www.genomes2fields.org/about/project-overview/#standards-and-methods.

[6] J. Y. Kim, "Roadmap to High Throughput Phenotyping for Plant Breeding," *Journal of Biosystems Engineering*, vol. 45, no. 1, pp. 43–55, Mar. 2020.

[7] J. Barker *et al.*, "Development of a field-based high-throughput mobile phenotyping platform," *Computers and Electronics in Agriculture*, vol. 122, pp. 74–85, Mar. 2016.

[8] J. L. Crain *et al.*, "Development and Deployment of a Portable Field Phenotyping Platform," *Crop Science*, vol. 56, no. 3, pp. 965–975, May 2016.

[9] I. R. Braun and C. J. Lawrence-Dill, "Automated Methods Enable Direct Computation on Phenotypic Descriptions for Novel Candidate Gene Prediction," *Frontiers in Plant Science*, vol. 10, 2020.

[10] I. R. Braun, C. F. Yanarella, and C. J. Lawrence-Dill, "Computing on Phenotypic Descriptions for Candidate Gene Discovery and Crop Improvement," *Plant Phenomics*, vol. 2020, May 2020.

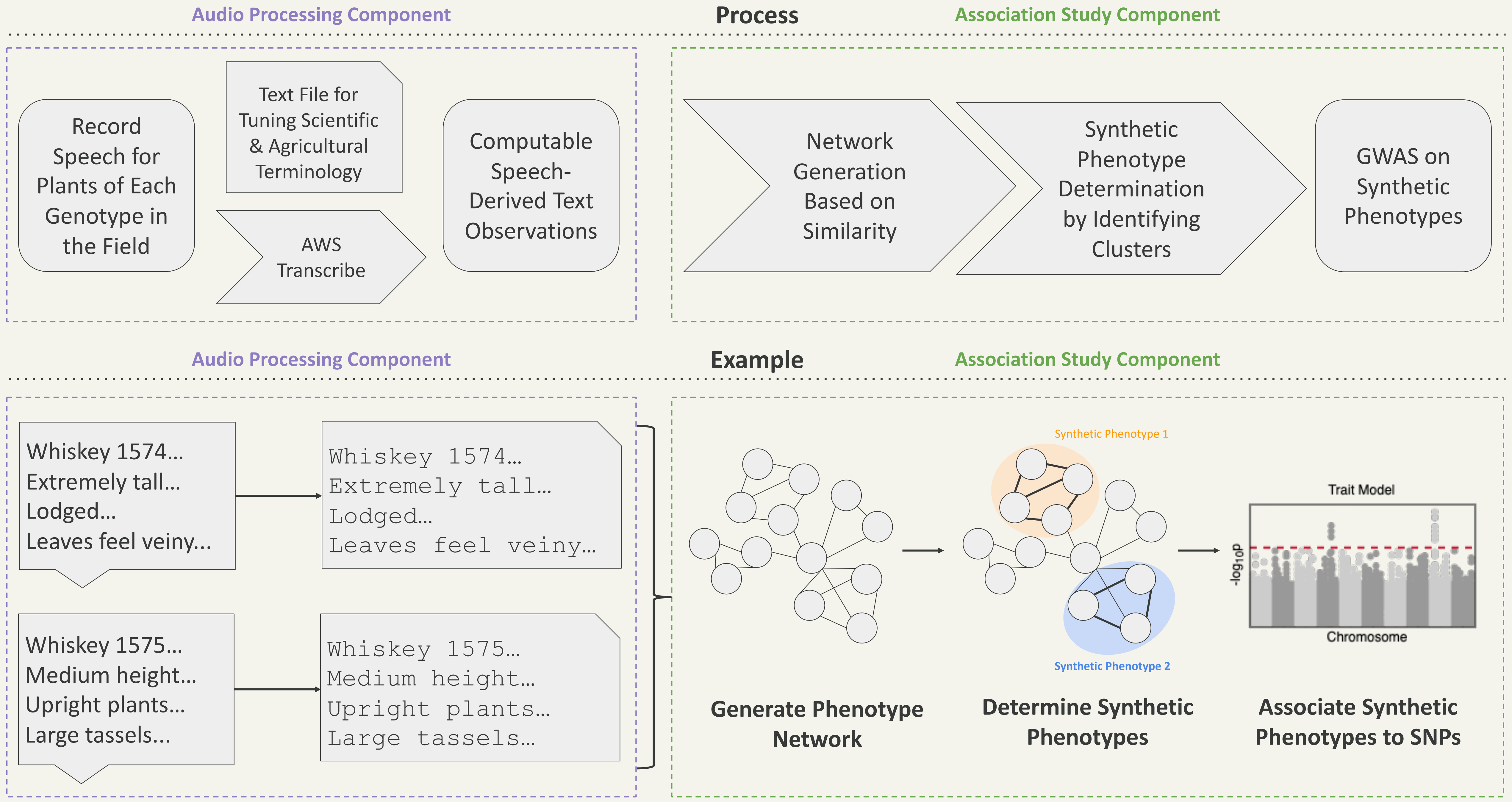
[11] I. R. Braun, D. C. Bassham, and C. J. Lawrence-Dill, "The Case for Retaining Natural Language Descriptions of Phenotypes in Plant Databases and a Web Application as Proof of Concept," *bioRxiv*, p. 2021.02.04.429796, Jan. 2021.

[12] Gary H. Kranz, "Amazon Transcribe Developer Guide." Amazon. https://docs.aws.amazon.com/transcribe/latest/dg/transcribe-dg.pdf#transcribe-whatis.

[13] F. Privé, H. Aschard, A. Ziyatdinov, and M. G. B. Blum, "Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr," *Bioinformatics*, vol. 34, no. 16, pp. 2781–2787, Aug. 2018.

[14] J. Wang and Z. Zhang, "GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction," *Genomics, Proteomics & Bioinformatics*, Sep. 2021.

[15] X. Liu, M. Huang, B. Fan, E. S. Buckler, and Z. Zhang, "Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies," *PLOS Genetics*, vol. 12, no. 2, p. e1005767, Feb. 2016.


Audio Processing Component — Process — Association Study Component

Record Speech for Plants of Each Genotype in the Field

Text File for Tuning Scientific & Agricultural Terminology

AWS Transcribe

Computable Speech-Derived Text Observations

Network Generation Based on Similarity

Synthetic Phenotype Determination by Identifying Clusters

GWAS on Synthetic Phenotypes

Audio Processing Component — Example — Association Study Component

Whiskey 1574…
Extremely tall…
Lodged…
Leaves feel veiny...

```
Whiskey 1574…
Extremely tall…
Lodged…
Leaves feel veiny…
```

Whiskey 1575…
Medium height…
Upright plants…
Large tassels...

```
Whiskey 1575…
Medium height…
Upright plants…
Large tassels…
```

Generate Phenotype Network

Determine Synthetic Phenotypes

Associate Synthetic Phenotypes to SNPs

←←←

Figure 1: Conceptual workflow of processing audio phenotypic description data for use in association studies.

Top represents the workflow process: *chevrons* indicate computational actions, *notched rectangles* indicate data files, and *rectangular bubbles* indicate expansive methods. *Bottom* represents a visual representation of a pictural example of the workflow process: *rectangular boxes with triangles on the bottom* indicate spoken descriptions*, notched rectangles* indicate text files, *networks* indicate semantic similarity networks. *Purple dashed* boxes indicate the "Audio Processing Component" and the *green dashed boxes* indicate the "Association Study Component."

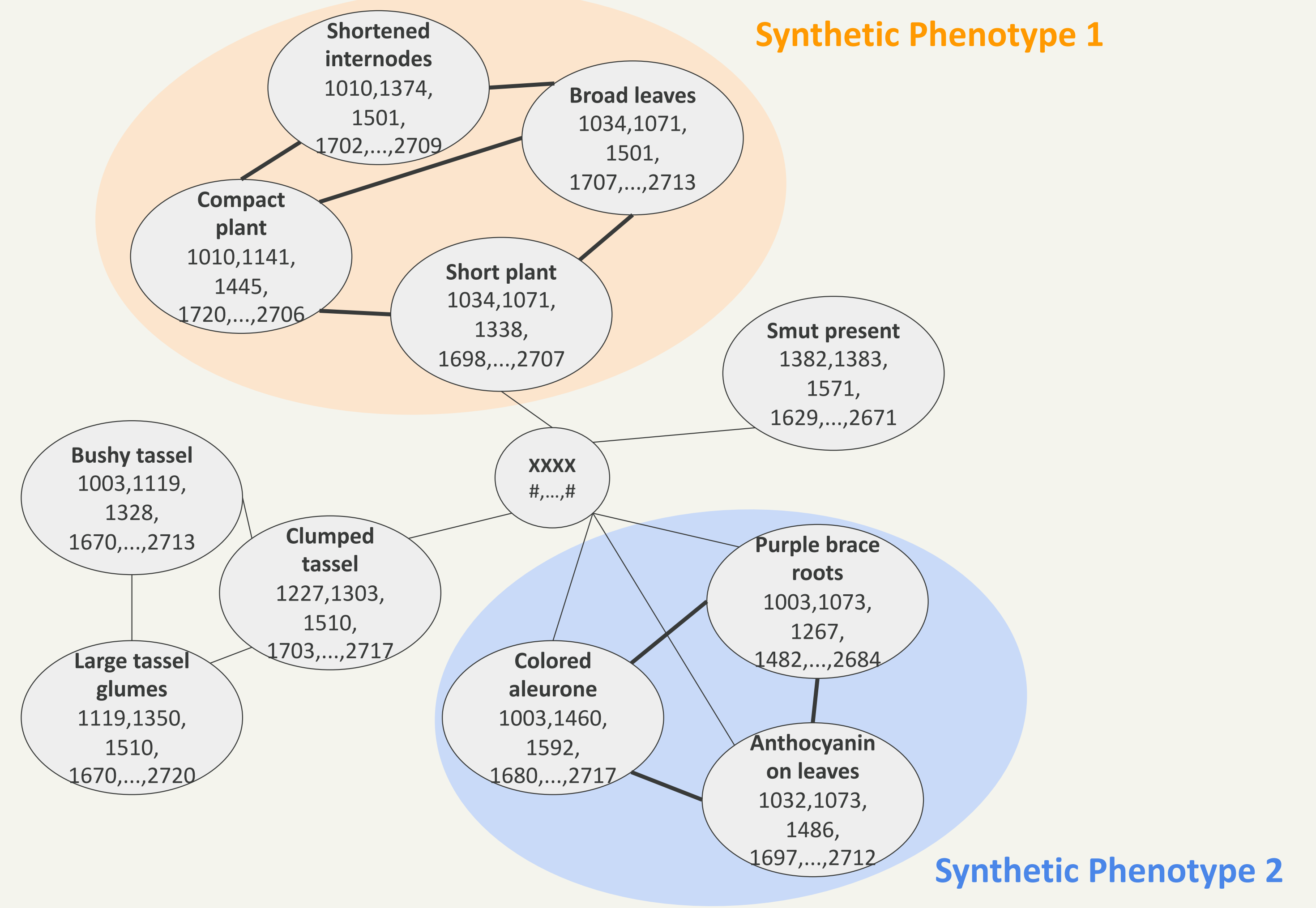
←←←

→→→

Figure 2: Network representing a hypothetical synthetic phenotype.

Gray ovals represent phenotype descriptions and genotype rows from field audio. *Light edges* indicate semantic similarity between terms, and *bold edges* indicate similarity scores above a threshold. *Orange oval* indicates hypothetical "Synthetic Phenotype 1", and *blue oval* indicates hypothetical "Synthetic Phenotype 2."

→→→



Synthetic Phenotype 1
Shortened internodes 1010,1374, 1501, 1702,…,2709
Broad leaves 1034,1071, 1501, 1707,…,2713
Compact plant 1010,1141, 1445, 1720,…,2706
Short plant 1034,1071, 1338, 1698,…,2707
Smut present 1382,1383, 1571, 1629,…,2671
XXXX #,…,#
Bushy tassel 1003,1119, 1328, 1670,…,2713
Clumped tassel 1227,1303, 1510, 1703,…,2717
Large tassel glumes 1119,1350, 1510, 1670,…,2720
Purple brace roots 1003,1073, 1267, 1482,…,2684
Colored aleurone 1003,1460, 1592, 1680,…,2717
Anthocyanin on leaves 1032,1073, 1486, 1697,…,2712
Synthetic Phenotype 2

←←←

Image 1: July 2021 Wisconsin Diversity panel workers recording speech observations.

←←←

I'M AN AGRONOMIST

IOWA STATE UNIVERSITY
Department of Agronomy

Colleen F. Yanarella[1], Darwin A. Campbell[2], Leila Fattel[1], Amanpreet Kaur[3,4], Rajdeep S. Khangura[3,4], Ásrún Ý. Kristmundsdóttir[2], Miriam D. Lopez[5], Summer 2021 Student Workers[1], Brian P. Dilkes[3,4], Carolyn J. Lawrence-Dill[1,2,6]

[1] Department of Agronomy, Iowa State University, Ames, IA
[2] College Agriculture and Life Sciences, Iowa State University, Ames, IA
[3] Department of Biochemistry, Purdue University, West Lafayette, IN
[4] Center for Plant Biology, Purdue University, West Lafayette, IN
[5] USDA-ARS Corn Insects and Crop Genetics Research Unit
[6] Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA

NSF DESE
1545453

NSF & USDA-NIFA
2021-67021-35329

IRB Study 21-179-00
Exempt Status