AN ABSTRACT OF THE DISSERTATION OF

Cedar Dale Warman for the degree of Doctor of Philosophy in Botany & Plant Pathology presented on June 10, 2020.

Title: Seeing Genes: Multiscale Phenotyping Reveals Gene Functions in Maize Pollen.

Abstract approved: _____

John E. Fowler

Plant sexual reproduction requires a broad array of molecular mechanisms to proceed successfully. Some of these mechanisms are well-studied, but our knowledge of them as a whole is fundamentally incomplete. Pollen tube growth is a key part of this process, facilitating the delivery of the sperm cells to the female ovule. In this work we examine pollen function on several scales in the model organism *Zea mays*. First, we characterize two homeologous genes in the *nop* family that promote pollen tube growth. These genes are highly expressed in maize pollen. The proteins encoded by these genes contain predicted functional domains involved in calcium and phosphoinositide binding and membrane organization, key mechanisms in pollen tube growth. Mutations in *nop* genes led to a reduction in pollen fitness, demonstrated by reduced transmission of mutant alleles when heterozygous mutant plants were crossed through the male, but not through the female. Examinations of *nop* mutant pollen tubes revealed shorter tubes in mutants than in wild type pollen. In one *nop* mutant, tube lengths were particularly sensitive to chemicals that interfered with phosphoinositide signaling, suggesting a role for *nop* intimately connected to those pathways.

Next, we studied a larger set of genes that were highly expressed in maize pollen. To identify these genes, we sequenced the maize pollen transcriptome at four developmental stages. Highly expressed genes in some of these stages were tested for their contributions to pollen fitness. Mutations in these genes were linked to fluorescent seed markers that could be tracked using a phenotyping system that we developed. Transmission rates of these mutant alleles were quantified, leading to the identification of several mutants that had negative effects on pollen fitness, suggesting functional roles for the genes they disrupted. One of these genes was *gex2*, which we linked to problems at fertilization. In this experiment, we demonstrated the

utility of phenotyping screens to track kernel markers, linking high expression to contributions of genes to pollen fitness.

The final section of this dissertation describes the method we developed to track kernels on a large scale. We created a novel high-throughput phenotyping system to scan maize ears. The system spins a maize ear while capturing a video, which is then processed into a flat projection of the surface of the ear. This platform creates a permanent record of the ear, which can then be measured to describe a variety of phenotypes. We designed and trained a deep-learning-based computer vision pipeline to rapidly identify kernel phenotypes. Our system identifies fluorescent and non-fluorescent kernels accurately, expanding our ability to study the contributions of many genes to pollen function.

This dissertation describes an effort to understand the function of genes, molecules that are invisible to the human eye. By building on the work of many scientists, we present several ways to "see" gene functions through genetic assays, phenotyping, and sequencing. These methods describe several scales of molecular function, from single genes to entire sets of genes. Ultimately, we hope to lay the groundwork for future studies that take advantage of multiscale phenotyping to uncover the molecular mechanisms of plant sexual reproduction.

©Copyright by Cedar Dale Warman
June 10, 2020
All Rights Reserved

Seeing genes: Multiscale Phenotyping Reveals Gene Functions in Maize Pollen

by
Cedar Dale Warman

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented June 10, 2020
Commencement June 2021

Doctor of Philosophy dissertation of Cedar Dale Warman presented on June 10, 2020

APPROVED:

Major Professor, representing Botany & Plant Pathology

Head of the Department of Botany & Plant Pathology

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Cedar Dale Warman, Author

ACKNOWLEDGEMENTS

To family, friends, colleagues, and strangers: without you things would be different, and probably worse.

This work was supported by the National Science Foundation grants IOS-1340050 (M. Evans PI), MCB-1832186 (John E. Fowler PI), and IOS-1340112 (Pankaj Jaiswal PI), as well as by the Department of Botany & Plant Pathology at Oregon State University.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF TABLES

Chapter 1

Deep-learning-based high-throughput phenotyping systems can drive future discoveries in plant reproductive biology

Cedar Warman, John E. Fowler

Abstract

High-throughput phenotyping systems are becoming critical for answering biological questions on a large scale. These systems have historically relied on simple computer vision techniques. However, neural networks, and specifically deep learning, are rapidly becoming more powerful and easier to implement. Here, we examine how deep learning can drive phenotyping systems and be used to answer fundamental questions in reproductive biology. We describe previous applications of deep learning, provide general recommendations for applying these methods to the study of plant reproduction, and present a case study in phenotyping maize ears. Finally, we highlight several cases where deep learning helped enable research that was previously out of reach and discuss the future outlook of these methods.

Introduction

Linking genotype to phenotype is a powerful method to understand complex biological systems. However, the increasing availability of genotypic data has not been matched by increasing high-throughput phenotyping capabilities. This is particularly the case in the field of plant reproduction, where relevant biological processes are heterogenous and difficult to observe, often taking place deep within vegetative and reproductive tissues. Until recently, computer vision based plant phenotyping methods have been dominated by easily observable structures, such as leaves (Arvidsson et al. 2011; Zhang et al. 2012, 2017; Junker et al. 2014; Choudhury et al. 2016; Awlia et al. 2016) and roots (Yazdanbakhsh and Fisahn 2009; Clark et al. 2013; Slovak et al. 2014; Passot et al. 2018; Jiang et al. 2019). These methods have relied on traditional computer vision, which leverages mathematical approaches such as thresholding (Otsu 1979), morphological opening and closing (Haralick et al. 1987), and k-means clustering (Lloyd 1982) to segment the object of interest from the background on a pixel level. These approaches are generally computationally efficient, but struggle when faced with complex backgrounds, closely packed objects, and objects with inconsistent features. Such challenging aspects are often found when phenotyping reproductive mutants, limiting the utility of traditional image processing approaches.

Over the past several years, neural network based deep learning methods have been increasingly applied to address a wide range of biological problems (reviewed in (Ching et al. 2018)). Neural-network-based computer vision approaches are fundamentally different from the previous generation of image processing techniques because they typically rely on large sets of training data to function. Neural networks work by combining input and output layers with at least one hidden layer. Each hidden layer is composed of a number of nodes (or neurons), which perform simple mathematical operations. Nodes are variously connected to adjacent layers, and have adjustable parameters that can be trained computationally to accurately predict patterns in the input layer. Data flow from the input layer, through the hidden layers, and finally to the output layer. Neural networks with multiple hidden layers are called deep neural networks.

Use of these types of networks has become a versatile solution to many phenotyping challenges. Recent improvements in ease of use, model robustness, and available computational resources make deep learning strategies well situated to drive future discoveries in plant reproductive biology. Here we examine potential applications of deep-learning-based image analysis to phenotyping challenges. We highlight current plant phenotyping systems taking advantage of deep learning, as well as useful deep learning frameworks, advice for implementing models, and the future outlook for these methods.

A gap in high-throughput phenotyping for plant reproduction

To date, deep learning approaches in plant phenotyping have been primarily utilized with large, easy to image structures. As with less complex computer vision approaches, these targeted structures include leaves (Ubbens and Stavness 2017; Ziamtsov and Navlakha 2019; Hüther et al. 2020) and roots (Pound et al. 2017; Douarre et al. 2018; Wang et al. 2019). Other applications include disease identification (Wang et al. 2017; Polder et al. 2019), inflorescence motion tracking (Gibbs et al. 2019), and fruit shape (Feldmann et al. 2020). These implementations of deep learning methods have performed well on images that challenge traditional computer vision approaches, particularly those with complex target objects and backgrounds. However, published tools for phenotyping reproductive tissues are notably scarce.

Deep learning approaches first require digital representations of the phenotype(s) of interest, generally in the form of images. Conceptually, there are three general targets for plant reproductive phenotyping (Figure 1A). For the first, whole plant or individual organs such as

inflorescences and fruit can be imaged using existing methods. Whole plant or organ imaging is generally simpler than other approaches, but observations are largely limited to secondary effects of reproductive processes. Such imaging has been used to track mutant inheritance through maize kernel markers (Warman and Fowler 2019), estimate seed yield (Uzal et al. 2018), and measure fruit characteristics (Feldmann et al. 2020). Key considerations for whole plant or organ imaging include imaging systems, background selection, and development of methods for discerning the secondary reproductive effects of interest. A second target for reproductive phenotyping is microscopic reproductive tissues, specifically the male and female gametophytes, the zygote and embryo, and surrounding sporophytic tissues. These tissues are experimentally useful because they offer the ability to directly observe reproductive processes *in vivo*. However, obtaining microscopic images of these tissues on a large scale can be challenging. Often the gametophytes are located deep within other tissues, and require special staining or fluorescent markers to visualize structures of interest. In addition, complicated backgrounds can make computer vision approaches challenging, even with the use of deep learning approaches. A final type of phenotyping target is gametophytes or zygotes imaged *in vitro* or semi-*in vivo* (Higashiyama et al. 2001; Palanivelu and Preuss 2006). These approaches have the potential to allow high-throughput imaging of gametophytes in a controlled environment, making them amenable to deep learning approaches, but may fail to capture some phenotypes only observed *in vivo*.

A variety of general purpose deep learning frameworks can be applied to reproductive phenotyping

Deep learning for image analysis can be broadly divided into three categories based on how objects are identified in the image: classification, object detection, and semantic segmentation (Figure 1B). In image recognition problems, classification answers what class, or type of object, describes a single image. This type of analysis can be helpful when individual images contain only one object to be identified, such as single ungerminated or germinated pollen grains. If more than one object is present in an image, or localization data are necessary, an object detection approach may be more suitable. Object detection uses deep learning (often convolutional neural networks, or CNNs) to identify both the classes and locations of multiple

objects in an image, signified by labeled bounding boxes. Some types of experiments, such as shape or area measurements, may require semantic segmentation. This approach identifies the class of every pixel in an image, outputting a mask over each object that shows not only its location, but its shape. The choice of deep learning strategy will vary by phenotyping task, and is important to consider before choosing a model, as models are optimized for specific strategies. An additional consideration is the ease of generating training data. While all deep learning approaches require training data, creating these data for classification tasks is faster than for object detection and semantic segmentation.

Several programming frameworks are available for neural network creation. Most require some knowledge of the Python or R programming languages, as well as a basic familiarity with command line interfaces and Linux/Unix operating systems. Many machine learning libraries exist, but here we highlight several freely available, open source frameworks that are a useful starting point. Currently, the most popular, and also the most versatile, library is TensorFlow (Abadi et al. 2016). TensorFlow has a well-developed set of machine learning tools, as well as thorough documentation and an active community of users. Many of the newest deep learning models are implemented in TensorFlow, and it remains at the forefront of machine learning. Taking full advantage of TensorFlow requires a deep understanding of neural networks and a high degree of coding ability. However, the Keras API (Chollet and Others 2015) simplifies many of TensorFlow's functions, and increases accessibility to researchers who are not primarily focused on machine learning. Several well-established and generalizable TensorFlow methods exist for computer vision tasks, including object detection and semantic segmentation.

Alternatives to TensorFlow include PyTorch (Paszke et al. 2019) and YOLO (Redmon and Farhadi 2018). While the user base and available resources are not as large as with TensorFlow, PyTorch offers a powerful alternative. As with TensorFlow, PyTorch provides versatile building blocks with which to construct neural networks. Similar network architectures can be created with both TensorFlow and PyTorch, so the best library for biological image analysis will depend on the researcher's preferences. In contrast, YOLO is a more specialized framework, offering one of the fastest object detection models available. This network is useful for applications in which speed is essential, such as real time analysis of video.

Implementing deep learning methods

A general workflow for implementing deep learning models can be broken down into a number of conceptual steps (Figure 1C). Before pursuing deep learning, it is important to ensure the research question is best answered by this approach. Because generating large amounts of high-quality training data requires a large initial time investment, traditional computer vision methods may be better suited to certain challenges, such as the segmentation of a green plant from a solid background. In addition, hand annotation might be a more efficient method when image sets are small. Deep learning approaches only start to see returns in truly high-throughput cases, with the analysis of hundreds or thousands of images.

If deep learning is well suited to the research question at hand, the first step is to choose which tissues or structures to image, and to develop a process to generate a large quantity of images. Many commercial and academic phenotyping platforms exist for image generation, but the specificity of reproductive structures may require the development of new technologies. Alongside image generation, the researcher should choose the category of image analysis (classification, object detection, or semantic segmentation) and deep learning model that can most effectively answer the research question. For example, counting fertilized ovules may only require object detection, whereas measuring the size of developing seeds could benefit from semantic segmentation.

Next, labeled training images must be generated, and subsequently split into training, validation, and test sets. Training and validation sets allow the model to learn and evaluate itself during the training process, while test sets are reserved for unbiased final evaluations of the trained model. A key consideration when training any deep learning model is to avoid overfitting the data. Overfitting results when the model predicts the training data too closely, and is consequently not generalizable to new data. Reserving an exclusive test dataset aside from training and validation datasets provides a final check that the model is generalizable beyond the datasets used to create the model. The number of training images required for an accurate model will vary by task, but expect to generate between 100-1000 training images, at a minimum. One commonly used strategy to leverage small amounts of training data is to employ transfer learning (Douarre et al. 2018; Hüther et al. 2020). Transfer learning begins with a well-trained model from a previous deep learning task, often a generalized task such as the identification of common objects, then retrains the top layers of the model with a new training

set for the task at hand. Another strategy is to use semi-automated methods to increase the size of training sets, a particularly useful method for semantic segmentation training data (Adams et al. 2020).

After sufficient training data are generated, the model is trained. In deep learning models, training is an iterative process in which parameters of the network's nodes are adjusted to optimize predictions. After each iteration, or epoch, the model adjusts to better predict outputs. Adjustments are typically made to minimize a "loss function," which is a measurement of how accurately the model predicted validation data. Model training time depends on the size of the training dataset and model complexity, and can vary widely, from minutes to days or even longer. The process can be vastly sped up with high performance GPUs, which are available through local computation resources or cloud based resources. Many cloud services offer computational resources optimized for deep learning frameworks.

Once a model is trained, predictions, or inference, can be made on experimental datasets. Inference is typically less computationally intensive than model training, and can often be run on a consumer grade machine. It is important that the researcher defines how the performance of the model is evaluated throughout the deep learning process. For image classification tasks, performance metrics typically aim to quantify the number of true positive, false positive, false negative, and true negative predictions (often termed the "confusion matrix"). Specific combinations of these quantities include accuracy, precision, recall, and specificity. Area Under the Receiver Operating Characteristics (AUROC) is a more thorough formulation of the previous measures. For object detection and semantic segmentation, the accuracy of object location and size predictions can be evaluated using intersection over union (IoU) and mean average precision (mAP). More accessible metrics can also be used. For example, with regression tasks, coefficients of determination comparing model predictions to ground truth are useful metrics, as well as mean absolute deviations.

Maize ear phenotyping: a case study

Our recent work to create a high-throughput phenotyping system for maize ears can serve as a practical example of one implementation of deep learning methods for plant reproductive phenotyping (Warman and Fowler 2019). Our goal was to track changes in Mendelian inheritance caused by mutations in genes highly expressed in the male

gametophyte. To do so, we used a collection of 56 GFP-tagged transposable element insertion mutants (Li et al. 2013; Warman et al. 2020). Kernels carrying mutant alleles express the Green Fluorescent Protein (GFP), allowing quantification of mutant transmission by counting fluorescent kernels. To facilitate high-throughput phenotyping, our method targeted the maize ear as a proxy, as functional defects in the male gametophyte result in altered transmission frequencies. We developed a custom rotational scanning system to capture images of the ear for downstream analysis (Figure 2A). We chose to use a deep learning approach to quantify kernels because traditional computer vision methods failed to overcome variations in our images, such as kernel shape and fluorescence intensity. For our deep learning model, we used the TensorFlow Object Detection API (Huang et al. 2016) implementation of Faster R-CNN with Inception Resnet v2 (Ren et al. 2015; Szegedy et al. 2016). This approach combined a powerful model with a simple API. We chose object detection because our primary goal was to count kernels to track mutant transmission. Segmentation-based approaches would have provided more detailed descriptions of each kernel, but bounding boxes gave us the information we needed, while minimizing the time required to produce training data. For training the model, we generated 300 images annotated with bounding boxes (Figure 2B), evenly distributed across ears generated in two field seasons. Each image took approximately 20 minutes to label, for a total of 100 hours of labor to create the entire set.

After image annotation, we split the dataset into training and validation sets. We used a transfer learning approach, taking advantage of a network pre-trained on the Collection Of Common Objects (COCO) dataset, which consists of ~200,000 labeled images of everyday objects. Transfer learning reduced both the amount of training data that was required, as well as the time required to train our model. Training took place on an Nvidia V100 GPU, and took approximately one hour. After the model was trained, we performed inference on the testing set, a set of 320 manually counted images that were not used in the training process (Figure 2C). We measured mAP, as well as calculated linear regressions and mean absolute deviations comparing hand counts to model predictions. The outcome was a highly accurate model, with adjusted R^2 for kernel counts between 0.96-0.98, depending on field season and phenotypic marker (Figure 2D).

The main challenges to implementing deep learning methods were imaging and training data generation. We found that using consistent imaging conditions (e.g., the same camera) was important, as networks trained on data from one camera were not easily transferable to data collected on another camera. In addition, the large number of kernels present in each

image reached the limits of our GPU's memory capacity, requiring a strategy of image subdivision and recombination. Ultimately, deep learning methods enabled us to both 1) survey a larger number of ears than would have been possible by manual counting, leading to the discovery of new mutants with quantitative phenotypic effects; and 2) generate larger datasets for each allele, enabling a more accurate estimation of transmission rate and increased statistical power to identify significant effects.

The future of deep learning in reproductive phenotyping

Deep learning is a rapidly advancing field. Many deep learning models are general purpose, allowing advancements in unrelated fields to be rapidly applied to specific cases like plant reproduction phenotyping. A thorough understanding of general purpose machine learning libraries like TensorFlow and PyTorch will enable biologists to take advantage of new research in image analysis. While classification and object detection models are relatively mature, semantic segmentation models are an area of highly active research. These models will gradually improve, allowing for improved identification of complex objects in noisy images, such as pollen tubes growing down the pistil or ovules at fertilization.

Large, high-quality training datasets improve the accuracy and versatility of models. Generating training datasets remains a major barrier to implementing these models. However, various methods of data augmentation can help researchers make the most out of available data. Traditional image augmentations include transformations like random cropping, horizontal and vertical flips, hue adjustments, and random noise. Recent work has developed new methods for increasing the size of training sets. One method is through generative adversarial networks, or GANs (Zhu et al. 2018). GANs are a type of machine learning that uses two neural networks to compete against each other to generative realistic data. One network generates data (in this case images), while the other network evaluates the data to determine which examples are real and which are artificial. The network learns from these comparisons, and can use this knowledge to create novel images that have the characteristics of the input set. These types of networks have been used to increase the size of training datasets to improve accuracy.

Another potential future direction is an expansion of what constitutes a phenotype. Historically, phenotypes have included simple concepts like malformed or undeveloped structures, growth defects, or color changes. Neural networks allow for the measurement of

more subtle, multidimensional phenotypes, such as abstractions for how images of plants change in response to treatments like drought or heat. An application of this concept, called latent space phenotyping (LSP), shows strong potential at characterizing changes in images of plants unrecognizable by the human eye (Ubbens et al. 2020). This approach compares sets of images from treated and control samples by using a neural network to create compressed representations of visual characteristics of the images, termed latent space. This process does not require labeled data beyond experimental and control sets, relying on the neural network to determine multidimensional phenotypes characteristic of the experimental set. These LSPs can then be interpreted using additional networks, ultimately uncovering complex phenotypes that, for example, can be used for genome-wide association studies (GWAS). This approach circumvents the need for training sets, and also has the potential to incorporate temporal changes into such phenotypes (Taghavi Namin et al. 2018). Additional methods can take more explicit approaches, such as multitask learning, which optimizes networks based on several predetermined phenotypes simultaneously (Dobrescu et al. 2020).

Implementing deep learning methods for phenotyping can be intimidating, as the time and effort necessary to develop a successful model is significant, and therefore not appropriate for all experiments. However, the power and versatility of these methods makes them particularly useful for high-throughput phenotyping, allowing for increased throughput, precision, and the measurement of complex phenotypes that may be difficult to describe using more traditional methods. As their use spreads, deep learning methods are becoming more accessible, making now an ideal time for applications in high-throughput phenotyping in plant reproductive biology.

References

Abadi M, Agarwal A, Barham P, et al (2016) TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv [cs.DC]

Adams J, Qiu Y, Xu Y, Schnable JC (2020) Plant segmentation by supervised machine learning methods. Plant phenome j 3:6980. https://doi.org/10.1002/ppj2.20001

Arvidsson S, Pérez-Rodríguez P, Mueller-Roeber B (2011) A growth phenotyping pipeline for Arabidopsis thaliana integrating image analysis and rosette area modeling for robust quantification of genotype effects. New Phytol 191:895–907.

https://doi.org/10.1111/j.1469-8137.2011.03756.x

Awlia M, Nigro A, Fajkus J, et al (2016) High-Throughput Non-destructive Phenotyping of Traits that Contribute to Salinity Tolerance in Arabidopsis thaliana. Front Plant Sci 7:1414. https://doi.org/10.3389/fpls.2016.01414

Ching T, Himmelstein DS, Beaulieu-Jones BK, et al (2018) Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 15.: https://doi.org/10.1098/rsif.2017.0387

Chollet F, Others (2015) Keras. https://keras.io

Choudhury SD, Stoerger V, Samal A, et al (2016) Automated vegetative stage phenotyping analysis of maize plants using visible light images. In: KDD workshop on data science for food, energy and water, San Francisco, California, USA. researchgate.net

Clark RT, Famoso AN, Zhao K, et al (2013) High-throughput two-dimensional root system phenotyping platform facilitates genetic analysis of root growth and development. Plant Cell Environ 36:454–466. https://doi.org/10.1111/j.1365-3040.2012.02587.x

Dobrescu A, Giuffrida MV, Tsaftaris SA (2020) Doing More With Less: A Multitask Deep Learning Approach in Plant Phenotyping. Front Plant Sci 11:141. https://doi.org/10.3389/fpls.2020.00141

Douarre C, Schielein R, Frindel C, et al (2018) Transfer Learning from Synthetic Data Applied to Soil–Root Segmentation in X-Ray Tomography Images. Journal of Imaging 4:65. https://doi.org/10.3390/jimaging4050065

Feldmann MJ, Hardigan MA, Famula RA, et al (2020) Multi-dimensional machine learning approaches for fruit shape phenotyping in strawberry. Gigascience 9.: https://doi.org/10.1093/gigascience/giaa030

Gibbs JA, Burgess AJ, Pound MP, et al (2019) Recovering Wind-Induced Plant Motion in Dense Field Environments via Deep Learning and Multiple Object Tracking. Plant Physiol 181:28–42. https://doi.org/10.1104/pp.19.00141

Haralick RM, Sternberg SR, Zhuang X (1987) Image analysis using mathematical morphology. IEEE Trans Pattern Anal Mach Intell 9:532–550. https://doi.org/10.1109/tpami.1987.4767941

Higashiyama T, Yabe S, Sasaki N, et al (2001) Pollen tube attraction by the synergid cell. Science 293:1480–1483. https://doi.org/10.1126/science.1062429

Huang J, Rathod V, Sun C, et al (2016) Speed/accuracy trade-offs for modern convolutional object detectors. arXiv [cs.CV]

Hüther P, Schandry N, Jandrasits K, et al (2020) araDEEPopsis: From images to phenotypic traits using deep transfer learning. bioRxiv 2020.04.01.018192

Jiang N, Floro E, Bray AL, et al (2019) Three-Dimensional Time-Lapse Analysis Reveals Multiscale Relationships in Maize Root Systems with Contrasting Architectures. Plant Cell

31:1708–1722. https://doi.org/10.1105/tpc.19.00015

Junker A, Muraya MM, Weigelt-Fischer K, et al (2014) Optimizing experimental procedures for quantitative evaluation of crop plant performance in high throughput phenotyping systems. Front Plant Sci 5:770. https://doi.org/10.3389/fpls.2014.00770

Li Y, Segal G, Wang Q, Dooner HK (2013) Gene Tagging with Engineered Ds Elements in Maize. In: Peterson T (ed) Plant Transposable Elements: Methods and Protocols. Humana Press, Totowa, NJ, pp 83–99

Lloyd S (1982) Least squares quantization in PCM. IEEE Trans Inf Theory 28:129–137. https://doi.org/10.1109/TIT.1982.1056489

Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9:62–66

Palanivelu R, Preuss D (2006) Distinct short-range ovule signals attract or repel Arabidopsis thaliana pollen tubes in vitro. BMC Plant Biol 6:7. https://doi.org/10.1186/1471-2229-6-7

Passot S, Moreno-Ortega B, Moukouanga D, et al (2018) A New Phenotyping Pipeline Reveals Three Types of Lateral Roots and a Random Branching Pattern in Two Cereals. Plant Physiol 177:896–910. https://doi.org/10.1104/pp.17.01648

Paszke A, Gross S, Massa F, et al (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, et al. (eds) Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp 8026–8037

Polder G, Blok PM, de Villiers HAC, et al (2019) Potato Virus Y Detection in Seed Potatoes Using Deep Learning on Hyperspectral Images. Front Plant Sci 10:209. https://doi.org/10.3389/fpls.2019.00209

Pound MP, Atkinson JA, Townsend AJ, et al (2017) Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. Gigascience 6:1–10. https://doi.org/10.1093/gigascience/gix083

Redmon J, Farhadi A (2018) YOLOv3: An Incremental Improvement. arXiv [cs.CV]

Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Cortes C, Lawrence ND, Lee DD, et al. (eds) Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp 91–99

Slovak R, Göschl C, Su X, et al (2014) A Scalable Open-Source Pipeline for Large-Scale Root Phenotyping of Arabidopsis. Plant Cell 26:2390–2403. https://doi.org/10.1105/tpc.114.124032

Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2016) Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv [cs.CV]

Taghavi Namin S, Esmaeilzadeh M, Najafi M, et al (2018) Deep phenotyping: deep learning for temporal phenotype/genotype classification. Plant Methods 14:66.

https://doi.org/10.1186/s13007-018-0333-4

Ubbens J, Cieslak M, Prusinkiewicz P, et al (2020) Latent Space Phenotyping: Automatic Image-Based Phenotyping for Treatment Studies. Plant Phenomics 2020:5801869. https://doi.org/10.34133/2020/5801869

Ubbens JR, Stavness I (2017) Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks. Front Plant Sci 8:1190. https://doi.org/10.3389/fpls.2017.01190

Uzal LC, Grinblat GL, Namías R, et al (2018) Seed-per-pod estimation for plant breeding using deep learning. Comput Electron Agric 150:196–204. https://doi.org/10.1016/j.compag.2018.04.024

Wang G, Sun Y, Wang J (2017) Automatic Image-Based Plant Disease Severity Estimation Using Deep Learning. Comput Intell Neurosci 2017:2917536. https://doi.org/10.1155/2017/2917536

Wang T, Rostamza M, Song Z, et al (2019) SegRoot: A high throughput segmentation method for root image analysis. Comput Electron Agric 162:845–854. https://doi.org/10.1016/j.compag.2019.05.017

Warman C, Fowler JE (2019) Custom built scanner and simple image processing pipeline enables low-cost, high-throughput phenotyping of maize ears. bioRxiv 780650

Warman C, Panda K, Vejlupkova Z, et al (2020) High expression in maize pollen correlates with genetic contributions to pollen fitness as well as with coordinated transcription from neighboring transposable elements. PLoS Genet 16:e1008462. https://doi.org/10.1371/journal.pgen.1008462

Yazdanbakhsh N, Fisahn J (2009) High throughput phenotyping of root growth dynamics, lateral root formation, root architecture and root hair development enabled by PlaRoM. Funct Plant Biol 36:938–946. https://doi.org/10.1071/FP09167

Zhang X, Hause RJ Jr, Borevitz JO (2012) Natural Genetic Variation for Growth and Development Revealed by High-Throughput Phenotyping in Arabidopsis thaliana. G3 2:29–34. https://doi.org/10.1534/g3.111.001487

Zhang X, Huang C, Wu D, et al (2017) High-Throughput Phenotyping and QTL Mapping Reveals the Genetic Architecture of Maize Plant Growth. Plant Physiol 173:1554–1564. https://doi.org/10.1104/pp.16.01516

Zhu Y, Aoun M, Krijn M, Vanschoren J (2018) Data Augmentation using Conditional Generative Adversarial Networks for Leaf Counting in Arabidopsis Plants. BMVC

Ziamtsov I, Navlakha S (2019) Machine Learning Approaches to Improve Three Basic Plant Phenotyping Tasks Using Three-Dimensional Point Clouds. Plant Physiol 181:1425–1440. https://doi.org/10.1104/pp.19.00524

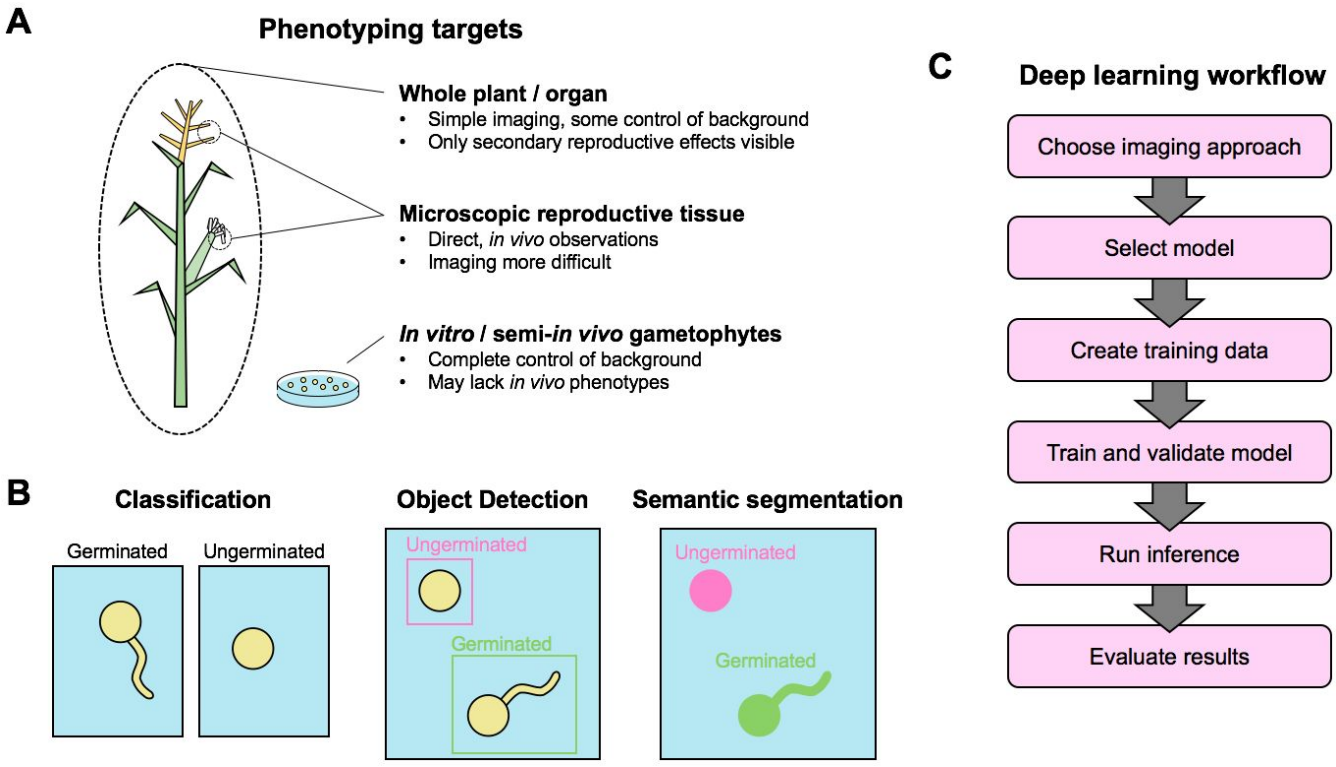


Figure 1: Leveraging deep learning for high-throughput phenotyping in plant reproduction

(A) Potential imaging targets for high-throughput phenotyping of reproductive systems can range from the whole plant to individual organs to microscopic structures, both *in vivo* and *in vitro*. **(B)** General deep learning strategies for image analysis. In classification, the class of each image is described, here germinated versus ungerminated pollen. When a single image contains more than one object, object detection can be used, a method that identifies objects and classes by bounding boxes. Semantic segmentation identifies the class of objects in an image on a pixel level, allowing for the identification of object attributes like shape and area. **(C)** Conceptual steps for implementing deep learning models.

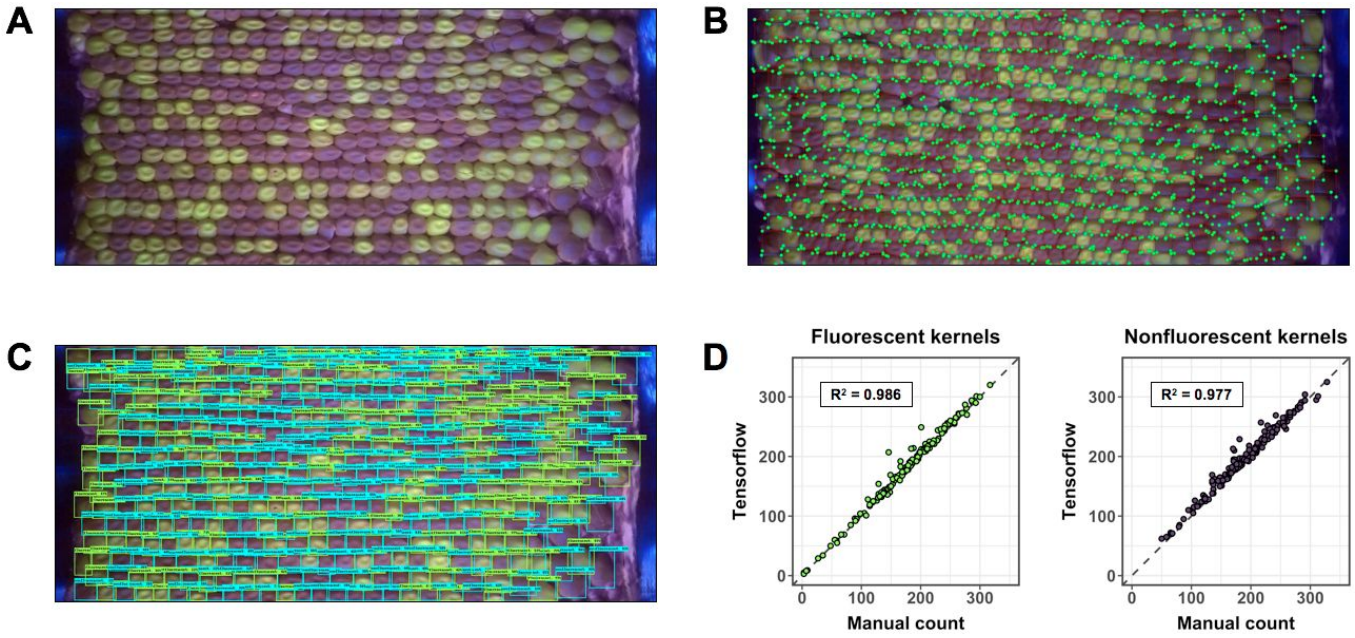


Figure 2: Maize ear scanning as an example deep learning workflow

(A) As an example, here we highlight a high-throughput maize ear phenotyping system. For our imaging approach, we developed a rotational scanning system that creates a flat projection of the surface of an ear. Fluorescent kernel markers are visible in this projection, signifying the presence of a genetically engineered transposable element insertion of interest. The ratio of fluorescent (mutant) kernels to non-fluorescent (wild-type) kernels can be tracked to screen for non-Mendelian inheritance of the mutant alleles. **(B)** To train our model, we created a training set from 300 images, with bounding boxes labeling each kernel (corners marked by green circles). We used a transfer learning approach and the Tensorflow Object Detection API to create our model based on the training dataset. **(C)** After training our model, we ran inference on our test set of images. Bounding boxes shown here are model predictions, with blue boxes signifying non-fluorescent kernels and green boxes signifying fluorescent kernels. **(D)** To evaluate the results of our model inference, we plotted model predictions versus manual counts for fluorescent and non-fluorescent kernels (2018 field season results shown here). Our model predicted fluorescent and non-fluorescent kernel counts with a high degree of accuracy (linear regression, adjusted R^2 = 0.986, 0.977, respectively).

Chapter 2

The *nop* genes promote pollen tube growth in the maize male gametophyte

Cedar Warman, Sean Colebrook, John E. Fowler

Abstract

Pollen tube growth is fundamental to plant reproduction, transporting the sperm cells to the embryo sac for double fertilization. Maize genes Zm00001d046540 and Zm00001d015789, designated *nop1* and *nop2*, are paralogous genes highly and specifically expressed in pollen. NOP proteins contain C2, GRAM, and VASt domains, which have predicted roles in calcium and phosphoinositide binding and membrane organization. Independent lines with transposable elements inserted in exons of either *nop1* (*Ds*) or *nop2* (*Mu*) were found to be linked to gametophytically-expressed male-specific transmission defects (16.1% and 2.3% transmission, respectively). Derivative *nop1* alleles with non-frameshifted footprints and a second mutant allele of *nop2* confirmed *nop1* and *nop2* male transmission defects are associated with reduced *nop1* and *nop2* function. Anthers and pollen grains showed no visible sporophytic phenotypes in homozygous *nop1::Ds*, *nop2::Mu*, and double homozygous mutant lines. *Waxy1* (*wx1*)-linked marker lines were generated for *nop1* and *nop2*, allowing *in vitro* pollen tube growth to be measured. *nop1::Ds* and *nop2::Mu* pollen tubes were significantly shorter than wild type pollen tubes (12% and 34% tube length reduction, respectively), indicating that NOP1 and NOP2 promote pollen tube growth and/or germination, and suggesting a possible explanation for the associated male transmission defects. Chemical trials were performed to examine tube length sensitivity in *wx1*-linked *nop2::Mu* pollen. *nop2::Mu* pollen tubes showed increased sensitivity to Wortmannin and LY294002, inhibitors of phosphoinositide 3-kinases, but not to Latrunculin B, an inhibitor of actin polymerization. Our data support the hypothesis that *nop* genes have an important role in the pollen tube, where they could help link phosphoinositides to membrane organization.

Introduction

The majority of the human food supply is the direct result of sexual reproduction, both for the plants and the seeds that it produces, and for the genetic diversity that allows crops to adapt to environmental changes. In flowering plants, sexual reproduction culminates in the delivery of two sperm cells to the ovule, where they fuse with the egg cell and central cell to complete double fertilization. For this process to take place, the non-motile sperm cells must be delivered to the ovule by the haploid vegetative cell through pollen tube growth (reviewed in (Johnson et al., 2019)). As the number of pollen tubes is far greater than the number of ovules, pollen tube

growth can be highly competitive (Hormaza and Herrero, 1996; Erbar, 2003; Williams and Reese, 2019). Likely as a consequence, pollen tubes are among the fastest growing eukaryotic cells (Stone et al., 2004).

Pollen tubes grow from the tip, exhibiting a characteristic "clear zone," the site of plasma membrane expansion and cell wall construction through secretory pathways (reviewed in (Hepler and Winship, 2015)). Tip growth is dominated by a tightly regulated interplay of endo- and exocytosis (Moscatelli and Idilli, 2009; Guo and Yang, 2020), delivering materials to the growing cell wall and subsequently modifying those materials to increase structural integrity during rapid expansion. A range of intracellular signaling pathways control this process, with a tip-focused calcium gradient playing a central role (reviewed in (Steinhorst and Kudla, 2013)). Calcium ion concentrations have been linked to the organization of pollen tip actin filaments (Zhang et al., 2010), membrane organization (Camacho and Malhó, 2003), and Rho GTPase (ROP) signaling pathways (Li et al., 1999; Gu et al., 2005). In turn, ROP signaling pathways are associated with a number of downstream effectors which can modulate pollen tube growth (Johnson et al., 2019), including actin remodeling, cell wall modifications, and feedback regulation of calcium concentrations. Because of the importance of membrane dynamics in pollen tube growth, phosphoinositides are also thought to play a key signaling role, with evidence for involvement in trans-golgi network organization, membrane trafficking, and vacuolar functions (reviewed in (Heilmann and Ischebeck, 2016)).

A full model linking calcium, ROP, and phosphoinositide signaling to the structures necessary for pollen tube growth is incomplete, but several key proteins have been identified, particularly in the model plant Arabidopsis. Progress towards understanding how these various pathways interact, and identifying the driving molecular components underlying pollen tube growth, have also been made in other species, with some highly conserved proteins appearing to have common functions across Embryophyta (e.g., *rop2* in maize (Arthur et al., 2003) and pollen receptor-like kinases in several plant species (Muschietti and Wengier, 2018)). Here, we describe two paralogous genes, Zm00001d046540 and Zm00001d015789 (*nop1* and *nop2*, respectively), that have roles in promoting competitive pollen tube growth in maize. Pollen with transposable element insertion mutations in these genes show male-specific transmission defects as well as diminished tube growth when compared to those of wild-type pollen. Based on the predicted domain structure of these proteins, we hypothesize that the proteins interact with phosphoinositide signaling pathways, and we provide evidence to support this possibility.

Results

nop genes are highly expressed in maize pollen and encode putative C2, VASt, and GRAM domains

An analysis of highly expressed genes in the maize male gametophyte (pollen) identified the gene designated GRMZM2G372877/Zm00001d046540 as associated with a male-specific transmission defect when mutant (Chettoor et al., 2014), suggesting an important function for the gene at this developmental stage. Maize has a second gene, GRMZM2G470666/Zm00001d015789, with a high level of sequence similarity. The translated sequences of two genes have 75.6% amino acid similarity and are in syntenic regions of maize chromosomes 9 and 5, respectively, and thus are likely homeologous genes. Based on transcriptomic profiling across maize development, the genes were found to be highly and specifically expressed in mature pollen (Figure 1A, left panel) (Walley et al., 2016). An independent study of gene expression focusing on maize male reproductive development found both genes to be highly expressed in mature pollen, and to a lesser extent sperm cells, but detectable only at very low levels in the tassel primordia and microspore (Figure 1A, right panel) (Warman et al., 2020).The translated sequences of the two maize genes are similar to the translated sequences of the rice ortholog *Osnop*, and have 84.1% and 78.3% amino acid similarity, respectively. In rice, a 65 kb deletion that affected *Osnop* and 13 other genes resulted in a sporophytic defect of undeveloped pollen and anthers, and the loss of *Osnop* (‘*no pollen*’) function was identified as the most likely causal candidate for this defect (Jiang et al., 2005). Due to this orthology, we will hereafter refer to Zm00001d046540 and Zm00001d015789 as *nop1* and *nop2*, respectively.

Both *nop1* and *nop2* contain predicted C2, VASt, and GRAM domains (Figure 1B). This domain structure is conserved in homologous proteins throughout Embryophyta (Supplemental Figure 1). C2 domains bind calcium and have multiple roles in signaling and membrane trafficking, and are generally thought to target proteins to cell membranes in a calcium-dependent manner (Coussens et al., 1986; Nalefski and Falke, 1996). VASt domains (also known as StART-like domains) are associated with sterol binding and membrane contact sites (Khafif et al., 2014; Gatta et al., 2018), whereas GRAM domains can bind phosphoinositides and may be involved in signaling pathways (Doerks et al., 2000; Jiang et al., 2008). Due to the potential linkage to signaling pathways known to influence the pollen tube,

and to the preliminary data linking mutation in *nop1* to a quantitative phenotype, we investigated *nop1* and *nop2* for potential roles in maize male gametophyte development and function.

A line containing a *Ds* element insertion in the second exon of *nop1* was acquired from the Brutnell/Vollbrecht collection (Figure 1C) (Vollbrecht et al., 2010). Independent lines containing a *Mu* element insertion in the second exon of *nop2* and a GFP-tagged *Ds* element insertion (*Ds-GFP*) in the 5' UTR of *nop2* were acquired from the Dooner/Du and UniformMu collections, respectively (Settles et al., 2007; Li et al., 2013; McCarty et al., 2013). Due to the transposable element insertions in coding regions, these lines were predicted to contain loss-of-function alleles of *nop1* and *nop2*. Insertion locations were verified by Sanger sequencing, and compared to the locations predicted from the high-throughput flanking sequence data used to originally characterize these insertions. The *nop1::Ds* insertion was in the predicted location, whereas *nop2::Mu* and *nop2::Ds-GFP* were 1 bp downstream and 7 bp upstream of their predicted locations, respectively. PCR assays were developed to reliably genotype plants with *nop1::Ds* and *nop2::Mu* insertions (Supplemental Figure 2).

Homozygous lines were generated containing the coding sequence insertion alleles in *nop1* and *nop2*, as well as a double homozygous mutant line. All lines were screened for sporophytic defects in male gametophyte development (Figure 1D). We observed no sporophytic defects in the anther for *nop1* and *nop2* homozygous mutants, as well as for the *nop1*/*nop2* double homozygous mutant. Pollen obtained from these anthers had no obvious phenotypic difference from wild type.

Mutations in *nop* genes cause male-specific transmission defects that are partially recovered by reduced competition

We tested haploid mutant pollen for gametophytic defects. Transmission of *nop1::Ds* and *nop2::Mu* mutant alleles was measured by reciprocal heterozygous outcrosses both through the female and the male to wild-type testers (Figure 2A, left). In addition, the effect of pollen competition was measured by crossing heterozygous mutant plants through the male with heavy and sparse pollen loads (Figure 2A, right). *nop1::Ds* mutant alleles were tracked with the *waxy1* (*wx1*) kernel marker (Collins, 1909). Both *nop1* and *wx1* are located on maize chromosome 9, with a measured recombination rate of 25.5%. Crossing *nop1::Ds* to inversion 9b reduced the recombination rate to 8.3%, enabling efficient tracking of mutant progeny using *wx1* (Figure 2B, Supplemental Table 1). *nop2::Mu* transmission rates were determined by PCR genotyping (Supplemental Figure 2).

When crossed through the female, the *nop1::Ds* allele showed no significant transmission difference from expected Mendelian (50%) inheritance (Figure 2C, top, χ^2 test p-value = 0.269). However, when crossed through the male, *nop1::Ds* showed a significant transmission defect, at 16.1% transmission (χ^2 test p-value < 10^{-16}). Reciprocal crosses were performed with heterozygous *wx1* plants linked to wild type *nop1* to control for potential transmission defects caused by the *wx1* mutation. *wx1* mutant alleles showed a slight but significant transmission defect, at 44.6% transmission (Supplemental Table 2). This transmission defect was significantly less than that observed when *wx1* was linked to *nop1::Ds* (Fisher's exact test p-value < 10^{-16}), indicating that *nop1::Ds* is associated with a significant gametophytic effect.

As with *nop1::Ds*, *nop2::Mu* showed no significant difference from Mendelian inheritance when crossed through the female (Figure 2C, bottom, χ^2 test p-value = 0.784), but showed a severe transmission defect when crossed through the male, with 2.3% transmission (χ^2 test p-value < 10^{-16}). Taken together, these results, detailed in Supplemental Table 2, suggest an important functional role for *nop* genes in the male gametophyte.

We next investigated the effect of pollen tube competition on *nop* mutant allele transmission by performing male outcrosses of heterozygous *nop1::Ds* and *nop2::Mu* mutants lines with heavy and sparse pollen loads (Figure 1D, results detailed in Supplemental Table 3). *nop1::Ds* transmission rates, as assessed via the linked *wx1*, significantly increased from 16.1% to 32.0% when competition was reduced with sparse pollination (Figure 1D, top, Cochran-Mantel-Haenszel test p-value = 0.00909). As previously, the experiment was repeated with the *wx1* heterozygous mutation alone to control for any possible transmission defects related to *wx1*. In this case, *wx1*$^-$ mutant pollen showed no significant increase in transmission rates when competition was reduced (Supplemental Table 3, Cochran-Mantel-Haenszel test p-value = 0.751). Similarly to *nop1::Ds*, *nop2::Mu* mutant alleles had significantly higher transmission rates with reduced competition in sparsely pollinated crosses, increasing from 4.1% to 11.8% transmission (Figure 1D, bottom, Cochran-Mantel-Haenszel test p-value = 0.0301). The increase in transmission rates for both *nop1::Ds* and *nop2::Mu* in an environment of reduced competition implies that the function associated with these genes has a role in pollen fitness during the competitive phase of gametophyte development following pollination.

Excision and second allele provide evidence that *nop* insertions cause transmission defects

To determine whether transmission defects are due to *nop1::Ds* insertion, we generated derivative alleles by first crossing lines with *nop1::Ds* mutants to an *Ac* active line. The presence of *Ac* causes excision of the non-autonomous *Ds* (McClintock, 1950), creating derivative alleles at a low frequency with added DNA 'footprints', typically ranging in size from 6 to 9 bp (Bai et al., 2007). By using PCR to screen individuals following an outcross of *Ac*-active *nop1::Ds* through the male, we identified several derivative alleles resulting from *Ds* excision (Figure 3A, top). Sanger sequencing identified four derivative alleles, *nop1-d2* (8 bp footprint), *nop1-d3* (9 bp footprint), *nop1-d4* (8 bp footprint), and *nop1-d6* (6 bp footprint). Heterozygous derivative alleles were outcrossed to wild type testers as males, with derivative transmission tracked by *wx1* kernel markers (Supplemental Table 4). Both *nop1-d2* and *nop1-d4* alleles were predicted to cause frameshift mutations in *nop1* due to their 8 bp footprints. Transmission rates for these alleles were 25.7% and 21.7%, respectively, retaining the male transmission defect observed in *nop1::Ds* (Figure 3A, bottom, χ^2 test p-value < 10^{-16}). *nop1-d3* and *nop1-d6* contained footprints of 9 and 6 bp, respectively, so were expected to cause small amino acid insertions that might not substantially disrupt NOP1 function. Transmission rates of these alleles through male outcrosses were not significantly different from Mendelian inheritance (χ^2 test p-value = 0.157, 0.276, respectively), indicating reversions of the mutant phenotype to wild-type, with small amino acid insertions that maintain the translation frame.

To verify transmission defects in *nop2::Mu* alleles were caused by defects in *nop2*, a second independent *nop2* allele was acquired from the Dooner/Du collection of GFP-tagged *Ds* element insertions. This second allele, *nop2::Ds-GFP*, contained an insertion in the predicted 5' UTR region of the gene (Figure 1C). GFP expression in this allele enabled the tracking of mutant transmission by quantifying fluorescent kernels (Figure 3B). Heterozygous male *nop2::Ds-GFP* plants were outcrossed in the field to wild type testers to track transmission (Supplemental Table 5). *nop2::Ds-GFP* showed a significant transmission defect when crossed through the male, at 46.25% transmission (χ^2 test p-value < 10^{-8}). Collectively, recovery of *nop1* transmission with a revertant excision allele and the presence of a transmission defect in an independent *nop2::Ds-GFP* allele confirm that the observed transmission defects are due to mutations in *nop1* and *nop2*.

Pollen from *nop1::Ds nop2::Mu* double heterozygotes has a germination defect, and transmits *in vivo* as predicted based on the single mutants

Double heterozygous *nop1::Ds nop2::Mu* pollen was germinated *in vitro* to determine if lower transmission rates of *nop* mutants were due to defects at germination. Double heterozygous mutant pollen germinated at a significantly lower rate than pollen from wild type siblings (Figure 4A, t-test p-value = 0.00542). In addition, double heterozygous pollen showed increased, but not significantly so, levels of ungerminated and burst pollen (t-test p-values = 0.139, 0.423, respectively). After meiosis, pollen grains from double heterozygous plants are 25% single mutant *nop1::Ds*, 25% single mutant *nop2::Mu*, 25% double mutant, and 25% wild type. Therefore, with this assay it is unclear which genotype(s) were responsible for reduced pollen germination. However, we can conclude that *nop* mutant genotypes can have a negative effect on pollen germination when compared to wild type.

If *nop1* and *nop2* are redundant or involved in similar pathways, a double mutant may lead to transmission defects that are greater than the combined transmission defects of individual mutants, known as a synergistic transmission defect. To determine if *nop1::Ds* and *nop2::Mu* alleles caused a synergistic transmission defect in double mutant pollen, we outcrossed double heterozygous plants to wild-type testers through the male with heavy and sparse pollination, as well as through the female as a control (Figure 4B, Supplemental Table 6). Based on individual *nop1::Ds* and *nop2::Mu* transmission defects, we predicted 0.37% transmission of both alleles together through the male. We observed one doubly heterozygote out of 119 progeny assessed by PCR (0.84% transmission), suggesting no significant interaction effect of *nop1::Ds* and *nop2::Mu* double mutants in pollen (χ^2 test p-value = 0.125). The distribution of single heterozygous, double heterozygous, and wild-type alleles was not significantly different from expected Mendelian ratios when crossed through the female (χ^2 test p-value = 0.298). As expected from individual *nop1::Ds* and *nop2::Mu* results, sparse pollination was associated with an increased transmission of both alleles together, from 0.84% to 3.18%.

nop mutants display *in vitro* pollen tube growth defects

We used an *in vitro* assay to examine *nop* pollen tube growth to further investigate one potential cause of mutant *nop* transmission defects (Figure 5). In addition to providing a useful kernel marker, *wx1* mutant pollen grains can be stained with I_2/KI to differentiate mutant pollen from wild type (Demerec, 1924; Brink and MacGillivray, 1924). In the presence of I_2/KI, mutant

wx1^{-} pollen grains stain white, whereas *Wx1*$^{+}$ pollen grains stain black (Figure 5A). By using previously described *nop1::Ds wx1*$^{-}$; *Nop1*$^{+}$ *Wx1*$^{+}$ *Inv9b* lines, haploid mutant and wild type *nop1* pollen grains could be identified in an *in vitro* assay where both genotypes originate from the same plant and are exposed to equivalent culture conditions. A similar experiment was designed with *nop2* heterozygous mutants by crossing a line containing mutant *nop2::Mu* by a reciprocal translocation line (Figure 5B). This line contained a reciprocal translocation between chromosomes 5 (the location of *nop2*) and 9 (the location of *wx1*), linking *nop2::Mu* to *Wx1*$^{+}$.

nop1 pollen tube lengths were measured from pollen collected from four individual plants, with *Nop1*$^{+}$ tube lengths compared to *nop1*$^{-}$ tube lengths by linked *wx1* markers (Figure 5C, left). *nop1*$^{-}$ mutant tube lengths were significantly shorter across all four plants (t-test p-values < 0.001), with a mean *nop1*$^{-}$ tube length reduction of 12%. To test for a possible effect of *wx1* mutation on tube length, sibling *wx1* heterozygous, *Nop1*$^{+}$ plants were also measured (Figure 5C, right). In control plants, there was no significant difference between *Wx1*$^{+}$ and *wx1*$^{-}$ tube lengths (t-test p-values > 0.5). For *nop2* heterozygous mutant plants, *nop2*$^{-}$ mutant tube lengths were also significantly shorter (t-test p-values < 10^{-15}), with a mean tube length reduction of 34% (Figure 5D, left). As with *nop1*, sibling plants that were heterozygous for *wx1* and wild type for *nop2* were measured as controls for the effect of *wx1* on tube length (Figure 5D, right). *nop2* control plants showed no difference between *Wx1*$^{+}$ and *wx1*$^{-}$ tube lengths (t-test p-values > 0.1). Tube length defects for *nop1* and *nop2* mutants offer a possible explanation for reduced transmission through the male, with more severe tube length reduction in *nop2* mutants associated with lower transmission rates for mutant alleles.

nop2::Mu pollen tubes show increased sensitivity to Wortmannin and LY294002, but not to Latrunculin B

To better understand the potential function of *nop* genes, we tested *nop2::Mu* pollen tube sensitivity to three chemicals: Latrunculin B, Wortmannin, and LY294002. We chose *nop2::Mu* for these experiments, as we reasoned that the increased tube length defect, compared to *nop1::Ds*, would likely make any quantitative effects more measurable. Wortmannin and LY294002 are phosphoinositide 3-kinase (PI3K) inhibitors, which have been shown to disrupt endocytosis in pollen tubes (Wang et al., 2013). Latrunculin B inhibits actin polymerization, pollen tube growth, and endocytosis (Gibbon et al., 1999; Vidali et al., 2001; Moscatelli et al., 2012). To test the effects of these chemicals on *nop2::Mu* pollen tubes, we used a modified version of the previous I_2/KI staining method in which pollen from heterozygous,

wx1-linked *nop2* plants was exposed to various concentrations of the chemicals during 30 minutes of *in vitro* growth. At increased concentrations of these chemicals, both mutant and wild type pollen grains experienced reduced tube lengths. In these trials, we tested for increased sensitivity in mutant pollen tube lengths relative to wild type when exposed to increased chemical concentrations (Figure 6A).

nop2::Mu mutant pollen tubes showed no increased sensitivity compared to wild type tubes when exposed to increased Latrunculin B concentrations (linear regression p-value = 0.317, Figure 6B). As a control, sibling heterozygous *wx1* plants were subjected to the same trial, with *wx1* mutants showing no increased sensitivity to Latrunculin B (linear regression p-value = 0.142). In contrast, *nop2::Mu* mutants showed greater sensitivity to increased concentrations of Wortmannin compared to wild type pollen (linear regression p-value = $< 10^{-4}$), whereas control plants showed no increased sensitivity (linear regression p-value = 0.476, Figure 6C). The second PI3K inhibitor tested, LY294002, led to similar results, with *nop2::Mu* mutants showing increased sensitivity (linear regression p-value $< 10^{-6}$) and controls remaining unaffected (linear regression p-value = 0.229, Figure 6D). These results suggest a functional relationship between *nop2* and PI3K inhibitors.

Discussion

nop genes have a role in pollen tube growth

nop genes are highly and specifically expressed in the maize male gametophyte and conserved across Embryophyta. Our study has established a role in pollen tube growth for both *nop1* and *nop2*. Mutations in *nop* genes lead to reduced transmission through the male (*nop1::Ds* 16.1%, *nop2::Mu* 2.3%), but female transmission rates are unaffected. Both *nop1::Ds* and *nop2::Mu* show reduced tube lengths when compared to wild type siblings (*nop1::Ds* 12%, *nop2::Mu* 33.9%), offering a possible explanation for reduced transmission rates. In addition, *nop1::Ds* and *nop2::Mu* show partial recovery of transmission defects when competition is reduced through sparse pollination. These transmission defects place *nop1* and *nop2* in the company of a potentially large number of genes that are highly expressed in mature pollen and have roles in pollen fitness (Warman et al., 2020).

Germination rates of *nop* mutants could also affect transmission rates of the mutant alleles by reducing the number of pollen grains producing viable pollen tubes. While double

heterozygous *nop1::Ds*; *nop2::Mu* pollen germinated at a slightly lower rate than wild type sibling pollen, the relative contributions of *nop1::Ds* and *nop2::Mu* to this germination defect are unknown. Despite this, the relatively low germination defect in the population segregating for both single and the double mutants suggests that pollen tube growth is likely the dominant determinant of low transmission rates. Beyond their shorter length, *nop* mutant pollen tubes showed no other obvious defects (e.g., aberrant shapes).

No evidence for sporophytic defect in maize *nop* mutants

Previous work on the orthologous rice gene *Osnop* described a severe sporophytic defect, including collapsed anthers and undeveloped pollen in mutant lines (Jiang et al., 2005). The described mutant was the result of a 65 kb deletion in a region that contained 14 predicted genes. Based on redundancy, tissue specific expression, and previous studies, the authors concluded that *Osnop* was responsible for the observed phenotype. Surprisingly, we found no evidence for a sporophytic defect in maize *nop* mutants. Anthers and pollen appeared to develop normally in homozygous mutants for *nop1*, *nop2*, and combined *nop1*; *nop2* double homozygous mutants. There are several possible explanations for this discrepancy. It remains a formal possibility that the two maize insertional mutants retain some residual function, when compared to the deletion associated with the rice phenotype. However, given that the insertions interrupt the *nop* coding sequences, we favor two other possibilities. First, the roles of *Osnop* and maize *nop* may have diverged evolutionarily in the time since their speciation. However, highly conserved sequences between *Osnop* and *nop1*/*nop2* (84.1% and 78.3% amino acid similarity, respectively), similar expression patterns, and *nop* conservation across diverse plant lineages, argue for a common function in these genes. Second, one or more of the remaining 13 rice genes deleted in the mutant line may have been responsible for the sporophytic anther phenotype. The interesting nature of the observed rice phenotype merits additional study, including additional examinations of the 13 other genes. Furthermore, a broader survey of *nop* mutants in other plant species could provide additional evidence for its role in sporophytic and/or gametophytic development.

Domain structure of *nop* offers insights into possible functions

The predicted NOP domain structure of C2-VASt-C2-GRAM-VASt is highly conserved in homologous proteins across Embryophyta (Supplemental Figure 1). C2 domains can bind

phospholipids in a calcium-dependent manner, and proteins containing C2 domains are responsible for a wide variety of functions in plants. In Arabidopsis, C2-containing proteins have been shown to play a role in defense responses (Yang et al., 2006), abscisic acid sensitivity (Rodriguez et al., 2014), flowering time (Liu et al., 2018), and pollen germination and tube growth (Wang et al., 2015). In rice, C2-domain proteins have been linked to defense responses (Kim et al., 2003), pollen fertility (Yang et al., 2008), and stress tolerance (Yokotani et al., 2009). C2 domains appear to serve as general purpose Ca^{+}-dependent membrane binding motifs, typically localized to plasma membranes, and often playing roles in the integration of signaling pathways.

Like C2, GRAM and VASt are typically associated with cellular membranes and lipid binding (Doerks et al., 2000; Khafif et al., 2014). These domains have been identified more recently than C2, but significant progress has been made in determining their functions. GRAM domains have been well-characterized in humans, and are thought to function in membrane organization and phosphatidylinositol signaling pathways (Berger et al., 2003; Tsujita et al., 2004; Choudhury et al., 2006; Besprozvannaya et al., 2018). In plants, GRAM-containing proteins have been shown to function in abscisic acid signaling pathways (Mauri et al., 2016), defense response (Li et al., 2016), and downstream inflorescence and seed development (Baron et al., 2014). The VASt domain, also designated the StART-like domain due to its structural similarity to the sterol-binding StART domain (González Montoro and Ungermann, 2015), is often present in the same protein with the GRAM domain, as in the Arabidopsis defense and cell death regulator VAD1 (Lorrain et al., 2004; Khafif et al., 2017) and yeast sterol transporter Ltc1 (Murley et al., 2015). In these contexts, VASt domains played a role in cellular localization that was both essential and distinct from GRAM domains (VAD1), and may assist in organizing or providing sterol transport function at membrane contact sites between the endoplasmic reticulum (ER), mitochondria, and vacuole (Ltc1).

Based on pollen tube growth defects, *nop* is likely to be involved in one or more of these functional roles. Pollen tube tip growth requires active vesicle transport, membrane organization, and interactions between the ER, Golgi apparatus, and cellular membranes, all in an environment that is highly dependent on calcium concentrations and phosphoinositide signaling (Johnson et al., 2019). These processes make *nop* well-situated to play a role in integrating signaling pathways, facilitating membrane contact sites, or targeting proteins to the cell membrane. Defects in any of these processes could result in the observed phenotype of reduced pollen tube growth. Likely *nop* plays a role in redundant or non-essential pathways, as

double homozygous *nop1::Ds* / *nop2::Mu* mutants can be generated. Cellular localization would further clarify the role of *nop*, and could be carried out in an experimental system more amenable to transformation than maize, such as lily, tobacco, or Arabidopsis.

Chemical sensitivity assays suggest a role for *nop* connected to phosphoinositide signaling

In this study, we performed chemical sensitivity assays to provide evidence for possible *nop2* functions. We found that *nop2::Mu* mutant pollen tubes were particularly sensitive to two PI3K inhibitors, Wortmannin and LY294002, but not sensitive to the actin polymerization inhibitor Latrunculin B. In this experiment, Latrunculin B served as a control for general sensitivity of *nop2::Mu* pollen, as none of the predicted functional domains in *nop2* had direct roles in actin dynamics. While secondary roles could be possible in the tightly linked processes governing pollen tube growth, the lack of sensitivity to Latrunculin B strongly supports the possibility of functional relevance in *nop2::Mu*'s sensitivity to PI3K inhibitors.

Phosphoinositide signaling is essential for pollen tube growth (Heilmann and Ischebeck, 2016). Phosphatidylinositol 3-phosphate has been localized to the pre-vacuolar compartment (PVC) in plant cells (Vermeer et al., 2006), including pollen tubes (Wang et al., 2013), and PI3K inhibitors have been shown to disrupt membrane trafficking between the PVC and Golgi apparatus in tobacco cell suspension cultures (Tse et al., 2004; Lam et al., 2007). Based on this evidence, *nop2* could be involved in these processes. As an alternative, *nop2* could be responding to indirect effects of PI3K disruption, as disruption in membrane trafficking affects a wide variety of cellular processes in growing pollen tubes. The visualization of phospholipid dynamics with label probes could be a useful approach to uncover more about *nop*'s role in these processes (Vermeer et al., 2006; Hammond and Balla, 2015).

Conclusions

nop genes have important roles in maize pollen tube growth, but how these roles fit into the mechanisms that govern pollen function is still unclear. Male transmission defects, reduced tube lengths, and increased sensitivity to PI3K inhibitors point to possible functions in the integration of phosphoinositide signaling with membrane organization in the dynamic process of tip growth, but further study is required to demonstrate this function. Cellular localizations of NOP proteins would be a useful first step, as well as targeted mutagenesis of proposed

domains. In addition, visualization of membrane dynamics, as well as experimentation with other inhibitors of calcium and phosphoinositide signaling, during pollen tube growth using labeled biomarkers in *nop* mutants could provide further insight into their functional significance.

As there are likely many highly expressed pollen genes that contribute to pollen fitness (Warman et al., 2020), developing systems to describe phenotypes in these genes will be a powerful tool to understanding pollen function. Here, we used an application of the *wx1* pollen phenotype (first described by (Brink and MacGillivray, 1924; Demerec, 1924), see (Phillips and Evans, 2011)) to track pollen tube growth in heterozygous plants. Because of large variations in *in vitro* growth of pollen tubes derived from different plants, this system enables phenotyping that would not otherwise be feasible. With the addition of automated image analysis, this method could provide a future approach to screen for abnormal pollen tubes in many different mutants, limited only by the ability to generate experimental lines with *wx1*-linked alleles.

Materials and Methods

Expression and sequence analysis

Zm00001d046540 and Zm00001d015789 (*nop1*, *nop2*) expression data were retrieved from qTeller (https://qteller.maizegdb.org/, (Walley et al., 2016)). Maize male gametophyte expression data was retrieved from (Warman et al., 2020). *nop1* and *nop2* sequences were retrieved from the Maize B73 reference genome version 4, B73 RefGen_v4 (Jiao et al., 2017) via MaizeGDB (Portwood et al., 2019). *Osnop* sequence was retrieved from GenBank, sequence AY702084.1 (Jiang et al., 2005). Pairwise amino acid sequence alignments were performed using EMBOSS Needle (Madeira et al., 2019). Functional domain predictions were carried out with EMBL SMART (Letunic and Bork, 2018). Synteny between *nop1* and *nop2* was examined using CoGe GEvo (Lyons and Freeling, 2008).

Transposable element insertion lines

Transposable element insertion lines were acquired for both *nop1* and *nop2* from the Maize Genetics Cooperation Stock Center. These included a *nop1* line containing *Ds* insertion 229619944 from the Brutnell and Vollbrecht collection of *Ds* insertion mutants (Vollbrecht et al., 2010), a *nop2* line containing *Mu* insertion mu1038778 from the UniformMu collection (McCarty et al., 2013), and a *nop2* line containing *Ds-GFP* insertion tdsgR95D10 from the Dooner and Du

collection of GFP-tagged *Ds* insertions (Li et al., 2013). All insertion sites were verified by PCR and Sanger sequencing. *nop1::Ds* primers were 372877_F2 (GATGCCAACCCCGTCGTCAG) and ACDS_JSR01 (GTTCGAAATCGATCGGGATA). *nop2::Mu* primers were 470666_F1 (TGTGGGACGAGGAGTTCGGG) and MuEnd (AGAGAAGCCAACGCCAWSGCCTCYATTTCGTC). *nop2::Ds-GFP* primers were 470666_F5 (CGAGCATATCGGAACAGGCCA) and GFP_3UTR (TGCAAGCTCGAGTTTCTCCA).

Growth conditions

Plants were grown at the Oregon State University Department of Botany & Plant Pathology Field Lab (Corvallis, OR) for sporophytic phenotypes, reciprocal cross transmission assays, and heavy/sparse crosses. Plants were grown in a controlled greenhouse environment (16 hrs light, 8 hrs dark, 80 ºF day/70 ºF night) for tube length and chemical trials.

Transposable element insertion line DNA extraction and genotyping

DNA extractions for genotyping were carried out from leaf tissue following the protocol outlined in (Vejlupkova and Fowler, 2003) (https://mnl.maizegdb.org/mnl/77/57vejlupkova.html). PCR reactions to assay transposable element insertion lines used three primers: two primers in the forward and reverse orientations in the coding sequence of the gene, and one primer within the transposable element insertion. Plants containing wild-type alleles of the gene produced PCR fragments amplified from the region between the forward and reverse coding sequence primers. Plants containing mutant alleles of the gene produced PCR fragments amplifying the region spanning the transposable element insertion primer and one or both of the coding sequence primers, depending on the transposable element's orientation and degree of repetitiveness in start and end regions. PCR product lengths were designed to be distinguishable by agarose gel electrophoresis, allowing wild type, heterozygous, and homozygous genotypes to be identified (Supplemental Figure 2). *nop1::Ds* primers were 372877_F4 (ACCAGCTCCAGCCCAAGAGC), B.S05.0827RA (CGACGACTGGATGGTCGGTG), and ACDS_JSR01 (GTTCGAAATCGATCGGGATA). *nop2::Mu* primers were 470666_F1 (TGTGGGACGAGGAGTTCGGG), 470666_R1 (TGGTGTCTGTTTCGACTGCTTTGA), and MuEnd (AGAGAAGCCAACGCCAWSGCCTCYATTTCGTC). *nop2::Ds-GFP* primers were 470666_F6 (ATTAGCCATTGATATATGTGATGCAA), 470666_R4 (CCGAAGTCCTCGCCATCCGC), and Ac5prime-178 (GTGAAACGGTCGGGAAACTAG CTCTAC).

Generation of *nop1::Ds* and *nop2::Mu* homozygous lines and sporophytic phenotyping

Homozygous *nop1::Ds* and *nop2::Mu* lines were created by self-crossing heterozygous plants and screening progeny for homozygotes via PCR. Because of strong transmission defects through the male, *nop1::Ds* and *nop2::Mu* double homozygous lines were generated in a two-step process. First, a homozygous mutant *nop1::Ds* line was crossed to a *nop2::Mu* heterozygous mutate line and progeny were screened by PCR to obtain a homozygous *nop1::Ds*, heterozygous *nop2::Mu* line. This line was then self-crossed, and progeny were screened by PCR to obtain the double homozygous line. Individual and double homozygotes, along with wild type relatives, were screened for abnormal anther and pollen phenotypes with a dissecting microscope.

Measuring recombination rates between *nop1* and *wx1*

Recombination rates for *nop1::Ds* linked in repulsion phase to *waxy1* (*wx1*) were measured by outcrossing a heterozygous line with genotype *nop1::Ds* $Wx1^+$ / $Nop1^+$ $wx1^-$ to a homozygous $wx1^-$ / $wx1^-$ tester. A total of 47 kernels from each kernel phenotype, $Wx1^+$ and $wx1^-$, were selected for genotyping. Kernels were germinated and DNA was extracted from leaf tissue as described in a previous section. *nop1* genotypes were assayed by PCR using previously described primers. Inversion 9b *nop1::Ds* lines were created by crossing a heterozygous *nop1::Ds* mutant line in a homozygous $wx1^-$ background (*nop1::Ds* $wx1^-$ / $Nop1^+$ $wx1^-$) to a homozygous line containing inversion 9b, acquired from the Maize Genetics Cooperation Stock Center. The resulting line with genotype *nop1::Ds* $wx1^-$ / $Nop1^+$ $Wx1^+$ *Inv9b* was crossed to a homozygous $wx1^-$ / $wx1^-$ tester. A total of 48 kernels were selected for genotyping, with DNA extraction and PCR as previously described. Results are summarized in Supplemental Table 1.

nop1::Ds and *nop2::Mu* reciprocal and heavy/sparse outcrosses

Reciprocal outcrosses of *nop1::Ds* and *nop2::Mu* plants were performed in the field under conditions described in a previous section. For *nop1::Ds*, heterozygous plants with genotype *nop1::Ds* $Wx1^+$ / $Nop1^+$ $wx1^-$ were reciprocally outcrossed to a homozygous $wx1^-$ / $wx1^-$ tester, with *nop1::Ds* mutant transmission tracked by *wx1* kernel markers. As a control for

potential *wx1* transmission defects independent of *nop1::Ds*, sibling plants with genotype *Nop1*$^+$ *Wx1*$^+$ / *Nop1*$^+$ *wx1*$^-$ were reciprocally outcrossed to a homozygous *wx1*$^-$ / *wx1*$^-$ tester. For *nop2::Mu*, heterozygous plants with genotype *nop2::Mu* / *Nop2*$^+$ were reciprocally outcrossed to wild-type testers, with *nop2::Mu* mutant transmission tracked by a previously described PCR assay. The significance of transmission defects was determined by χ^2 test, with p-value threshold of < 0.05.

Heavy and sparse male outcrosses for *nop1::Ds*, *nop1*$^+$ *wx1* control, and *nop2::Mu* plants were performed as described in the preceding section. Heavy crosses were performed after trimming maize ears the previous day, allowing the silks to emerge 1-3 cm, then saturating the silks with pollen. Sparse crosses were performed by lightly administering pollen to silks, with a goal of <5 pollen grains per silk (approximately 50-100 µl of pollen per ear, by volume). Changes in transmission rates between heavy and sparse crosses were quantified with the Cochran-Mantel-Haenszel test, with a significance at p-value < 0.05.

Generation of *nop1::Ds* derivative alleles and outcrosses

nop1::Ds derivative alleles were generated by crossing homozygous *nop1::Ds* lines to homozygous *Ac-im Ac* active lines (Conrad and Brutnell, 2005) acquired from E. Vollbrecht. Progeny were screened by PCR for the presence of DNA footprints resulting from *Ds* excision. Primers 372877_F2 (GATGCCAACCCCGTCGTCAG) and 372877_R2 (GAGCTCGCGTTTGGCTCTT) were selected in locations upstream and downstream of the *nop1::Ds* insertion site. Wild type PCR product length was expected to be 156 bp, while PCR products from alleles containing footprints were 6-9 bp longer. Derivative alleles were identified by PCR product size difference, as determined by gel electrophoresis on a 2% agarose gel. *nop-d2*, *-d3*, and, *-d4* were identified in this screen and verified by Sanger sequencing.

A second screen targeted 6 bp non-frameshifted derivative alleles using forward PCR primer 372877_F3 (GGTGGAGGTCCTGTTCCTGTC). The sequence of 372877_F3 was designed to exactly compliment the *nop1::Ds* insertion site with a 6 bp footprint. A detectable PCR product could be amplified with the addition of reverse primer 372877_R4 (GCTCAGTGTCCTCTGCACGAAG). Lines tested in this screen originated from the same genetic material as previous, but were crossed in an additional generation to homozygous *nop1::Ds* plants. Progeny were planted and arranged in a 15x15 grid (225 total plants). Pooled DNA extractions were performed for each column and row of the grid and tested for the presence of 6 bp derivative alleles using the previously described primers. A second round of

PCR was carried out for individual plants at intersections of rows and columns with positive results. *nop1-d6* was identified in this screen and verified by Sanger sequencing.

All derivative alleles were outcrossed to homozygous *wx1*$^{-}$ / *wx1*$^{-}$ testers as heterozygous males in the field (Supplemental Table 4). Inheritance of derivative alleles was tracked by linkage to *Wx1*$^{+}$ kernel markers. The significance of transmission defects was determined by χ^2 test, with a significant transmission defect defined as p-value < 0.05.

nop2::Ds-GFP outcrosses

Multiple sibling lines of the previously described *nop2::Ds-GFP* allele were outcrossed as heterozygous males to wild type testers in the field (Supplemental Table 5). Fluorescent kernel markers were quantified from dried ears using a rotational scanning system (Warman and Fowler, 2019). In brief, videos of rotating ears were captured, then videos were computationally projected into two-dimensional images. Fluorescent and non-fluorescent kernels were then counted using the Cell Counter plugin in the FIJI distribution of ImageJ (Schindelin et al., 2012), with significance of transmission defects determined by χ^2 test (significant transmission defect defined as p-value < 0.05).

Double heterozygous *nop* mutant pollen germination

Pollen was collected from *nop1::Ds* / *Nop1*$^{+}$; *nop2::Mu / Nop2*$^{+}$ double heterozygous and wild-type siblings from field grown plants. Pollen was germinated *in vitro* for 30 minutes in pollen growth medium (PGM) (Schreiber and Dresselhaus, 2003). Images of pollen were captured using a dissecting scope and digital camera. Pollen grains, categorized as germinated, ungerminated, or burst, were quantified using the FIJI distribution of ImageJ.

Generation of *nop2::Mu* reciprocal translocation line

nop2::Mu reciprocal translocation lines were generated to allow the use of *wx1* iodine staining to distinguish *nop2*$^{-}$ mutant from wild-type pollen *in vitro* through *wx1* iodine staining. A homozygous line containing maize reciprocal translocation T5-9(4817) was acquired from the Maize Genetics Cooperation Stock Center. To generate *nop2::Mu* reciprocal translocation lines, a *nop2::Mu* / *Nop2*$^{+}$; *wx1*$^{-}$ / *Wx1*$^{+}$ line was crossed by the homozygous *wx1*$^{-}$ T5-9(4817) / *wx1*$^{-}$ T5-9(4817) line. For tube length experiments, F1 progeny was selected with *Wx1*$^{+}$ kernel phenotypes to ensure the *Wx1*$^{+}$ allele was inherited, allowing for differentiation of *nop2*$^{-}$ and

Nop2^{+} alleles. We selected T5-9(4817) because of the close proximity of the chromosome 5 breakpoint (genetic map coordinate 80.0) to *nop2* (approximate genetic map coordinate 79.10), in order to minimize recombination (MaizeGDP Genetic 5, https://maizegdb.org/data_center/map?id=1203641).

I_2/KI staining to measure tube lengths

Heterozygous plants containing *nop1* and *nop2* linked to *wx1* as previously described, as well as sibling controls heterozygous for *wx1* and homozygous for *Nop1*$^{+}$ and *Nop2*$^{+}$, were grown in the greenhouse until anther dehiscence. Pollen was collected, germinated *in vitro*, and stained with I_2/KI following the protocol outlined in Supplemental Methods 1. In brief, pollen was collected for 1 hour, beginning between 10:00-11:00 am, using standard methods. Pollen was then germinated in PGM for 30 minutes. After germination and tube growth, pollen was first fixed in ethanol-acetic acid fix, then stained with I_2/KI. A dissecting scope with an attached digital camera was used to capture images of stained pollen.

Pollen tubes were measured using the FIJI distribution of ImageJ. A minimum of 104 and a mean of 292 pollen tubes were measured per plant. Significant differences between *wx1*$^{-}$ and *Wx1*$^{+}$ tube lengths were determined by two-sample t-test, with significance defined by p-value < 0.05.

nop2::Mu tube length chemical trials

Reciprocal translocation lines for both *nop2::Mu wx1* and wild type *Nop2*$^{+}$ *wx1* siblings were grown in the greenhouse for pollen chemical trials. Pollen was collected at anther dehiscence as previously described. Pollen from a single plant was used for 11 concentrations of each chemical in separate culture plates. Pollen germination and staining was carried out according to Supplemental Methods 1, with pollen added every 2 minutes to a pre-prepared chemical-PGM solution, mixed, and plated. After staining, pollen tubes for each chemical concentration were imaged and measured using the FIJI distribution of ImageJ. Latrunculin B was prepared as a 1 μM stock solution in DMSO, with experimental concentrations between 0-20 nM in 2 nM steps. Wortmannin (Sigma-Aldrich W1628) was prepared as a 1 mM stock solution in DMSO, with experimental concentrations between 0-10 μM in 1 μM steps. LY294002 (Sigma-Aldrich L9908) was prepared as a 2 mM stock solution in DMSO, with experimental concentrations between 0-20 μM in 2 μM steps.

Acknowledgements

Thank you to Zuzana Vejlupkova, Mendi Antisdel, Olyvia Childress, Rogue Hartman, Brooke Hamilton, Matthew Borchers, Peter de Roos, and Quinn DeYoung for technical assistance. This work was funded by National Science Foundation grant IOS-1340050 (M. Evans PI) and the Department of Botany & Plant Pathology at Oregon State University.

Works Cited

Arthur, K.M., Vejlupkova, Z., Meeley, R.B., and Fowler, J.E. (2003). Maize ROP2 GTPase provides a competitive advantage to the male gametophyte. Genetics **165**: 2137–2151.

Bai, L., Singh, M., Pitt, L., Sweeney, M., and Brutnell, T.P. (2007). Generating novel allelic variation through Activator insertional mutagenesis in maize. Genetics **175**: 981–992.

Baron, K.N., Schroeder, D.F., and Stasolla, C. (2014). GEm-Related 5 (GER5), an ABA and stress-responsive GRAM domain protein regulating seed development and inflorescence architecture. Plant Sci. **223**: 153–166.

Berger, P., Schaffitzel, C., Berger, I., Ban, N., and Suter, U. (2003). Membrane association of myotubularin-related protein 2 is mediated by a pleckstrin homology-GRAM domain and a coiled-coil dimerization module. Proc. Natl. Acad. Sci. U. S. A. **100**: 12177–12182.

Besprozvannaya, M., Dickson, E., Li, H., Ginburg, K.S., Bers, D.M., Auwerx, J., and Nunnari, J. (2018). GRAM domain proteins specialize functionally distinct ER-PM contact sites in human cells. Elife **7**.

Brink, R.A. and MacGillivray, J.H. (1924). Segregation for the Waxy Character in Maize Pollen and Differential Development of the Male Gametophyte. Am. J. Bot. **11**: 465–469.

Camacho, L. and Malhó, R. (2003). Endo/exocytosis in the pollen tube apex is differentially regulated by Ca2+ and GTPases. J. Exp. Bot. **54**: 83–92.

Chettoor, A.M., Givan, S.A., Cole, R.A., Coker, C.T., Unger-Wallace, E., Vejlupkova, Z., Vollbrecht, E., Fowler, J.E., and Evans, M.M. (2014). Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. Genome Biol. **15**: 414.

Choudhury, P., Srivastava, S., Li, Z., Ko, K., Albaqumi, M., Narayan, K., Coetzee, W.A., Lemmon, M.A., and Skolnik, E.Y. (2006). Specificity of the myotubularin family of phosphatidylinositol-3-phosphatase is determined by the PH/GRAM domain. J. Biol. Chem. **281**: 31762–31769.

Collins, G.N. (1909). A new type of Indian corn from China (US Government Printing Office).

Conrad, L.J. and Brutnell, T.P. (2005). Ac-immobilized, a stable source of Activator transposase that mediates sporophytic and gametophytic excision of Dissociation elements in maize. Genetics **171**: 1999–2012.

Coussens, L., Parker, P.J., Rhee, L., Yang-Feng, T.L., Chen, E., Waterfield, M.D., Francke, U., and Ullrich, A. (1986). Multiple, distinct forms of bovine and human protein kinase C suggest diversity in cellular signaling pathways. Science **233**: 859–866.

Demerec, M. (1924). A Case of Pollen Dimorphism in Maize. Am. J. Bot. **11**: 461–464.

Doerks, T., Strauss, M., Brendel, M., and Bork, P. (2000). GRAM, a novel domain in glucosyltransferases, myotubularins and other putative membrane-associated proteins. Trends Biochem. Sci. **25**: 483–485.

Erbar, C. (2003). Pollen Tube Transmitting Tissue: Place of Competition of Male Gametophytes. Int. J. Plant Sci. **164**: S265–S277.

Gatta, A.T., Sauerwein, A.C., Zhuravleva, A., Levine, T.P., and Matthews, S. (2018). Structural insights into a StART-like domain in Lam4 and its interaction with sterol ligands. Biochem. Biophys. Res. Commun. **495**: 2270–2274.

Gibbon, B.C., Kovar, D.R., and Staiger, C.J. (1999). Latrunculin B has different effects on pollen germination and tube growth. Plant Cell **11**: 2349–2363.

González Montoro, A. and Ungermann, C. (2015). StARTing to understand membrane contact sites. Trends Cell Biol. **25**: 497–498.

Guo, J. and Yang, Z. (2020). Exocytosis and endocytosis: coordinating and fine-tuning the polar tip growth domain in pollen tubes. J. Exp. Bot. **71**: 2428–2438.

Gu, Y., Fu, Y., Dowd, P., Li, S., Vernoud, V., Gilroy, S., and Yang, Z. (2005). A Rho family GTPase controls actin dynamics and tip growth via two counteracting downstream pathways in pollen tubes. J. Cell Biol. **169**: 127–138.

Hammond, G.R.V. and Balla, T. (2015). Polyphosphoinositide binding domains: Key to inositol lipid biology. Biochim. Biophys. Acta **1851**: 746–758.

Heilmann, I. and Ischebeck, T. (2016). Male functions and malfunctions: the impact of phosphoinositides on pollen development and pollen tube growth. Plant Reprod. **29**: 3–20.

Hepler, P.K. and Winship, L.J. (2015). The pollen tube clear zone: clues to the mechanism of polarized growth. J. Integr. Plant Biol. **57**: 79–92.

Hormaza, J.I. and Herrero, M. (1996). Dynamics of pollen tube growth under different competition regimes. Sex. Plant Reprod. **9**: 153–160.

Jiang, S.Y., Cai, M., and Ramachandran, S. (2005). The Oryza sativa no pollen (Osnop) gene plays a role in male gametophyte development and most likely encodes a C2-GRAM domain-containing protein. Plant Mol. Biol. **57**: 835–853.

Jiang, S.-Y., Ramamoorthy, R., and Ramachandran, S. (2008). Comparative transcriptional

profiling and evolutionary analysis of the GRAM domain family in eukaryotes. Dev. Biol. **314**: 418–432.

Jiao, Y. et al. (2017). Improved maize reference genome with single-molecule technologies. Nature **546**: 524–527.

Johnson, M.A., Harper, J.F., and Palanivelu, R. (2019). A Fruitful Journey: Pollen Tube Navigation from Germination to Fertilization. Annu. Rev. Plant Biol. **70**: 809–837.

Khafif, M., Balagué, C., Huard-Chauveau, C., and Roby, D. (2017). An essential role for the VASt domain of the Arabidopsis VAD1 protein in the regulation of defense and cell death in response to pathogens. PLoS One **12**: e0179782.

Khafif, M., Cottret, L., Balagué, C., and Raffaele, S. (2014). Identification and phylogenetic analyses of VASt, an uncharacterized protein domain associated with lipid-binding domains in Eukaryotes. BMC Bioinformatics **15**: 222.

Kim, C.Y. et al. (2003). Rice C2-domain proteins are induced and translocated to the plasma membrane in response to a fungal elicitor. Biochemistry **42**: 11625–11633.

Lam, S.K., Siu, C.L., Hillmer, S., Jang, S., An, G., Robinson, D.G., and Jiang, L. (2007). Rice SCAMP1 defines clathrin-coated, trans-golgi-located tubular-vesicular structures as an early endosome in tobacco BY-2 cells. Plant Cell **19**: 296–319.

Letunic, I. and Bork, P. (2018). 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. **46**: D493–D496.

Li, H., Lin, Y., Heath, R.M., Zhu, M.X., and Yang, Z. (1999). Control of pollen tube tip growth by a Rop GTPase-dependent pathway that leads to tip-localized calcium influx. Plant Cell **11**: 1731–1742.

Liu, L., Li, C., Liang, Z., and Yu, H. (2018). Characterization of Multiple C2 Domain and Transmembrane Region Proteins in Arabidopsis. Plant Physiol. **176**: 2119–2132.

Li, Y., Kabbage, M., Liu, W., and Dickman, M.B. (2016). Aspartyl Protease-Mediated Cleavage of BAG6 Is Necessary for Autophagy and Fungal Resistance in Plants. Plant Cell **28**: 233–247.

Li, Y., Segal, G., Wang, Q., and Dooner, H.K. (2013). Gene Tagging with Engineered Ds Elements in Maize. In Plant Transposable Elements: Methods and Protocols, T. Peterson, ed (Humana Press: Totowa, NJ), pp. 83–99.

Lorrain, S., Lin, B., Auriac, M.C., Kroj, T., Saindrenan, P., Nicole, M., Balagué, C., and Roby, D. (2004). Vascular associated death1, a novel GRAM domain-containing protein, is a regulator of cell death and defense responses in vascular tissues. Plant Cell **16**: 2217–2232.

Lyons, E. and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences: How to usefully compare plant genomes. Plant J. **53**: 661–673.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., and Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. **47**: W636–W641.

Mauri, N., Fernández-Marcos, M., Costas, C., Desvoyes, B., Pichel, A., Caro, E., and Gutierrez, C. (2016). GEM, a member of the GRAM domain family of proteins, is part of the ABA signaling pathway. Sci. Rep. **6**: 22660.

McCarty, D.R., Latshaw, S., Wu, S., Suzuki, M., Hunter, C.T., Avigne, W.T., and Koch, K.E. (2013). Mu-seq: sequence-based mapping and identification of transposon induced mutations. PLoS One **8**: e77172.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. U. S. A. **36**: 344–355.

Moscatelli, A. and Idilli, A.I. (2009). Pollen tube growth: a delicate equilibrium between secretory and endocytic pathways. J. Integr. Plant Biol. **51**: 727–739.

Moscatelli, A., Idilli, A.I., Rodighiero, S., and Caccianiga, M. (2012). Inhibition of actin polymerisation by low concentration Latrunculin B affects endocytosis and alters exocytosis in shank and tip of tobacco pollen tubes: Actin-dependent endo/exocytosis in pollen tubes. Plant Biol. **14**: 770–782.

Murley, A., Sarsam, R.D., Toulmay, A., Yamada, J., Prinz, W.A., and Nunnari, J. (2015). Ltc1 is an ER-localized sterol transporter and a component of ER-mitochondria and ER-vacuole contacts. J. Cell Biol. **209**: 539–548.

Muschietti, J.P. and Wengier, D.L. (2018). How many receptor-like kinases are required to operate a pollen tube. Curr. Opin. Plant Biol. **41**: 73–82.

Nalefski, E.A. and Falke, J.J. (1996). The C2 domain calcium-binding motif: structural and functional diversity. Protein Sci. **5**: 2375–2390.

Phillips, A.R. and Evans, M.M.S. (2011). Analysis of stunter1, a maize mutant with reduced gametophyte size and maternal effects on seed development. Genetics **187**: 1085–1097.

Portwood, J.L., 2nd et al. (2019). MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic Acids Res. **47**: D1146–D1154.

Rodriguez, L. et al. (2014). C2-domain abscisic acid-related proteins mediate the interaction of PYR/PYL/RCAR abscisic acid receptors with the plasma membrane and regulate abscisic acid sensitivity in Arabidopsis. Plant Cell **26**: 4802–4820.

Schindelin, J. et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods **9**: 676–682.

Schreiber, D.N. and Dresselhaus, T. (2003). In vitro pollen germination and transient transformation of Zea mays and other plant species. Plant Mol. Biol. Rep. **21**: 319–319.

Settles, A.M. et al. (2007). Sequence-indexed mutations in maize using the UniformMu

transposon-tagging population. BMC Genomics **8**: 116.

Steinhorst, L. and Kudla, J. (2013). Calcium - a central regulator of pollen germination and tube growth. Biochim. Biophys. Acta **1833**: 1573–1581.

Stone, L.M., Seaton, K.A., Kuo, J., and McComb, J.A. (2004). Fast pollen tube growth in Conospermum species. Ann. Bot. **93**: 369–378.

Tse, Y.C., Mo, B., Hillmer, S., Zhao, M., Lo, S.W., Robinson, D.G., and Jiang, L. (2004). Identification of multivesicular bodies as prevacuolar compartments in Nicotiana tabacum BY-2 cells. Plant Cell **16**: 672–693.

Tsujita, K., Itoh, T., Ijuin, T., Yamamoto, A., Shisheva, A., Laporte, J., and Takenawa, T. (2004). Myotubularin regulates the function of the late endosome through the gram domain-phosphatidylinositol 3,5-bisphosphate interaction. J. Biol. Chem. **279**: 13817–13824.

Vejlupkova, Z. and Fowler, J.E. (2003). Maize DNA preps for undergraduate students: a robust method for PCR genotyping. Maize Genetics Cooperation Newsletter **77**: 24–25.

Vermeer, J.E.M., van Leeuwen, W., Tobeña-Santamaria, R., Laxalt, A.M., Jones, D.R., Divecha, N., Gadella, T.W.J., Jr, and Munnik, T. (2006). Visualization of PtdIns3 P dynamics in living plant cells : PtdIns3 P dynamics in plant cells. Plant J. **47**: 687–700.

Vidali, L., McKenna, S.T., and Hepler, P.K. (2001). Actin polymerization is essential for pollen tube growth. Mol. Biol. Cell **12**: 2534–2545.

Vollbrecht, E. et al. (2010). Genome-wide distribution of transposed Dissociation elements in maize. Plant Cell **22**: 1667–1685.

Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., and Briggs, S.P. (2016). Integration of omic networks in a developmental atlas of maize. Science **353**: 814–818.

Wang, H. et al. (2015). Arabidopsis Synaptotagmin 2 Participates in Pollen Germination and Tube Growth and Is Delivered to Plasma Membrane via Conventional Secretion. Mol. Plant **8**: 1737–1750.

Wang, H., Zhuang, X., Cai, Y., Cheung, A.Y., and Jiang, L. (2013). Apical F-actin-regulated exocytic targeting of NtPPME1 is essential for construction and rigidity of the pollen tube cell wall. Plant J. **76**: 367–379.

Warman, C. and Fowler, J.E. (2019). Custom built scanner and simple image processing pipeline enables low-cost, high-throughput phenotyping of maize ears. bioRxiv: 780650.

Warman, C., Panda, K., Vejlupkova, Z., Hokin, S., Unger-Wallace, E., Cole, R.A., Chettoor, A.M., Jiang, D., Vollbrecht, E., Evans, M.M.S., Slotkin, R.K., and Fowler, J.E. (2020). High expression in maize pollen correlates with genetic contributions to pollen fitness as well as with coordinated transcription from neighboring transposable elements. PLoS Genet. **16**: e1008462.

Williams, J.H. and Reese, J.B. (2019). Evolution of development of pollen performance. Curr. Top. Dev. Biol. **131**: 299–336.

Yang, H., Li, Y., and Hua, J. (2006). The C2 domain protein BAP1 negatively regulates defense responses in Arabidopsis. Plant J. **48**: 238–248.

Yang, W.-Q., Lai, Y., Li, M.-N., Xu, W.-Y., and Xue, Y.-B. (2008). A novel C2-domain phospholipid-binding protein, OsPBP1, is required for pollen fertility in rice. Mol. Plant **1**: 770–785.

Yokotani, N., Ichikawa, T., Kondou, Y., Maeda, S., Iwabuchi, M., Mori, M., Hirochika, H., Matsui, M., and Oda, K. (2009). Overexpression of a rice gene encoding a small C2 domain protein OsSMCP1 increases tolerance to abiotic and biotic stresses in transgenic Arabidopsis. Plant Mol. Biol. **71**: 391–402.

Zhang, H., Qu, X., Bao, C., Khurana, P., Wang, Q., Xie, Y., Zheng, Y., Chen, N., Blanchoin, L., Staiger, C.J., and Huang, S. (2010). Arabidopsis VILLIN5, an actin filament bundling and severing protein, is necessary for normal pollen tube growth. Plant Cell **22**: 2749–2767.

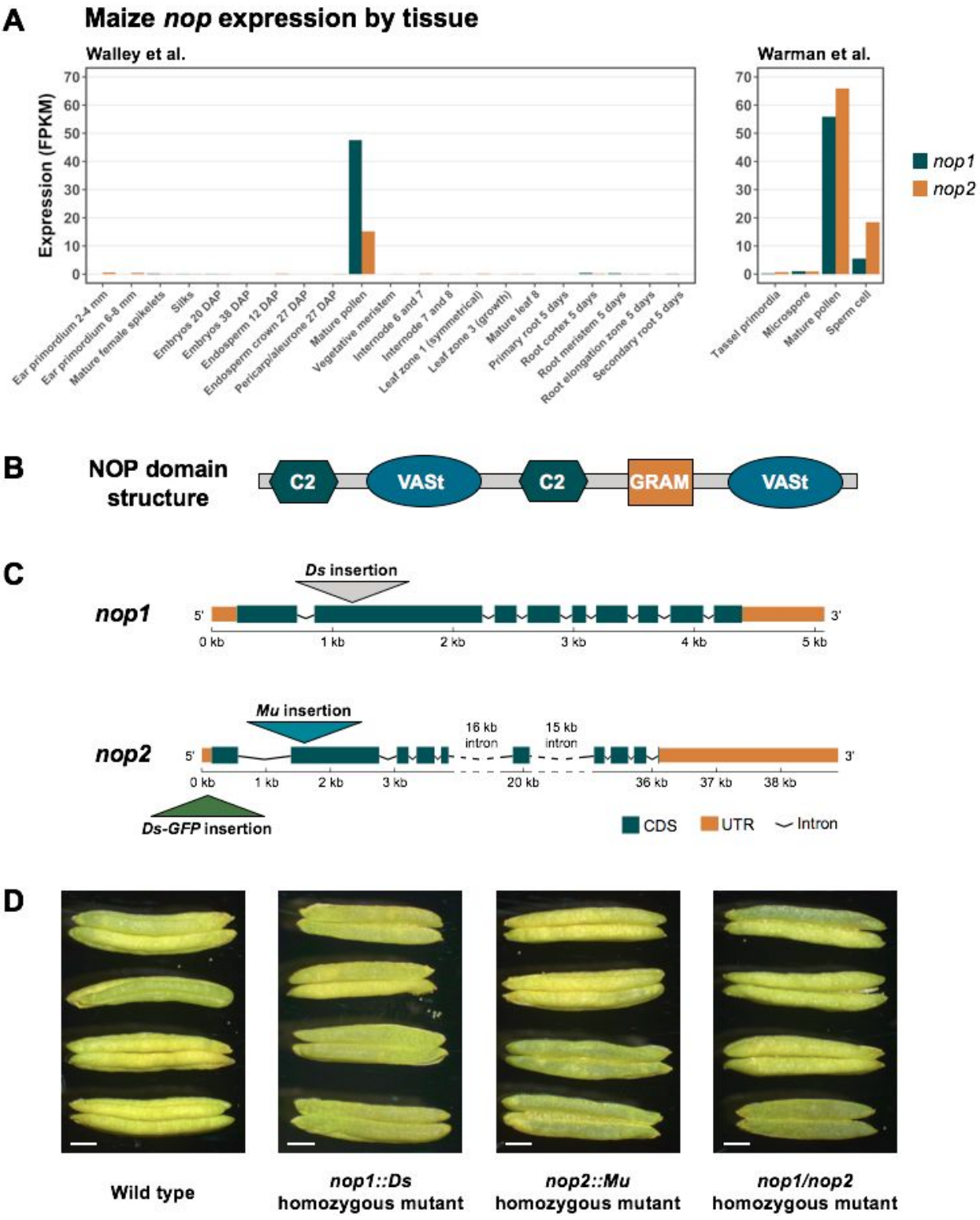

A
Maize nop expression by tissue
Walley et al.
Warman et al.
Expression (FPKM)
nop1
nop2
Ear primordium 2-4 mm
Ear primordium 6-8 mm
Mature female spikelets
Silks
Embryos 20 DAP
Embryos 38 DAP
Endosperm 12 DAP
Endosperm crown 27 DAP
Pericarp/aleurone 27 DAP
Mature pollen
Vegetative meristem
Internode 6 and 7
Internode 7 and 8
Leaf zone 1 (symmetrical)
Leaf zone 3 (growth)
Mature leaf 8
Primary root 5 days
Root cortex 5 days
Root meristem 5 days
Root elongation zone 5 days
Secondary root 5 days
Tassel primordia
Microspore
Mature pollen
Sperm cell
B
NOP domain structure
C2
VASt
C2
GRAM
VASt
C
Ds insertion
nop1
5'
3'
0 kb
1 kb
2 kb
3 kb
4 kb
5 kb
Mu insertion
nop2
16 kb intron
15 kb intron
20 kb
36 kb
37 kb
38 kb
Ds-GFP insertion
CDS
UTR
Intron
D
Wild type
nop1::Ds homozygous mutant
nop2::Mu homozygous mutant
nop1/nop2 homozygous mutant

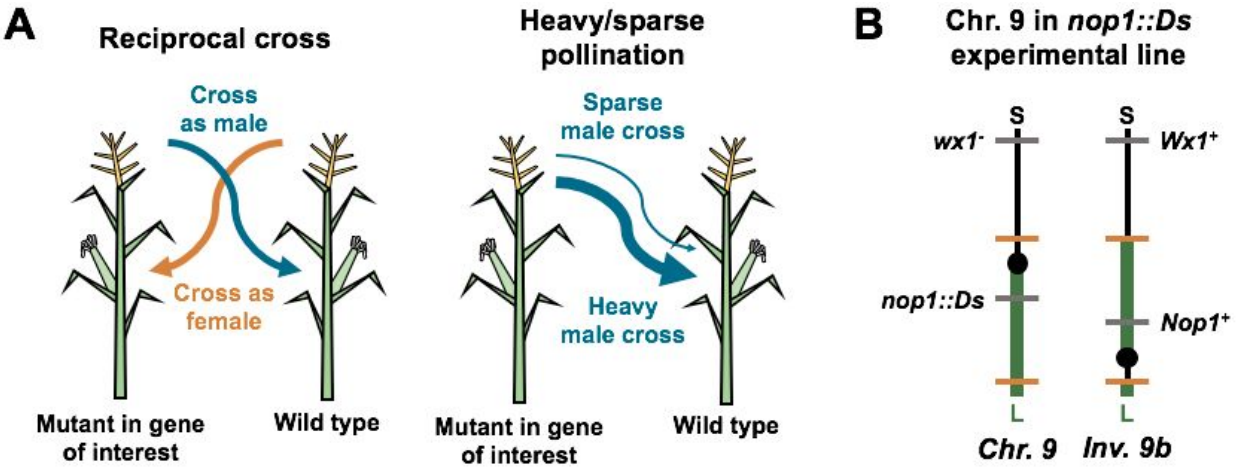
Figure 1: Maize *nop* gene expression is highly specific to the mature male gametophyte

A) Zm00001d046540 and Zm00001d015789 are homeologous genes that are highly and specifically expressed in maize pollen (Walley et al., 2016). Across maize male reproductive development, *nop1* and *nop2* are highly expressed in mature pollen, as well as detected at a lower level in sperm cells (Warman et al., 2020).

B) Both *nop* genes encode proteins with a shared predicted domain structure, with two C2 domains, two VASt domains and a GRAM domain.

C) Transposable element insertions in coding sequences of *nop1* (*Ds*) and *nop2* (*Mu*, *Ds-GFP*). Insertion locations were confirmed by Sanger sequencing. Gene models are predicted from single molecule transcript sequencing (Wang et al., 2016). Dashed lines in the *nop2* gene model indicate large predicted introns.

D) Images of representative mature anthers from homozygous *nop1* and *nop2* insertion mutants, double *nop1/nop2* homozygous mutants, and a closely related wild type control. Anthers are from different individual plants. Scale bar = 1 mm.



C

***nop1* reciprocal crosses**

	# of crosses	# of progeny	% *wx1⁻ nop1⁻*	% *Wx1⁺ Nop1⁺*	p-value (χ^2)
Mendelian	-	-	50.0%	50.0%	-
Nop1⁺/nop1⁻ as female	3	1060	48.3%	51.7%	0.269
Nop1⁺/nop1⁻ as male	3	681	16.1%	83.9%	~0

***nop2* reciprocal crosses**

	# of crosses	# of progeny	% *nop2⁻*	% *Nop2⁺*	p-value (χ^2)
Mendelian	-	-	50.0%	50.0%	-
Nop2⁺/nop2⁻ as female	10	212	49.1%	50.9%	0.784
Nop2⁺/nop2⁻ as male	10	214	2.3%	97.7%	~0

D

***nop1* heavy/sparse pollination**

	# of crosses	# of progeny	% *wx1⁻ nop1⁻*	% *Wx1⁺ Nop1⁺*
Mendelian	-	-	50.0%	50.0%
Nop1⁺/nop1⁻ as male (Heavy)	3	681	16.1%	83.9%
Nop1⁺/nop1⁻ as male (Sparse)	3	75	32.0%	68.0%

***nop2* heavy/sparse pollination**

	# of crosses	# of progeny	% *nop2⁻*	% *Nop2⁺*
Mendelian	-	-	50.0%	50.0%
Nop2⁺/nop2⁻ as male (Heavy)	3	144	4.1%	95.9%
Nop2⁺/nop2⁻ as male (Sparse)	3	144	11.8%	88.2%

Figure 2: Mutant *nop1*/*nop2* alleles show male-specific transmission defects that are partially recovered by reduced competition

A) Left: Heterozygous mutants in *nop1* and *nop2* were reciprocally crossed to wild type testers both through the male and the female to test for gametophytic transmission defects. Right: To test whether the transmission defect is competitive in nature, pollen collections were crossed with either heavy or sparse pollen loads to a series of wild-type females.

B) A line containing a *nop1::Ds* allele was crossed to a line containing inversion 9b to reduce *nop1-wx1* recombination rates. Wild-type maize chromosome 9 is shown of the left, and chromosome 9 with inversion 9b is shown on the right. Long and short arms of the chromosomes are marked "L" and "S," respectively, and inversion breakpoints are marked by orange bars. Chromosome centromeres are marked by black points.

C) *nop1::Ds* transmission rates were measured using the *wx1* kernel marker on inversion 9b lines. Heterozygous *nop1::Ds* plants were outcrossed to wild-type testers both through the male and through the female. While *wx1*$^{-}$ alone showed a slight transmission defect through the male (44.6%), *nop1::Ds* mutants showed a significantly greater male-specific transmission defect, at 16.1% (χ^2 test p-value < 0.00001). *nop2::Mu* transmission rates were tracked by PCR genotyping. *nop2::Mu* heterozygous mutants showed significant male-specific transmission defects when crossed through the male, with 2.3% transmission (χ^2 test p-value < 0.00001).

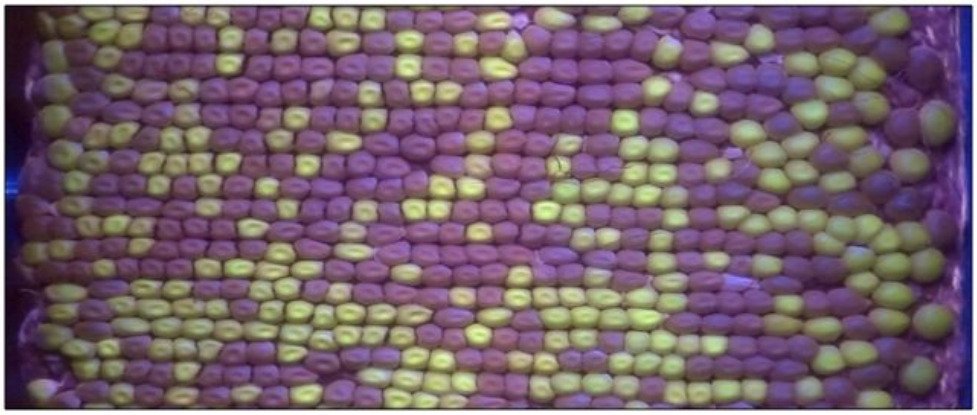
D) The effect of reduced pollen competition was measured for *nop1::Ds* and *nop2::Mu* heterozygous mutants crossed through the male with heavy and sparse pollen loads. Both *nop1::Ds* and *nop2::Mu* transmission defects were reduced by reduced pollen competition, with *nop1::Ds* transmission increasing from 16.1% to 32.0% (Cochran-Mantel-Haenszel test p-value = 0.00909) and *nop2::Mu* transmission increasing from 4.1% to 11.8% (Cochran-Mantel-Haenszel test p-value = 0.0301)

A

nop1 WT	CGTCGTCAGTGGTGGAGG---------TCCTGTCCCG	no footprint
nop1-d2	CGTCGTCAGTGGTGGAGGTCCTGTGA-TCCTGTCCCG	8 bp footprint
nop1-d4	CGTCGTCAGTGGTGGAGGTCCTGTGA-TCCTGTCCCG	8 bp footprint
nop1-d3	CGTCGTCAGTGGTGGAGGTCCTGTGACTCCTGTCCCG	9 bp footprint
nop1-d6	CGTCGTCAGTGGTGGAGGTCCTGT---TCCTGTCCCG	6 bp footprint

Cross (as male)	# of crosses	# of progeny	% *Wx1⁺* *nop1-d*	% *wx1-* *Nop1⁺*	p-value (χ^2)
Mendelian	-	-	50.0%	50.0%	-
nop1-d2 (8 bp footprint)	3	296	25.7%	74.3%	~0
nop1-d4 (8 bp footprint)	3	341	21.7%	78.3%	~0
nop1-d3 (9 bp footprint)	10	1737	48.3%	51.7%	0.157
nop1-d6 (6 bp footprint)	2	570	52.3%	47.7%	0.276

B



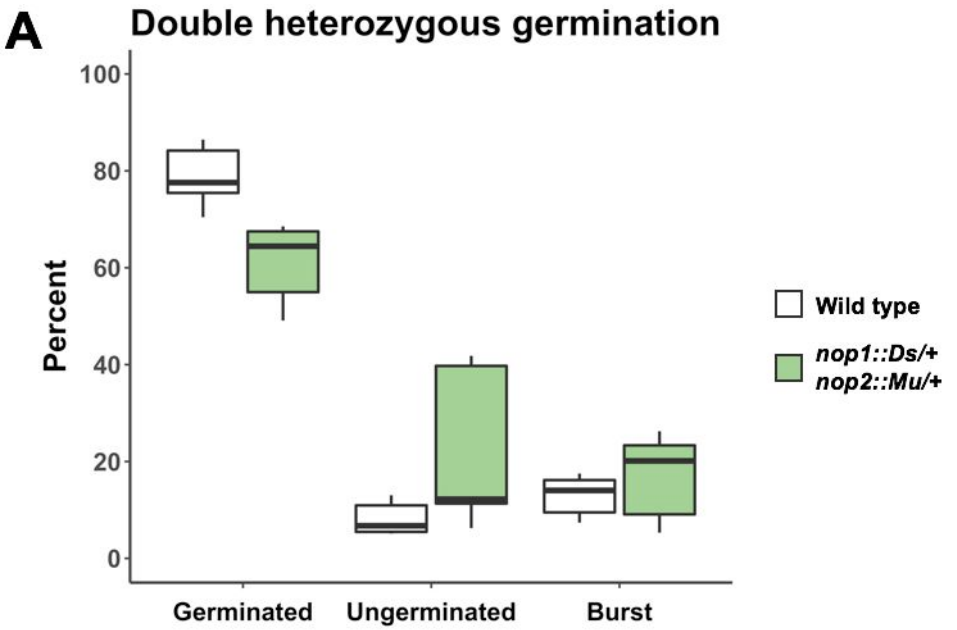
Male family	# of crosses	# of progeny screened	% fluorescent (*nop2::Ds-GFP*)	% non-fluorescent (*Nop2⁺*)	p-value (χ^2)
X3-6	5	1978	45.35%	54.65%	-
X4-1	4	1840	46.61%	53.39%	-
X4-2	4	1626	48.51%	51.49%	-
X4-4	1	376	46.54%	53.46%	-
X4-6	1	528	45.08%	54.92%	-
Summary					
-	15	6348	46.25%	53.75%	$<10^{-6}$

Figure 3: Additional *nop* alleles confirm the transmission phenotypes are due to altering *nop* function

A) Derivative *nop1* alleles with distinct footprints (linked to *Wx1⁺*) were recovered in *Ac* active lines. Both *nop1-d3*, a 9 bp footprint allele, and *nop1-d6*, a 6 bp footprint allele,

showed no significant difference from Mendelian transmission rates when crossed through the male (measured by *Wx1*$^{+}$ transmission, χ^2 test p-value = 0.157, 0.276, respectively). For comparison, two independent 8-bp derivative alleles (*-d2* and *-d4*) were associated with significantly reduced transmission when crossed through the male (χ^2 test p-values < 10^{-6}).

B) Transmission of a second allele of *nop2*, *nop2::Ds-GFP*, was tracked by fluorescent kernel markers after scanning ears and projecting the surface into a two-dimensional image, shown here. In this example image, fluorescent kernels are transmitted at 46.4%. Across all male outcrosses, *nop2::Ds-GFP* transmission was 46.25%, significantly below Mendelian inheritance (χ^2 test p-values < 10^{-8}).



B

Cross	*nop1* het (predicted)	*nop2* het (predicted)	Double het (predicted)	Wild type (predicted)	χ^2 p-value Mendelian inheritance	χ^2 p-value predicted inheritance
nop1::Ds/+ nop2::Mu/+ as female	30.65% (25%)	16.13% (25%)	30.65% (25%)	22.58% (25%)	0.298	–
nop1::Ds/+ nop2::Mu/+ as male (Heavy)	23.53% (15.73%)	2.52% (1.95%)	0.84% (0.37%)	73.11% (81.94%)	~0	0.085
nop1::Ds/+ nop2::Mu/+ as male (Sparse)	22.29%	9.55%	3.18%	64.97%	~0	–

Figure 4: *nop1::Ds/+ nop2::mu/+* plants show reduced germination, but no synergistic transmission defects

(A) A box plot showing rates of germinated, ungerminated, and burst pollen in *nop1::Ds/+ nop2::Mu/+* double heterozygous plants compared to sibling wild-type controls *in vitro*. The percentage of germinated pollen was significantly lower in double heterozygous plants compared to wild-type controls (t-test p-value = 0.00542). Double heterozygous plants showed increased, but not significantly so, rates of ungerminated and burst pollen (t-test p-values = 0.139, 0.423, respectively).

(B) Double heterozygous plants were outcrossed to wild type testers. Plants were first outcrossed as females, and showed no significant difference from Mendelian transmission ratios (χ^2 test p-value = 0.298). When crossed through the male with heavy pollination, double heterozygous plants showed no significant differences from

experimentally predicted transmission rates (χ^2 test p-value = 0.085). Transmission of single and double *nop* mutants increased when double heterozygous mutant plants were crossed as a male with sparse pollination.

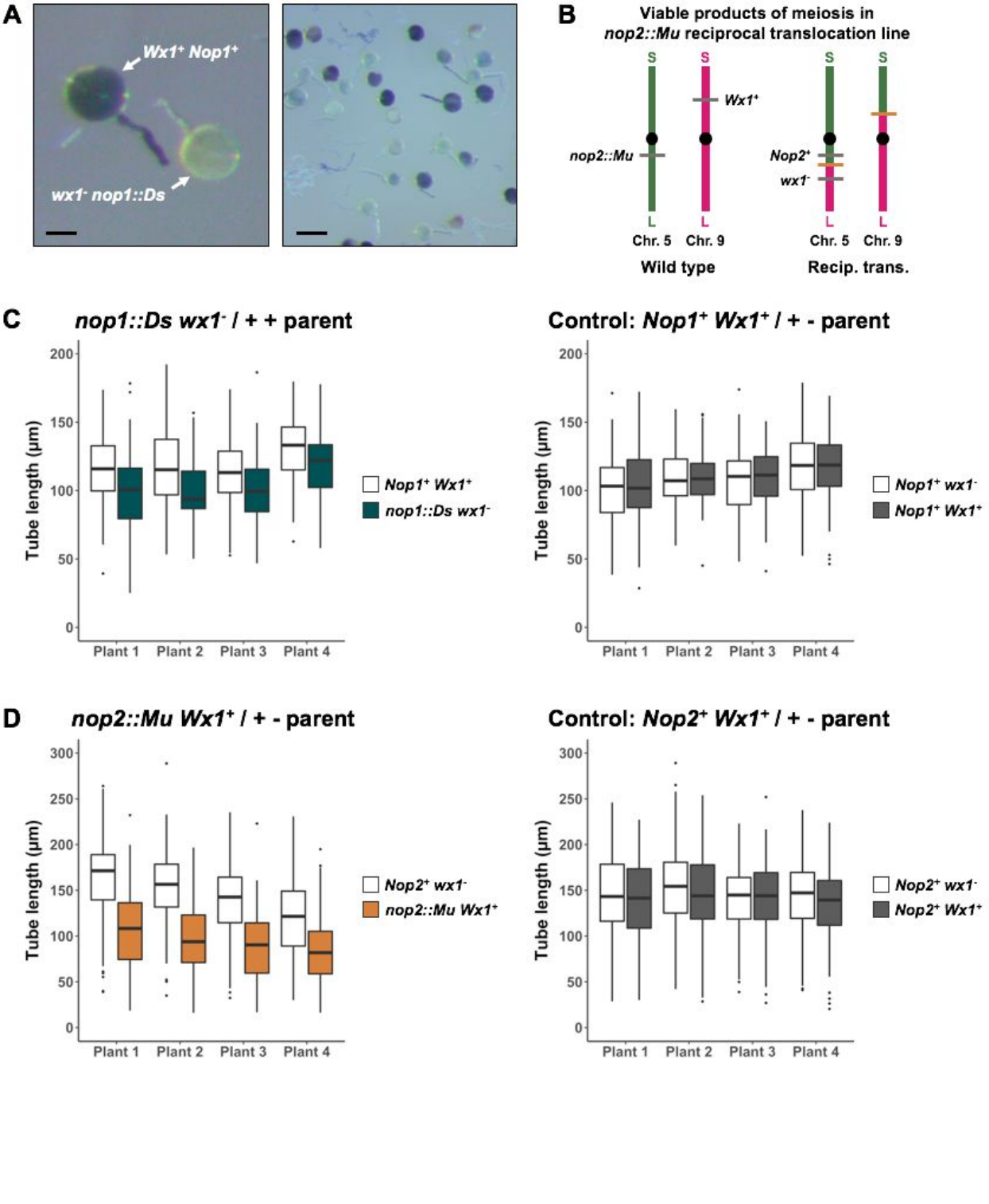

A
Wx1+ Nop1+
wx1- nop1::Ds
B
Viable products of meiosis in
nop2::Mu reciprocal translocation line
S
L
Wx1+
nop2::Mu
Nop2+
wx1-
Chr. 5
Chr. 9
Wild type
Recip. trans.
C
nop1::Ds wx1- / + + parent
Control: Nop1+ Wx1+ / + - parent
Tube length (µm)
Plant 1
Plant 2
Plant 3
Plant 4
Nop1+ Wx1+
nop1::Ds wx1-
Nop1+ wx1-
Nop1+ Wx1+
D
nop2::Mu Wx1+ / + - parent
Control: Nop2+ Wx1+ / + - parent
Nop2+ wx1-
nop2::Mu Wx1+
Nop2+ wx1-
Nop2+ Wx1+

Figure 5: *nop* mutant pollen tubes show growth defects when germinated *in vitro*

A) Representative pollen grains from *in vitro* tube length trials. Mutant *nop1::Ds* and *nop2::Mu* pollen tubes were measured *in vitro* using a linked *wx1* marker. $wx1^-$ pollen stains white in the presence of an iodine solution, whereas $Wx1^+$ pollen stains black. Closely linking *nop1* and *nop2* mutants to *wx1* markers (inversion 9b in *nop1*, reciprocal translocation in *nop2*) allows pollen tubes of mutant and wild type pollen from a single heterozygous parent to be distinguished and measured in the same experiment. Tube lengths were measured after fixing and iodine staining. Left scale bar = 50 μm, right scale bar = 200 μm.

B) Heterozygous *nop2::mu / Nop2*$^+$ plants with wild type $Wx1^+$ were crossed to reciprocal translocation line T5-9(4817) with mutant $wx1^-$. Meiosis in this cross results in 50% unviable progeny due to abnormal chromosomes. Chromosomes 5 and 9 in the remaining, viable progeny are illustrated here. Due to the reciprocal translocation, in these plants *nop2::Mu* is linked to $Wx1^+$, whereas $Nop2^+$ is linked to $wx1^-$. Recombination is minimized by the close proximity of the translocation breakpoint to *nop2* on chromosome 5 (predicted ~1 map unit).

C) After 30 minutes of growth *in vitro*, *nop1::Ds* mutant pollen grains had significantly shorter tubes than wild type controls, with an average reduction in tube length of 12% (Welch's two sample t-test p-value < 0.001).

D) After 30 minutes of growth *in vitro*, *nop2::Mu* mutant pollen grains had significantly shorter tubes than wild type controls, with an average reduction in tube length of 33.9% (Welch' two sample t-test p-values < 10^{-15}).

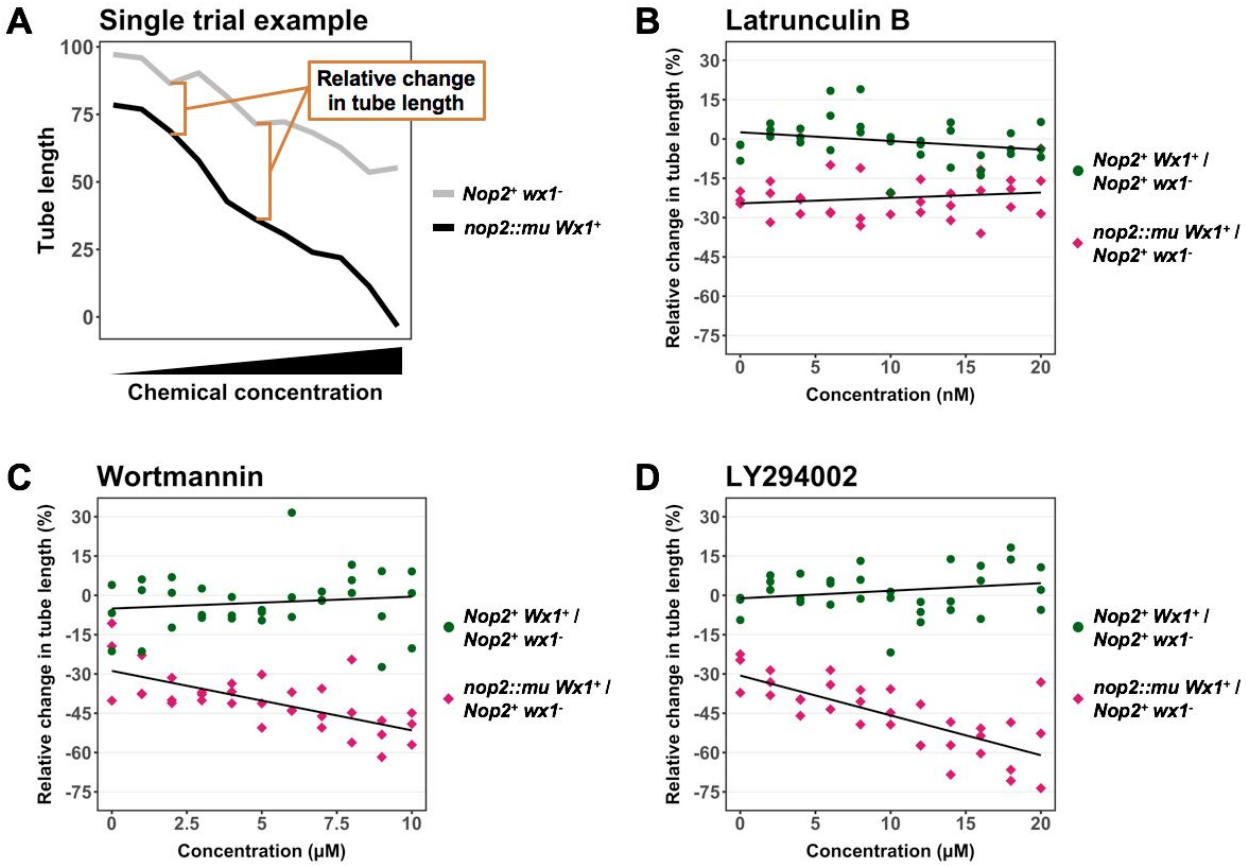
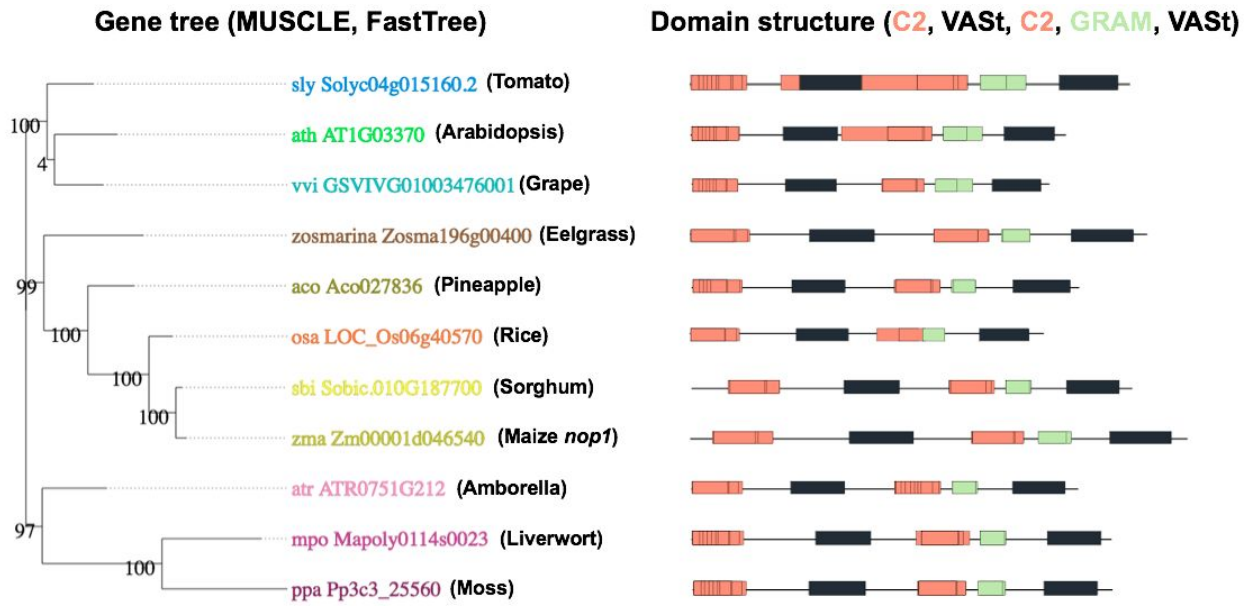


Figure 6: *nop2::Mu* pollen tubes show increased sensitivity to Wortmannin and LY294002, inhibitors of PI3K, but not to Latrunculin B, an inhibitor of actin polymerization

A) Heterozygous *nop2::Mu* pollen was subjected to various concentrations of several chemicals known to inhibit pollen tube growth. B-D show relative changes in tube lengths in *nop2::Mu* tubes at these different chemical concentrations. Here, the method for calculating percent tube length reduction in a single hypothetical experimental trial of a single chemical is illustrated. *nop2::Mu* pollen grains are linked to $Wx1^{+}$. If *nop2::Mu* pollen is equally sensitive as wild type pollen to higher chemical concentrations, the slopes of the two lines will remain equal, and the percent tube length reduction will have a slope of zero. If *nop2::Mu* is more sensitive than wild type (illustrated here using example data), the slope of *nop2::Mu* will be more negative than the wild type slope ($wx1^{-}$), and the relative change in tube length will have a negative slope in plots B-D.

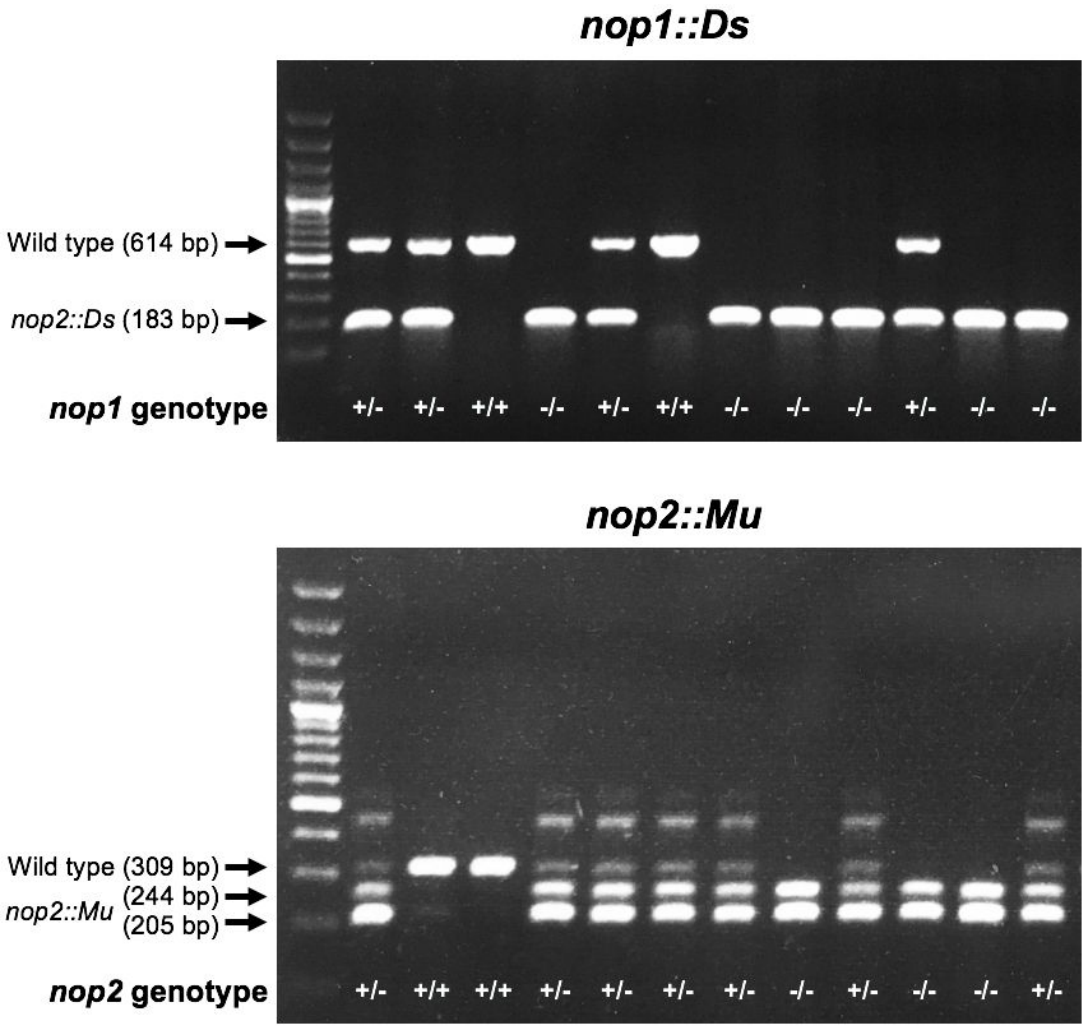
B) Latrunculin B dose-response measurements for *wx1*-marked pollen from a heterozygous *nop2::Mu* parent. Relative tube lengths do not change for *nop2::Mu* pollen or *wx1* heterozygous $Nop2^{+}$ homozygous controls at increased concentrations (linear regression p-values = 0.317, 0.142 respectively, lines represent linear regression best fit).

C) Wortmannin dose-response measurements for *wx1*-marked pollen from a heterozygous *nop2::Mu* parent. Here, *nop2::Mu* pollen tubes show increased tube length reduction at higher Wortmannin concentrations (linear regression p-value < 10^{-4}). Controls show no increased tube length reduction (linear regression p-value = 0.476).
D) As with Wortmannin, *nop2::Mu* pollen exposed to increasing LY294002 concentrations showed increased tube length reduction at higher concentrations (linear regression p-value < 10^{-6}). Controls remained unaffected (linear regression p-value = 0.229).



Supplemental Figure 1: *nop* genes are conserved throughout Embryophyta

The *nop* genes, including their predicted C2, VASt, and GRAM domains, are conserved throughout Embryophyta. Sequences of *nop* genes in maize are more closely related to sequences of *nop* genes in other monocot species such as sorghum and rice, but are also conserved in basal lineages like moss and liverwort.



Supplemental Figure 2: PCR assays were developed to detect *nop1::Ds* and *nop2::Mu* transposable element insertions

Top panel: *nop1::Ds* transposable element insertions were identified in wild-type, heterozygous, and homozygous plants by PCR. Wild-type alleles had one 614 bp fragment. Heterozygous alleles had both 614 and 183 bp fragments, while homozygous alleles had only a 183 bp fragment. Bottom panel: *nop2::Mu* transposable element insertion PCR. Wild-type plants had 309 bp amplified fragments. Heterozygous plants had 309, 244, and 205 bp fragments, with additional bands likely resulting from repetitive regions in the transposable element sequence. Homozygous mutants had bands at 244 and 205 bp.

Supplemental Table 1: nop1/wx1 recombination

Recombination frequency, no inversion

	Parental genotype		**Recombinant genotype**	
	nop1::Ds Wx+	*Nop1+ wx-*	*nop1::Ds wx-*	*Nop1+ Wx+*
Cross 1	27	21	11	6
Cross 2	4	9	0	2
Cross 3	7	2	2	3
Total	**38**	**32**	**13**	**11**

Recombination frequency = 24/94, 25.5%

Recombination frequency, inversion 9b

	Parental genotype		**Recombinant genotype**	
	nop1::Ds wx-	*Nop1+ Wx+*	*nop1::Ds Wx+*	*Nop1+ wx-*
Cross 1	22	22	2	2
Total	**22**	**22**	**2**	**2**

Recombination frequency = 4/48, 8.3%

Supplemental Table 2: nop reciprocal outcrosses

nop1::Ds transmission

Cross name	Mutant parent	Number wx1- kernels (linked to nop1::Ds)	Number Wx1+ kernels (linked to Nop1+)	% wx1-	% wx1+	p-value (chi-square)
V134E-1	female	124	131	48.63%	51.37%	-
V134E-5	female	135	137	49.63%	50.37%	-
V134E-8	female	253	280	47.47%	52.53%	-
nop1 female sum		**512**	**548**	**0.4830188679**	**0.5169811321**	**0.269**
V134E-1	male	28	112	20.00%	80.00%	-
V134E-5	male	45	286	13.60%	86.40%	-
V134E-8	male	37	173	17.62%	82.38%	-
nop1 male sum		**110**	**571**	**0.1615271659**	**0.8384728341**	**0**

Nop1+ Wx1 control transmission

Cross name	Mutant parent	Number wx1- kernels	Number Wx1+ kernels	% wx1-	% wx1+	p-value (chi-square)
V134E-3	female	144	140	50.70%	49.30%	-
V134E-4	female	198	191	50.90%	49.10%	-
V134E-7	female	157	153	50.65%	49.35%	-
Nop1+ wx1 control female sum		**499**	**484**	**0.507629705**	**0.492370295**	**0.632**
V134E-3	male	113	131	46.31%	53.69%	-
V134E-4	male	172	229	42.89%	57.11%	-
V134E-7	male	150	181	45.32%	54.68%	-
Nop1+ wx1 control male sum		**435**	**541**	**0.4456967213**	**0.5543032787**	**0.000691**

nop2::Mu transmission

Cross name	Mutant parent	Number nop2::Mu kernels	Number Nop2+ kernels	% nop2::Mu	% Nop2+	p-value (chi-square)
V198-4	female	21	24	46.67%	53.33%	-
W7	female	11	3	78.57%	21.43%	-
W13X	female	5	3	62.50%	37.50%	-
W13Y	female	3	5	37.50%	62.50%	-
W13Z	female	1	6	14.29%	85.71%	-
W14X	female	6	1	85.71%	14.29%	-
W14Y	female	2	5	28.57%	71.43%	-
W14Z	female	2	6	25.00%	75.00%	-
W7L	female	3	2	60.00%	40.00%	-
W13LX	female	4	3	57.14%	42.86%	-
W13LY	female	2	5	28.57%	71.43%	-
W13LZ	female	3	3	50.00%	50.00%	-
W14LX	female	2	6	25.00%	75.00%	-
W14LY	female	3	4	42.86%	57.14%	-
W14LZ	female	5	2	71.43%	28.57%	-
W7M	female	8	6	57.14%	42.86%	-
W13MX	female	4	4	50.00%	50.00%	-
W13MY	female	5	3	62.50%	37.50%	-
W13MZ	female	5	3	62.50%	37.50%	-
W14MX	female	1	6	14.29%	85.71%	-
W14MY	female	5	3	62.50%	37.50%	-

W14MZ	female	3	5	37.50%	62.50%	-
nop2 female sum		**104**	**108**	**0.4905660377**	**0.5094339623**	**0.784**
V198-4	male	1	41	2.38%	97.62%	-
W10X	male	0	5	0.00%	100.00%	-
W10Y	male	0	5	0.00%	100.00%	-
W10Z	male	0	8	0.00%	100.00%	-
W11X	male	0	5	0.00%	100.00%	-
W11Y	male	0	7	0.00%	100.00%	-
W11Z	male	0	8	0.00%	100.00%	-
W12X	male	0	8	0.00%	100.00%	-
W12Y	male	0	8	0.00%	100.00%	-
W12Z	male	0	7	0.00%	100.00%	-
W10LX	male	0	3	0.00%	100.00%	-
W10LY	male	1	2	33.33%	66.67%	-
W10LZ	male	1	11	8.33%	91.67%	-
W11LX	male	0	6	0.00%	100.00%	-
W11LY	male	1	5	16.67%	83.33%	-
W11LZ	male	0	5	0.00%	100.00%	-
W12LX	male	0	6	0.00%	100.00%	-
W12LY	male	0	7	0.00%	100.00%	-
W12LZ	male	0	5	0.00%	100.00%	-
W10MZ	male	0	10	0.00%	100.00%	-
W11MX	male	0	8	0.00%	100.00%	-
W11MY	male	0	8	0.00%	100.00%	-
W11MZ	male	0	8	0.00%	100.00%	-
W12MX	male	0	8	0.00%	100.00%	-
W12MY	male	1	7	12.50%	87.50%	-
W12MZ	male	0	8	0.00%	100.00%	-
nop2 male sum		**5**	**209**	**0.02336448598**	**0.976635514**	**0**

Supplemental Table 3: nop heavy and sparse outcrosses

nop1::Ds heavy/sparse

Cross name	Allele	Heavy kernel #	Sparse kernel #
V134E-1	nop1- wx1-	28	14
	Nop1+ Wx1+	112	14
	% nop1- wx1-	**20.00%**	**50.00%**
V134E-5	nop1- wx1-	45	2
	Nop1+ Wx1+	286	6
	% nop1- wx1-	**13.60%**	**25.00%**
V134E-8	nop1- wx1-	37	8
	Nop1+ Wx1+	173	31
	% nop1- wx1-	**17.62%**	**20.51%**
CMH test p-value		**0.00909**	

Nop1+ Wx1 control heavy/sparse

Cross name	Allele	Heavy kernel #	Sparse kernel #
V134E-3	wx1-	105	13
	Wx1+	111	10
	% wx1-	**48.61%**	**56.52%**
V134E-4	wx1-	172	17
	Wx1+	229	27
	% wx1-	**42.89%**	**38.64%**
V134E-7	wx1-	150	15
	Wx1+	181	23
	% wx1-	**45.32%**	**39.47%**
CMH test p-value		**0.751**	

nop2::Mu heavy/sparse

Cross name	Allele	Heavy kernel #	Sparse kernel #
W14X-2	nop2-	4	6
	nop2+	44	42
	% nop2-	**8.33%**	**12.50%**
W14X-3	nop2-	1	7
	nop2+	47	41
	% nop2-	**2.08%**	**14.58%**
W14X-7	nop2-	1	4
	nop2+	47	44
	% nop2-	**2.08%**	**8.33%**
CMH test p-value		**0.0301**	

Supplemental Table 4: nop1::Ds excision allele transmission

nop1-d2 transmission (8 bp footprint)

Cross name	Mutant parent	Number Wx1+ kernels (linked to nop1-d2)	Number wx1- kernels (linked to Nop1+)	% Wx1+	% wx1-	p-value (chi-square)
V186-4	male	21	48	30.43%	69.57%	-
V186-5	male	33	73	31.13%	68.87%	-
V186-6	male	22	99	18.18%	81.82%	-
nop1-d2 sum		**76**	**220**	**0.2567567568**	**0.7432432432**	**0**

nop1-d3 transmission (9 bp footprint)

Cross name	Mutant parent	Number Wx1+ kernels (linked to nop1-d3)	Number wx1- kernels (linked to Nop1+)	% Wx1+	% wx1-	p-value (chi-square)
V415-3a	male	82	88	48.24%	51.76%	-
V415-3b	male	114	141	44.71%	55.29%	-
V415-4a	male	64	29	68.82%	31.18%	-
V415-4b	male	60	109	35.50%	64.50%	-
V416-1a	male	123	108	53.25%	46.75%	-
V416-1b	male	100	126	44.25%	55.75%	-
V416-2a	male	102	101	50.25%	49.75%	-
V416-3b	male	24	12	66.67%	33.33%	-
V416-4a	male	106	123	46.29%	53.71%	-
V416-4b	male	64	61	51.20%	48.80%	-
nop1-d3 sum		**839**	**898**	**0.4830166955**	**0.5169833045**	**0.157**

nop1-d4 transmission (8 bp footprint)

Cross name	Mutant parent	Number Wx1+ kernels (linked to nop1-d4)	Number wx1- kernels (linked to Nop1+)	% Wx1+	% wx1-	p-value (chi-square)
V188-1	male	31	98	24.03%	75.97%	-
V188-2	male	21	78	21.21%	78.79%	-
V188-3	male	22	91	19.47%	80.53%	-
nop1-d4 sum		**74**	**267**	**0.2170087977**	**0.7829912023**	**0**

nop1-d6 transmission (6 bp footprint)

Cross name	Mutant parent	Number Wx1+ kernels (linked to nop1-d6)	Number wx1- kernels (linked to Nop1+)	% Wx1+	% wx1-	p-value (chi-square)
W28M-8	male	145	124	53.90%	46.10%	-
W28M-11	male	153	148	50.83%	49.17%	-
nop1-d6 sum		**298**	**272**	**0.5228070175**	**0.4771929825**	**0.276**

Supplemental Table 5: nop1::Ds-GFP transmission

nop2::Ds-GFP transmission

Cross name	Mutant parent	Number fluorescent kernels (linked to nop2::Ds-GFP)	Number nonfluorescent kernels (linked to Nop2+)	% fluorescent	% nonfluorescent	p-value (chi-square)
X3-6m1	male	183	231	44.20%	55.80%	-
X3-6m2	male	235	272	46.35%	53.65%	-
X3-6m3	male	91	87	51.12%	48.88%	-
X3-6m4	male	183	251	42.17%	57.83%	-
X3-6m5	male	191	254	42.92%	57.08%	-
X4-1m1	male	217	231	48.44%	51.56%	-
X4-1m2	male	201	224	47.29%	52.71%	-
X4-1m3	male	235	301	43.84%	56.16%	-
X4-1m4	male	202	229	46.87%	53.13%	-
X4-2m1	male	237	235	50.21%	49.79%	-
X4-2m2	male	159	143	52.65%	47.35%	-
X4-2m3	male	211	239	46.89%	53.11%	-
X4-2m4	male	178	224	44.28%	55.72%	-
X4-4m1	male	175	201	46.54%	53.46%	-
X4-6m1	male	238	290	45.08%	54.92%	-
nop2::Ds-GFP sum		**2936**	**3412**	**0.4625078765**	**0.5374921235**	**0.0000000**

Supplemental Table 6: nop1::Ds/nop2::Mu double heterogeneous male outcrosses

nop1::Ds / nop2::Mu double het transmission (by parent line)

Cross name	Mutant parent	Pollen amount	Total crosses	*nop1::Ds* single het	*nop2::Mu* single het	Double het	Wild type	% *nop1::Ds het*	% *nop2::Mu* het	% double het	% wild type
W602	F	HVY	8	0	5	1	2	0.00%	62.50%	12.50%	25.00%
W603	F	HVY	8	3	0	5	0	37.50%	0.00%	62.50%	0.00%
W605	F	HVY	16	7	1	5	3	43.75%	6.25%	31.25%	18.75%
W608	F	HVY	8	1	3	0	4	12.50%	37.50%	0.00%	50.00%
W609	F	HVY	8	4	0	3	1	50.00%	0.00%	37.50%	12.50%
W611	F	HVY	14	4	1	5	4	28.57%	7.14%	35.71%	28.57%
W601	M	HVY	24	14	0	0	10	58.33%	0.00%	0.00%	41.67%
W606	M	HVY	23	0	0	0	23	0.00%	0.00%	0.00%	100.00%
W607	M	HVY	24	9	1	1	13	37.50%	4.17%	4.17%	54.17%
W610	M	HVY	48	5	2	0	41	10.42%	4.17%	0.00%	85.42%
W619	M	SPS	18	3	1	0	14	16.67%	5.56%	0.00%	77.78%
W620	M	SPS	28	7	3	0	18	25.00%	10.71%	0.00%	64.29%
W621	M	SPS	8	0	3	0	5	0.00%	37.50%	0.00%	62.50%
W622	M	SPS	9	3	0	0	6	33.33%	0.00%	0.00%	66.67%
W623	M	SPS	13	0	1	0	12	0.00%	7.69%	0.00%	92.31%
W624	M	SPS	5	4	0	0	1	80.00%	0.00%	0.00%	20.00%
W625	M	SPS	2	0	0	0	2	0.00%	0.00%	0.00%	100.00%
W626	M	SPS	52	14	5	3	30	26.92%	9.62%	5.77%	57.69%
W627	M	SPS	8	0	0	1	7	0.00%	0.00%	12.50%	87.50%
W628	M	SPS	7	1	1	1	4	14.29%	14.29%	14.29%	57.14%
W643	M	SPS	7	3	1	0	3	42.86%	14.29%	0.00%	42.86%

nop1::Ds / nop2::Mu double het transmission (summary and statistical tests)

Pollen amount	Mutant parent	Total crosses	*nop1::Ds* single het	*nop2::Mu* single het	Double het	Wild type	% *nop1::Ds het*	% *nop2::Mu* het	% double het	% wild type	Chi-square p-value for Mendelian inheritance	Chi-square p-value for experimentally predicted inheritance (HVY cross)
HVY	F	62	19	10	19	14	30.65%	16.13%	30.65%	22.58%	2.98E-01	-
HVY	M	119	28	3	1	87	23.53%	2.52%	0.84%	73.11%	6.43E-35	8.50E-02
SPS	M	157	35	15	5	102	22.29%	9.55%	3.18%	64.97%	2.29E-31	2.93E-19

Chapter 3

High expression in maize pollen correlates with genetic contributions to pollen fitness as well as with coordinated transcription from neighboring transposable elements

Cedar Warman, Kaushik Panda, Zuzana Vejlupkova, Sam Hokin, Erica Unger-Wallace, Rex A. Cole, Antony M. Chettoor, Duo Jiang, Erik Vollbrecht, Matthew M. S. Evans, R. Keith Slotkin, John E. Fowler

PLOS Genetics. 16: e1008462. https://doi.org/10.1371/journal.pgen.1008462

This paper covers a transcriptome-driven investigation into the function of genes and transposable elements in the maize male gametophyte. Cedar Warman made substantive contributions to the following Results sections: "Experimental design and gene expression during maize male reproductive development," "Large-scale insertional mutagenesis supports a relationship between transcript level and fitness contribution for vegetative cell-expressed genes," and "Insertional mutants in the sperm cell-expressed Zm gex2 cause paternally triggered aberrant seed development." In addition, Cedar Warman coordinated the writing and editing of the manuscript.

Abstract

In flowering plants, gene expression in the haploid male gametophyte (pollen) is essential for sperm delivery and double fertilization. Pollen also undergoes dynamic epigenetic regulation of expression from transposable elements (TEs), but how this process interacts with gene expression is not clearly understood. To explore relationships among these processes, we quantified transcript levels in four male reproductive stages of maize (tassel primordia, microspores, mature pollen, and sperm cells) via RNA-seq. We found that, in contrast with vegetative cell-limited TE expression in Arabidopsis pollen, TE transcripts in maize accumulate as early as the microspore stage and are also present in sperm cells. Intriguingly, coordinate expression was observed between highly expressed protein-coding genes and their neighboring TEs, specifically in mature pollen and sperm cells. To investigate a potential relationship between elevated gene transcript level and pollen function, we measured the fitness cost (male-specific transmission defect) of GFP-tagged coding sequence insertion mutations in over 50 genes identified as highly expressed in the pollen vegetative cell, sperm cell, or seedling (as a sporophytic control). Insertions in seedling genes or sperm cell genes (with one exception) exhibited no difference from the expected 1:1 transmission ratio. In contrast, insertions in over 20% of vegetative cell genes were associated with significant reductions in fitness, showing a positive correlation of transcript level with non-Mendelian segregation when mutant. Insertions in maize *gamete expressed2* (*Zm gex2*), the sole sperm cell gene with measured contributions to fitness, also triggered seed defects when crossed as a male, indicating a conserved role in double fertilization, given the similar phenotype previously demonstrated for the Arabidopsis ortholog *GEX2*. Overall, our study demonstrates a developmentally programmed and coordinated transcriptional activation of TEs and genes in pollen, and further identifies maize pollen as a model in which transcriptomic data have predictive value for quantitative phenotypes.

Author Summary

In flowering plants, pollen is essential for delivering sperm cells to the egg and central cell for double fertilization, initiating the process of seed development. In plants with abundant pollen like maize, sperm cell delivery can be highly competitive. In an added layer of complexity, growing evidence indicates expression of transposable elements (TEs) is more dynamic in pollen than in other plant tissues. How these elements impact pollen function and gene

regulation is not well understood. We used transcriptional profiling to generate a framework for detailed analysis of TE expression, as well as for quantitative assessment of gene function during maize pollen development. TEs are expressed early and persist, many showing coordinated activation with highly-expressed neighboring genes in the pollen vegetative cell and sperm cells. Measuring fitness costs for a set of over 50 mutations indicates a correlation between elevated transcript level and gene function in the vegetative cell. We also establish a role in fertilization for the maize *gamete expressed2* (*Zm gex2*) gene, identified based on its specific expression in sperm cells. These results highlight maize pollen as a powerful model for investigating the developmental interplay of TEs and genes, as well as for measuring fitness contributions of specific genes.

Introduction

Sexual reproduction enables the segregation and recombination of genetic material, which increases genetic diversity in populations and contributes to the vast diversity of eukaryotes. In flowering plants, sexual reproduction requires the development of reduced, haploid gametophytes from sporophytic, diploid parents. The mature female gametophyte, the embryo sac, includes the binucleate central cell and the egg cell (reviewed in (Yang et al., 2010; Zhou et al., 2017)), each of which is fertilized by a sperm cell to generate the triploid endosperm and diploid embryo, respectively. The mature male gametophyte, pollen, consists of a vegetative cell harboring two sperm cells (reviewed in (McCormick, 1993; Hafidh et al., 2016)). In maize, male gametophytes arise from microspore mother cells in the tassel primordium. The transition from diploid sporophyte to haploid gametophyte occurs when these cells undergo meiosis, each resulting in four haploid microspores. Each microspore then undergoes two rounds of mitosis to produce the pollen grain, first generating the large vegetative cell and a smaller generative cell via asymmetric division, and then producing the two sperm from the generative cell. After the arrival of the pollen grain on the floral stigma, the vegetative cell transports the two sperm cells to the female gametophyte via pollen tube growth (reviewed in (Dresselhaus et al., 2016; Zhou and Dresselhaus, 2019)). Accurate navigation of the pollen tube as it grows down the style is dependent on the architecture of the style's transmitting tract (Lausser et al., 2010) and additional signaling and recognition mechanisms that are poorly understood (Mizukami et al., 2016). The final stages of pollen tube growth depend on a complex

interplay of signals to guide the pollen tube to the micropyle of the ovule for sperm delivery to the embryo sac (reviewed in (Higashiyama and Takeuchi, 2015)).

In maize, a pollen tube must grow up to 30 cm through the silk to reach the female gametophyte, often competing with multiple pollen tubes to eventually enter the embryo sac and release its sperm cells for fertilization (reviewed in (Zhou et al., 2017; Dresselhaus et al., 2016)). Across the angiosperms, this competitive context for pollen tube development differs, depending on the pollen population as well as sporophytic characters (reviewed in (Williams and Reese, 2019)). In a highly competitive environment, successful fertilization is likely enhanced by pollen tubes functioning at full capacity (Arthur et al., 2003; Cole et al., 2005; Huang et al., 2017), as generally only the first tube to reach the micropyle is permitted to enter the female gametophyte. The mechanisms preventing entry of multiple pollen tubes, known as the polytubey block, are not well-understood, but presumably act to reduce polyspermy, which typically leads to sterile offspring (Zhou and Dresselhaus, 2019). In maize, mutations in the genes *MATRILINEAL*/*NLD*/*ZmPLA1* and *ZmDMP* have been linked to pollen-induced production of haploid embryos and other seed defects, which are likely associated with aberrant events at fertilization (Kelliher et al., 2017; Gilles et al., 2017; Liu et al., 2017; Zhong et al., 2019) or soon after (Kelliher et al., 2019). Thus, many mechanisms associated with both pollen tube growth and fertilization remain enigmatic.

Given their specialized biological functions and well-defined developmental stages, gametophytes are prime targets for transcriptome analysis. Initial studies of plant gametophytic transcriptomes in Arabidopsis pollen (Honys and Twell, 2003; Becker et al., 2003) and embryo sacs (Steffen et al., 2007; Jones-Rhoades et al., 2007) described a limited and specialized set of transcripts and identified numerous candidate genes for gametophytic function. In maize, the first RNA-seq study of male and female gametophyte transcriptomes (mature pollen and embryo sacs) similarly identified subsets of developmentally specific genes, with pollen showing the most specialized transcriptome relative to other tissues assessed (Chettoor et al., 2014). More recently, RNA-seq has been carried out on additional stages of maize reproductive development, including pre-meiotic and meiotic anther cells (Zhai et al., 2015; Nelms and Walbot, 2019; Begcy et al., 2019), as well as sperm cells, egg cells, and early stages of zygotic development (Chen et al., 2017).

Gametophytic tissues are known to show dynamic expression of transposable elements (TEs). In Arabidopsis, global TE expression is derepressed at the late stages of pollen development, occurring in the pollen vegetative nucleus only after pollen mitosis II (Slotkin et al.,

2009). The pollen vegetative nucleus undergoes a programmed loss of heterochromatin, resulting in TE activation, TE transposition and subsequent increased RNA-directed DNA methylation (Schoft et al., 2009; Slotkin et al., 2009; Calarco et al., 2012; Dooner et al., 2019; He et al., 2019). A variety of functions have been ascribed to this male gametophytic "developmental relaxation of TE silencing" (DRTS) event (Martínez and Slotkin, 2012), including the generation of TE small interfering RNAs that are mobilized to the sperm cells (Martínez et al., 2016), and control of imprinted gene expression after fertilization (Martinez et al., 2018). However, the dynamics of TE expression during gametophytic development in a transposable element-rich species such as maize have not been investigated.

To provide a more full description of transcriptome dynamics across maize male reproductive development, including TE transcriptional activity, we generated RNA-seq datasets from tassel primordia, microspores, mature pollen, and isolated sperm cells. Using these data, we describe differential expression patterns of genes and TEs across these stages, uncovering a coordinated regulation of TEs and their neighboring genes in pollen grains. Then, within a framework provided by the transcriptome data, we conducted a functional validation of highly expressed genes by testing over fifty insertional mutations for male-specific fitness effects. Finally, the same transcriptome data guided the discovery of mutant alleles in the sperm cell-enriched *gex2*, which induces seed development defects when present in the pollen parent, implying a role in fertilization.

Results

Experimental design and gene expression during maize male reproductive development

RNA-seq was performed on four tissues representing integral stages in maize male gametophyte development: immature tassel primordia (TP), isolated unicellular microspores (MS), mature pollen (MP), and isolated sperm cells (SC) (Fig 1A). Techniques were developed to efficiently isolate RNA from TP, MS, and SC (see Methods). RNA was extracted from the inbred maize line B73, with four biological replicates for each tissue. In addition, a single RNA replicate was isolated for the bicellular stage of pollen development (MS-B). Libraries were sequenced using Illumina sequencing (100 bp paired-end reads) and mapped to the B73 AGPv4 reference genome (Jiao et al., 2017). Principal Component Analysis (PCA) showed samples from each tissue clustering together along PC1 and PC2, which together explained

49.8% of the variance between samples (Fig 1B). One sample, SC1, had significant levels of ribosomal RNA (rRNA) contamination, as well as the fewest number of mapped reads (approximately 1 million). However, to maintain a balanced experimental design with a consistent false discovery rate (FDR), we chose to include SC1 in our analysis of gene expression patterns.

Differential gene expression was defined in two ways: in the first, gene expression in later developmental stages was compared to the premeiotic, diploid tassel primordia (TP vs MS, TP vs MP, and TP vs SC); in the second, gene expression was compared between all adjacent developmental stages (TP vs MS, MS vs MP, MS vs SC, MP vs SC) (S2 and S11 Tables). Enriched GO terms highlighted the differences in gene expression among developmental stages and suggested consistency with the established functions of each tissue (Honys and Twell, 2003; Chettoor et al., 2014; Chen et al., 2017). GO terms in MS were consistent with a post-meiotic tissue still at an early stage of development, with terms related to protein synthesis and transport, morphogenesis, and reproduction showing enrichment. MP showed more specific enriched GO terms, including those related to pollen tube growth, signaling, and actin filament-based movement. SC shared many GO terms with MP when compared to MS, but was uniquely enriched for GO terms related to epigenetic regulation of gene expression, such as histone H3K9 demethylation and gene silencing by RNA, potential mechanisms involved in differential regulation of TEs.

Comparison of the most highly expressed genes from all four sample types showed that, generally, such transcripts were highly enriched at a single developmental stage (Fig 1C). Notably, no single gene was highly expressed in all four tissues, and fewer than twenty were highly expressed in three out of the four stages assessed. These data suggest a simple hypothesis in which the expression level for protein-coding genes reflects, at some level, functional importance, i.e., a high expression level at a specific developmental stage implies an increased contribution by the gene's encoded function at that particular stage. Alternatively, or in addition, high expression could be reflective of regulatory mechanisms specific to each stage, each primarily influencing specific subsets of genes. We were interested to explore the possibility of some regulatory linkage between those genes and TEs highly expressed in the male gametophyte.

A subset of transposable elements in the maize genome show developmentally dynamic expression

To obtain a baseline view of TE expression throughout maize development, our RNA-seq data for maize male reproductive development (samples with asterisks, S1 Fig) was combined with publicly available datasets from nine-day old above-ground seedlings, juvenile leaves, ovules, another set of independently isolated sperm cells, and three independent studies of pollen RNA-seq (Chettoor et al., 2014; Chen et al., 2017; Lunardon et al., 2016; Walley et al., 2016) & SRP067853). The complete list of samples, their sequencing statistics, references and data availability can be found in S3 Table. All of the raw data were remapped using the same parameters (see Methods). Principal component analysis demonstrates that replicates of the same tissue and growth state typically group together (S1 Fig).

We aimed to identify the set of dynamically expressed TEs within the tissues sampled, and thus calculated expression levels for each individual TE in the genome located more than 2 kb away from annotated non-TE genes. Our rationale was to avoid false positive signals of TE expression due to a TE residing within a gene, and to minimize the influence of read-through transcription from a nearby gene, which could not be distinguished from TE-initiated transcription. Because we concentrated on individual elements, and not TE families, the majority of annotated TEs were not assessed in this analysis (55%; Fig 2A 'not covered'), either because no expression was detected in any dataset, or because their sequence lacks the polymorphisms necessary for mapping reads to a specific TE. To relate TE expression comparatively across development, we used seedling tissue as a baseline against which other tissues were measured. Seedling was chosen for several reasons: it is not a reproductive tissue, it has low to average levels of TE expression, and a large number of TEs show no evidence of expression in this tissue (S2 Fig).

Apart from the 18.3% of annotated TEs that are near genes and analyzed separately (see below), we calculated the number of TEs with statistically significant expression differences in each tissue compared to the seedling reference. This identified the subset of TEs that are developmentally dynamic, meaning that they show differential expression in at least one tissue in our dataset compared to the seedling reference. Only 4.4% of all maize annotated TEs are developmentally dynamic, whereas 22.2% of TEs have detectable expression but do not change in our dataset and therefore are developmentally 'static' (Fig 2A). Each TE category was interrogated for feature overrepresentation. Both dynamic and static TEs are longer than the genome average, and longer than the sets of TEs 'not covered' or 'near genes' (Fig 2B). To

determine if one long family of TEs was contributing this difference, we performed this analysis for each TE superfamily and found that for dynamic TEs, this observation is not specific to one TE type (S3 Fig). The finding that expressed TEs as a group are longer correlates with Arabidopsis data where longer TE transcripts are overrepresented and differentially regulated when epigenetic repression is lost (Panda et al., 2016).

Expressed TEs show an under-representation for DNA transposon and SINE families, which are mainly within the 'near genes' set (Fig 2C). In contrast, the 'LTR unknown' TE annotation is over-represented in the dynamic TE set (Fig 2C). Since some LTR retrotransposons are enriched in the pericentromere (Wolfgruber et al., 2009), we tested if the dynamic TE set is enriched in the pericentromere compared to the genome average, but did not detect any correlation (Fig 2D). Therefore, we conclude that expressed TEs are generally longer elements, and the subset of developmentally dynamic TEs are enriched for uncharacterized LTR retrotransposons located throughout the genome.

Transposable element transcript levels are up-regulated in the post-meiotic male reproductive lineage

From the developmentally dynamic TE set, we calculated the number of differentially expressed TEs in each tissue/stage compared to the seedling reference. In some tissues, such as tassel primordia and ovules, we observed a similar number of TEs up-regulated and down-regulated (Fig 3A), demonstrating that while there are shifts in which TEs are expressed, a genome-scale change in TE expression does not occur. In other tissues, such as juvenile leaves, there is a skew towards increased TE expression. The largest TE up-regulation occurs in the tissues of the male reproductive lineage, including unicellular and bicellular microspores, mature pollen and isolated sperm cells (Fig 3A). Our data confirm the recent finding that the tissue with the largest number of TE families activated is mature pollen (Anderson et al., 2019). The number of up-regulated TEs compared to down-regulated TEs in these tissues suggest that there is a genome-wide activation of TE expression, similar to the DRTS event that occurs in Arabidopsis pollen (Slotkin et al., 2009; Martínez and Slotkin, 2012). One important distinction is that TE expression is present in maize sperm cells (Fig 3A), whereas it is not detected in Arabidopsis sperm cells (Slotkin et al., 2009). To verify this finding, we compared our sperm cell RNA-seq data to an independent maize sperm cell dataset (Chen et al., 2017). We found that TEs are also significantly expressed in this independent dataset, and 70% of those expressed TEs are also detected in our dataset ($p<0.001$) (Fig 3B). This shared set of 810 sperm

cell-expressed TEs (38% of those detected in our dataset), supports the conclusion that significant expression of TEs occurs in maize sperm cells. Of the sperm cell-expressed TEs, 36% were not observable in total pollen, but rather required the isolation and enrichment of sperm cells for detection (Fig 3B). Overall, we detect 157 TEs expressed in both sperm cell datasets that are not expressed throughout development, but specifically in the sperm cells (sperm-cell exclusive).

A second notable difference between maize and Arabidopsis is the activation of TE expression early in the male gametophytic phase of maize. A genome-wide increase in TE transcript levels is detected at the earliest post-meiotic stage tested, the microspore, in contrast to low TE expression in the sporophytic tassel primordia (Fig 3A). Arabidopsis TE expression occurs only late in pollen development, after pollen mitosis I when the vegetative cell is generated (Slotkin et al., 2009). To determine if TEs were indeed activated early in maize male reproductive development, we asked if the same TEs that we identified as expressed in the unicellular microspore remain active throughout the male reproductive lineage. We used the set of differentially expressed up-regulated TEs in unicellular microspores (3,335) and found that 62% are still expressed in bicellular microspores and 54% in mature pollen (Fig 3C), demonstrating that once TEs are activated early in development, expression and/or steady-state mRNA frequently remains through pollen maturation. Only some of these male-lineage expressed TEs continue to be expressed in sperm cells (32%), raising the possibility that many TEs with active expression in the early gametophytic stages are under negative/repressive regulation in the gametes. This large-scale developmental activation is potentially limited to the male lineage, as ovules express relatively few TEs (Fig 3A) and only 14% of the male lineage-expressed elements (Fig 3C). Together, our data demonstrate conserved activation of TE expression in the male gametophytes of maize and Arabidopsis, with key differences such as the developmental timing and localization of TE expression in the gamete cells.

We determined what types of TEs activate in the male reproductive lineage and sperm cells and compared these to the whole-genome distribution of TEs analyzed. Overall, both male lineage-expressed TEs and sperm cell-expressed TEs reflect the genome-wide TE distribution (Fig 3D). This suggests that TE family type does not have a determining role in the developmental regulation of TE expression. One notable exception is the enrichment of *Mutator* family TE expression in sperm cells (Fig 3D). When normalized for genome-wide TE distribution, *Mutator* element expression is highly enriched across the male lineage, including in sperm cells (Fig 3E). The expression of some *Mutator* TEs in sperm cells is both high confidence (present in

both sperm cell datasets) and specific to only that tissue (high confidence sperm cell specific, Fig 3E). LINE L1 elements are also expressed throughout the male lineage and sperm cells, but their expression is general and not specific to these cell types (Fig 3E). Our data demonstrate that there is a general (TE family-independent) activation of TE expression in the male reproductive lineage, with one observable bias towards *Mutator* family expression in both the male lineage and sperm cells.

Mature pollen and sperm cells display coexpression of highly expressed genes and their neighboring TEs

To determine if TEs have an effect on neighboring gene expression, or vice versa, we next analyzed the set of 36,945 assayable TEs within 2kb of genes (Fig 2A). We calculated the absolute expression level of each genic isoform and categorized them into 100 bins of expression levels for each developmental stage (Fig 4). We found no significant relationship between highly expressed genes and the number of up- or down-regulated TEs in tassel primordia or microspores (grey bars, top row, Fig 4). In contrast, in both mature pollen and isolated sperm cells there is a statistically significant ($p < 1E-6$) positive association between highly expressed genes and the number of up-regulated TEs within 2kb of those genes (grey bars, bottom row, Fig 4). Similarly, there is a negative correlation ($p < 1E-6$) between high gene expression and the number of down-regulated TEs in the same samples (grey bars, bottom row, Fig 4). This relationship is not due to the fact that pollen or sperm cell-expressed genes are more likely to be located nearby a TE (S4A Fig), nor due to sample contamination between these two datasets (S4B Fig). To determine if our analysis is biased by the presence of multiple TEs close to just a few highly expressed genes, we also counted the number of genes with at least one differentially expressed TE within 2kb (black dots, Fig 4 and S4 Fig). We found that a similar number of genes were adjacent to differentially expressed TEs (compared grey bars to black dots, Fig 4 and S4 Fig), demonstrating that a small number of genes was not biasing our dataset. We conclude that specifically in the mature male gametophyte the most highly expressed genes tend to be near actively expressing TEs. However, it remains unclear whether gene expression is influencing TE expression, or TE expression is affecting gene regulation, or alternatively, some global regulatory mechanism is influencing both.

Large-scale insertional mutagenesis supports a relationship between transcript level and fitness contribution for vegetative cell-expressed genes

Using the quantitative framework provided by our transcriptome dataset, we next tested the simple hypothesis that highly expressed genes contribute to male gametophytic function – i.e, to reproductive success (pollen fitness). The functional validation approach we used relied on a large, sequence-indexed collection of green fluorescent protein (GFP)-marked transposable element (*Ds-GFP*) insertion mutants (Li et al., 2013), enabling assessment of the effects of mutations in select genes (Fig 5). We focused on expression data from the MP and SC stages, as these have distinctive cell fates and roles in reproduction: the vegetative cell generates the pollen tube for competitive delivery of gametes, and the sperm cells accomplish double fertilization. Expression data from seedlings (Chettoor et al., 2014) was used to design a sporophytic control. Highly expressed genes, operationally defined as in the top 20% for a tissue by FPKM, were grouped into three mutually exclusive classes: Seedling, Sperm Cell, and Vegetative Cell. The seedling group also excluded any gene highly expressed in either MP or SC. Due to the significant overlap among genes highly expressed in both MP and SC, we compared expression values to assign each of these genes to a single class. Vegetative Cell genes were not only highly expressed in MP, but were also associated with an FPKM greater in MP than in SC, and vice versa for Sperm Cell genes (S5 Table). All genes in these classes were then cross-referenced with *Ds-GFP* insertion locations to identify potential mutant alleles for study, restricting the search to insertions in coding sequence (CDS), as these were rationalized as most likely to generate loss-of-function effects. Finally, to insure our results were as generalizable as possible, each class list was randomized to identify the specific subset of *Ds-GFP* lines for study. Insertion locations were verified by PCR for 64 of 83 alleles obtained (S6 Table) (see Methods), of which 56, representing mutations in 52 genes, generated sufficient transmission data to include in our final analysis.

Mendelian inheritance predicts 50% transmission of mutant and wild-type alleles when a heterozygous mutant is outcrossed to a wild-type plant. However, a mutation that alters the function of a gene expressed during the haploid gametophytic phase can result in a reduced transmission rate if that gene contributes to the fitness of the male gametophyte – i.e., to its ability to succeed in the highly competitive process of pollen tube growth, given that 50% of the pollen population will be wild-type for the same gene. Thus, reduced transmission of a mutant through the male (a male transmission defect) provides not only evidence for gene function in the gametophyte, but also a measure of the mutated gene's contribution to fitness.

Transmission rates through the female serve as a control, as 50% transmission through the female would confirm both a single *Ds-GFP* insertion in the genome and male-specificity for any defect identified. To measure the fitness cost of each *Ds-GFP* insertion, heterozygous mutant plants were reciprocally outcrossed with a heavy pollen load to a wild-type plant, maximizing pollen competition within each silk. Transmission rates were then quantified by assessing the ratio of the non-mutant to mutant progeny using a novel scanning system and image analysis pipeline (Fig 5) (see Methods) (Warman and Fowler, 2019). Mutant alleles were tracked using linked endosperm markers: either the GFP encoded by the inserted transposable element (Fig 5A-B), or, in ~10% of the lines, a tightly linked *C1*$^{+}$ anthocyanin transgene (present due to the initial *Ds-GFP* generation protocol) (Fig 5C, S7 Table).

Transmission rates for all groups were tested through quasi-likelihood tests on generalized linear models with a logit link function for binomial counts (see Methods, S8 Table). When crossed through the female, no genes showed significant differences from Mendelian inheritance (Fig 6A). When crossed through the male, no genes with insertion alleles in the Seedling category (n=10) showed evidence of abnormal transmission rates (Fig 6B). Most Sperm Cell genes (n=10, 90%) showed no statistically significant transmission defects, with one notable exception (two independent alleles of the *gex2* gene, described in detail below) (Fig 6C). However, among Vegetative Cell genes tested (n=32), a larger proportion of insertion alleles (7 out of 32 or 21.9%) showed significant male transmission defects (quasi-likelihood test, adjusted p-value threshold < 0.05) (Fig 6D). The proportions of genes with transmission defects in the three classes were not significantly different by Fisher's exact test (Seedling vs Sperm Cell p-value = 0.500, Seedling vs Vegetative Cell p-value = 0.125, Vegetative Cell vs Sperm Cell p-value = 0.374), likely due to the small number of mutations assessed in the Seedling and Sperm Cell classes. For the insertion alleles tested, a summary description of genes showing non-Mendelian inheritance can be found in Table 1, whereas a description of those showing Mendelian inheritance can be found in Table 2.

The majority of transmission defects in the Vegetative Cell class genes (six of the seven with significant effects) were mild, at approximately 45% transmission, with only one reducing transmission by a moderate amount, to ~30%. Notably, six of the genes associated with significant defects were measured at a $\log_2$(FPKM) > 8 (i.e., in the top 5% of Vegetative Cell genes by FPKM). Given that twelve genes above this threshold were tested, these most highly expressed Vegetative Cell genes were significantly more likely to be associated with non-Mendelian transmission (6 out of 12) than the group of Vegetative Cell genes below this

expression threshold (1 out of 20) (Fisher's exact test, p-value = 0.00572). Consistent with this observation, an increase in $\log_2$(FPKM) was associated with both reduced transmission rate and an increase in $-\log_{10}$(p-value) (linear regression, p-value = 0.0120, 0.0255, respectively). Thus, our data indicate that transcript level in the Vegetative Cell does provide some limited predictive power for identifying gene-specific contributions to male gametophytic fitness (adjusted R^2 = 0.151, 0.116, respectively). Vegetative Cell genes associated with non-Mendelian inheritance had a range of predicted cellular functions, including cell wall modification, cell signaling, protein folding, vesicle trafficking, and actin binding (Table 1).

To ensure the experimental design was robust, we examined two potential confounding variables: the presence of the *wx1-m7::Ac* allele in a subset of lines tested and the potential for epigenetic silencing of GFP transgenes (see S1 Methods). We found no evidence that the presence of *wx1-m7::Ac* significantly impacted the overall conclusions drawn from the dataset, nor evidence of epigenetic silencing of GFP transgenes.

Insertional mutants in the sperm cell-expressed *Zm gex2* cause paternally triggered aberrant seed development

The male-specific transmission defect for the sole affected gene in the Sperm Cell class, Zm00001d005781 (GRMZM2G036832), was notably more severe than the average defect across all *Ds-GFP* mutants identified with decreased transmission through the male (Table 1). This gene is hereafter referred to as *Zm gex2* or *gex2*, for reasons detailed below. The two independent alleles assessed, *gex2-tdsgR82A03* and *gex2-tdsgR84A12*, were associated with transmission rates of 33.4% and 23.1%, respectively. Sequencing confirmed that these *Ds-GFP* elements were inserted into their predicted CDS locations (Fig 7A). In addition to the transmission defect, both alleles, when crossed through the male, conditioned unusual phenotypes: underdeveloped or aborted seeds, as well as ovules with no apparent seed development despite heavy pollination (Fig 7B). These features motivated further investigation of this gene.

Across maize tissues, *gex2* is highly and specifically expressed in sperm cells (Walley et al., 2016) (Fig 7C). Like many highly expressed genes in mature pollen, it is within 2kb of a transcriptionally active TE, a downstream RLG retrotransposon that displays sperm cell-specific activation (Fig 7C). The *Zm gex2* gene was first identified via EST sequencing of maize sperm cells (Engel et al., 2003), and subsequently used to isolate the Arabidopsis ortholog, named *GAMETE EXPRESSED2* (*GEX2*) (Engel et al., 2005). In Arabidopsis, *GEX2* is necessary for

effective double fertilization, causing seed abortion and empty spaces in the silique when a mutant allele is inherited through the male (Mori et al., 2014). *Zm gex2* encodes a protein with similar structure and amino acid sequence to its Arabidopsis ortholog (Fig 7D).

Small and aborted seeds were quantified for both *gex2* insertion alleles when outcrossed to wild-type plants (S10 Table, S5A Fig). Two different *Ds-GFP* insertion lines that were not associated with transmission defects (*tdsgR12H07*, *tdsgR46C04*), as well as the *Ds-GFP* associated with the strongest male transmission defect in the Vegetative Cell class (*tdsgR96C12*), were used as controls. Pollination with both *gex2::Ds-GFP* insertion alleles was associated with increased percentages of small or aborted seeds, significantly so in *gex2-tdsgR84A12* (pairwise t-test against *Ds-GFP* controls, all p-values < 0.05), and pollination from *gex2-tdsgR84A12* homozygotes approximately doubled the percentage of aberrant seeds from heterozygotes (pairwise t-test against *Ds-GFP* controls, all p-values < 0.01) (Fig 7E). From the heterozygous *gex2::Ds-GFP* crosses, small seeds with endosperm large enough for DNA preparation were genotyped, and 79.2% were found to harbor the *gex2* mutation, whereas the *tdsgR46C04* control showed Mendelian segregation in small seeds (S5B Fig). These data support the hypothesis that aberrant seed development is induced by fertilization by *gex2::Ds-GFP* sperm.

If the Zm GEX2 protein acts to promote double fertilization similarly to its Arabidopsis ortholog, the arrival of a *gex2::Ds-GFP* pollen tube at the embryo sac could lead to failure of one or both fertilization events. Given an active polytubey block, this could produce the observed gaps between seeds on the ear, resulting from ovules associated with completely failed fertilization, or with very early seed abortion due to single fertilization. Consistent with this possibility, pollination with both heterozygous and homozygous *gex2-tdsgR82A03* alleles resulted in increased seedless area relative to controls (S6 Fig). To test for aberrant fertilization more directly, seed development was assessed at 4 days post-pollination with either wild-type or *gex2-tdsgR84A12* homozygous pollen (Fig 8, Table 3). Typical embryo and endosperm development, as well as indication of the polytubey block (i.e., arrival of only single pollen tubes at the embryo sac), was observed in all ovules assessed from wild-type pollination. In contrast, half of the ovules assessed following pollination with *gex2::Ds-GFP* showed significant evidence of abnormal double fertilization, demonstrating single fertilization of either embryo or endosperm or indication of arrival of more than one pollen tube at the embryo sac (Fisher's exact test, p-value = 0.000241). We conclude that in maize, similarly to Arabidopsis, Zm GEX2 is part of the sperm cell machinery that helps ensure proper double fertilization.

Discussion

The *Zm gex2* gene has a conserved role in promoting double fertilization

The generation of a well-replicated developmental time course of transcriptomic data enabled the targeting of highly expressed genes in vegetative and sperm cells for mutational screening. Two independent insertions in the highly and specifically expressed maize sperm cell gene *gex2* led not only to severely reduced transmission through the male, but also, in contrast to other mutations analyzed in this study, to paternally triggered post-fertilization defects. *Zm gex2* was first identified in maize by sperm cell EST sequencing (Engel et al., 2003), which led to identification of the orthologous gene in Arabidopsis, *GEX2*, and its sperm cell-specific promoter (Engel et al., 2005). In Arabidopsis, single fertilization events were observed at increased frequency in *GEX2* mutant-pollinated plants, both for the egg cell and the central cell, leading to an observed increase in aberrant seed development and abortion (Mori et al., 2014). Our results in maize are similar, with *gex2* mutant pollen resulting in unfilled ovules, single fertilization events in embryo sacs, and aberrant early seed development from embryo sacs fertilized by *gex2::Ds-GFP* sperm cells. In Arabidopsis, the interaction between the plasma membrane-localized GEX2 and either the female egg or central cells has been suggested to contribute to gamete attachment. The two orthologues share a predicted domain structure, including a large N-terminal non-cytoplasmic region containing filamin repeat domains potentially contributing to this interaction (Mori et al., 2014), raising the possibility that Zm GEX2 acts similarly during double fertilization. With conserved *GEX2*-like genes widely distributed throughout the currently sequenced Embryophyta taxa, our results support the idea that, in flowering plants, these genes play a crucial role in double fertilization.

Maize pollen provides a powerful model for quantifying gene-specific contributions to fitness

The transcriptome dataset also provided a framework to ask broader questions regarding potential relationships between elevated expression and gene function, i.e., utilizing *Ds-GFP* insertions not merely in a genetic screen, but in a mutational interrogation of gene function guided by quantitative hypotheses. Despite the explosion of omic-scale methods to characterize genomes and to measure molecular characters (e.g., transcript levels), our ability to predict phenotypic relevance for specific genes is limited, particularly in multicellular

organisms. A simple hypothesis is that a high transcript level at a particular developmental stage implies functional relevance for the associated gene at that stage, and thus potential for phenotypic influence, a hypothesis supported by observations in mice (Liao and Weng, 2015). This study addresses this hypothesis in plants with a systematic assessment of the functional relevance of highly expressed genes in maize pollen, taking advantage of the ease of reciprocal outcross pollination in maize, the availability of a sufficient number of marked and likely null mutations, and the development of an imaging technique that enables sensitive quantitation.

In the post-pollination progamic stage, pollen grains, as independent, genetically distinct organisms, compete to be the first to deliver the sperm cells to the embryo sac for double fertilization. In an outcrossing plant with an extensive stigma and style like maize, there is likely a heightened context for competition among these individuals (Williams and Reese, 2019), thus providing a milieu that may be particularly sensitive to genetic perturbation. We reasoned that genes highly expressed in the vegetative cell would tend to contribute to a competitive advantage at this stage, which is responsible for pollen tube germination and growth. Indeed, we found that CDS-insertion alleles for 7 out of 32 (21.9%) tested genes in this class are associated with mild to moderate male-specific transmission defects, with the majority of these defects classified as mild (~45% transmission) and thus detectable only by assessing large populations. In this class, transcript level was significantly correlated with both reduced transmission rate and an increased likelihood of significant non-Mendelian transmission, although the explanatory power is limited (R^2 <0.2), as expected for a complex biological system.

Genetic redundancy could also be predicted to influence the phenotypic outcome of mutating single genes, and there is a suggestive trend consistent with this idea in our dataset (Fisher's exact test, p-value = 0.2117), with 6/7 non-Mendelian alleles classified as singletons in the maize genome (86%), compared to 14/25 singletons in genes harboring Mendelian alleles (56%) (Tables 1 and 2). In addition, the variety of biological processes predicted for genes with documented fitness contributions (Table 1) is consistent with the idea that competitive pollen tube growth requires an array of cellular functions. It should be noted that our approach relies on the availability of *Ds-GFP*-marked insertion alleles, and it seems likely that such availability is biased against genes with severe transmission phenotypes, as these would be selected against in a transposon-mutagenized population. In fact, two large *Mutator* transposon populations show a statistically significant deficit in insertions in genes associated with gametophyte-enriched expression (Chettoor et al., 2014). Notably, both previously described

maize genes associated with severe male transmission defect mutants (*rop2*, *apt1*; transmission <15% (Arthur et al., 2003; Xu and Dooner, 2006)) also would be classified as highly expressed in the vegetative cell ($\log_2$(FPKM) > 8), suggesting that the trend we found is applicable even to the types of genes most likely to be absent from the *Ds-GFP* insertion population.

Few studies have investigated the relationship between transcriptomic data and quantitative mutant phenotypes, particularly in multicellular organisms. Large-scale screening of the Arabidopsis T-DNA mutant collection for leaf or reproduction-related phenotypes has identified effects in ~4% of lines assessed (Wilson-Sánchez et al., 2014; Rutter et al., 2019), but these efforts were not guided by transcriptome data. Our use of a sensitive phenotypic assay, combined with a focus on sampling mutations in genes that are most highly expressed at a developmental stage relevant to the phenotype tested, seem likely to have contributed to the higher frequency of phenotypic effects we found. In mice, a meta-analysis found a relationship similar to the one we observed, with aberrant phenotypes more likely to be associated with genes highly transcribed in the tissue exhibiting the phenotype (Liao and Weng, 2015). In contrast to these results, genome scale measurement of the fitness costs of gene knockouts via competitive assays in yeast (Giaever et al., 2002; Berry et al., 2011) and bacteria (Price et al., 2013; Helmann et al., 2019) found that, for particular environmental conditions, there was little to no correlation between the expression level of a gene and its impact on fitness in that condition. This could be indicative of differences between the mechanisms underlying single-celled organisms' response to the environment versus those underlying developmental complexity in multicellular organisms. Interestingly, our results are consistent with a recent meta-analysis that identified higher mRNA expression levels as a feature distinguishing gene models with known mutant phenotypes from the overall population of gene models defined by molecular approaches (Schnable, 2019).

Transposable element dynamics in the maize male gametophyte

The transcriptomic time course enabled exploration of the dynamic relationships among developmental progression, gene expression levels, and transcriptional activation of TEs. Understanding TE expression during maize male reproductive development provides an informative comparison to similar analyses in Arabidopsis, an evolutionarily distant plant with a genome landscape that is distinct from maize. Although maize has a higher number and percentage of its genome occupied by TEs compared to Arabidopsis, we found that only a fraction of maize TEs are developmentally dynamic with regards to transcript accumulation.

These ‘dynamic’ TEs tend to be longer elements than average, which suggests that they have protein coding and transpositional potential. From this dynamic TE set, we were able to identify individual elements that are expressed in a number of specific tissues. However, more globally, there is a trend towards activation of TE transcription over the course of the development of the male gametophyte. This finding confirms that both monocots and eudicots have developmental activation of TE expression in pollen. Consistent with our findings, a recent study found that spontaneous retrotransposon mutations are much more frequent through the male than the female in certain maize lines (Dooner et al., 2019). This conservation suggests that the roles of TE and TE-induced small RNAs during reproductive development may also be conserved between monocots and eudicots (Wang et al., 2018; Martinez et al., 2018; Borges et al., 2018).

Although TE activation is conserved in maize and Arabidopsis pollen, we have identified key differences in the timing and location. Maize TE activation is detected earlier (in the unicellular microspore) compared to when it is thought to occur in Arabidopsis (Slotkin et al., 2009). Transcripts from these early-activated TEs in the microspores typically remain detectable through pollen development and in the mature pollen grain, which may be due to continued expression or transcript stability. A second distinction is the location of activation, which in Arabidopsis is confined to the pollen vegetative cell nucleus (Slotkin et al., 2009; He et al., 2019), whereas in maize also occurs in sperm cells. *Mutator* family TEs are overrepresented in the pool of sperm-cell transcripts, suggesting that this family of TEs may have evolved (or co-opted) specific regulatory mechanism(s) such as an enhancer element that confers expression in this cell type.

Given our results indicating a linkage between elevated gene expression levels and functional relevance, we also assessed whether similar correlations exist between gene and TE expression locally in the maize genome. We found that in mature pollen and sperm cells there is a positive correlation: the more highly expressed a gene is, the more likely it is to have an up-regulated TE nearby. This tissue-specific correlation is a developmentally-specific co-regulation of gene and TE expression. Notably, there does not appear to be any strong trend linking this co-regulation to gene function. We find instances of local gene/TE coordinate regulation are present in similar proportions in genes with documented transmission defects vs. those showing Mendelian segregation when mutant (14% vs 28% respectively, S5 Table). Therefore, although our data indicate the population of highly expressed maize pollen genes has a tendency to contribute to pollen fitness, and a tendency to be adjacent to pollen-expressed TEs, these two characteristics appear to be independent.

Several potential mechanisms may account for the coordinated gene/TE expression. First, programmed activation of TE expression may influence chromatin, enhancer, or other regulatory functions that influence neighboring genes. Second, the genes and TEs may be directly controlled by the same mechanism of large-scale epigenetic activation, limiting the expression of both to a specific tissue or developmental time point. Third, gene activation may influence the expression of the neighboring TE via read-through transcription. Future studies using alternative transcriptomic approaches (e.g., CAGE or long-read RNA sequencing) will enable the dissection of these possible mechanisms.

Methods

Plant materials

Maize inbred line B73 was used for all RNA isolations. Plants were grown in a controlled greenhouse environment (16 hrs light, 8 hrs dark, 80 F day/70 F night) and in the field at the Botany & Plant Pathology Field Lab (Oregon State University, Corvallis, OR) using standard practices. Lines containing *Ds-GFP* insertion alleles were acquired from the Maize Genetics Cooperation Stock Center.

RNA isolation, library preparation and sequencing

Detailed methods are available in S1 Methods. Briefly, tissue was isolated either by dissection (TP), differential density centrifugation (MS, MS-B and SC), or collection at anthesis (MP). Total RNA from TP, MS, and MP was extracted using a modified Trizol Reagent (Life Technologies) protocol; SC total RNA was extracted via a phenol/chloroform protocol. Poly-A RNA (mRNA) was isolated using streptavidin magnetic beads (New England Biolabs, # S1420S) and a biotin-linked poly-T primer. RNA libraries were prepared and sequenced by the Central Services Lab (CSL) at the Center for Genome Research and Biocomputing (CGRB, Oregon State University) using WaferGen robotic strand specific RNA preparation (WaferGen Biosystems) with an Illumina TruSeq RNA LT (single index) prep kit and run on an Illumina HiSeq 3000 with 100 bp paired-end reads.

Mapping reads to genes, differential expression assessment and GO enrichment analysis

Ribosomal reads (rRNA) were removed from all samples using STAR, version 2.5.1b (Dobin et al., 2013) to map reads to a repository of maize rRNA sequences (parameters: --outSAMunmapped Within --outSAMattributes NH HI AS NM MD --outSAMstrandFieldintronMotif --limitBAMsortRAM 50000000000 --outReadsUnmapped Fastx). The number of mappable reads generated from each sample after rRNA removal ranged from approximately 1 million to approximately 41 million, with an average mappable reads of approximately 18 million per sample. Total reads, mapped reads, rRNA contamination, and other statistics are summarized in S1 Table.

rRNA-filtered sequences were mapped to the maize reference genome, version B73 RefGen_v4.33 (Jiao et al., 2017) using STAR, keeping only unique alignments (parameters: --outSAMunmapped Within --outSAMattributes NH HI AS NM MD --outSAMstrandField intronMotif --outFilterMultimapNmax 1 --limitBAMsortRAM 50000000000). Transcript levels of annotated gene isoforms were measured using Cufflinks, version 2.2.1 (Trapnell et al., 2010). FPKM (fragments per kilobase of transcript per million mapped reads) values are shown in S4 Table. Differential expression was calculated between each tissue with Cuffdiff, version 1.0.2, using default parameters. FPKM counts were normalized using the geometric library normalization method. A pooled dispersion method was used by Cuffdiff to model variance. Differential expression results are summarized in S11 Table.

Gene ontology (GO) terms were found for enriched genes in each tissue using the AgriGO 2: GO Analysis Toolkit (Tian et al., 2017). Enriched genes were defined as the top 300 significantly differentially expressed genes (q-value) from Cuffdiff output, with ties broken by $\log_2$ fold change. Enriched sets were split into up- and down-expressed genes. GO term enrichment was calculated using the singular enrichment analysis method with a Fisher test and Yekutieli multi-test adjustment. GO annotations were based off the maize-GAMER annotation set (Wimalanathan et al., 2018).

Mapping reads to transposable elements

The rRNA-filtered reads were quality trimmed (QC30) and adapter sequences were removed using BBDUK2 (Bushnell, 2014). The remaining sequences were mapped to the whole genome using STAR, allowing mapping to at most 100 ‘best’ matching loci. (parameters:

--outMultimapperOrder Random --outSAMmultNmax -1 --outFilterMultimapNmax 100). For paired-end reads, the unmapped reads were re-mapped using single-end approach to maximize the number of mappable reads. The mapping percentage is reported in S3 Table. Because 19% of the total reads in the dataset mapped to more than one location, such reads were mapped to only their best match in the genome, and when multiple best matches existed, they were mapped to all of these loci, and then counted fractionally. For example, if one read maps to 4 TE locations equally well, each TE would receive a weighted value of 0.25 mapped reads. Because the TE expression of the aberrant SC1 biological replicate did not cluster with the other three SC replicates (S1 Fig), it was removed from all subsequent analysis of TE expression.

Principal component analysis (PCA)

Using the maize gene and TE annotation file available from Ensembl Genomes (v38) (Jiao et al., 2017), a combined annotation file was generated for both genes and TEs to run PCA for all samples. FeatureCounts (Liao et al., 2014) was used to calculate the accumulation of each gene and TE in all samples following fractional assignment of reads (parameters: -O --largestOverlap -M --fraction -p -C). This counts file was used in DESeq-2 (Love et al., 2014) to generate the PCA plot.

Analysis of transposable elements

The featureCount file (described above) was used as input for differential expression analysis using DESeq2. Since DESeq2 accepts only integers as raw counts, we used ‘round’ function of R to round the counts to their nearest integers. For differential expression using default parameters for normalization in DESeq2, we only included TEs with a sum total of >=10 read counts across samples; the rest were categorized as ‘not covered’ TEs. First, normalized read counts for all TEs were obtained (data in S2 Fig) and then, only TEs (farther than 2kb from genes) were filtered to be investigated further whereas TEs less than 2kb away from a gene were categorized as ‘near genes’.

After selecting seedling as the reference tissue, pairwise volcano plots were generated for all samples against the reference seedling tissue. Each pairwise comparison with the seedling tissue yielded set of TEs with adjusted p-value of either ‘NA’ or a real number. The set of TEs with a p-value of ‘NA’ in all pairwise comparisons was added to the count of ‘not covered’ category since there was not enough statistical power to call differential expression in any of the tissues. The number of TEs statistically significantly up- and down-regulated (adjusted p-value <

0.05) in each tissue was calculated, categorized as ‘dynamic TEs’ and plotted (Fig 3). Additionally, the number of TEs with adjusted p-value >=0.05 were categorized as developmentally ‘static TEs’ as no evidence of TE expression was observed over different developmental time points analyzed. For all categories, the length, family or distance from centromere was calculated based on the published TE annotation file.

Validation of *Ds-GFP* insertion sites

A FASTA file containing 2 kb of genomic sequence surrounding each *Ds-GFP* insertion site was used as input to a primer3-based tool to generate a pair of specific primers to genotype individual plants from each line (https://vollbrechtlab.gdcb.iastate.edu/tools/primer-server/). The primers used for each *Ds-GFP* line are listed in S6 Table.

To genotype the plants, two 7 mm discs of leaf tissue were collected from each plant using a modified paper punch. The samples were collected in 1.2 ml tubes that fit within a labeled 96 well plate/rack (https://vollbrechtlab.gdcb.iastate.edu/tools/tissue-sample-plate-mapper/) (Phenix Research Products, Candler, NC; M845 and M845BR or equivalent). Genomic DNA was isolated from the leaf punches (Gao et al., 2010) with the following modifications. An additional centrifugation (3,000 *g* for 10 min.) was added to clear the leaf extracts prior to loading onto a 96-well glass fiber filter plate (Pall, 8032). DNA was eluted from filter plates in 125 µL water, and 2 µL was used as template for PCR. Amplification followed standard PCR conditions using GoTaq Green Master Mix (Promega) with 4% DMSO (v/v) and amplicons were resolved using agarose gel electrophoresis. Lines were genotyped using the pair of *Ds-GFP* line gene-specific primers plus one *Ds*-specific primer (JSR01 GTTCGAAATCGATCGGGATA or JGP3 ACCCGACCGGATCGTATCGG). All lines were also screened by PCR for the presence of *wx1-m7::Ac* using primers for *wx1* (CACAGCACGTTGCGGATTTC) and *Ac* (CCGGATCGTATCGGTTTTCG). Followup PCR to test for co-segregation of GFP fluorescence with the presence of the insertion used the appropriate set of three PCR primers (two gene-specific and one *Ds*-specific) and DNA prepared either from endosperm or seedling leaves (Vejlupkova and Fowler, 2003).

Insertional mutagenesis transmission quantification and statistics

Heterozygous lines with PCR-validated *Ds-GFP* insertion alleles were planted in the Botany & Plant Pathology Field Lab (Oregon State University, Corvallis, OR). All insertions were

in coding sequence (CDS) sites. Heterozygous *Ds-GFP* plants were outcrossed to tester plants (*c1*/*c1 wx1*/*wx1* or *c1*/*c1* genetic background) through both the female and the male, with male pollinations made with a heavy pollen load on extended silks (silks that had been allowed to grow for at least two days following cutback). Following harvest, resulting ears were imaged using a custom rotational scanner in the presence of a blue light source and orange filter for GFP seed illumination (Warman and Fowler, 2019). Briefly, videos were captured of rotating ears, which were then processed to generate flat cylindrical projections covering the surface of the ear (for examples, see Figs 5 and 7). Seeds were manually counted using the Cell Counter plugin of the Fiji distribution of ImageJ (Schindelin et al., 2012). Ears showing evidence of more than a single *Ds-GFP* insertion (~75% GFP transmission) were excluded from further analysis. For an allele to be included in the final dataset, we required a minimum of three independent male outcrosses from two different plants.

Seed transmission rates of remaining ears were quantified using a generalized linear model with a logit link function for binomial counts and a quasi-binomial family to correct for overdispersion between parent lines. By incorporating overdispersion, we allowed for the possibility that seeds on the same ear were not completely independent, and for varying transmission rates between ears associated with a given mutation (e.g. by environmental or maternal effects). A quasi-likelihood approach is more realistic than a simple chi-square test, which assumes that all seeds are independent and transmission rates between ears associated with a given mutation are the same. The dispersion parameter for Sperm Cell and Vegetative Cell categories was approximately 1.8, indicating substantially more heterogeneity among seeds on different ears than is expected in a model which assumes independence. Non-Mendelian inheritance was assessed with a quasi-likelihood test with p-values corrected for multiple testing using the Benjamini-Hochberg procedure to control the false discovery rate at 0.05. Significant non-Mendelian segregation was defined with an adjusted p-value < 0.05. Proportions of genes with male-specific transmission defects in the Seedling, Sperm Cell, and Vegetative Cell categories were compared using a two-sided Fisher's exact test, with significance defined as a p-value < 0.05. A two-sided Fisher's exact test was also used to compare the proportions of male-specific transmission defects in the most highly expressed genes and the less highly expressed in the vegetative cell category. A two-sided test for equality of proportions with continuity correction was used to compare transmission rates in families with partial *Ac* presence. A Git repository containing statistical tests and plotting information for this

portion of the study can be found at https://github.com/fowler-lab-osu/maize_gametophyte_transcriptome.

Zm gex2 sequence analysis and phenotype characterization

Zm gex2 protein sequence (Zm00001d005781_T002) was retrieved from the Maize Genetics and Genomics Database (MaizeGDB) hosting of the B73 v4 genome (Portwood et al., 2019; Jiao et al., 2017). Arabidopsis *GEX2* protein sequence (AT5G49150.3) was retrieved from the Arabidopsis Information Portal (ARAPORT) Col-0 Araport11 release (Krishnakumar et al., 2015; Cheng et al., 2017). Maize and Arabidopsis GEX2 protein domains were predicted by InterPro (Mitchell et al., 2019), with transmembrane helix predictions by TMHMM (Krogh et al., 2001). Prediction of land plant species *GEX2* conservation was retrieved from PLAZA, gene family HOM04M006791 (Van Bel et al., 2018). *Zm gex2* gene duplication searches were performed using BLAST (Altschul et al., 1997) and the B73 v4 genome. To confirm the predicted insertion sites for the two *gex2::Ds-GFP* alleles, flanking insertion site fragments were PCR-amplified with a gene-specific primer and a *Ds-GFP*-specific primer (DsGFP_3UTR – TGCAAGCTCGAGTTTCTCCA) and sequenced via Sanger sequencing.

To quantify small seed phenotype, mature, dried down maize ears were imaged prior to seed removal from the ear. For small seeds selection, the ear was first visually scanned row by row from the top to the bottom of the ear. Seeds that were noticeably smaller than their surrounding (regular-sized) seeds are carefully removed from the ear using a pin tool. This sometimes required the removal of regular-sized surrounding seeds, which were saved for later counting. A second visual inspection of the ear often resulted in additional small seeds and is recommended. All remaining seeds were then removed from the ear by hand or using a hand corn sheller tool (Seedburo Equip. Co., Chicago, IL). The ear was screened again for any small (flat/tiny) seeds that could have been missed previously. The cob was inspected prior to discarding, and if any small seed was left behind it was removed and accounted for. Small/smaller seeds and regular-sized seeds were counted and counts were recorded (S10 Table). To measure seedless area, ears were scanned as previously described to create flat surface projections. "Seedless area" was defined as ear surface area that lacked mature or partially developed seeds. Seedless area was quantified as a percentage of total area, as measured with the "Freehand selection" tool of the Fiji distribution of ImageJ (Schindelin et al., 2012). A Git repository containing statistical tests and plotting information for this portion of the study can be found at https://github.com/fowler-lab-osu/maize_gametophyte_transcriptome.

For analysis of embryo sacs by confocal microscopy, tissues were stained with acriflavine, followed by propidium iodide staining (Vollbrecht and Hake, 1995; Running et al., 1995). After staining, samples were dehydrated in an ethanol series and cleared in methyl salicylate. Samples were visualized on a Leica SP8 point-scanning confocal microscope using excitations of 436 nm and 536 nm and emissions of 540 ± 20 nm and 640 ± 20 nm.

Acknowledgements

We thank O. Childress, H. Fowler, B. Galardi, B. Hamilton, R. Hartman, and C. Lambert for their tireless seed counting, genotyping, field work, and other technical assistance; and Dr. Lian Zhou for her contributions to maize field genetics. We also thank K. Wimalanathan and T. Shibamoto for computational support at ISU, and M. Dasenko and the Center for Genome Research and Biocomputing for library preparation, sequencing and computational support at OSU. We thank D. Auger for reading the manuscript.

Supporting information

Supplemental data can be found at the original publication: https://doi.org/10.1371/journal.pgen.1008462

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389–3402.

Anderson, S.N., Stitzer, M.C., Zhou, P., Ross-Ibarra, J., Hirsch, C.D., and Springer, N.M. (2019). Dynamic Patterns of Transcript Abundance of Transposable Element Families in Maize. G3 **9**: 3673–3682.

Arthur, K.M., Vejlupkova, Z., Meeley, R.B., and Fowler, J.E. (2003). Maize ROP2 GTPase provides a competitive advantage to the male gametophyte. Genetics **165**: 2137–2151.

Becker, J.D., Boavida, L.C., Carneiro, J., Haury, M., and Feijó, J.A. (2003). Transcriptional profiling of Arabidopsis tissues reveals the unique characteristics of the pollen transcriptome. Plant Physiol. **133**: 713–725.

Begcy, K., Nosenko, T., Zhou, L.-Z., Fragner, L., Weckwerth, W., and Dresselhaus, T. (2019). Male Sterility in Maize after Transient Heat Stress during the Tetrad Stage of Pollen Development. Plant Physiol.

Berry, D.B., Guan, Q., Hose, J., Haroon, S., Gebbia, M., Heisler, L.E., Nislow, C., Giaever, G., and Gasch, A.P. (2011). Multiple means to the same end: the genetic basis of acquired stress resistance in yeast. PLoS Genet. **7**: e1002353.

Borges, F., Parent, J.-S., van Ex, F., Wolff, P., Martínez, G., Köhler, C., and Martienssen, R.A. (2018). Transposon-derived small RNAs triggered by miR845 mediate genome dosage response in Arabidopsis. Nat. Genet. **50**: 186–192.

Bushnell, B. (2014). BBTools software package. URL http://sourceforge.net/projects/bbmap.

Calarco, J.P., Borges, F., Donoghue, M.T.A., Van Ex, F., Jullien, P.E., Lopes, T., Gardner, R., Berger, F., Feijó, J.A., Becker, J.D., and Martienssen, R.A. (2012). Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. Cell **151**: 194–205.

Cheng, C.-Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., and Town, C.D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. **89**: 789–804.

Chen, J., Strieder, N., Krohn, N.G., Cyprys, P., Sprunck, S., Engelmann, J.C., and Dresselhaus, T. (2017). Zygotic Genome Activation Occurs Shortly after Fertilization in Maize. Plant Cell **29**: 2106–2125.

Chettoor, A.M., Givan, S.A., Cole, R.A., Coker, C.T., Unger-Wallace, E., Vejlupkova, Z., Vollbrecht, E., Fowler, J.E., and Evans, M.M. (2014). Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. Genome Biol. **15**: 414.

Cole, R.A., Synek, L., Zarsky, V., and Fowler, J.E. (2005). SEC8, a subunit of the putative Arabidopsis exocyst complex, facilitates pollen germination and competitive pollen tube growth. Plant Physiol. **138**: 2005–2018.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**: 15–21.

Dooner, H.K., Wang, Q., Huang, J.T., Li, Y., He, L., Xiong, W., and Du, C. (2019). Spontaneous mutations in maize pollen are frequent in some lines and arise mainly from retrotranspositions and deletions. Proc. Natl. Acad. Sci. U. S. A.

Dresselhaus, T., Sprunck, S., and Wessel, G.M. (2016). Fertilization Mechanisms in Flowering Plants. Curr. Biol. **26**: R125–39.

Engel, M.L., Chaboud, A., Dumas, C., and McCormick, S. (2003). Sperm cells of Zea mays have a complex complement of mRNAs. Plant J. **34**: 697–707.

Engel, M.L., Holmes-Davis, R., and McCormick, S. (2005). Green sperm. Identification of male gamete promoters in Arabidopsis. Plant Physiol. **138**: 2124–2133.

Gao, H., Smith, J., Yang, M., Jones, S., Djukanovic, V., Nicholson, M.G., West, A., Bidney,

D., Falco, S.C., Jantz, D., and Lyznik, L.A. (2010). Heritable targeted mutagenesis in maize using a designed endonuclease. Plant J. **61**: 176–187.

Giaever, G. et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature **418**: 387–391.

Gilles, L.M. et al. (2017). Loss of pollen-specific phospholipase NOT LIKE DAD triggers gynogenesis in maize. EMBO J. **36**: 707–717.

Hafidh, S., Fíla, J., and Honys, D. (2016). Male gametophyte development and function in angiosperms: a general concept. Plant Reprod. **29**: 31–51.

Helmann, T.C., Deutschbauer, A.M., and Lindow, S.E. (2019). Genome-wide identification of Pseudomonas syringae genes required for fitness during colonization of the leaf surface and apoplast. Proc. Natl. Acad. Sci. U. S. A.

He, S., Vickers, M., Zhang, J., and Feng, X. (2019). Natural depletion of histone H1 in sex cells causes DNA demethylation, heterochromatin decondensation and transposon activation. Elife **8**.

Higashiyama, T. and Takeuchi, H. (2015). The mechanism and key molecules involved in pollen tube guidance. Annu. Rev. Plant Biol. **66**: 393–413.

Honys, D. and Twell, D. (2003). Comparative analysis of the Arabidopsis pollen transcriptome. Plant Physiol. **132**: 640–652.

Huang, J.T., Wang, Q., Park, W., Feng, Y., Kumar, D., Meeley, R., and Dooner, H.K. (2017). Competitive Ability of Maize Pollen Grains Requires Paralogous Serine Threonine Protein Kinases STK1 and STK2. Genetics **207**: 1361–1370.

Jiao, Y. et al. (2017). Improved maize reference genome with single-molecule technologies. Nature **546**: 524–527.

Jones-Rhoades, M.W., Borevitz, J.O., and Preuss, D. (2007). Genome-wide expression profiling of the Arabidopsis female gametophyte identifies families of small, secreted proteins. PLoS Genet. **3**: 1848–1861.

Kelliher, T. et al. (2017). MATRILINEAL, a sperm-specific phospholipase, triggers maize haploid induction. Nature **542**: 105–109.

Kelliher, T. et al. (2019). One-step genome editing of elite crop germplasm during haploid induction. Nat. Biotechnol. **37**: 287–292.

Krishnakumar, V. et al. (2015). Araport: the Arabidopsis information portal. Nucleic Acids Res. **43**: D1003–9.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. **305**: 567–580.

Lausser, A., Kliwer, I., Srilunchang, K.-O., and Dresselhaus, T. (2010). Sporophytic control

of pollen tube growth and guidance in maize. J. Exp. Bot. **61**: 673–682.

Liao, B.-Y. and Weng, M.-P. (2015). Unraveling the association between mRNA expressions and mutant phenotypes in a genome-wide assessment of mice. Proc. Natl. Acad. Sci. U. S. A. **112**: 4707–4712.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**: 923–930.

Liu, C. et al. (2017). A 4-bp Insertion at ZmPLA1 Encoding a Putative Phospholipase A Generates Haploid Induction in Maize. Mol. Plant **10**: 520–522.

Li, Y., Segal, G., Wang, Q., and Dooner, H.K. (2013). Gene Tagging with Engineered Ds Elements in Maize. In Plant Transposable Elements: Methods and Protocols, T. Peterson, ed (Humana Press: Totowa, NJ), pp. 83–99.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**: 550.

Lunardon, A., Forestan, C., Farinati, S., Axtell, M.J., and Varotto, S. (2016). Genome-Wide Characterization of Maize Small RNA Loci and Their Regulation in the required to maintain repression6-1 (rmr6-1) Mutant and Long-Term Abiotic Stresses. Plant Physiol. **170**: 1535–1548.

Martínez, G., Panda, K., Köhler, C., and Slotkin, R.K. (2016). Silencing in sperm cells is directed by RNA movement from the surrounding nurse cell. Nat Plants **2**: 16030.

Martínez, G. and Slotkin, R.K. (2012). Developmental relaxation of transposable element silencing in plants: functional or byproduct? Curr. Opin. Plant Biol. **15**: 496–502.

Martinez, G., Wolff, P., Wang, Z., Moreno-Romero, J., Santos-González, J., Conze, L.L., DeFraia, C., Slotkin, R.K., and Köhler, C. (2018). Paternal easiRNAs regulate parental genome dosage in Arabidopsis. Nat. Genet. **50**: 193–198.

McCormick, S. (1993). Male Gametophyte Development. Plant Cell **5**: 1265–1275.

Mitchell, A.L. et al. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. **47**: D351–D360.

Mizukami, A.G. et al. (2016). The AMOR Arabinogalactan Sugar Chain Induces Pollen-Tube Competency to Respond to Ovular Guidance. Curr. Biol. **26**: 1091–1097.

Mori, T., Igawa, T., Tamiya, G., Miyagishima, S.-Y., and Berger, F. (2014). Gamete attachment requires GEX2 for successful fertilization in Arabidopsis. Curr. Biol. **24**: 170–175.

Nelms, B. and Walbot, V. (2019). Defining the developmental program leading to meiosis in maize. Science **364**: 52–56.

Panda, K., Ji, L., Neumann, D.A., Daron, J., Schmitz, R.J., and Slotkin, R.K. (2016). Full-length autonomous transposable elements are preferentially targeted by

expression-dependent forms of RNA-directed DNA methylation. Genome Biol. **17**: 170.

Portwood, J.L., 2nd et al. (2019). MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic Acids Res. **47**: D1146–D1154.

Price, M.N., Deutschbauer, A.M., Skerker, J.M., Wetmore, K.M., Ruths, T., Mar, J.S., Kuehl, J.V., Shao, W., and Arkin, A.P. (2013). Indirect and suboptimal control of gene expression is widespread in bacteria. Mol. Syst. Biol. **9**: 660.

Running, M.P., Clark, S.E., and Meyerowitz, E.M. (1995). Chapter 15 Confocal Microscopy of the Shoot Apex. In Methods in Cell Biology, D.W. Galbraith, H.J. Bohnert, and D.P. Bourque, eds (Academic Press), pp. 217–229.

Rutter, M.T., Murren, C.J., Callahan, H.S., Bisner, A.M., Leebens-Mack, J., Wolyniak, M.J., and Strand, A.E. (2019). Distributed phenomics with the unPAK project reveals the effects of mutations. Plant J.

Schindelin, J. et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods **9**: 676–682.

Schnable, J.C. (2019). Genes and gene models, an important distinction. New Phytol.

Schoft, V.K., Chumak, N., Mosiolek, M., Slusarz, L., Komnenovic, V., Brownfield, L., Twell, D., Kakutani, T., and Tamaru, H. (2009). Induction of RNA-directed DNA methylation upon decondensation of constitutive heterochromatin. EMBO Rep. **10**: 1015–1021.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdzić, M., Becker, J.D., Feijó, J.A., and Martienssen, R.A. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell **136**: 461–472.

Steffen, J.G., Kang, I.-H., Macfarlane, J., and Drews, G.N. (2007). Identification of genes expressed in the Arabidopsis female gametophyte: Female gametophyte-expressed genes. Plant J. **51**: 281–292.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., and Su, Z. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. **45**: W122–W129.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28**: 511–515.

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., Coppens, F., and Vandepoele, K. (2018). PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. **46**: D1190–D1196.

Vejlupkova, Z. and Fowler, J.E. (2003). Maize DNA preps for undergraduate students: a robust method for PCR genotyping. Maize Genetics Cooperation Newsletter **77**: 24–25.

Vollbrecht, E. and Hake, S. (1995). Deficiency analysis of female gametogenesis in maize.

Dev. Genet. **16**: 44–63.

Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., and Briggs, S.P. (2016). Integration of omic networks in a developmental atlas of maize. Science **353**: 814–818.

Wang, G., Jiang, H., Del Toro de León, G., Martinez, G., and Köhler, C. (2018). Sequestration of a Transposon-Derived siRNA by a Target Mimic Imprinted Gene Induces Postzygotic Reproductive Isolation in Arabidopsis. Dev. Cell **46**: 696–705.e4.

Warman, C. and Fowler, J.E. (2019). Custom built scanner and simple image processing pipeline enables low-cost, high-throughput phenotyping of maize ears. bioRxiv: 780650.

Williams, J.H. and Reese, J.B. (2019). Evolution of development of pollen performance. Curr. Top. Dev. Biol. **131**: 299–336.

Wilson-Sánchez, D., Rubio-Díaz, S., Muñoz-Viana, R., Pérez-Pérez, J.M., Jover-Gil, S., Ponce, M.R., and Micol, J.L. (2014). Leaf phenomics: a systematic reverse genetic screen for Arabidopsis leaf mutants. Plant J. **79**: 878–891.

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**: e00052.

Wolfgruber, T.K. et al. (2009). Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons. PLoS Genet. **5**: e1000743.

Xu, Z. and Dooner, H.K. (2006). The maize aberrant pollen transmission 1 gene is a SABRE/KIP homolog required for pollen tube growth. Genetics **172**: 1251–1261.

Yang, W.-C., Shi, D.-Q., and Chen, Y.-H. (2010). Female gametophyte development in flowering plants. Annu. Rev. Plant Biol. **61**: 89–108.

Zhai, J., Zhang, H., Arikit, S., Huang, K., Nan, G.-L., Walbot, V., and Meyers, B.C. (2015). Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. Proc. Natl. Acad. Sci. U. S. A. **112**: 3146–3151.

Zhong, Y. et al. (2019). Mutation of ZmDMP enhances haploid induction in maize. Nat Plants **5**: 575–580.

Zhou, L.-Z. and Dresselhaus, T. (2019). Chapter Seventeen - Friend or foe: Signaling mechanisms during double fertilization in flowering seed plants. In Current Topics in Developmental Biology, U. Grossniklaus, ed (Academic Press), pp. 453–496.

Zhou, L.-Z., Juranić, M., and Dresselhaus, T. (2017). Germline Development and Fertilization Mechanisms in Maize. Mol. Plant **10**: 389–401.

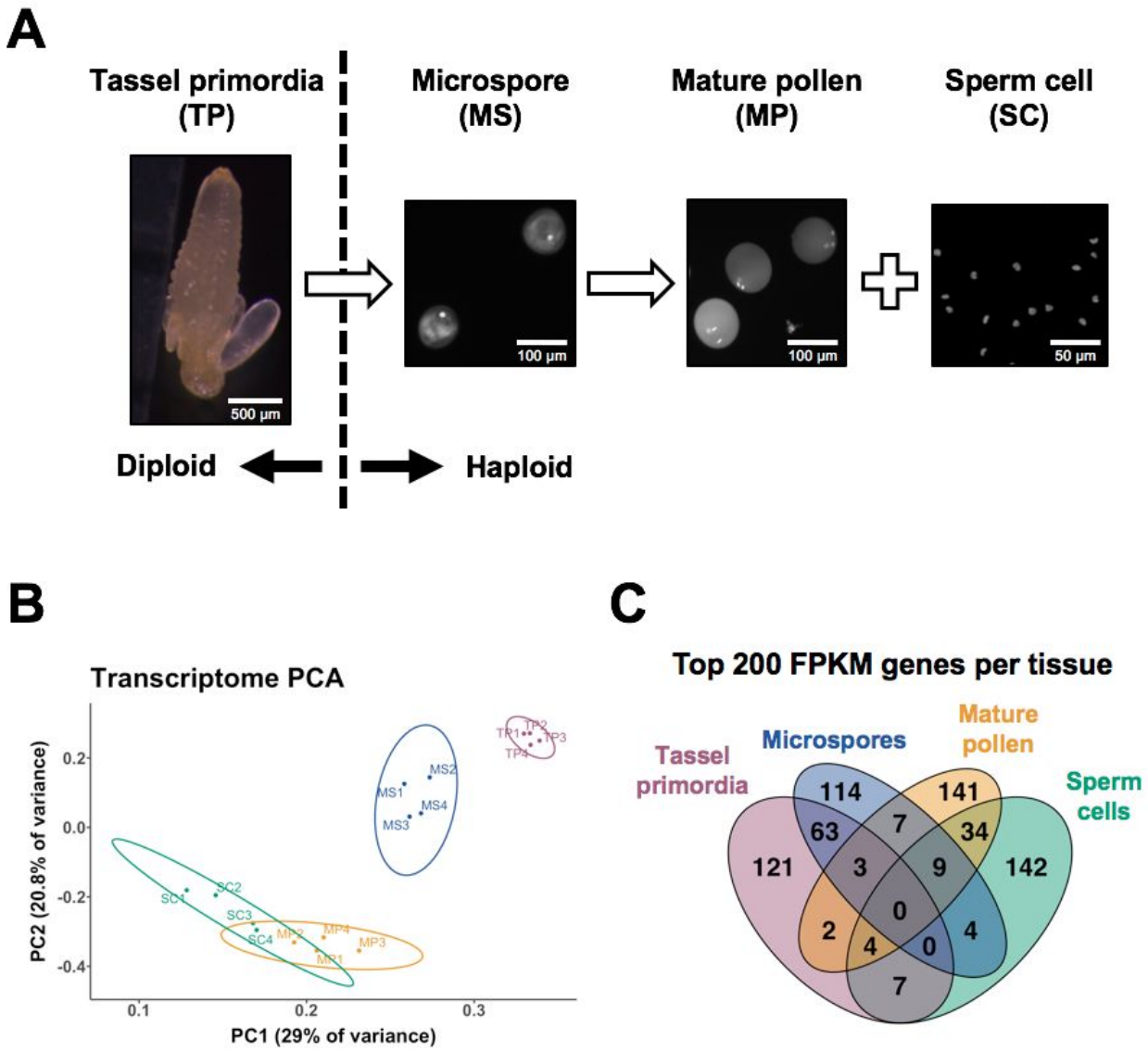


Fig 1. Experimental design and characteristics of maize male reproductive transcriptomes.

(A) mRNA was isolated from four developmental stages of maize male reproductive development, with four biological replicates for each: pre-meiotic tassel primordia (TP), post-meiotic, unicellular microspores (MS), mature pollen (MP), and sperm cells (SC). A single biological replicate of mRNA from the bicellular stage of pollen development was also isolated and sequenced (MS-B, not shown). Nuclei were stained with either DAPI or Dyecycle green. **(B)** Principal component analysis of genic transcriptomic data generated by this study, showing the 2 major components (explaining 49.8% of the variance) on x- and y-axis. The four biological replicates of each of the four sequenced tissues clustered with other replicates from the same tissue. TP and MS were clearly separated in principal component space, whereas SC and MP samples displayed less separation from each other. **(C)** High expression levels are associated

with developmental specificity: approximately 2/3 of the genes associated with the highest FPKM values in each of the four sample types are highly expressed in only that sample type.

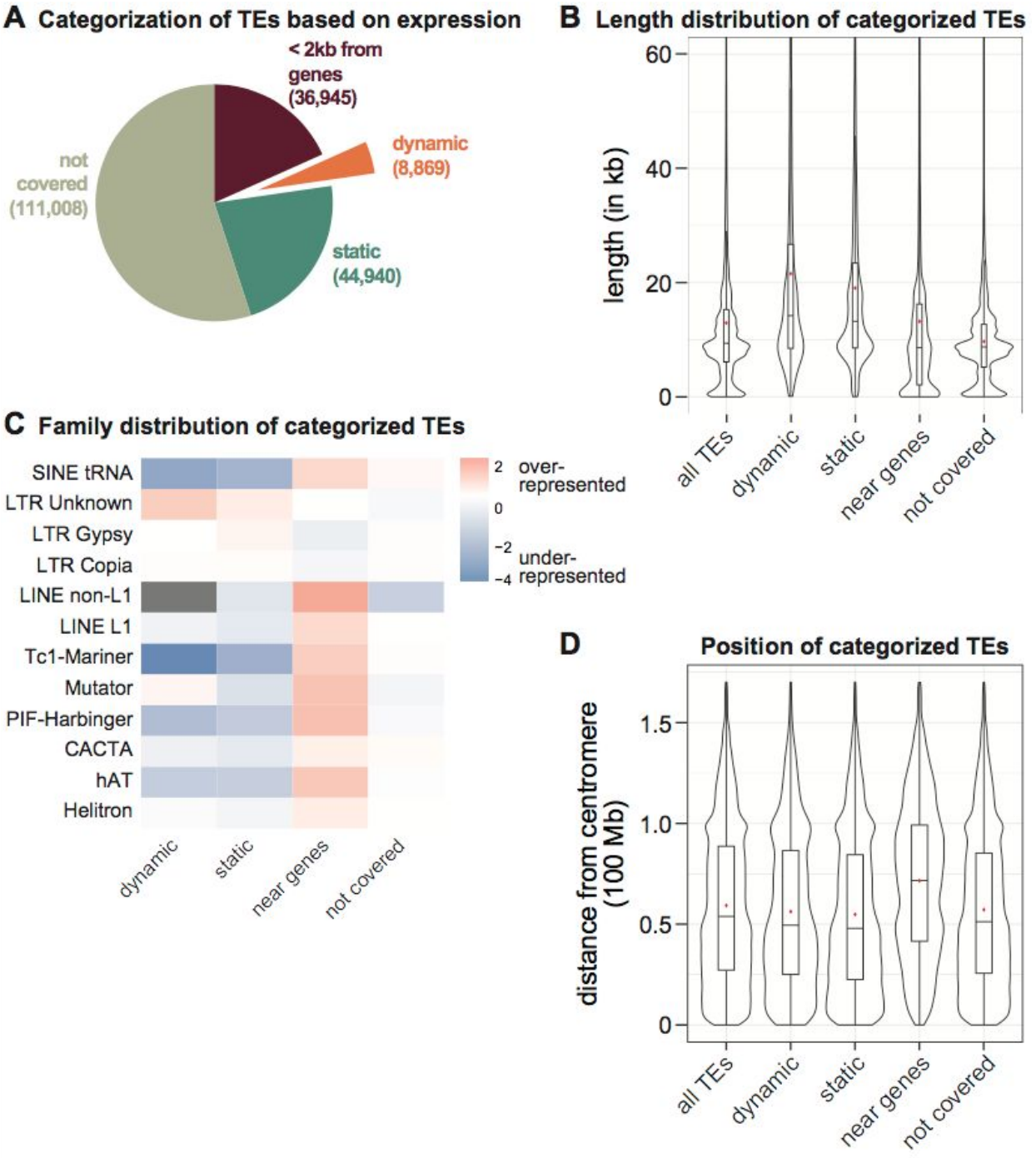


Fig 2. Characterization of developmentally dynamic transcription from transposable elements (TEs).

(A) Distribution of different categories of TEs based on their expression. Number of TEs are in parentheses. **(B)** Length of TEs in the different TE categories from part A. The violin plots around the box show the kernel probability density of the data. The box represents lower and upper quartiles, the line is the median, and the whiskers represent 10-90% range. Red asterisk denotes the mean. **(C)** Observed / expected Log2 ratios of TE family proportions in the different

TE categories from part A. Grey indicates no data. **(D)** Distance from the annotated centromere for different TE categories from panel A.

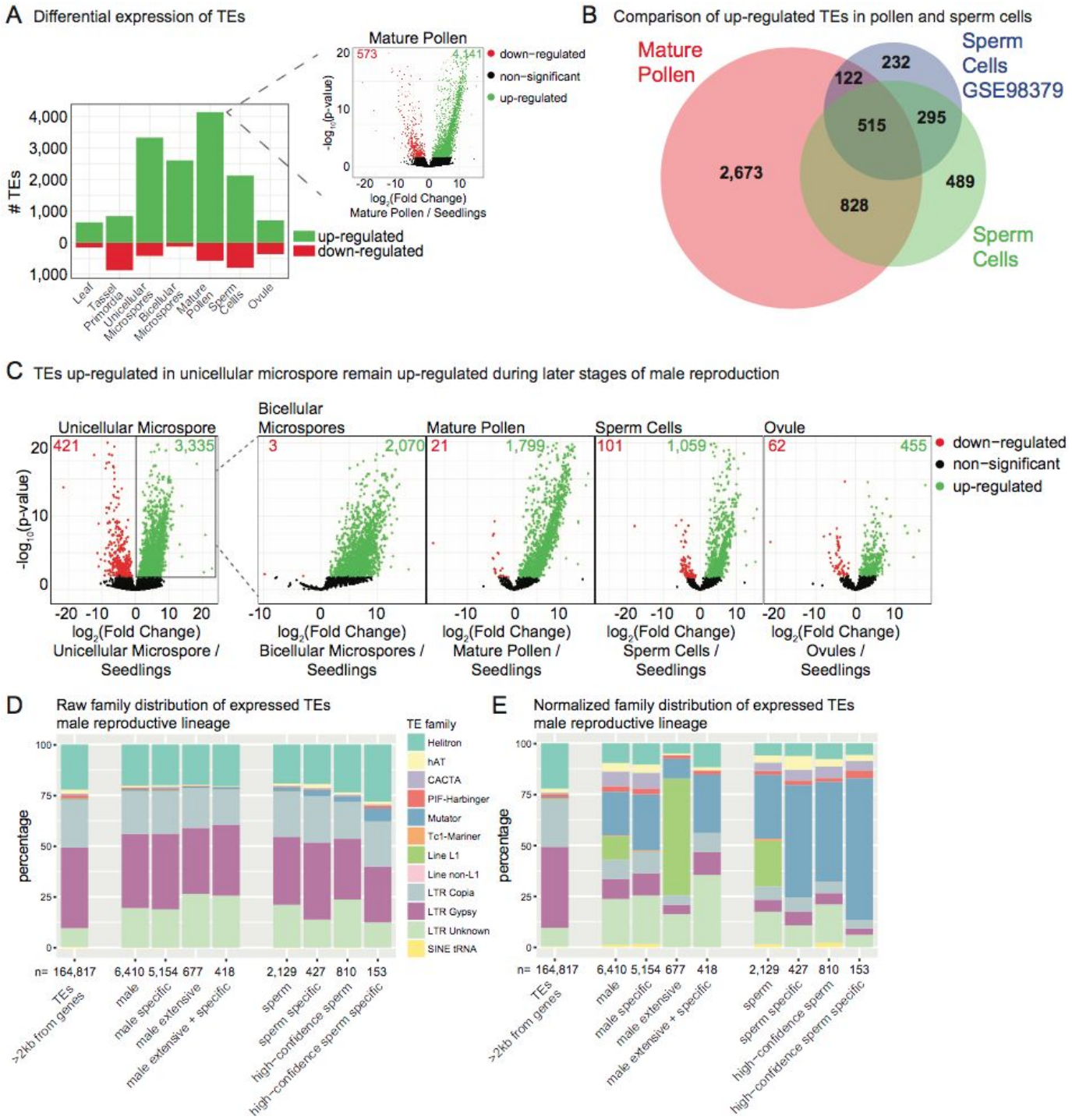


Fig 3. High TE expression in the maize male gametophyte lineage.
(A) Number of differentially expressed TEs in seven tissues compared to seedlings. The inset volcano plot shows for mature pollen how differentially expressed TEs were identified. Green and red numbers within the volcano plot indicate how many TEs were statistically up- or down-regulated, respectively. **(B)** Number of up-regulated TEs in mature pollen compared to isolated sperm cells from this study and a previously published distinct isolation and sequencing of sperm cell mRNA. **(C)** Starting with TEs differentially up-regulated in unicellular microspores (boxed, far left volcano plot), we determined how many of these same TEs are expressed at other developmental time points. **(D)** Raw distribution of expressed TE family annotations. 'Male' refers to the set of TEs expressed in any male lineage dataset (MS, MS-B, MP, SC). 'Male specific' are TEs expressed in only the male lineage (not other tissues / timepoints). 'Male extensive' TEs are expressed in all of the male lineage datasets, and 'male extensive + specific' refers to TEs expressed all male lineage datasets and not other tissues / timepoints. 'High-confidence sperm' refers to TEs identified as expressed in both analyzed sperm cell datasets from part B. **(E)** Expressed TE family annotations normalized to the genome-wide TE distribution of TEs >2 kb from genes. Categories are the same as part D.

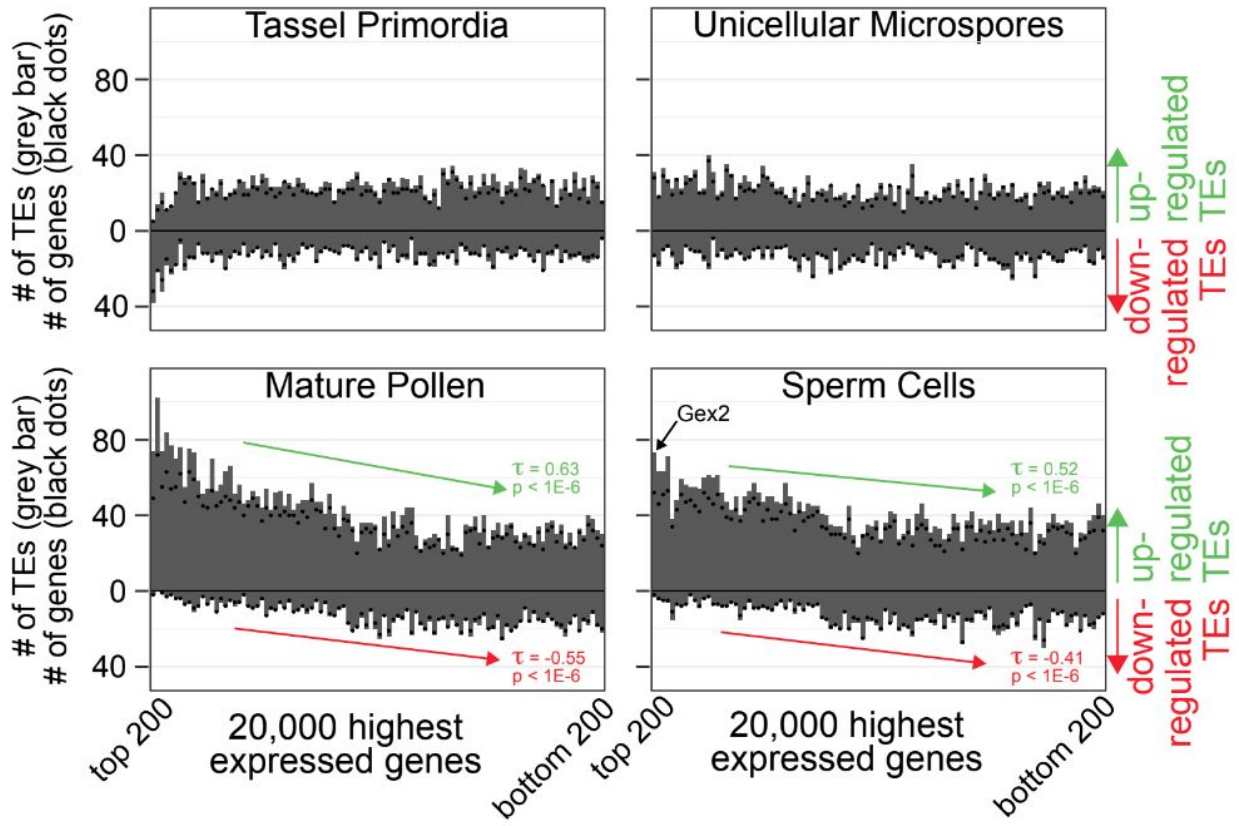


Fig 4. Co-regulation of TE and gene expression in the male gametophyte.
For each tissue type, the top 20,000 most highly expressed genes are distributed along the X-axis in bins of 200, with the most highly expressed bin on the far left. For each bin the number of up- and down-regulated TEs near (<2kb) that bin's genes is then summed on the Y-axis (shown in grey bar). For each bin, the number of genes with at least 1 up- or down-regulated TE within 2kb is displayed as black dots. In unicellular microspores (top right) there is little correlation, whereas in mature pollen and sperm cells (bottom panels) the most highly expressed genes are near primarily up-regulated TEs. To check if the perceived correlations are statistically significant, we performed Kendall's Tau rank correlation tests and found significant correlations ($p < 1E-6$) only for mature pollen and sperm cells for both number of TEs and number of genes. For mature pollen, gene expression has a positive correlation with up-regulated TEs (#TEs: τ=0.65, #genes: τ=0.64) and a negative correlation with down-regulated TEs (#TEs: τ=-0.55, #genes: τ=-0.55). Similarly for sperm cells, gene expression has a positive correlation with up-regulated TEs (#TEs: τ=0.52, #genes: τ=0.52) and a negative correlation with down-regulated TEs (#TEs: τ=-0.42, #genes: τ=-0.44). Only TE (not genes) rank correlation coefficients (τ) are displayed in the figure. The bin location of *gex2* (see Fig 7) is annotated in the sperm cell data.

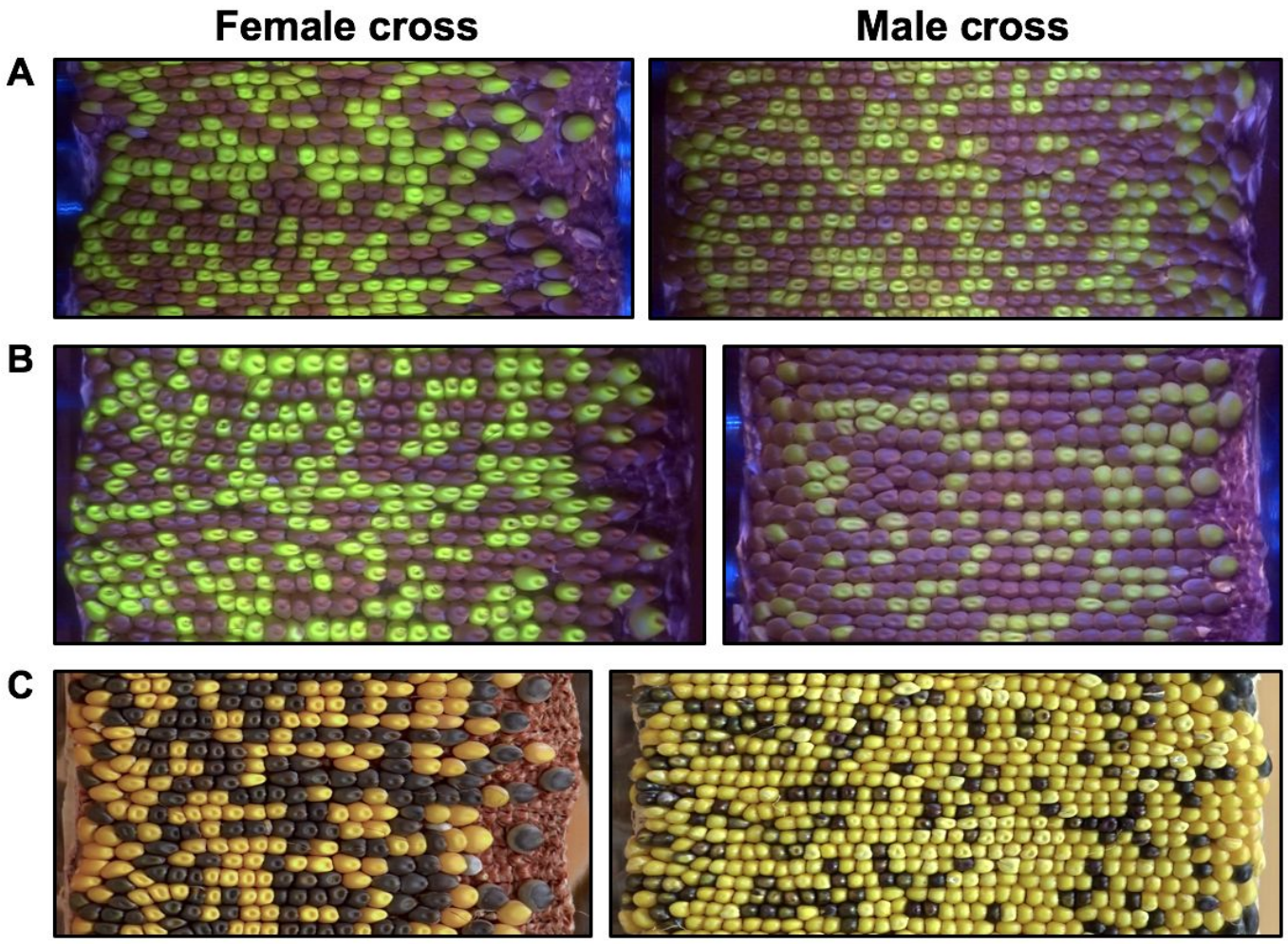


Fig 5. Large-scale tracking of seed marker transmission frequencies was accomplished by generating ear projections with a custom built rotational scanner.
(A) When crossed either through the male or the female, *Ds-GFP* mutant allele *tdsgR107C12* (in gene Zm00001d012382), marked by green fluorescent seeds, shows 1:1 Mendelian inheritance (50% transmission of the GFP seed marker). Images captured in blue light with an orange filter. **(B)** Mutant alleles in other genes, such as *tdsgR102H01* (Zm00001d037695), showed non-Mendelian segregation when crossed through the male (37.5% GFP transmission). Segregation through the female remained Mendelian, indicating a male-specific transmission defect. **(C)** For some mutant alleles (~10% of lines in this study), the anthocyanin transgene *C1* was tightly linked to the insertion mutant. In these cases, seeds carrying a mutant allele of a gene of interest could be tracked by their purple color. Here, insertion *tdsgR96C12* (Zm00001d015901) shows a strong male-specific transmission defect (24.8% *C1* transmission through the male). Images captured in full spectrum visible light.

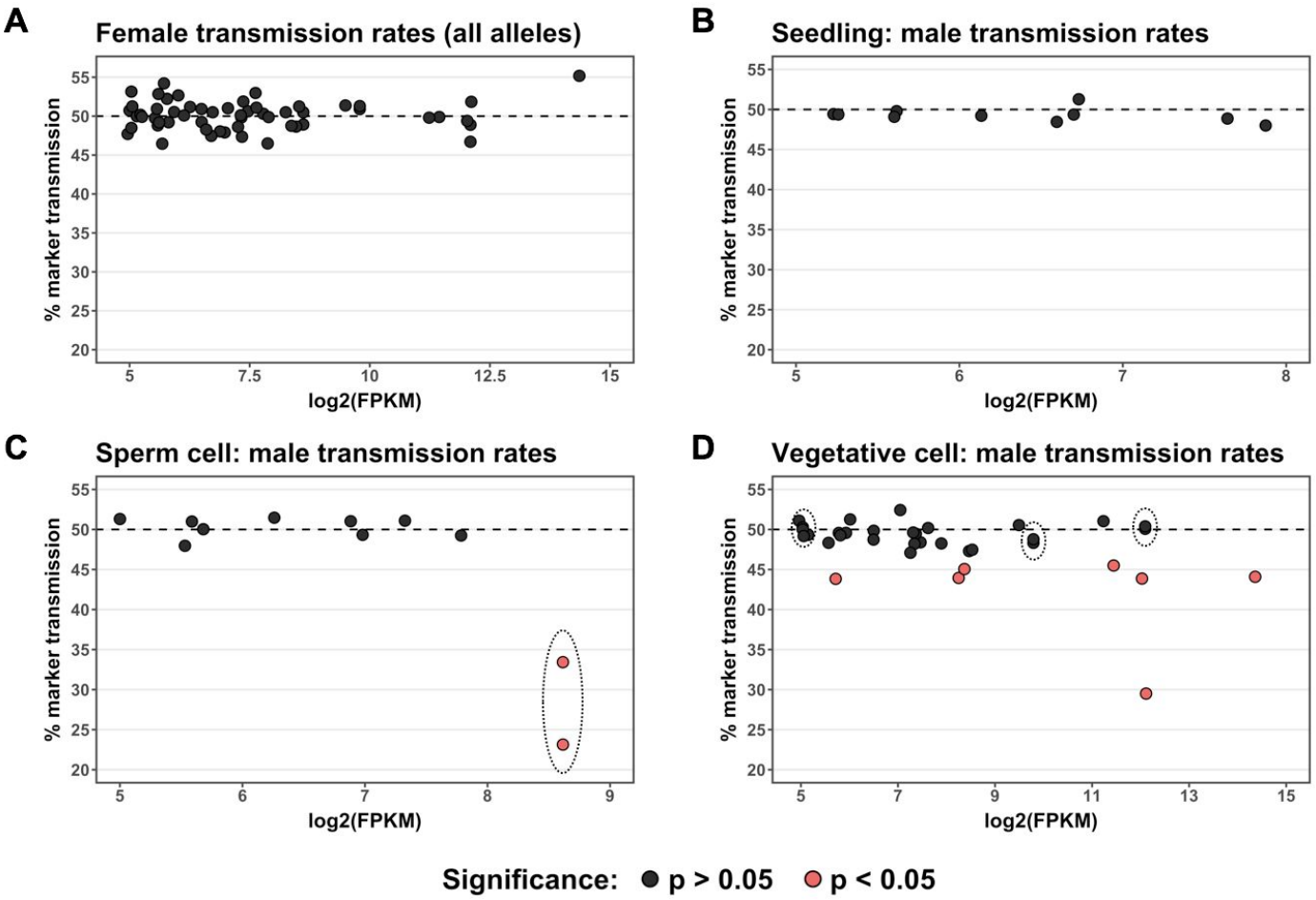


Fig 6. Functional validation of highly-expressed gametophyte genes by quantification of transmission rates in *Ds-GFP* insertional mutants.

Alleles with CDS insertions were tested for differences from Mendelian inheritance using a quasi-likelihood test, with p-values corrected for multiple testing using the Benjamini-Hochberg procedure; alleles with quasi-likelihood adjusted p-value < 0.05 are represented in pink. Alleles are plotted by the $\log_2$(FPKM) of their respective gene according to that gene's expression class (Seedling, Vegetative Cell, or Sperm Cell). Insertion alleles distributed among the classes as follows: Vegetative Cell, 35 alleles; Sperm Cell, 11 alleles; Seedling, 10 alleles. The number of seeds categorized for each allele ranged from 1,522 to 5,219, with an average of 2,807. Genes represented by two independent insertion alleles are enclosed by dotted lines. **(A)** Transmission rates of 56 mutant allele seed markers for heterozygous *Ds-GFP* mutant plants crossed through the female. **(B)** Transmission rates for alleles in the negative control Seedling class when crossed through the male. **(C)** For genes belonging to the Sperm Cell class, one out of ten (10%) was associated with significant non-Mendelian inheritance. The single gene with a male transmission defect in this group (*gex2*) showed a strong defect for both of the independent alleles tested. **(D)** An increased proportion of the genes in the Vegetative Cell class were associated with significant non-Mendelian inheritance when mutant (7/32 genes, 21.9%). In this class, an increase in $\log_2$(FPKM) was significantly associated with a decrease in marker transmission (linear regression, p = 0.0120).

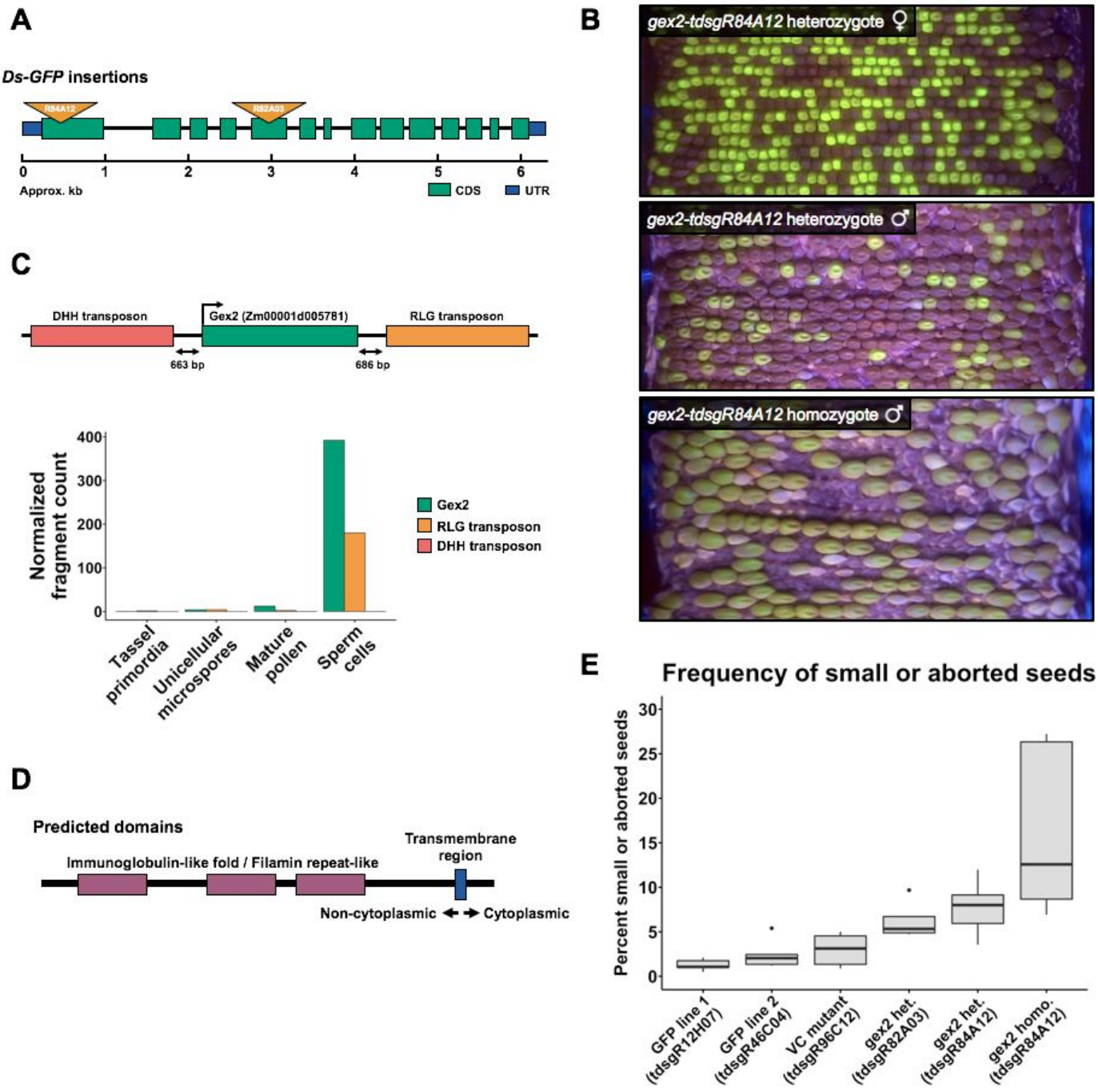


Fig 7. Mutations in the sperm cell-specific *gex2* gene cause aberrant seed development. **(A)** The exon/intron structure of *gex2* (Zm00001d005781/GRMZM2G036832), showing the locations of the two independent *Ds-GFP* insertion mutants. **(B)** Ear projections of *gex2* mutant outcrosses. Top: heterozygote outcrossed as female, showing 1:1 transmission of the GFP-tagged allele. Middle: heterozygote outcrossed as a male, with 26.1% transmission of the mutant allele. Additionally, small seeds and occasional, small gaps between seeds are visible. Bottom: homozygous mutant outcrossed as a male, with many small seeds and large gaps,

despite heavy pollination. **(C)** Genomic neighborhood of the GEX2 locus, with two nearby TEs, and their RNA-seq expression levels across male reproductive development. **(D)** Predicted domain structure of ZmGEX2; the amino acid sequence shows 44.2% similarity with Arabidopsis GEX2. **(E)** Quantification of small/aborted seeds resulting from pollination by *gex2* mutant plants and controls. Controls included two *Ds-GFP* lines that did not show transmission defects (*tdsgR12H07* and *tdsgR46C04*) and one *Ds-GFP* line that showed a strong transmission defect in the vegetative cell group (*tdsgR96C12*; 29.5% transmission). A higher percentage of small/aborted seeds was present following pollination by heterozygous *gex2* plants representing the two mutant alleles (*tdsgR82A03* and *tdsgR84A12*), and pollination by homozygous *gex2-tdsgR84A12* plants further increased the percentage of small/aborted seeds.

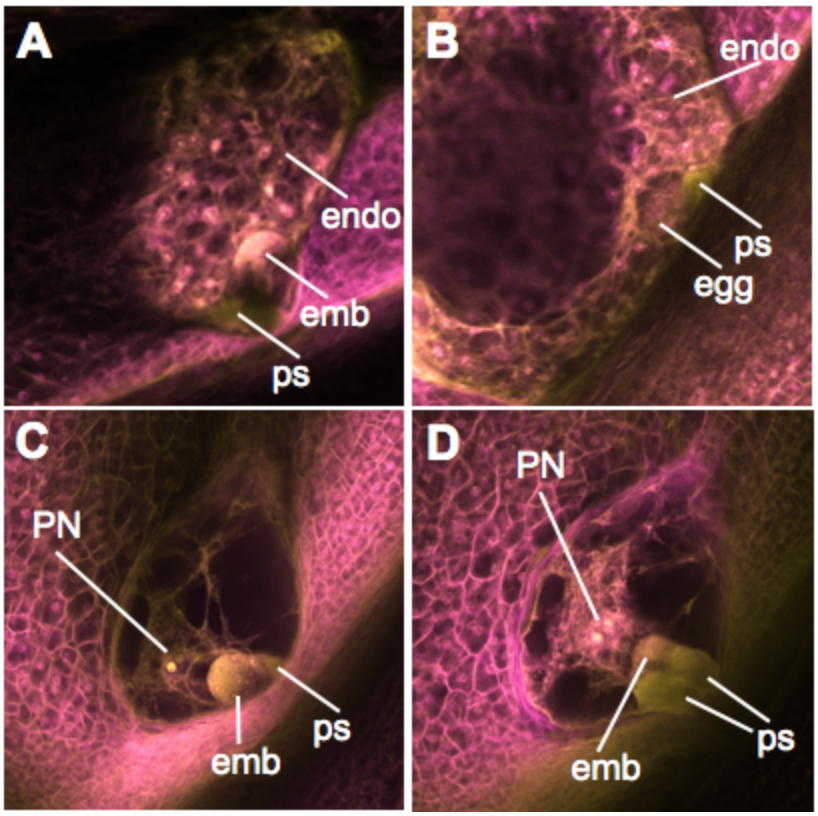
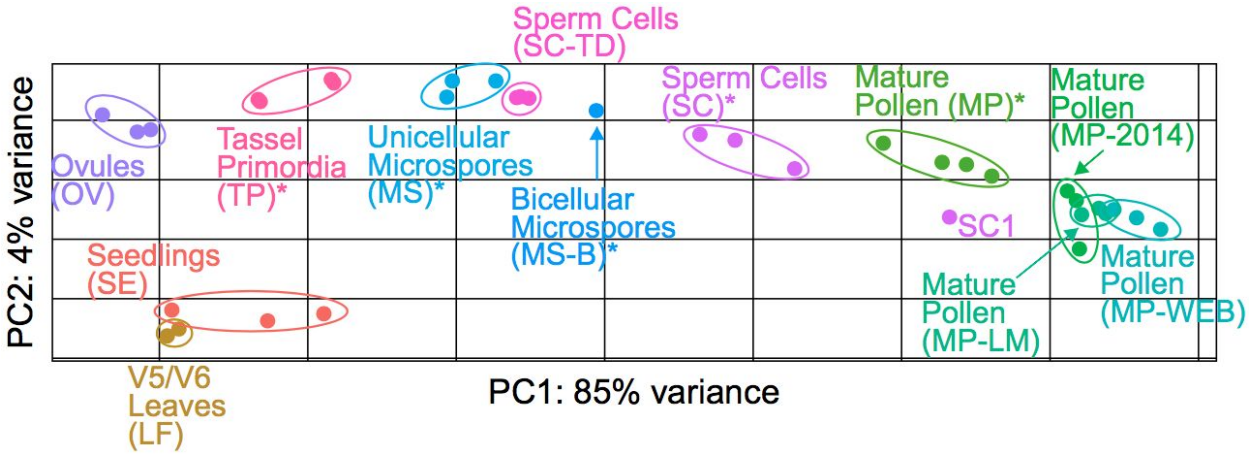


Fig 8. Pollination by *gex2-tdsgR84A12* leads to aberrant fertilization events and developing seed phenotypes.
(A) Seed development in a typical ovule pollinated by wild-type pollen, with one synergid penetrated by a pollen tube, and both embryo and endosperm development initiated. **(B-D)** Abnormal phenotypes seen following *gex2-tdsgR84A12* pollination. **(B)** Ovule with developing (cellularizing) endosperm but unfertilized egg cell. **(C)** Ovule with developing embryo but unfertilized central cell. **(D)** Ovule with both synergids penetrated by a pollen tube, and a developing embryo and unfertilized central cell. emb=embryo; endo=endosperm; ps=synergid penetrated by a pollen tube; PN=polar nuclei.

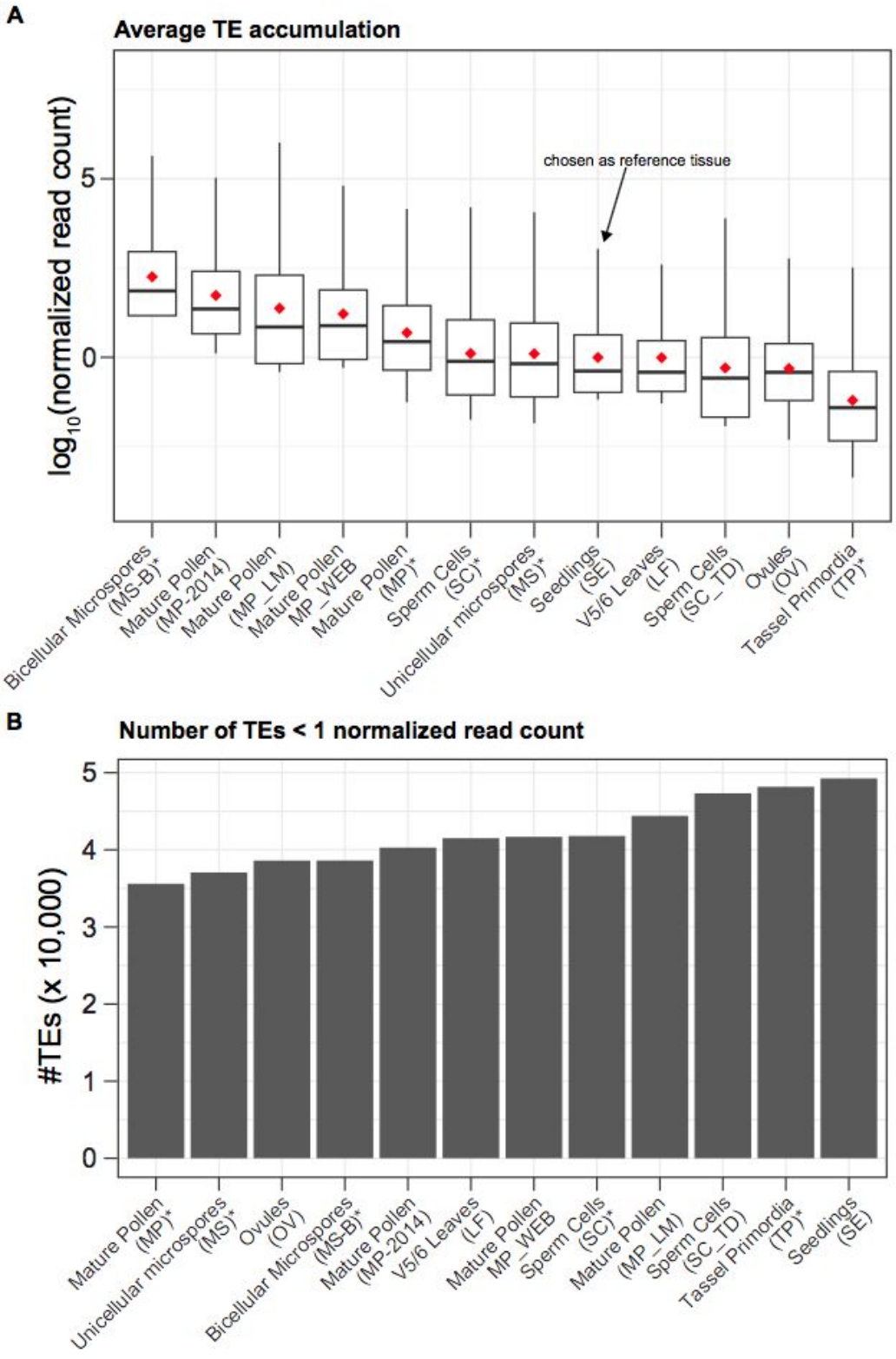
Table 3. Seed development at 4 days after pollination by wild-type or *gex2::Ds-GFP* pollen

Pollen parent	One synergid penetrated by a pollen tube			Both synergids penetrated by a pollen tube		
	Both embryo and endosperm	Endosperm without embryo	Embryo without endosperm	Both embryo and endosperm	Endosperm without embryo	Embryo without endosperm
gex2-tdsgR84A12/ gex2-tdsgR84A12	6	2	2	0	0	2
Wild-type	28	0	0	0	0	0

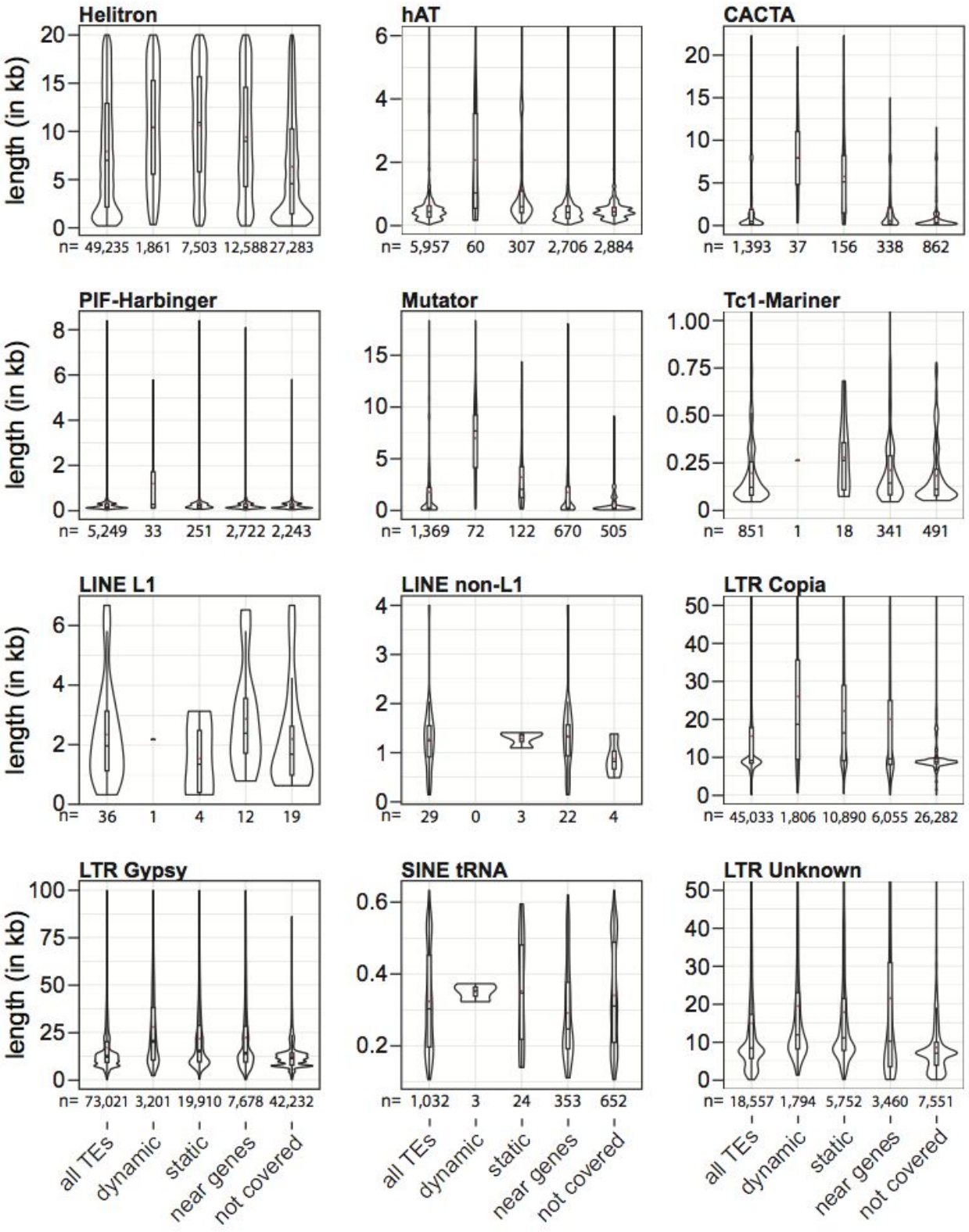


S1 Fig. Principal component analysis of gene and transposable element (TE) expression levels.

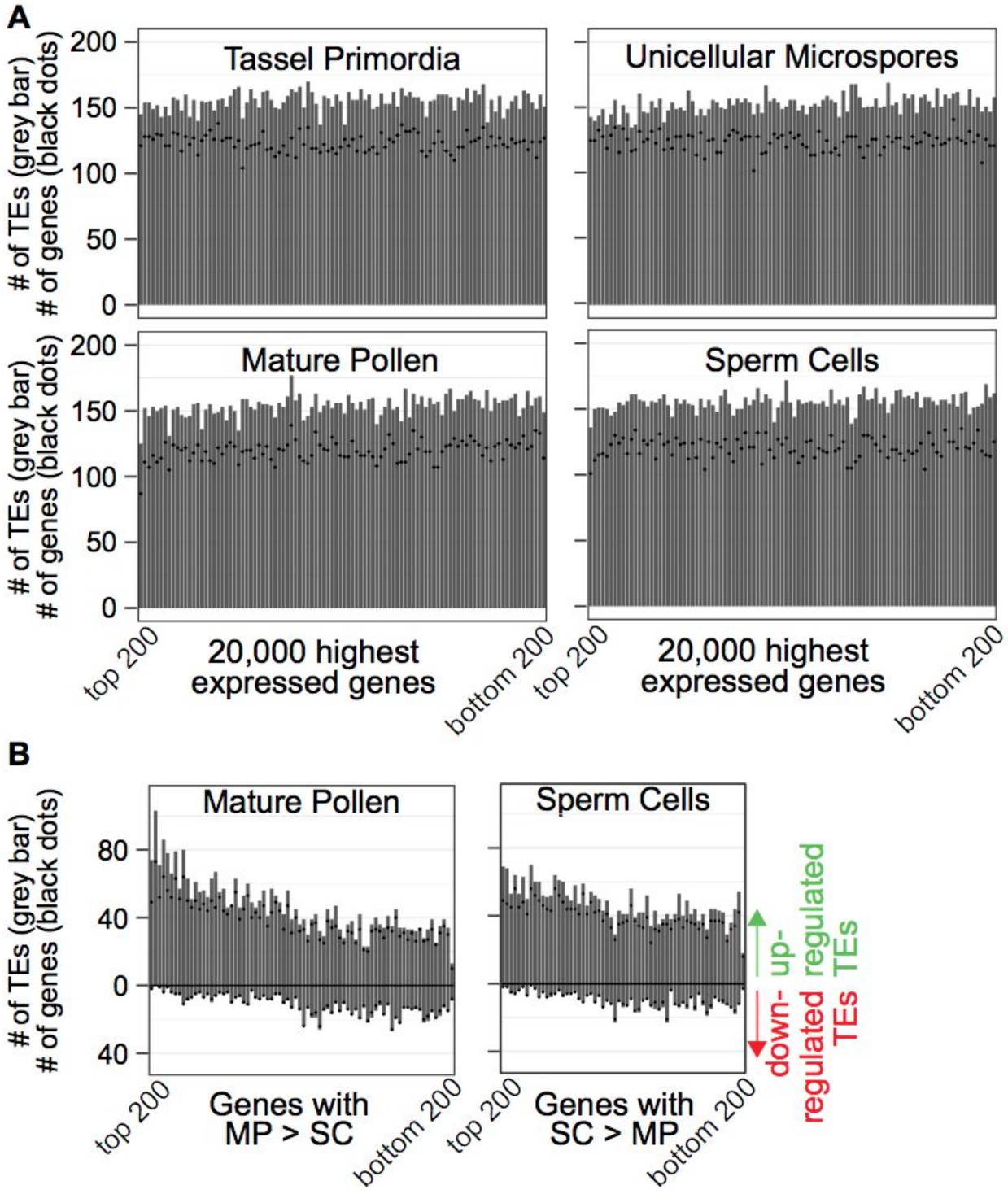
Two major components, on x- and y-axis, explain 89% of the variance in gene and TE expression levels. Asterisk (*) mark indicates the sample generated as part of this study, whereas other datasets are publicly available. For the sperm cells isolated in this study (SC), the TE expression of one biological replicate did not cluster with the other three (SC1), and therefore was removed from subsequent analyses of expression from TEs. MP-2014, SE, and OV are from [1]; MP-WEB is from [2]; LF is from [3]; MP-LM is from NCBI BioProject 306885 (2015); SC-TD is from [4].


A
Average TE accumulation
log10(normalized read count)
chosen as reference tissue
Bicellular Microspores (MS-B)*
Mature Pollen (MP-2014)
Mature Pollen (MP_LM)
Mature Pollen MP_WEB
Mature Pollen (MP)*
Sperm Cells (SC)*
Unicellular microspores (MS)*
Seedlings (SE)
V5/6 Leaves (LF)
Sperm Cells (SC_TD)
Ovules (OV)
Tassel Primordia (TP)*
B
Number of TEs < 1 normalized read count
#TEs (x 10,000)
Mature Pollen (MP)*
Unicellular microspores (MS)*
Ovules (OV)
Bicellular Microspores (MS-B)*
Mature Pollen (MP-2014)
V5/6 Leaves (LF)
Mature Pollen MP_WEB
Sperm Cells (SC)*
Mature Pollen (MP_LM)
Sperm Cells (SC_TD)
Tassel Primordia (TP)*
Seedlings (SE)

S2 Fig. Seedling tissue is the appropriate reference for comparison of TE activity.
(A) Steady-state mRNA accumulation of all TEs in different tissues. Datasets generated in this study are marked with an asterisk. **(B)** The number of TEs with zero or near-zero expression levels in different tissues. Seedlings (SE) have the most TEs with low expression levels.


Helitron
hAT
CACTA
PIF-Harbinger
Mutator
Tc1-Mariner
LINE L1
LINE non-L1
LTR Copia
LTR Gypsy
SINE tRNA
LTR Unknown
length (in kb)
n= 49,235 1,861 7,503 12,588 27,283
n= 5,957 60 307 2,706 2,884
n= 1,393 37 156 338 862
n= 5,249 33 251 2,722 2,243
n= 1,369 72 122 670 505
n= 851 1 18 341 491
n= 36 1 4 12 19
n= 29 0 3 22 4
n= 45,033 1,806 10,890 6,055 26,282
n= 73,021 3,201 19,910 7,678 42,232
n= 1,032 3 24 353 652
n= 18,557 1,794 5,752 3,460 7,551
all TEs
dynamic
static
near genes
not covered

S3 Fig. Length distribution of categorized TEs subdivided by superfamilies.
Length of TEs in the different TE categories from Figure 2A but further subcategorized by different superfamilies. The violin plots around the box show the kernel probability density of the data. The box represents lower and upper quartile, the line is the median, and the whiskers represent 10-90% range. Red asterisk denotes the mean. ‘n’ shows the number of TEs in each category for each superfamily.



S4 Fig. Abundance of TEs near genes in each tissue.

(A) For each tissue type, the top 20,000 expressed genes are distributed along the X-axis in bins of 200, with the highest expressed bin on the far left. The number of TEs near (<2kb) these genes is then counted on the Y-axis (shown in grey bar) and the number of genes with at least 1 TE within 2kb is displayed as black dots. **(B)** Genes filtered for either higher expression in pollen (MP) over sperm cells (SC) (left) or SC>MP (right) were used to determine if the association in

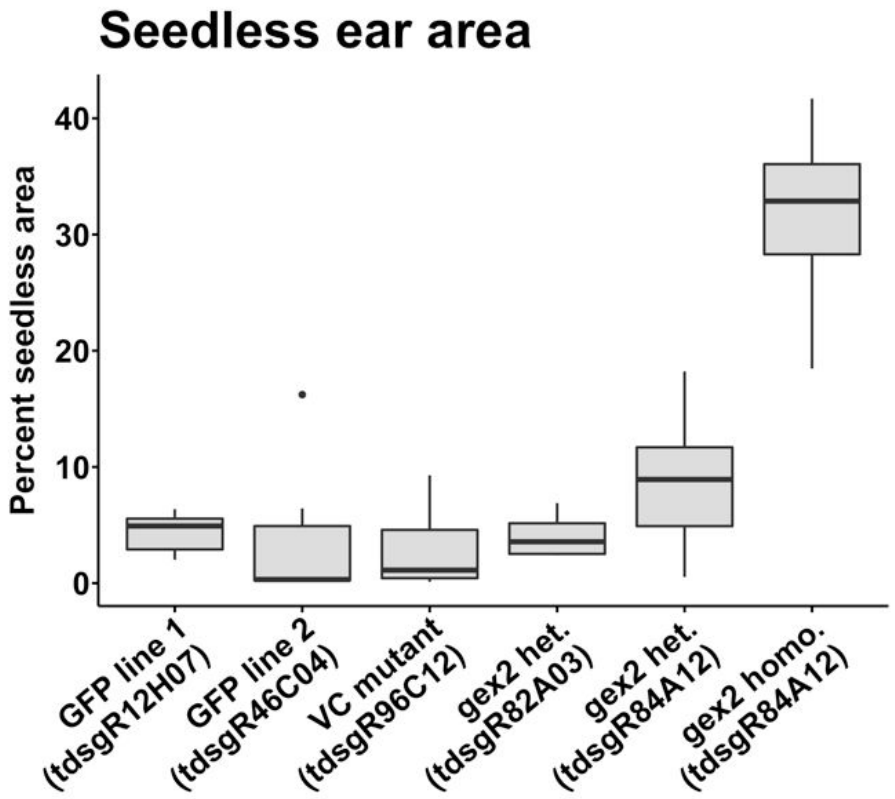
Figure 4 is due to sample contamination between SC and MP. Once genes were filtered, the top expressed genes in that tissue were distributed along the X-axis in bins of 200 based on their expression values, with the highest expressed bin on the far left. The number of up- and down-regulated TEs near (<2kb) these genes is then counted on the Y-axis (shown in grey bar) and the number of genes with at least 1 TE within 2kb is displayed as black dots.

B

Pollen parent genotype	Number of heterozygous small endosperm seeds (*tdsg*/+) by PCR	Number of wild-type small endosperm seeds (+/+) by PCR	% *tdsg*/+ heterozygotes in small seeds	% male transmission for *tdsg* allele (all outcrosses)
gex2-tdsgR82A03 /+	9	3	75%	33.4%
gex2-tdsgR84A12 /+	10	2	83%	23.1%
tdsgR46C04 /+ (control)	7	9	44%	50.5%

S5 Fig. *gex2* mutant pollen is associated with increased small and aborted seeds in outcross progeny.

(A) Seeds were removed from ears, arranged according to size, and counted. Images of representative seed populations are shown, with the top two rows in each image showing representative fully developed seeds. Rows below the top two contain all of the smaller or aborted seed from that particular ear. **(B)** PCR genotyping of small endosperm seeds from two independent crosses for the two *gex2* alleles show the majority of small seeds harbor the *gex2::Ds-GFP* allele, despite overall reduced transmission of the insertion alleles through the male. Small seeds from control *tsdgR46C04* crosses segregate in a Mendelian fashion.



S6 Fig. Characterization of gex2 seedless ear area.
Seedless area was quantified from scanned ear images for *gex2 Ds-GFP* alleles and *Ds-GFP* controls. Pollen from heterozygous *gex2* plants did not show significantly increased seedless area (*gex2-tdsgR82A03* pairwise t-test p-values relative to GFP line 1, GFP line 2, and VC mutant 0.95, 0.96, and 0.74, respectively; *gex2-tdsgR84A12* pairwise t-test p-values 0.19, 0.13, and 0.06, respectively), whereas pollen from homozygous *gex2-tdsgR84A12* plants had significantly increased seedless area (pairwise t-test against Ds-GFP controls separately, all p-value < 0.0001).

Chapter 4

A maize ear phenotyping system combines rotational scanning and deep learning to quantify fluorescent kernels

Cedar Warman, Justin Preece, Christopher M. Sullivan, Pankaj Jaiswal, John E. Fowler

Abstract

High-throughput phenotyping systems are powerful, dramatically changing our ability to document, measure, and detect biological phenomena. Here, we describe a cost-effective combination of a custom-built imaging platform and deep-learning-based computer vision pipeline. A minimal version of the maize ear scanner was built with low-cost and readily available parts. The scanner rotates a maize ear while a cellphone or digital camera captures a video of the surface of the ear. Videos are subsequently digitally flattened into two-dimensional ear projections. Segregating GFP and anthocyanin kernel phenotypes were clearly distinguishable in ear projections, and were manually annotated using ImageJ. Increased throughput was attained by designing and implementing an automated kernel counting system using transfer learning and a deep learning object detection model. The model was able to identify male-specific transmission defects across a wide range of mutant alleles, including a previously undescribed defect associated with mutation of Zm00001d002824, encoding a putative vacuolar processing enzyme. We show that by using this system, the quantification of transmission data can be accelerated and scaled to generate large datasets for robust analyses of phenotypes.

Introduction

High-throughput plant phenotyping is rapidly transforming crop improvement, disease management, and basic research (reviewed in (Fahlgren et al., 2015; Mahlein, 2016; Tardieu et al., 2017)). High-throughput phenotyping methods have been developed in several agricultural and model plant systems, including maize. There has been substantial progress towards deploying maize phenotyping systems, both in the private (Choudhury et al., 2016) and academic (Miller et al., 2017) realms. Many existing systems focus on phenotyping maize roots (Clark et al., 2013; Jiang et al., 2019) and above-ground vegetation (Chaivivatrakul et al., 2014; Junker et al., 2014; Choudhury et al., 2016; Zhang et al., 2017). Maize ears, with the kernels they carry, contain information about the plant and its progeny. Ears are easily stored, and do not require phenotyping equipment to be in place in the field or greenhouse at specific times during the growing season. Ears are a primary agricultural product of maize, which has led the majority of previous phenotyping efforts to focus on aspects of the ear that influence yield, such as ear size, row number, and kernel dimensions (Liang et al., 2016; Miller et al., 2017; Makanza et al., 2018). These studies have used techniques that varied from expensive and specialized

three-dimensional or line-scanning cameras (Liang et al., 2016; Wen et al., 2019) to relatively low-cost flatbed scanners and digital cameras (Miller et al., 2017; Makanza et al., 2018).

Beyond their agricultural importance, studying maize ears can answer fundamental questions about basic biology. The transmission of mutant genes can be easily tracked in maize kernels by taking advantage of a wide variety of visible endosperm markers (Neuffer et al., 1997; Li et al., 2013), which can be genetically linked to a mutant of interest (e.g. (Arthur et al., 2003; Phillips and Evans, 2011; Bai et al., 2016; Huang et al., 2017; Warman et al., 2020)). On the ear, kernels occur as an ordered array of progeny, which allows the transmission of mutant alleles to be tracked not only by total transmission for each individual cross, but within individual ears. Historically, transmission of markers has been quantified by manual counting. This approach has several limitations, among them a lack of a permanent record of the surface arrangement of kernels on the ear. The same disadvantages apply to most high-throughput kernel phenotyping methods, which generally rely on kernels being removed from the ear before scanning and do not typically include marker information.

Computer vision approaches to automated counting can improve throughput in phenotyping systems and improve the quality of data collected by including positional information for each kernel. One central challenge is successfully identifying which parts of an image contain the objects of interest and which parts contain the background, either through object detection (drawing a bounding box around the object) or segmentation (assigning each pixel in the image as "object" or "not object"). Previous systems have taken advantage of plant color or edges to algorithmically separate objects for quantification in some specialized contexts (Zhang et al., 2017; Makanza et al., 2018). These approaches can be computationally efficient, but are limited by variations in lighting conditions, image quality, and the distribution of objects in an image. Closely packed objects, such as kernels on a maize ear, can be difficult to separate using these methods, especially when the objects do not have consistent colors or clear edges.

Some of these obstacles to object detection can be overcome with deep learning approaches. These approaches have been applied to a variety of biological problems, and can show dramatic improvements over traditional methods (reviewed in (Angermueller et al., 2016; Ching et al., 2018)). Deep learning uses the fundamental concept of artificial neural networks, in which multiple nodes (sometimes referred to as neurons) are arranged in variously connected layers. Nodes have associated parameters that can be adjusted as the model is exposed to data. Data moves through the network from an input layer to at least one hidden layer, and finally to the output layer. Deep learning is characterized by a neural network with multiple

hidden layers, in which each layer describes features of the data being passed through the network (Ching et al., 2018). A subset of deep learning approaches called convolutional neural networks (CNNs) are particularly useful for image analysis. CNNs contain at least one convolutional layer, in which a filter moves (convolves) across an image to abstract information into the layer (Rawat and Wang, 2017). CNNs form the foundation of the object detection methods implemented in Tensorflow (e.g. Object Detection API, (Huang et al., 2016)) and Darknet (e.g. YOLO (Redmon and Farhadi, 2018)) that have seen widespread use across disciplines. Examples of such networks being used in biological contexts include plant disease detection (Mohanty et al., 2016), leaf quantification (Ubbens and Stavness, 2017), inflorescence movement tracking (Gibbs et al., 2019), and hypocotyl segmentation (Dobos et al., 2019).

Here we describe a novel maize ear phenotyping system and computer vision pipeline. The rotational ear scanner and image processing pipeline is a cost-effective method for high-throughput ear phenotyping. Flat projections of roughly cylindrical maize ears can be produced that provide a digital record of the surface of the ear, which can then be quantified in a variety of ways to track the locations and identities of kernel phenotypes, including marker genes. The design described here is built from easily acquired parts and a basic camera, making this approach accessible to most if not all labs. Projections can then be quantified for kernel identities and locations with a deep-learning-based computer vision pipeline implemented in Tensorflow, a free and open source framework. Finally, we use the system to analyze a large dataset of ears to assess mutant effects on transmission rate. We demonstrate substantial increases in phenotyping throughput, enabling more rapid biological discovery and more thorough quantitative analyses.

Results

Design and construction of the maize ear scanner

We designed a simple, custom-built scanner to efficiently phenotype maize ears (Maize Ear Scanner, MES). The MES rotates each ear 360º while a stationary camera records a video, which can then be processed into a cylindrical projection. Materials for constructing the scanner were limited to those that are widely available and affordable (Table 1). The frame of the scanner was built from dimensional lumber, with a movable mechanism built from drawer slides that enables a wide range of ear sizes to be accommodated (Figure 1A). A rotisserie motor

spins the ear at a constant speed while a USB camera or cell phone records a video of the rotating ear. The scanning process, including the insertion of the ear into the scanner and video capture, takes approximately 1 minute per ear.

We tested two configurations of the scanning system. In the first, a minimal configuration, a cell phone was used to capture movies of the rotating ear in full spectrum visible light (MES v1.0). This configuration cost less than $100 (Table 1), excluding the cost of the cell phone, and is capable of producing flat projections from a variety of ears in visible light. The second configuration uses a dedicated USB camera driven by a desktop computer (MES v2.0, Figure 1B). This configuration costs about $1,400 (Table 1), including a blue light and orange camera filter to capture GFP kernel markers present in a population of transgenic mutants (Li et al., 2013). The second configuration increases the scanner's efficiency by automating video processing, annotation, and distribution to cloud or local storage systems.

Processing videos into projections for manual quantification

The output of the scanner is a video of the rotating ear. This video can be directly quantified, but we found a ‘flat’ image projection most useful for visualizing the entire surface of the ear, as well as for quantifying the distribution of kernel phenotypes. To produce this projection with videos captured by an external camera or cell phone, videos were first uploaded to a local computer and annotated with identifying metadata. This process was streamlined in the second configuration of the scanner. In this configuration, videos were captured directly to the computer using the command line utility FFmpeg (version 3.4.6) to control a USB camera. Videos were automatically processed each night, with the resulting projections uploaded to cloud storage (Figure 2A).

Video processing consists of three steps (Figure 2B). In the first, FFmpeg is used to extract frames from the video into separate images. Next, images are cropped to the center horizontal row of pixels using the command line utility ImageMagick (version 6.9.7). Finally, all rows of pixels, one from each frame, are appended sequentially, resulting in the final image. Due to the scanner's consistent rotational speed, a fixed number of frames cover one complete rotation, resulting in no gaps or overlap in ear projections.

Images of a variety of maize ears representing several widely used kernel markers were captured using the scanner (Figure 3A). Both anthocyanin (*c1*, *a2*, and *pr1*) and fluorescent (*Ds-GFP*) kernel markers were clearly discernible in the final images, as well as the kernel morphology marker *brittle endosperm1* (*bt1*). Digital projections were manually quantified for

color and fluorescent kernel phenotypes using the FIJI distribution of ImageJ (Figure 3B). Using this approach, annotation of an entire ear could be completed in 5 to 10 minutes, depending on the size of the ear and the relative experience level of the annotator. In addition to producing total quantities of each kernel phenotype, manual annotations result in coordinates for each annotated kernel, which can be further analyzed if desired. Manual annotations of scanner images in ImageJ were compared to manually counting the kernels on the ear (Figure 3C). We observed a significant correlation between these two methods ($R^2 > 0.999$), validating the scanning method. To test the utility of the maize ear scanner, we scanned and manually counted over 400 ears with marker-linked mutations in >50 genes. With these methods, we were able to detect weak but significant transmission defects (~45% transmission of a marker-linked mutation) for a number of mutant alleles, using both anthocyanin and GFP kernel markers. Manually counted ear scanner image validation is described in detail in a previous study (Warman et al., 2020).

A traditional computer vision approach for automated discrimination of fluorescent and wild-type kernels

To increase throughput, we investigated computer vision methods to identify kernel locations and phenotypes from two-dimensional ear projections. We were most interested to establish a pipeline to automate counting of the *Ds-GFP* vs. wild-type non-fluorescent kernels, due to the broadly applicable use of this marker to quantify transmission rates for several thousand available mutations (Li et al., 2013; Warman et al., 2020). First, a traditional computer vision approach was assessed for its feasibility for quantification of images with GFP kernel markers. In this method, region-based segmentation of a two-dimensional ear projection was performed using a watershed transformation followed by morphological opening to segment individual kernels (Supplemental Figure 1). We found that extracting the blue channel of the RGB image for segmentation avoided inaccuracies resulting from varying intensities of kernel fluorescence in the green and red channels. After segmentation, segments were classified using k-means clustering into two groups for presence and absence of GFP. Fine-tuning of watershed parameters resulted in the accurate segmentation of individual images (Supplemental Figure 1A). However, because of variations in lighting, GFP intensity, kernel shape, and spacing on the ear, this method generalized poorly across a larger test dataset (Supplemental Figure 1B). This method was able to predict total fluorescent and non-fluorescent kernel numbers with some success (linear regression, adjusted $R^2 = 0.186, 0.205$, respectively), but failed to accurately

predict the percentage of kernels carrying the GFP marker (linear regression, adjusted R^2 = 0.000). Because marker-tagged mutants show Mendelian (50%) or near-Mendelian inheritance, accurate counts are required to measure abnormal inheritance with sufficient statistical significance.

Implementation of a deep learning model for automated kernel detection

To overcome the large variation in ear images, we turned to deep learning models, which are effective in detecting objects within heterogeneous images. Models using convolutional neural network architecture (CNNs) have dominated performance metrics in the computer vision field for several years (for example, Collection Of Common Objects (COCO) Object Detection Challenge [http://cocodataset.org/#detection-leaderboard], Open Images Object Detection Challenge [https://storage.googleapis.com/openimages/web/challenge2019.html#object_detection]). We used the TensorFlow library (Abadi et al., 2016) and Object Detection API (Huang et al., 2016) to implement a CNN-based model for our purposes. A tradeoff between speed and accuracy exists when implementing deep learning models for object detection. Some models, such as YOLO (Redmon and Farhadi, 2018) are significantly faster than the models available in the TensorFlow Object Detection API. However, because the image processing could be run independently of ear scanning, even several minutes of processing time per image would not present a bottleneck. For the pipeline, we chose to use the Faster R-CNN with Inception Resnet v2, with Atrous convolutions (Ren et al., 2015; Szegedy et al., 2016). This model was selected to balance speed and accuracy for our application based on its performance on the COCO dataset. (https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md).

CNNs require training data to generate effective models. To train the network, we generated a dataset of 300 scanned ear images with all kernels annotated with bounding boxes and marker classes, either fluorescent or non-fluorescent. Images were generated by scanning ears produced from heterozygous outcrosses of mutant alleles tagged with GFP fluorescent kernel markers, with 150 scanned ear images for each field season. The mean kernel number for training ear images was 349, resulting in >100,000 bounding boxes in the training set. We used a transfer learning approach because of the large amount of training data required to accurately train a neural network from scratch. Transfer learning takes advantage of a

well-trained network (in this case trained on the COCO dataset, >200,000 images with objects in 80 categories labeled with bounding boxes) to form the foundation for a new network optimized for a specific task. The weights of the inner layers of the network are updated based on the new training data, and the output layer is modified to reflect the new classes (fluorescent and non-fluorescent).

Our first attempts at training the network led to poor results (Figure 4A). Kernel bounding boxes were accurate in the top portion of test images, but these results failed to generalize across the entire image. Due to the large number of kernels on each image (over 600 on some ears), we suspected graphics processing unit (GPU) memory limitations may have caused incomplete annotations. Supporting this explanation, we gained incremental improvements by running the training and testing on a GPU with more memory (Nvidia V100 with 32 GB of memory versus an Nvidia M10 with 8 GB of memory) and a configuration that increased the number of initial bounding box proposals in the model.

One way to reduce the computational power necessary for a deep learning task is to subdivide the task into a series of simpler problems. In this case, we chose to subdivide each image into three sub-images, both for the training and for the testing of the model (Figure 4B). Images were subdivided vertically, with overlapping regions included between each division. After images were subdivided, the model was run on each sub-image individually. Bounding boxes near the vertical divisions were then removed to avoid partial bounding boxes for kernels that spanned two sub-images. Finally, annotations for the three sub-images were combined, and redundant bounding boxes in the overlapping areas were removed with non-maximum suppression, a process that resolves redundant bounding boxes by comparing their overlap and confidence scores. This method reduced the GPU memory required for inference, and resulted in accurate annotations across entire images.

Deep learning models trained on images from individual cameras improved detection of kernels and phenotypic classes

To test the deep learning models, we created a dataset of scanned ear images from the 2018 and 2019 field seasons, with 160 images from each season. Ears were generated from reciprocal outcrosses (heterozygous mutants crossed to wild-type lines both through the male and female), and were manually annotated with ImageJ to produce total fluorescent and non-fluorescent kernel counts for each ear. Testing set ear images were not used for training or validation of the model. Lines represented mutant alleles in a variety of genes highly expressed

in the maize male gametophyte (Warman et al., 2020). Kernels containing mutant alleles were marked with a GFP seed marker originating from *Ds-GFP* transposable element insertions. Projections generated across the two seasons represented a wide range of ears. Variations found in projections included differences in kernel size, shape, GFP intensity, and color. In addition, different cameras were used in each year, representing MES v1.0 and v2.0.

We first aimed to create a model with as much generalizability as possible, and thus included training images from both years. This first model, trained on images from two cameras, detected kernels in a test dataset with a moderate degree of accuracy (Supplemental Figure 2). We used adjusted R^2 values as the principal performance metric, comparing total fluorescent and non-fluorescent kernel counts between model predictions and manual annotations. The closer the R^2 value is to one, the closer the model predictions are to manual counts, indicating higher accuracy. The resulting percentage of fluorescence kernels was quantified with an adjusted R^2 of 0.930. To better visualize the model's accuracy, we calculated the mean absolute deviation in kernel count across the entire test dataset. The mean absolute deviation for fluorescent kernels was 5.87, while the mean absolute deviation for non-fluorescent kernels was 11.92. The mean absolute deviation for percent fluorescent kernel transmission was 1.85%.

A single model trained on a combined dataset from both years accurately identified kernels in scanned ear images. However, training separate models for each year substantially increased overall performance across a wide variety of images from both years of our test dataset (Figure 5). Individual models were robust to variations in kernel appearance, as well as to variations in ear size and kernel spacing. The models predicted total fluorescent and non-fluorescent kernels across the 2018 and 2019 test datasets with a high degree of accuracy (Figure 5B-C). The resulting transmission rate predictions were accurate across a wide range of inheritance values for both years (linear regression; adjusted R^2 = 0.984, 0.945, respectively). The mean absolute deviations for fluorescent kernels in individual models for 2018 and 2019 were 5.74 and 5.75, respectively, whereas the mean absolute deviations for non-fluorescent kernels were 8.58 and 6.81. The mean absolute deviations for percent fluorescent kernel transmission were 0.885% and 1.38%. Training individual models was substantially faster than training a single model (~100-fold faster training time on an Nvidia V100 GPU). Detailed metrics for model training can be found in the Materials and Methods section below. Because of their increased accuracy, we proceeded to use individual models for each camera/year to investigate transmission rates for *Ds-GFP* mutant alleles. While the variation introduced by using different

cameras for each year was likely responsible for the increased accuracy of individual models, we cannot rule out other potentially correlated factors in the two growing seasons.

Application of deep learning model to a large ear projection dataset

To test the deep learning models on a larger dataset, we quantified a set of 369 scanned ear images that had manually counted kernels from a previous study (Warman et al., 2020). The original dataset consisted of images of ears from maize plants grown during the 2018 field season. Ears were harvested from different plants with single *Ds-GFP* insertions in 44 genes. A total of 48 mutant alleles were examined, with four genes having two independent *Ds-GFP* insertions. Genes were selected because they are highly expressed in different developmental stages of the male gametophyte. Reciprocal outcrosses of heterozygous mutants were carried out to functionally validate sets of highly expressed genes by observing the transmission rates of mutant alleles. Observations of transmission rates showed several mutant alleles with reduced transmission through the male. We assessed the accuracy of the model's predictions by comparing transmission rates for manually counted images and model predictions (Supplemental Figure 3). For crosses through the female, the model predicted that mutations in all 44 genes had no significant difference from Mendelian (50%) inheritance, consistent with manual counts (Supplemental Figure 3A). For crosses through the male, the model successfully predicted 7/8 alleles that showed significant transmission defects when transmission was quantified manually, with the transmission of one of these alleles predicted as nonsignificant by the model (Supplemental Figure 3B). A generalized linear model showed no evidence of significant systematic differences between manual counts and model predictions (p-value > 0.8).

A second set of reciprocal crosses was carried out in the 2019 field season to increase the size of the ear image dataset. Crosses from the 2019 field season included plants with previously tested mutant alleles to determine whether transmission rates identified in 2018 remained consistent in the following year. Crosses also contained plants with additional alleles that were not included in the published analysis (Warman et al., 2020), either because of insufficient crosses in 2018 (5 alleles) or lack of PCR confirmation of *Ds* insertion location (7 alleles). In total, approximately 1000 ears from plants containing 60 mutant alleles were quantified using individual computer vision models for 2018 and 2019 field seasons. Combined 2018+2019 model estimates were largely aligned with 2018 manual counts for both male and female crosses (Figure 6). The data from the combined models correctly predicted no significant transmission defects through the female for 56/60 alleles in the combined dataset, whereas 4/60

alleles were assigned GFP transmission rates significantly increased over Mendelian inheritance (Figure 6A). These apparent false positives were likely the result of a systematic undercount of non-fluorescent kernels in a small subset of female crosses in the 2019 dataset (Supplemental Figure 4A-B). This is potentially due to the relatively strong GFP signal arising from doubled dosage of *Ds-GFP* in the endosperm, leading to reduced accuracy in the recognition of non-fluorescent kernels (Supplemental Figure 4C). The model correctly predicted all 8 alleles showing significant transmission defects as determined by 2018 manual counts (Figure 6B). In addition, the model predicted male transmission rates for the 12 alleles not present in the 2018 dataset, the majority of which (11/12) showed no evidence of non-Mendelian inheritance. However, the model identified a significant, previously undescribed, male-specific transmission defect associated with a *Ds* insertion in the maize gene Zm00001d002824 (Table 2). Zm00001d002824 codes for a predicted vacuolar-processing enzyme (VPE). VPEs have been shown to be involved in the maturation of vacuolar proteins as well as vacuolar-organized programmed cell death (Yamada et al., 2005), and its potential role in the male gametophyte is unexplored.

Discussion

Large amounts of information can be obtained from maize ears. Certain types of information, such as yield and kernel quality, have direct relevance for improving maize for agricultural purposes. Other types of information, such as mutant transmission, can be used to study fundamental biological processes. Our goal was to develop a methodology to capture some of this information via digital imaging and automated kernel detection. This methodology enables standardized, replicable measurement of kernel phenotype distribution, as well as provides a permanent digital record of ears. The scanner is fast and cost-effective in its minimal configuration (MES v1.0). A step-up from the minimal configuration (MES v2.0) enables a more automated system for file transfer and video-to-projection generation, and the addition of deep-learning-based kernel quantification of the resulting images dramatically scales up the amount of quantitative data that can be feasibly generated. In addition, automated quantification avoids variation introduced by multiple individuals manually quantifying images.

While the scanner provides useful phenotyping data, it has some notable limitations. Cylindrical projections are a convenient way of visualizing the entire surface of an ear in a single image. However, because maize ears are not perfect cylinders, the projections distort regions of

the ear that are not cylindrical, typically the top and bottom, resulting in kernels that appear larger than those in the middle of the ear (Figure 3A). Excessively curved ears, sometimes resulting from uneven pollination, can also lead to severe distortions. Because of these distortions, measuring qualities like kernel and ear dimensions can be challenging. While approximate values for these metrics can be calculated, in the future more precise measurements could potentially use the source video as input to model the ear in three dimensions, particularly with the addition of calibration objects (Feldmann et al., 2019). Thus, we have limited the use of the computer vision pipeline to relatively straight, uniform-thickness ears, whereas more distorted ear images can be quantified following manual annotation.

Capturing high-quality and standard images is crucial for the system. Differences in photography equipment, image quality, and variation in ear and kernel morphology can compromise accuracy. A small subset of kernels that were significantly outside the normal range of color variation were not identified by the model, particularly in images with non-optimal exposure (Supplemental Figure 4). These cases can likely be resolved with an improved imaging protocol and quality control, particularly of high contrast, strong GFP signal images. The overall impact of these weaknesses on the model's accuracy was low, due to their relative scarcity.

The computer vision pipeline is amenable to a wide range of kernel markers and phenotypes (for example, those found in Figure 3A). Additional phenotypes will require additional training data, but no technical limitations exist for the addition of kernel markers such as anthocyanin genes or kernel morphology markers like *bt1*. Agriculturally relevant aspects of maize ears, such as kernel size, row number, and other yield components, could also conceivably be measured with images generated by the ear scanner. The addition of calibration surfaces (such as a ruler or grid) to the scanner would likely improve the ability to estimate such ear or kernel dimensions.

Images of ears provide a convenient, long-lasting record of experiments, particularly if they are shared by researchers. These data can be used in a variety of ways, such as measuring patterns of kernel distribution, quantifying empty space on the ear, and recording other phenotypes such as abnormal or aborted kernels. For our experimental objectives, the system made it feasible to generate a nearly two-fold larger set of fluorescent kernel transmission data compared to our initial study (Warman et al. 2020). Manual quantification of the original dataset took approximately 50 hours. Automated quantification of the larger dataset took less than four hours when run on multiple GPUs, representing an approximately 25-fold

decrease in the time required to quantify the images. Not only did the larger dataset confirm the observations in that study, it also enabled the identification of a new male-specific gametophytic mutant, pointing toward a previously unknown function for a vacuolar processing enzyme. The ear scanning and computer vision system increases the scope of feasible experiments addressing maize reproductive biology and related agricultural traits by addressing a bottleneck in data acquisition and quantification, paving the way for high-throughput phenotyping in this area.

Materials and Methods

Building the maize ear scanner

The maize ear scanner was built from dimensional lumber and widely available parts. For detailed plans and three-dimensional models, see Supplemental File 1. The base of the scanner was built from a nominal 2x12 (38x286 mm) fir board, while the frame of the scanner was built from nominal 2x2 (38x38 mm) cedar boards. Boards were fastened together with screws. Strict adherence to materials and exact dimensions of the scanner frame is not necessary, as long as the scanner is structurally sound and large enough to accommodate ears of varying sizes.

A standard rotisserie motor (Minostar universal grill electric replacement rotisserie motor, 120 volt 4 watt), used to rotate the maize ear, was attached to the base of the scanner by way of a wood enclosure. A 5/16" (8 mm) steel rod was placed in the rotisserie motor to provide a point to fasten the lower portion of the ear. The top of the steel rod was ground to a flattened point with a bench grinder to allow it to be inserted into the pith at the center of the base of the ear.

The top of the ear was held in place with an adjustable assembly constructed from a nominal 2x4 board (38x89 mm) fastened to drawer slides (Liberty D80618C-ZP-W 18-inch ball bearing drawer slides) on either side of the scanner frame (Supplemental File 1). In the center of the 2x4, facing down towards the top of the ear, is a steel pin mounted on a pillow block bearing (Letool 12mm mounted housing self-aligning pillow flange block bearing). The steel pin (12mm) was sharpened to a point to penetrate the top of the ear as it is lowered, temporarily holding it in place while the ear is rotated during scanning. Because the pin can be moved up and down on the drawer slides, a variety of ear sizes can be accommodated in the scanner.

Ambient lighting was used for full spectrum visible light images. To capture GFP fluorescence, a blue light (Clare Chemical HL34T) was used to illuminate the ear. An orange filter (Neewer camera flash color gel kit) was placed in front of the camera lens to partially filter out non-GFP wavelengths.

Ear scanning workflow

Preparation for the scanning process begins by trimming the top and bottom of the ear to expose the central pith. Before scanning, ear dimensions (length and diameter at widest point) were recorded. Following measurement, the bottom pin is inserted into the bottom of the ear, after which the pin with ear attached is placed in the rotisserie motor. The top of the ear is secured by lowering the top pin into the pith at the top of the ear.

Ear scanning was divided into two configurations. In the first configuration (MES v1.0), a camera capable of capturing videos (such as a cell phone or point-and-shoot digital camera, a Sony DSCWX220 was used in our version) was mounted on a tripod approximately 60 cm in front of the rotating ear. Videos were captured by the camera, then manually imported to a computer for processing and downstream analysis. In the second configuration (MES v2.0), a USB camera (ELP USBFHD06H-SFV) capable of capturing 1080p resolution video at 30 fps is directly controlled by a desktop computer (Dell 3630) running the Ubuntu Linux distribution (version 18.04.3). The camera was placed approximately 60 cm in front of the ear for video capture. Videos are previewed using the command line utility qv4l2 (V4L2 Test Bench, version 1.10.0), then captured using a custom FFmpeg command (ffmpeg -t 27 -f v4l2 -framerate 30 -video_size 1920x1080 -i /dev/video1 /output.mov). The command captures the number of frames required for one complete rotation of the ear plus a small initial buffer. Videos are processed into flat images each night by running a custom script (see following section) with the Linux cron utility. After video processing, flat images are uploaded to Google cloud and local server space using the rclone (version 1.50.2) and rsync (version 3.1.2) utilities, respectively. A detailed protocol for scanning ears with the maize ear scanner using ears with GFP kernel markers and an ELP USBFHD06H-SFV USB camera can be found in Supplemental File 2.

Creating flat images

Videos were processed to flat images. Frames were first extracted from videos to png formatted images using FFmpeg with default options (ffmpeg -i ./"$file" -threads 4 ./maize_processing_folder/output_%04d.png). These images were then cropped to the central

row of pixels using ImageMagick (mogrify -verbose -crop 1920x1+0+540 +repage ./maize_processing_folder/*.png). The collection of single pixel row images was then appended in sequential order (convert -verbose -append +repage ./maize_processing_folder/*.png ./"$name.png"). Finally, the image was rotated and cropped (mogrify -rotate "180" +repage ./"$name.png"; mogrify -crop 1920x746+0+40 +repage ./"$name.png"). We chose the convention of a horizontal flattened image with the top of the ear to the right and the bottom of the ear to the left. Because the videos were captured vertically, a rotation was required after appending the individual frames. The vertical dimension of the final crop reflects the number of frames (746) required for one full rotation of the ear. A GitHub repository containing the script used to create flat images from videos is located at https://github.com/fowler-lab-osu/flatten_all_videos_in_pwd.

Manually quantifying kernels using flat images

Kernels were quantified from ear projections using the Cell Counter plugin of the FIJI distribution of ImageJ (version 2.0.0) (Schindelin et al., 2012). Ears were assigned counter types to correspond to different kernel markers, after which kernels on ear images were located and annotated manually. The Cell Counter plugin exports results in an xml file, which contains the coordinates and marker type of every annotated kernel. This file can be processed to create a map of kernel locations on the ear. A detailed protocol describing the quantification process can be found in Supplemental File 3.

Image segmentation and labeling by watershed transformation and k-means clustering

Two-dimensional projections of images containing GFP kernel markers were segmented using a watershed transformation implemented in the scikit-image Python library, version 0.16.1. The tutorials located at (https://scikit-image.org/docs/stable/auto_examples/applications/plot_coins_segmentation.html) and (https://scikit-image.org/docs/dev/auto_examples/color_exposure/plot_regional_maxima.html) were used as starting points. Images were first cropped by 15% along each side to remove distorted regions along the top and bottom of the ear. Next, regional maxima were isolated from the images using the scikit-image "reconstruction" function with the original image minus a fixed

h-value of 0.3 as the seed image. The resulting h-dome regional maxima were further processed using the Sobel operator (scikit-image "sobel" function). Extreme high values of the resulting image's histogram were used as seeds for the scikit-image "watershed" function. Finally, connections between adjacent kernel segments were reduced by morphological opening using the "binary_opening" scikit-image function.

Once segments identifying potential kernels were identified, segments were classified into either "fluorescent" or "non-fluorescent" categories. First, segment centers were identified using the "center_of_mass" function from the SciPy Multi-dimensional image processing package (version 1.4.1). Mean intensity in red, green, and blue channels was then calculated for each segment. Segments were divided into two clusters by channel intensity by k-means clustering using the "kmeans" function from the scikit-learn library (version 0.22.2). Clusters were collectively identified as the "fluorescent" or "non-fluorescent" cluster based on their relative mean segment intensity in the green channel. A python script used to run the previously described process can be found at https://github.com/cedarwarman/maize_cv. Fluorescent, non-fluorescent, and percent fluorescent metrics were calculated using this method for 320 images in the test dataset described in the following section. Adjusted R^2 values were calculated using a linear regression for each metric (see https://github.com/cedarwarman/maize_scanner_cv_paper).

Training, validation, and test dataset generation

Training and validation datasets were generated from 300 scanned ear images from the 2018 and 2019 field seasons (150 images each season), using a 0.7 training to validation ratio. Lines contained a selection of single mutations from the Dooner/Du collection of *Ds-GFP*-tagged transposable element insertions (Li et al., 2013). Kernels were manually annotated with bounding boxes and classes (fluorescent or non-fluorescent) using LabelImg (https://github.com/tzutalin/labelImg).

A test dataset was generated using 320 scanned images of *Ds-GFP*-tagged ears from the 2018 and 2019 field seasons (160 images each season). A Sony DSCWX220 camera was used to capture images in 2018 and an ELP USBFHD06H-SFV was used to capture images in 2019. Images used for training and validation of the model were excluded from the test dataset. Total fluorescent and non-fluorescent kernels were quantified using ImageJ as previously described (see section "Manually quantifying kernels using flat images").

Deep learning model selection and configuration

The deep learning pipeline used the Faster R-CNN with Inception Resnet v2 with Atrous convolutions model, implemented in the Tensorflow Object Detection API. To preserve GPU memory, images were resized to maximum dimensions of 600x1024 pixels. Training data was split into training and validation sets with a ratio of training to validation of 0.7. First stage RPN anchor proposals were limited to 3000, with 8 aspect ratios at each anchor point. Max total detections were set at 2000. Data augmentations were limited to a random horizontal flip.

Models were created using two approaches. In the first, a single model was trained with data from combined 2018 and 2019 field seasons. Separate cameras were used for each season (see description in the previous section). This model was trained for 940 epochs on an Nvidia V100 GPU, a process that took approximately 74 hours. The training length was determined by optimizing mAP (mean average precision) at 0.5 IOU (intersection over union). This parameter measures the average precision (true positives divided by the sum of true positives and false positives) over a range of recall values (true positives divided by the sum of true positives and false negatives). This metric summarizes the model's performance at correctly identifying bounding boxes and classes while minimizing false positives. At epoch 940, the model's mAP at 0.5 IOU was 0.790, with an average recall of 0.622.

In the second approach, two models were trained independently on 2018 and 2019 images. These models were trained for 2226 and 2468 epochs, respectively, for approximately 45 minutes on an Nvidia V100 GPU. Training lengths were optimized as with the single model described previously. The 2018 and 2019 mAP at 0.5 IOU were approximately 0.843 and 0.867, respectively, with average recalls of 0.601 and 0.693.

Image subdivision and bounding box confidence scores

Before training, images and annotations were divided into three sub-images using a custom script. For inference, input images were likewise divided into three sub-images. In both cases, images were divided along the horizontal axis (Figure 4B). Overlapping regions of 100 pixels in width (all pixel measurements based on non-scaled input images, generally 1920x746 pixels) were included in the left and right sub-images. Because empty margins on the left and right of the original image generally led to the center sub-image having the largest number of kernels, only the left and right sub-images included 100 pixels overlapping regions.

Inference was first run on each sub-image individually. Next, bounding boxes within 40 pixels of subdivision borders were removed. This process removed partial bounding boxes of

kernels located along the dividing lines between images. Because of image overlap, these kernels were still marked by complete bounding boxes after partial bounding boxes were removed. After this step, bounding boxes and annotations from the three sub-images were combined. Redundant bounding boxes in overlapping regions were removed by non-maximum suppression using the Tensorflow function "non_max_suppression". Non-maximum suppression calculates the intersection over union (IOU) value for all bounding box pairs. For pairs that exceed a defined intersection over union (IOU) value, in our case 0.5, the bounding box with the lowest confidence score is removed. Inference for each input image took approximately one minute on a Nvidia M10 GPU, with individual models performing slightly faster than the single model.

Optimal confidence score thresholds for final bounding box outputs were determined empirically by maximizing the R-squared value for total fluorescent and non-fluorescent kernel count across the test image dataset. R-squared values for confidence thresholds ranging from 0-1 in 0.01 increments were calculated for both fluorescent and non-fluorescent total kernel counts by comparing model predictions and manually validated data (Supplemental Figure 5). A single confidence threshold of 0.12 was chosen for the combined 2018/2019 model to maximize the combined R-squared value in both classes (Supplemental Figure 5A). Confidence thresholds of 0.08 and 0.12 were chosen for 2018 and 2019 individual models (Supplemental Figure 5B-C).

Statistical methods for deep learning model application to test dataset

Manually counted kernel totals were compared with deep learning model predictions for the 320 test images by fitting a linear regression using the "lm" function in R (see https://github.com/cedarwarman/maize_scanner_cv_paper). Adjusted R^2 values were calculated for fluorescent and non-fluorescent kernels, as well as for percent fluorescent kernel transmission. Mean absolute deviations were calculated for fluorescent and non-fluorescent total kernel counts, and percent fluorescent kernel transmission. Analysis was carried out using both a single model trained on 2018 and 2019 images, as well as individual models trained on each year alone.

Experimental design and statistical methods for deep learning model application to 2018 and 2019 field trials

Inference was run on 983 scanned images from the 2018 (369 images) and 2019 (614 images) field seasons (Supplemental Table 1). Scanned ear images were the result of reciprocal outcrosses of heterozygous plants carrying GFP-tagged *Ds* insertion alleles in a variety of genes highly expressed across maize gametophyte development. For detailed experimental description, see (Warman et al., 2020). In brief, alleles were chosen from highly expressed genes (top 20% by FPKM) in three categories: Vegetative Cell, Sperm Cell, and Seedling as a sporophytic control. A total of 56 alleles were quantified in (Warman et al., 2020), of which 48 displayed fluorescent seed markers and were analyzed in this study. Eight alleles were associated with anthocyanin seed markers, and were thus not included in this analysis. Ear images from the 2019 field season contained additional crosses from the alleles present in the 2018 field season, plus 12 additional alleles (summarized in Table 2), for a total of 60 alleles.

After model inference, total fluorescent and non-fluorescent seed counts were analyzed using a generalized linear model with a logit link function for binomial counts and a quasi-binomial family to correct for overdispersion between parent lines. Significant differences from expected 50% inheritance were assessed with a quasi-likelihood test with p-values corrected for multiple testing using the Benjamini-Hochberg procedure to control the false discovery rate at 0.05. Significant differences from 50% inheritance were defined with an adjusted p-value < 0.05. Separate generalized linear models were carried out for each year and cross category (female, Seedling male, Vegetative Cell male, Sperm Cell male). A combined generalized linear model with all 2018 manual counts and all 2018 computer vision predictions was also created in order to determine the significance of manual counts versus computer vision predictions as a factor. For a detailed description of statistical methods, see https://github.com/cedarwarman/maize_scanner_cv_paper.

Acknowledgements

We thank O. Childress, H. Fowler, B. Galardi, B. Hamilton, R. Hartman, K. Kress, C. Lambert, and M. Wesel for their annotation and kernel counting assistance. In addition, we thank O. Childress for protocol feedback and E. Vischulis for technical assistance. This work

was funded by National Science Foundation grants MCB-1832186 (John E. Fowler PI) and IOS-1340112 (Pankaj Jaiswal PI), as well as the Department of Botany & Plant Pathology at Oregon State University.

Works Cited

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. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). pp 265–283

Angermueller C, Pärnamaa T, Parts L, Stegle O (2016) Deep learning for computational biology. Mol Syst Biol **12**: 878

Arthur KM, Vejlupkova Z, Meeley RB, Fowler JE (2003) Maize ROP2 GTPase provides a competitive advantage to the male gametophyte. Genetics **165**: 2137–2151

Bai F, Daliberti M, Bagadion A, Xu M, Li Y, Baier J, Tseung C-W, Evans MMS, Settles AM (2016) Parent-of-Origin-Effect rough endosperm Mutants in Maize. Genetics **204**: 221–231

Chaivivatrakul S, Tang L, Dailey MN, Nakarmi AD (2014) Automatic morphological trait characterization for corn plants via 3D holographic reconstruction. Comput Electron Agric **109**: 109–123

Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow P-M, Zietz M, Hoffman MM, et al (2018) Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. doi: 10.1098/rsif.2017.0387

Choudhury SD, Stoerger V, Samal A, Schnable JC, Liang Z, Yu J-G (2016) Automated vegetative stage phenotyping analysis of maize plants using visible light images. KDD workshop on data science for food, energy and water, San Francisco, California, USA

Clark RT, Famoso AN, Zhao K, Shaff JE, Craft EJ, Bustamante CD, McCouch SR, Aneshansley DJ, Kochian LV (2013) High-throughput two-dimensional root system phenotyping platform facilitates genetic analysis of root growth and development. Plant Cell Environ **36**: 454–466

Dobos O, Horvath P, Nagy F, Danka T, Viczián A (2019) A Deep Learning-Based Approach for High-Throughput Hypocotyl Phenotyping. Plant Physiol **181**: 1415–1424

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

Feldmann M, Tabb A, Knapp SJ (2019) Cost-effective, high-throughput 3D reconstruction method for fruit phenotyping. Computer Vision Problems in Plant Phenotyping (CVPPP) 1

Gibbs JA, Burgess AJ, Pound MP, Pridmore TP, Murchie EH (2019) Recovering

Wind-Induced Plant Motion in Dense Field Environments via Deep Learning and Multiple Object Tracking. Plant Physiol **181**: 28–42

Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z, Song Y, Guadarrama S, et al (2016) Speed/accuracy trade-offs for modern convolutional object detectors. arXiv [cs.CV]

Huang JT, Wang Q, Park W, Feng Y, Kumar D, Meeley R, Dooner HK (2017) Competitive Ability of Maize Pollen Grains Requires Paralogous Serine Threonine Protein Kinases STK1 and STK2. Genetics **207**: 1361–1370

Jiang N, Floro E, Bray AL, Laws B, Duncan KE, Topp CN (2019) Three-Dimensional Time-Lapse Analysis Reveals Multiscale Relationships in Maize Root Systems with Contrasting Architectures. Plant Cell **31**: 1708–1722

Junker A, Muraya MM, Weigelt-Fischer K, Arana-Ceballos F, Klukas C, Melchinger AE, Meyer RC, Riewe D, Altmann T (2014) Optimizing experimental procedures for quantitative evaluation of crop plant performance in high throughput phenotyping systems. Front Plant Sci **5**: 770

Liang X, Wang K, Huang C, Zhang X, Yan J, Yang W (2016) A high-throughput maize kernel traits scorer based on line-scan imaging. Measurement **90**: 453–460

Li Y, Segal G, Wang Q, Dooner HK (2013) Gene Tagging with Engineered Ds Elements in Maize. *In* T Peterson, ed, Plant Transposable Elements: Methods and Protocols. Humana Press, Totowa, NJ, pp 83–99

Mahlein A-K (2016) Plant Disease Detection by Imaging Sensors - Parallels and Specific Demands for Precision Agriculture and Plant Phenotyping. Plant Dis **100**: 241–251

Makanza R, Zaman-Allah M, Cairns JE, Eyre J, Burgueño J, Pacheco Á, Diepenbrock C, Magorokosho C, Tarekegne A, Olsen M, et al (2018) High-throughput method for ear phenotyping and kernel weight estimation in maize using ear digital imaging. Plant Methods **14**: 49

Miller ND, Haase NJ, Lee J, Kaeppler SM, de Leon N, Spalding EP (2017) A robust, high-throughput method for computing maize ear, cob, and kernel attributes automatically from images. Plant J **89**: 169–178

Mohanty SP, Hughes DP, Salathé M (2016) Using Deep Learning for Image-Based Plant Disease Detection. Front Plant Sci **7**: 1419

Neuffer MG, Coe EH, Wessler SR (1997) Mutants of maize. Cold Spring Harbor Laboratory Press

Phillips AR, Evans MMS (2011) Analysis of stunter1, a maize mutant with reduced gametophyte size and maternal effects on seed development. Genetics **187**: 1085–1097

Rawat W, Wang Z (2017) Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Comput **29**: 2352–2449

Redmon J, Farhadi A (2018) YOLOv3: An Incremental Improvement. arXiv [cs.CV]

Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *In* C Cortes, ND Lawrence, DD Lee, M Sugiyama, R Garnett, eds, Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp 91–99

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, et al (2012) Fiji: an open-source platform for biological-image analysis. Nat Methods **9**: 676–682

Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2016) Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv [cs.CV]

Tardieu F, Cabrera-Bosquet L, Pridmore T, Bennett M (2017) Plant Phenomics, From Sensors to Knowledge. Curr Biol **27**: R770–R783

Ubbens JR, Stavness I (2017) Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks. Front Plant Sci **8**: 1190

Warman C, Panda K, Vejlupkova Z, Hokin S, Unger-Wallace E, Cole RA, Chettoor AM, Jiang D, Vollbrecht E, Evans MMS, et al (2020) High expression in maize pollen correlates with genetic contributions to pollen fitness as well as with coordinated transcription from neighboring transposable elements. PLoS Genet **16**: e1008462

Wen W, Guo X, Lu X, Wang Y, Yu Z (2019) Multi-scale 3D Data Acquisition of Maize. Computer and Computing Technologies in Agriculture XI. Springer International Publishing, pp 108–115

Yamada K, Shimada T, Nishimura M, Hara-Nishimura I (2005) A VPE family supporting various vacuolar functions in plants. Physiol Plant **123**: 369–375

Zhang X, Huang C, Wu D, Qiao F, Li W, Duan L, Wang K, Xiao Y, Chen G, Liu Q, et al (2017) High-Throughput Phenotyping and QTL Mapping Reveals the Genetic Architecture of Maize Plant Growth. Plant Physiol **173**: 1554–1564

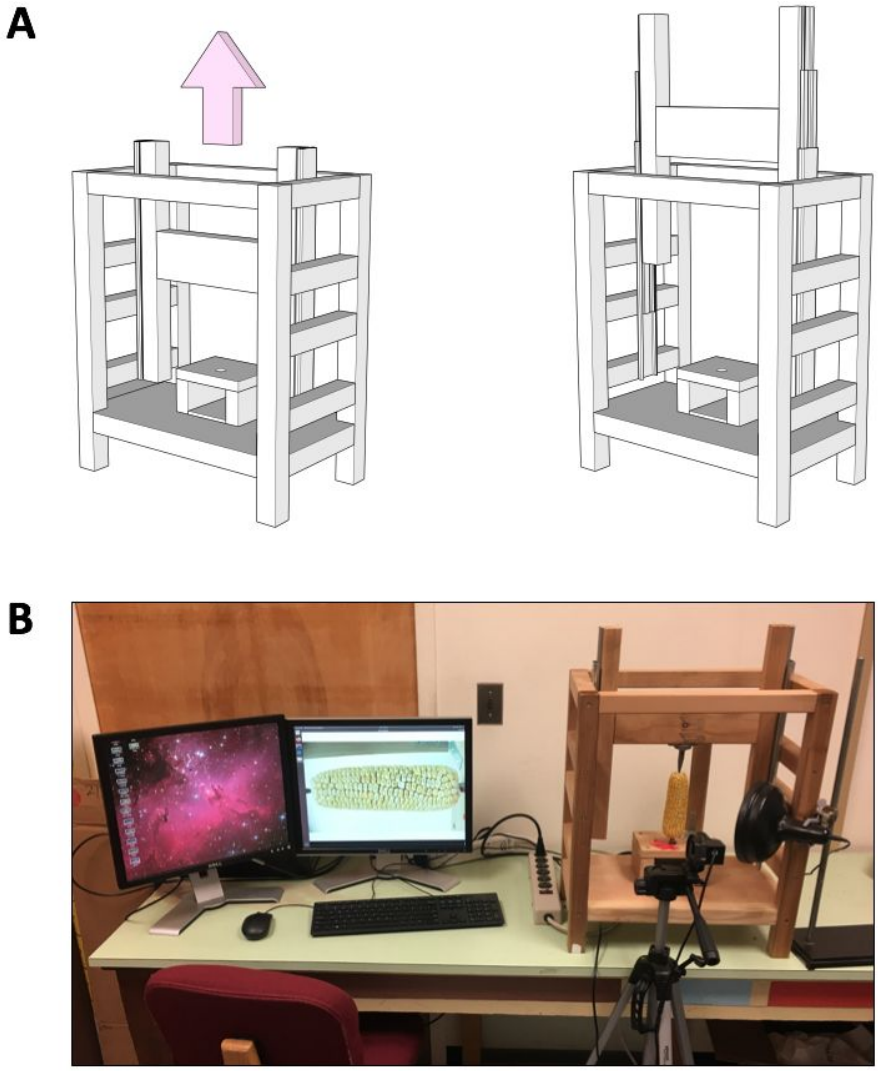


Figure 1. Efficient, cost-effective maize ear phenotyping with rotational scanner.

(A) Schematics of maize ear scanner in closed position (left) and open position (right). Full construction diagrams are available in Supplemental File 1.
(B) Image of scanner with ear in place. A dedicated USB camera is positioned in front of the ear as shown, with the ear centered in the frame. A video is captured as the ear spins through one full rotation, which is then processed to project the surface of the ear onto a single flat image. An optional blue light source for GFP imaging is shown on the right side of the scanner.

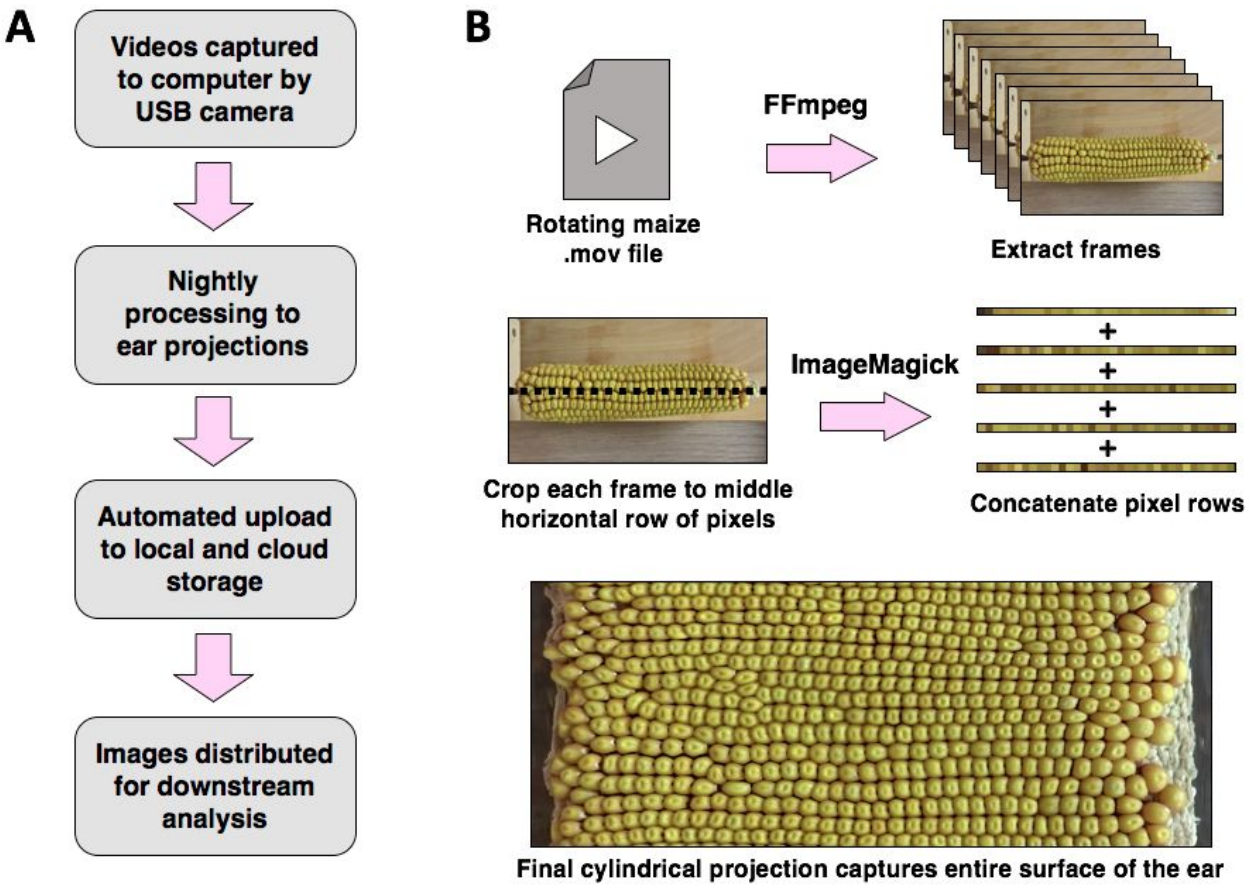


Figure 2. Processing videos into flat ear projections.

(A) Video annotation and processing workflow. **(B)** Processing videos to flat ear projections. The process of generating a projection from a video begins by extracting individual frames using FFmpeg. After frames are extracted, each frame is cropped to the middle horizontal row of pixels using the command line utility ImageMagick. The resulting collection of pixel rows, one per frame, is then concatenated into a single image depicting the entire surface of the ear.

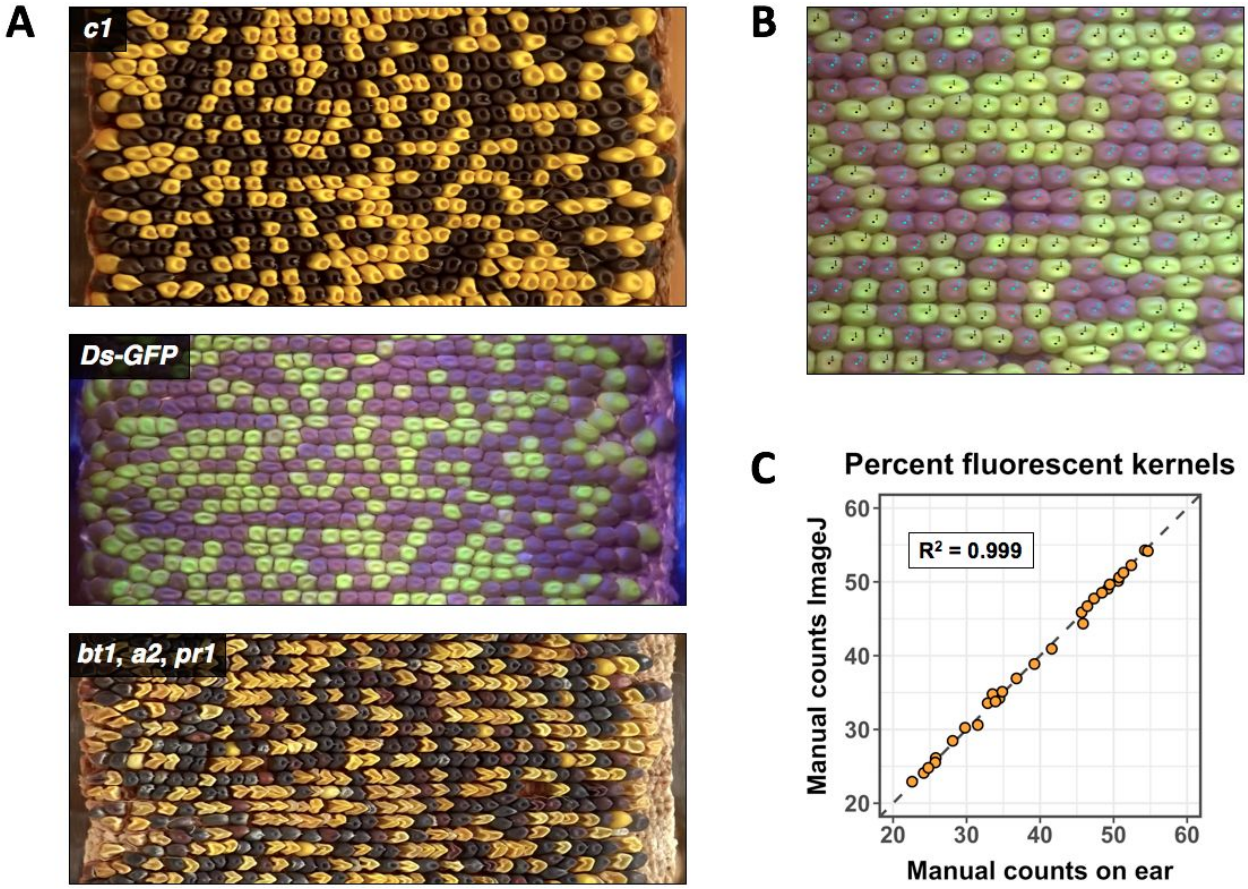


Figure 3. Examples of ear surface projections and manual quantification using ImageJ.

(A) Representative ear projections for several widely-used maize kernel markers. From top to bottom: anthocyanin gene *c1*; GFP fluorescent kernel marker *Ds-GFP*; anthocyanin and kernel morphology markers *bt1*, *a2*, and *pr1.*
(B) Representative example of manual annotation of fluorescent maize kernels. An ear of maize segregating for fluorescence was imaged. Fluorescent (black dots) and non-fluorescent (blue dots) kernels were manually identified using the ImageJ Cell Counter plugin.
(C) Comparison of manually counting kernels on ears vs manually counting kernels from ear projections using ImageJ. Fluorescent and non-fluorescent kernels were counted, with the percentage of kernels showing fluorescence shown here. The diagonal dashed line indicates equal values for manual counts on the ear and manual counts using ImageJ.

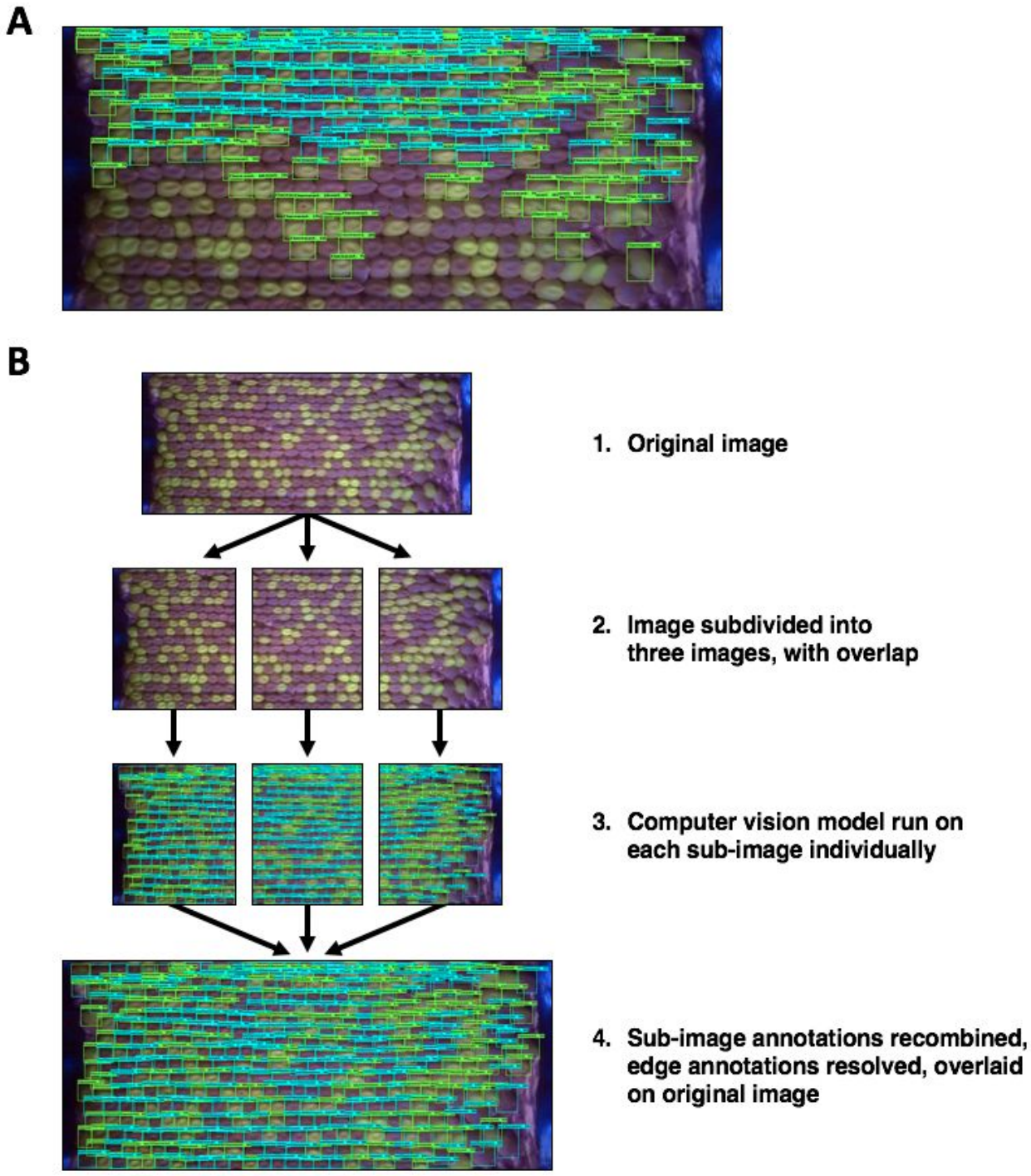
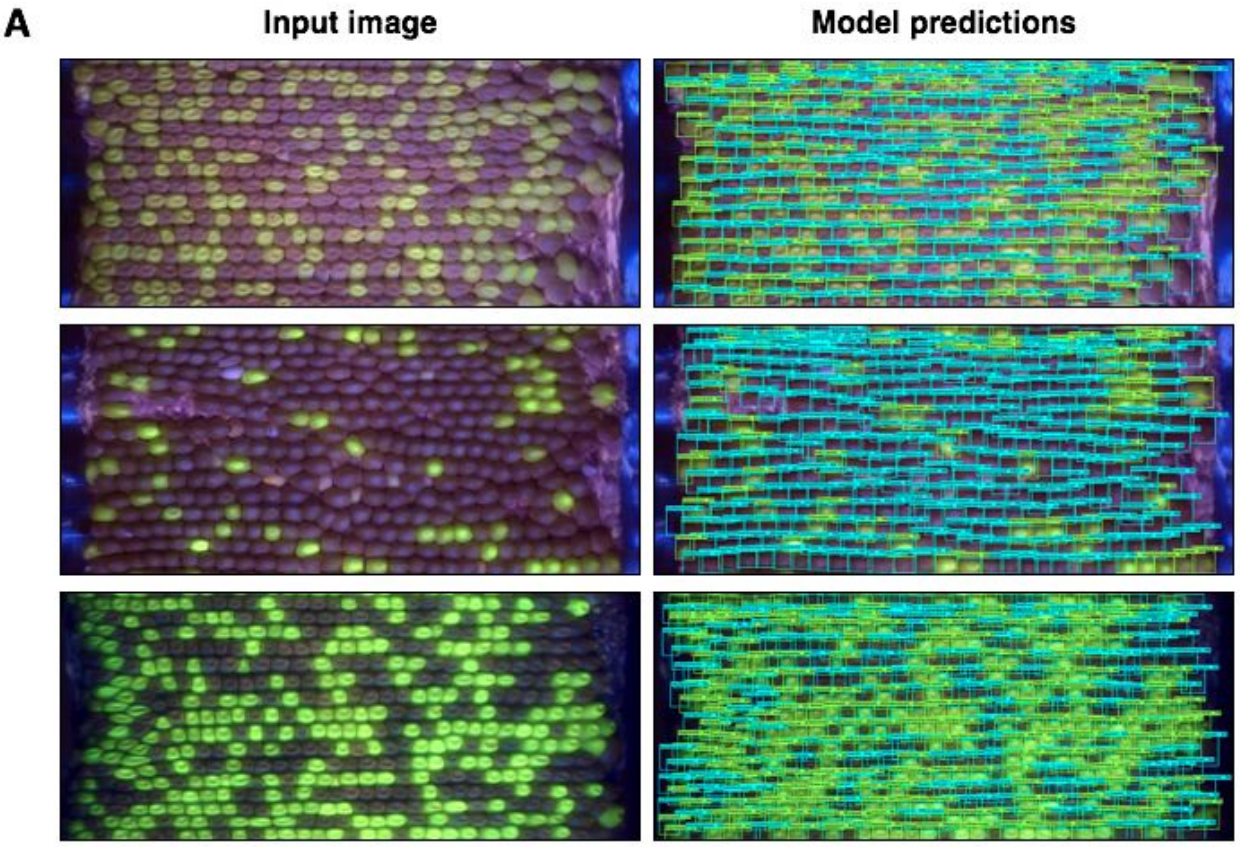


Figure 4. Workflow for subdividing images during model training and inference

(A) Representative image of initial object detection attempts showing incomplete bounding box annotations. Annotations were biased towards the top of the image, and failed to identify the majority of kernels.

(B) Image subdivision workflow. Images were first subdivided into three smaller images, with overlap between the images. The computer vision model was then run on each sub-image individually. Bounding boxes near the vertical borders between sub-images were removed to avoid split bounding boxes on single kernels. Annotations were recombined, then redundant boxes in overlapping sections were removed with non-maximum suppression. Finally, completed annotations were overlaid on the original image.


A
Input image
Model predictions

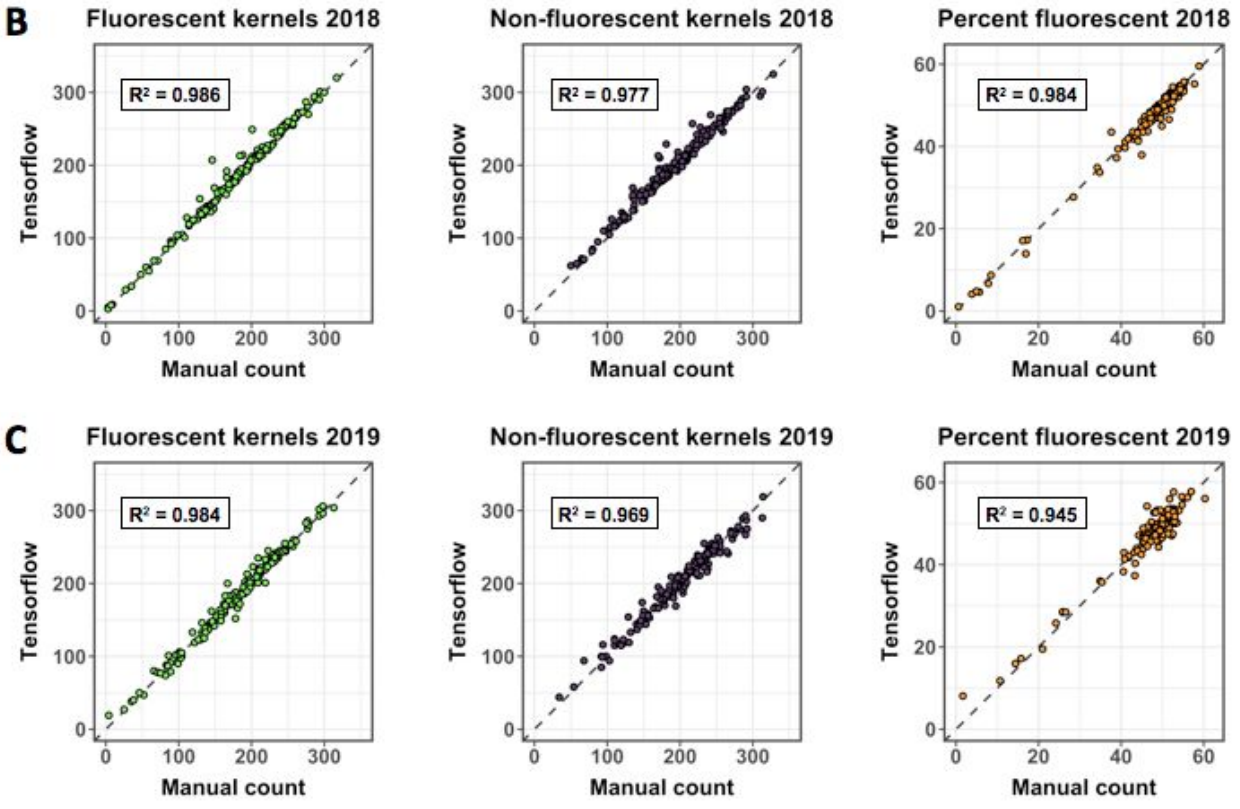

B
Fluorescent kernels 2018
R² = 0.986
Non-fluorescent kernels 2018
R² = 0.977
Percent fluorescent 2018
R² = 0.984
C
Fluorescent kernels 2019
R² = 0.984
Non-fluorescent kernels 2019
R² = 0.969
Percent fluorescent 2019
R² = 0.945
Tensorflow
Manual count

Figure 5. Deep learning models trained on image datasets from different field years and cameras accurately detected kernels and classes across a test dataset

(A) Example test images and annotations predicted by model. Top images: A typical ear from the 2018 field season showing Mendelian inheritance of GFP-marked kernels. The model predicted a 45% transmission rate, whereas manual count validation was 46.3% transmission. Middle images: a 2018 ear showing significant transmission defect, with few GFP-marker kernels. The model predicted a 16.9% transmission rate, while manual count validation was 16.2% transmission. Bottom images: 2019 ear showing Mendelian inheritance. The model predicted a 50.1% transmission rate, while manual count validation was 50.7% transmission.
(B) Total kernel counts and percent GFP across the 2018 test dataset (160 images). Adjusted R^2 values were calculated using a linear model comparing manual counts (x-axis) to deep learning model predictions (y-axis). Dashed diagonal lines represent equal values for both manual counts and model predictions. Adjusted R^2 values for total fluorescent, non-fluorescent, and percent fluorescent kernels were above 0.97.
(C) Total kernel counts and percent GFP across the 2019 test dataset (160 images). Adjusted R^2 values were again calculated using a linear model comparing manual counts to deep model predictions, and were above 0.94.

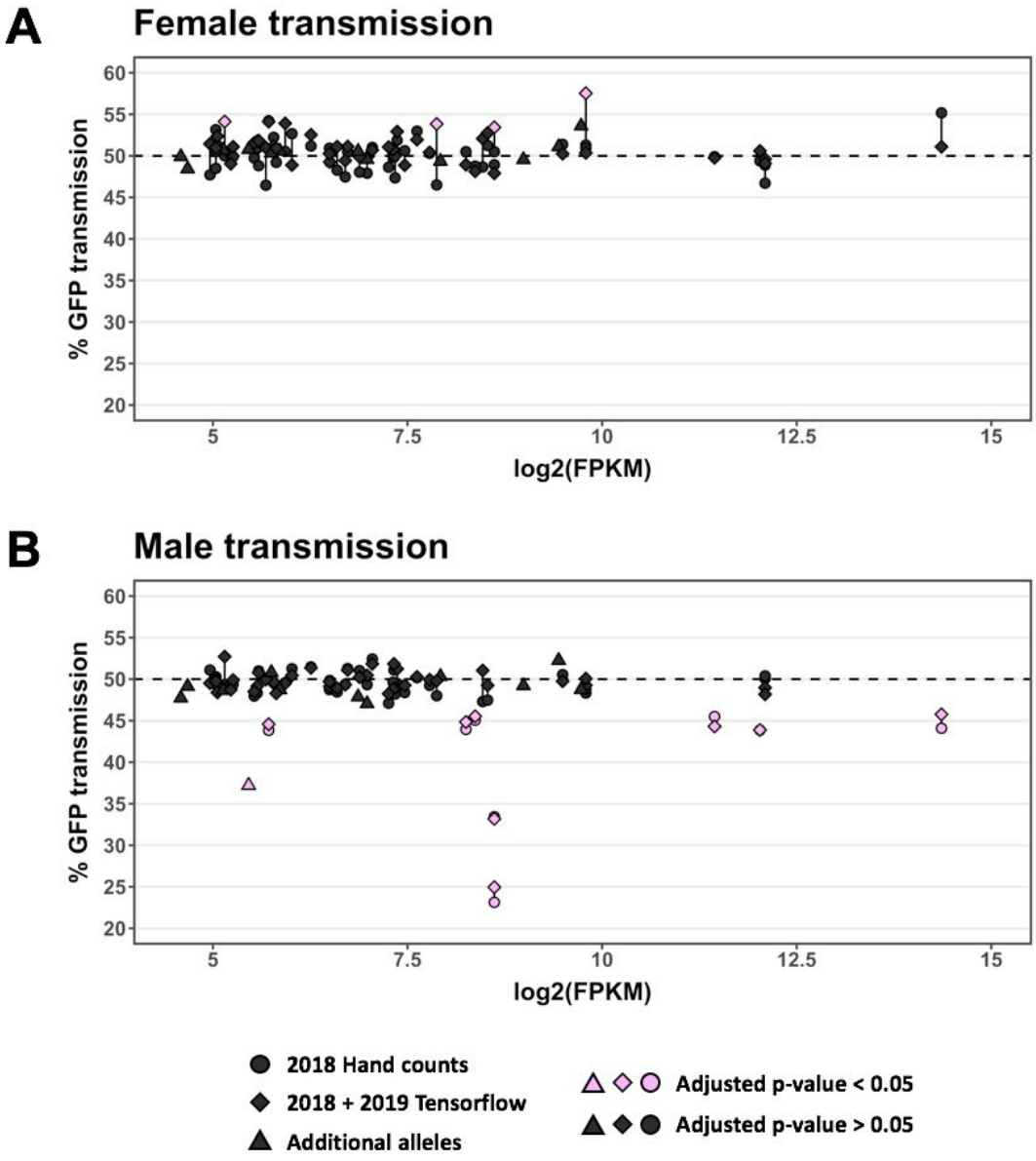


Figure 6. Comparison of model predictions and manual counts for GFP transmission data in two field seasons

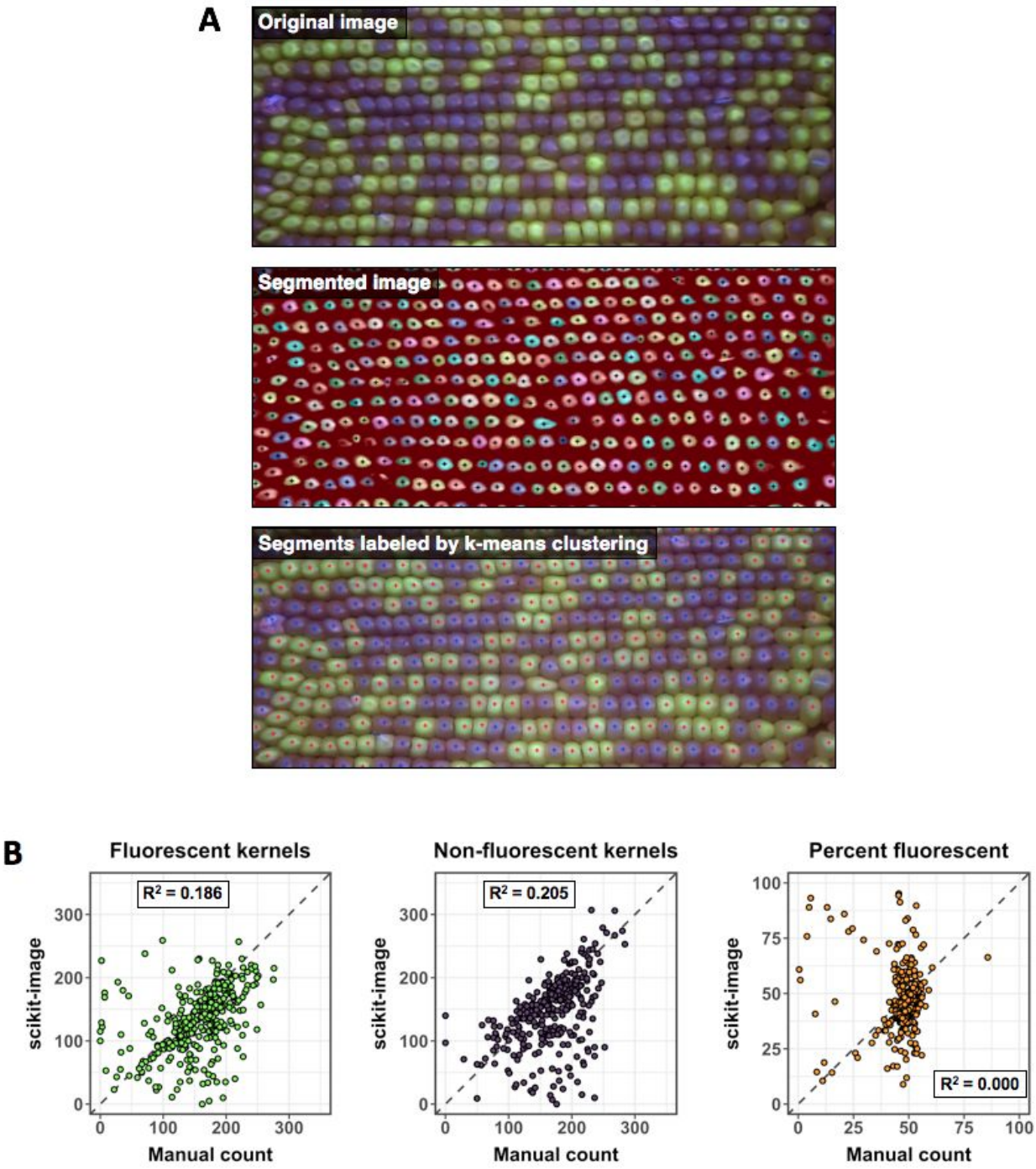
(A) Transmission results from plants containing heterozygous *Ds*-GFP insertion alleles outcrossed through the female. A total of 60 alleles were quantified in individual plants. Genes with mutant alleles belonged to three categories (Seedling, Vegetative Cell, Sperm Cell) based on high expression levels in those categories, shown on the x-axis in log2(FPKM).
(B) Transmission rates for all alleles assessed, when plants were crossed through the male. Expression levels in previously described categories are shown on the x-axis.

Table 1: Materials and costs for scanner construction

Material	Cost
Rotisserie motor (Minostar universal grill electric replacement rotisserie motor, 120 volt 4 watt)	$22.99
Drawer slides (Liberty D80618C-ZP-W 18-inch ball bearing drawer slides)	$11.94
Pillow block bearing (Letool 12mm mounted housing self-aligning pillow flange block bearing)	$3.75
Lumber	~$25.00
Screws	~$5.00
Metal rod	~$5.00
Tripod (AmazonBasics 60-Inch Lightweight Tripod)	$23.49
Computer (Dell 3630, Ubuntu Linux)	$616.00
Camera (ELP USBFHD06H-SFV)	$76.99
Blue light (for GFP, Clare Chemical HL34T)	$590.00
Orange filter set (for GFP, Neewer camera flash color gel kit)	$13.99
Total cost, alternative system 1 (smartphone, full visible spectrum images)	**$97.17**
Total cost, alternative system 2 (Computer, dedicated camera, light & filters for GFP imaging)	**$1,394.15**

Table 2: Characteristics of genes harboring newly-assessed Ds-GFP alleles

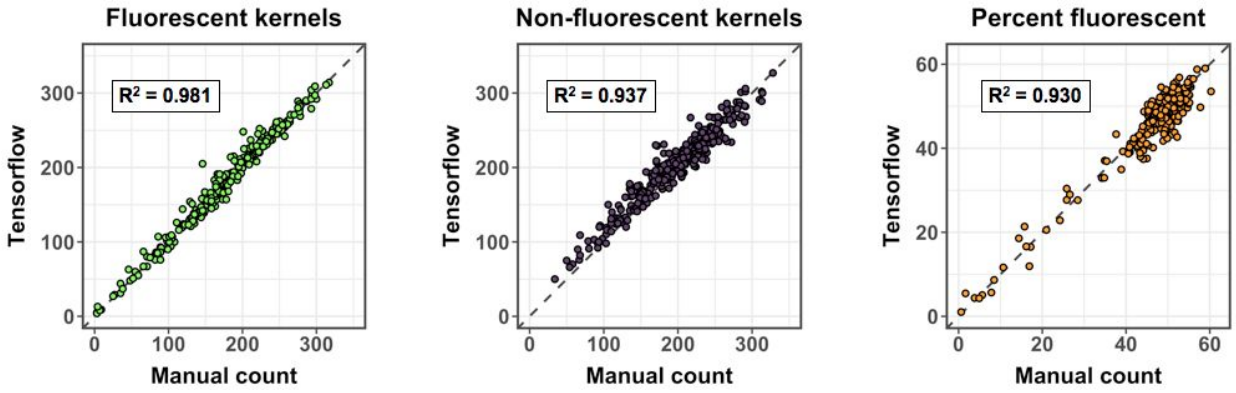
Category	Reason for exclusion from Warman et al. 2020	Gene designation (v3)	Gene designation (v4)	Ds-GFP allele	Male transmission rate	Adjusted p-value	Best BLAST Hit, A. thaliana	Annotation (B73v4 Gramene)
seedling_only	Insufficient crosses	GRMZM2G080724	Zm00001d031325	tdsgR106E07	48.64%	0.589	AT4G27670	Heat shock protein 21
seedling_only	Insufficient crosses	GRMZM2G148333	Zm00001d005798	tdsgR44E07	47.87%	0.503	AT3G14230	Ethylene-responsive transcription factor RAP2-2
seedling_only	Insufficient crosses	GRMZM2G374302	Zm00001d051194	tdsgR65A10	52.26%	0.258	AT2G16500	Arginine decarboxylase
sperm_cell_high	Insufficient crosses	AC194405.3_FG021	Zm00001d012575	tdsgR83A02	49.21%	0.796	AT1G19100	Protein MICRORCHIDIA 6
seedling_only	Insufficient crosses	GRMZM2G148387	Zm00001d017240	tdsgR91G06	50.80%	0.686	AT5G63030	Glutaredoxin-C1
vegetative_cell_high	non-PCR confirmed inser	GRMZM2G140107	Zm00001d042353	tdsgR02A05	47.43%	0.283	AT1G04920	Sucrose phosphate synthase2
vegetative_cell_high	non-PCR confirmed inser	GRMZM2G319167	Zm00001d039693	tdsgR108A02	47.05%	0.243	AT3G27670	Protein RST1
sperm_cell_high	non-PCR confirmed inser	GRMZM2G124365	Zm00001d012674	tdsgR29A11	49.13%	0.796	AT3G29200	Chorismate mutase1
vegetative_cell_high	non-PCR confirmed inser	GRMZM2G033828	Zm00001d031678	tdsgR67H12	48.73%	0.703	AT3G12280	Retinoblastoma family3
vegetative_cell_high	non-PCR confirmed inser	GRMZM2G051491	Zm00001d005053	tdsgR85A08	50.32%	0.890	AT3G10870	Methylesterase 17
sperm_cell_high	non-PCR confirmed inser	GRMZM2G062554	Zm00001d002824	tdsgR89B08	37.21%	<0.000001	AT1G62710	Vacuolar processing enzyme, beta-isozyme
sperm_cell_high	non-PCR confirmed inser	GRMZM2G038851	Zm00001d002570	tdsgR06C04	48.76%	0.716	AT3G57870	SUMO-conjugating enzyme SCE1



Supplemental Figure 1. A traditional computer vision approach generalized poorly for automated kernel detection.

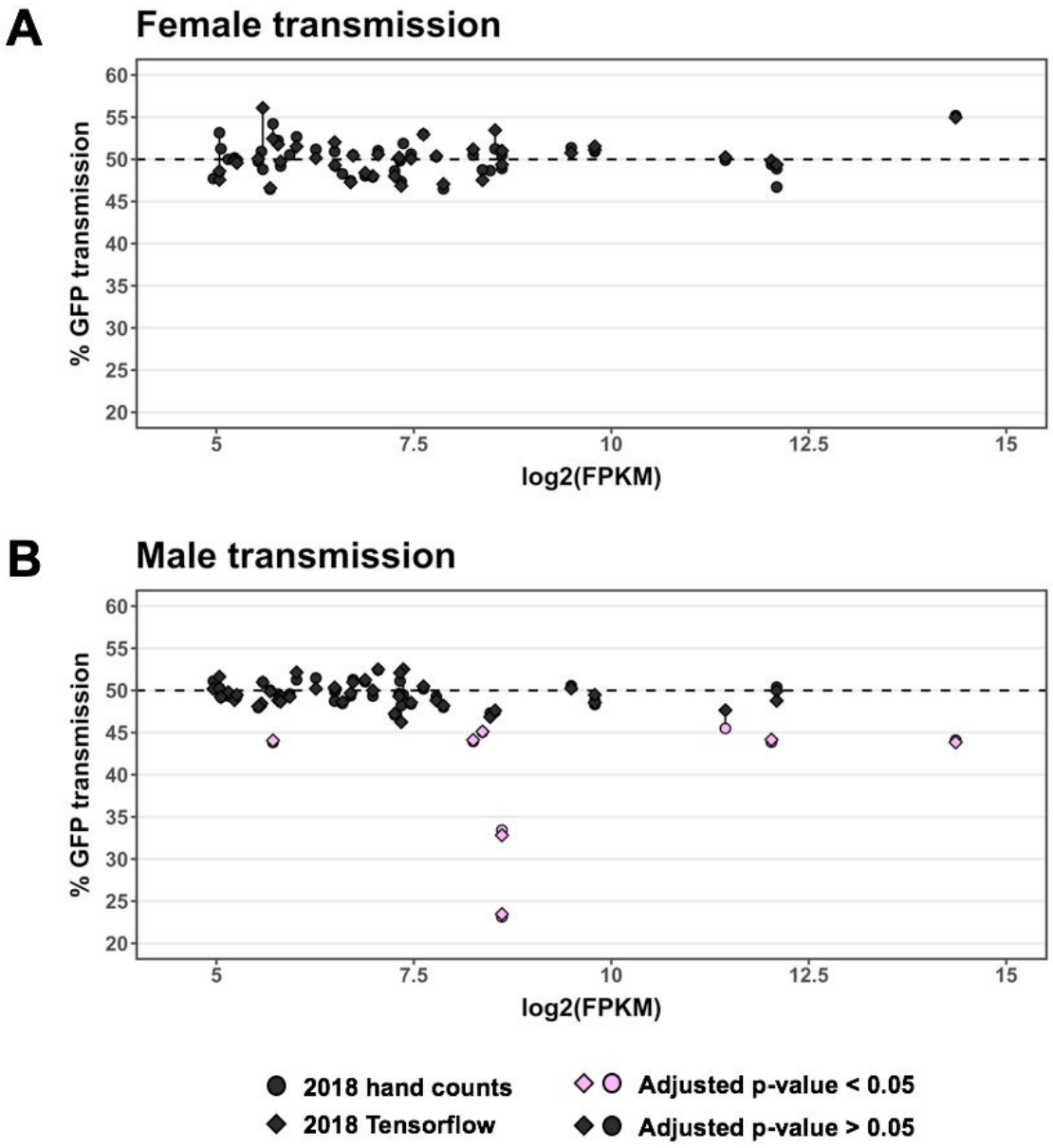
(A) Top panel: Two-dimension maize ear projection showing fluorescent and non-fluorescent kernels (GFP kernel marker). Middle panel: Watershed transformation followed by morphological opening segments individual kernels. Bottom panel: Mean intensity of RGB color

values within each segment are clustered by k-means clustering into two groups, fluorescent and non-fluorescent.
(B) Watershed computer vision approach applied to combined 2018 + 2019 ear image test sets. Diagonal dashed line indicates equal manual counts and computer vision predictions.



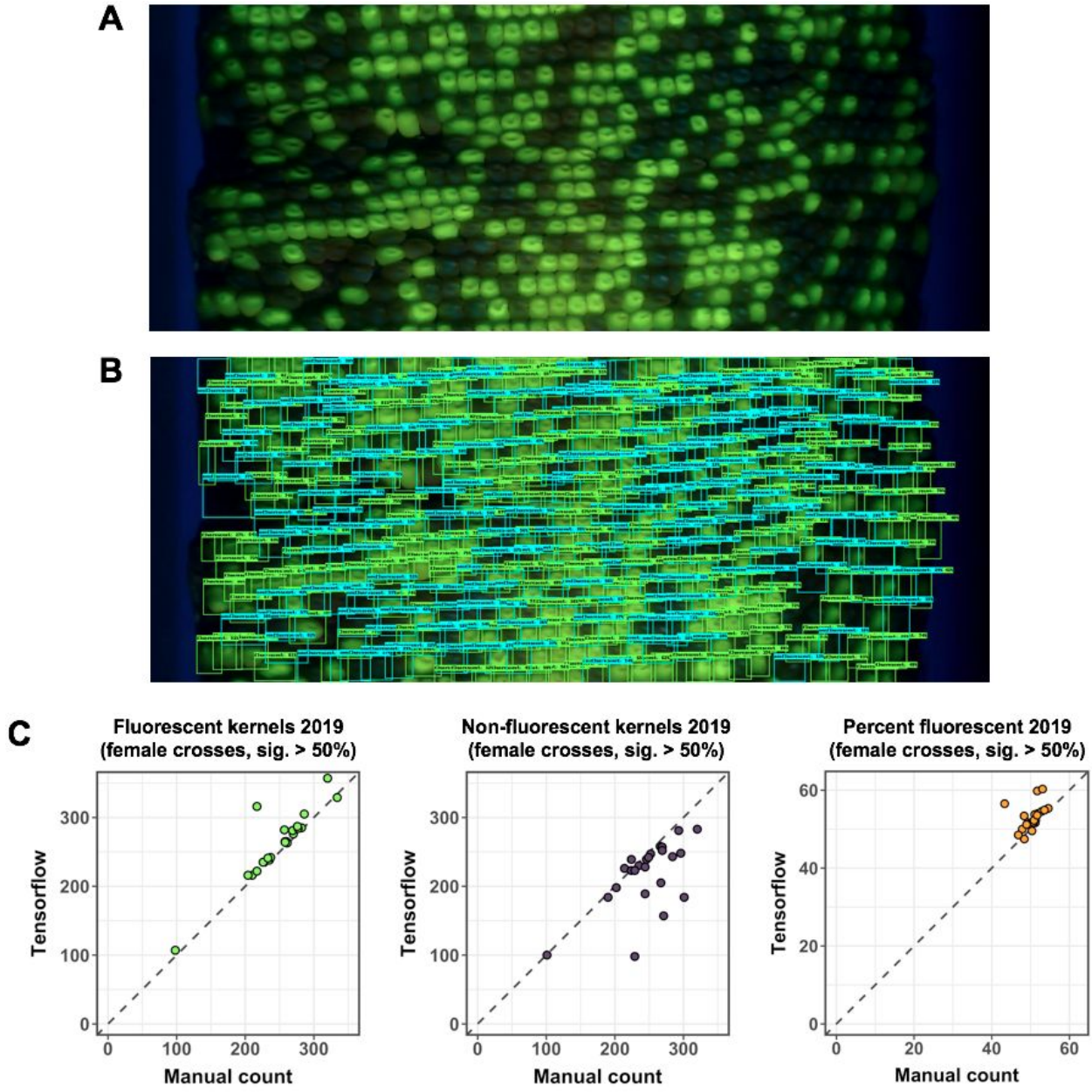
Supplemental Figure 2. A single deep learning model trained on images from two different cameras was able to detect kernels with moderate success.

Plots comparing model predictions and manual counts in the ear image test dataset. Model predictions were made using a model trained on images from two different cameras. Dashed diagonal lines indicate equal manual counts and model predictions.



Supplemental Figure 3. Comparing 2018 hand counts to 2018 Tensorflow model predictions.

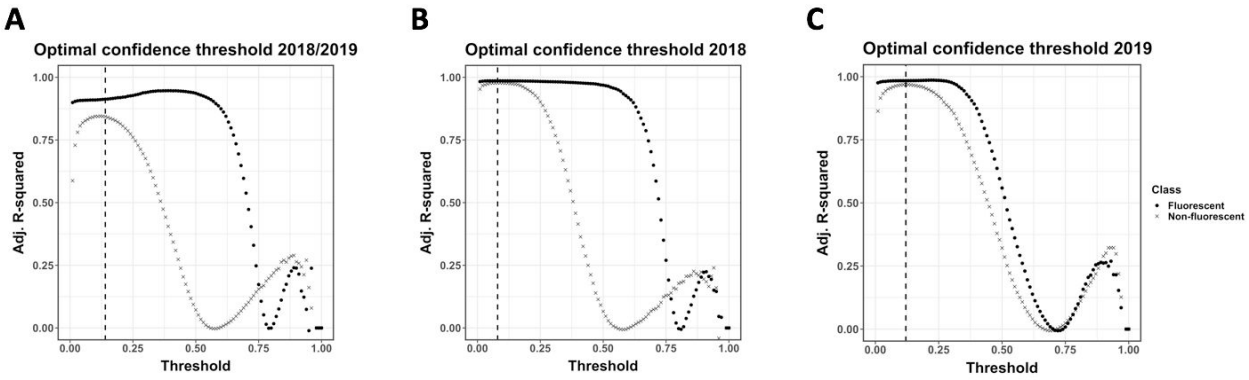
(A) GFP transmission in female crosses. No model-predicted transmission rates were significantly different from Mendelian inheritance, in agreement with hand counts. **(B)** GFP transmission in male crosses. While 7 out of 8 alleles with significant transmission defects by hand counts were identified by model predictions, 1 allele that was determined to be significant by hand counts was predicted to be non-significantly different than Mendelian by our model.



Supplemental Figure 4. Underexposure of a subset of 2019 images led to false positives for female transmission

(A) Original image for a female cross from *Ds-GFP* allele *tdsgR84A12*, an insertion in maize gene Zm00001d005781.
(B) Model predicted bounding boxes.
(C) Model predictions for female crosses in the four alleles that had significantly over 50% inheritance, compared to manual counts. Two images had substantial overestimations of GFP kernel totals (left), likely the result of strong fluorescence illuminating adjacent non-fluorescent

kernels in these images. Underestimations of non-fluorescent kernels were widespread (middle). Overestimation of percent fluorescent kernels was present in a small subset of female images (right).



Supplemental Figure 5. Empirically determining optimal bounding box output confidence thresholds.

(A) Optimal confidence thresholds were predicted from test image sets for the combined 2018/2019 model. R^2 values were calculated for fluorescent and non-fluorescent classes by comparing total predicted kernels to ground truth data across confidence thresholds from 0 to 1 in 0.01 intervals. The highest combined R^2 value was found at a confidence threshold of 0.12, marked with a vertical dashed line. **(B)** Optimal confidence thresholds for 2018 model. The highest combined R^2 value was found at a confidence threshold of 0.08 (vertical dashed line). **(C)** Optimal confidence thresholds for 2019 model. The highest combined R^2 value was found at a confidence threshold of 0.12 (vertical dashed line). Confidence thresholds from A-C were used to output kernel counts and bounding boxes in the final models.

Chapter 5

Conclusion

Cedar Warman

Phenotyping is a powerful method for understanding basic biological processes, particularly when coupled with quantitative analysis. Here, I have applied phenotyping approaches across three scales. In the first, I used low-throughput assays to describe the effects of mutations in two maize *nop* family genes. In the second, I employed medium-throughput methods to study the fitness costs of mutations in 52 genes. In the third, I developed a high-throughput deep learning computer vision pipeline to quantify transmission phenotypes in hundreds of maize ears. Each of these approaches has a role in answering biological questions about the mechanisms that enable plant reproduction.

The study of *nop* function provided a useful demonstration of the importance of individual gene studies in the era of big data. While high-throughput screens can place gene function in a larger context, understanding molecular mechanisms still requires the careful examination of individual genes. *nop* genes are highly and specifically expressed in the maize male gametophyte, offering the first hint of the functional role that they play. Although mutations in the two maize *nop* genes showed no sporophytic defects, I observed reduced transmission in crosses through the male, suggesting a gametophytic function. These transmission defects were partially recovered when competition was reduced by sparse pollen loads. After verifying that the observed transmission defects were caused by transposable element insertions in *nop1* and *nop2* using *Ac*-induced revertants (*nop1*) and a second mutant allele (*nop2*), I next measured tube length phenotypes using *wx1*-linked pollen staining. This assay showed that both *nop1* and *nop2* have shorter pollen tubes than wild type siblings, offering a potential explanation for transmission defects. I further examined *nop2* tube length responses to three chemicals by measuring pollen tubes exposed to different levels of each chemical. These studies showed that *nop* mutants are not particularly sensitive to Latrunculin B, an actin polymerization inhibitor, but are more sensitive than wild type pollen to two phosphoinositide 3-kinases (PI3K) inhibitors, Wortmannin and LY294002. Collectively, these results point to a role for *nop* genes in pollen tube growth. Predicted domain structures, combined with increased sensitivity to PI3K inhibitors, suggest a possible role in membrane binding and/or phospholipid signaling, two key processes in pollen tube growth.

In our second study, I scaled up the number of genes I examined from two to 52. This study was a collaborative effort between several research groups to study the maize male gametophyte transcriptome. My contribution to this project included an examination of highly expressed genes. To select these genes, I used transcriptome sequencing data that we generated from four stages of maize male gametophyte development: tassel primordia,

microspores, mature pollen, and sperm cells. We found that each of these developmental stages had unique sets of genes that were highly expressed. To determine if high expression in a developmental stage was correlated with function importance, I tested the transmission of mutant alleles of these highly expressed genes in three categories: vegetative cell, sperm cell, and a seedling control. I used a medium-throughput approach to test these alleles, taking advantage of *Ds-GFP* fluorescent seed markers to quantify transmission from digital ear images by hand. I identified several novel mutants, and showed that in the vegetative cell category, higher expression is correlated with increased fitness cost of mutant alleles. In addition, I found that many of the defective alleles in this category showed subtle transmission defects, at around 45% transmission of the mutant allele. These results point to a wide range of genes contributing to vegetative cell fitness. In the sperm cell category, the single transmission defect observed was severe, and caused by a mutation in the *gex2* gene. I contributed to research showing that mutant *gex* alleles are associated with incomplete fertilizations of the female gametophyte, leading to gaps and undeveloped kernels on the ear.

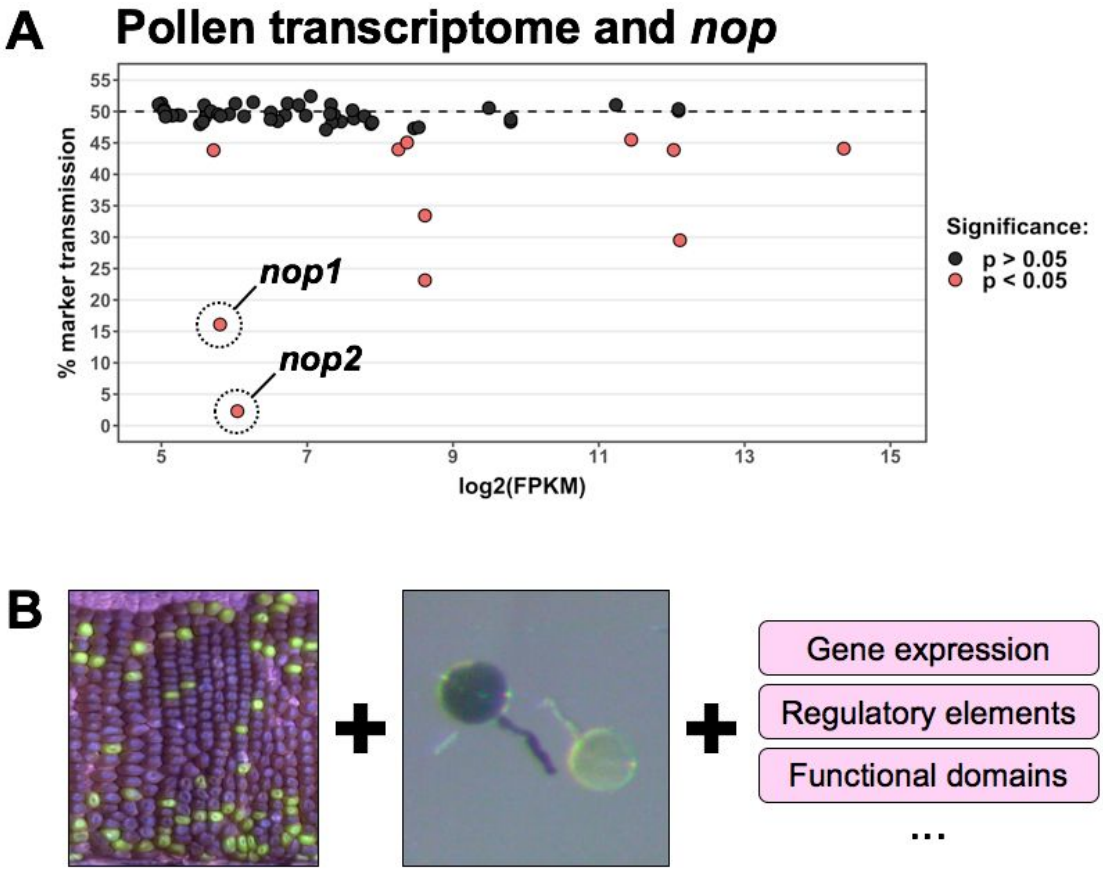
Our study of highly expressed maize pollen genes was enabled by the development of an ear scanning system. This system uses a custom-built platform to capture a video of a rotating maize ear. After the video is captured, software that I developed extracts frames from the video, crops each frame to a single row of pixels across the length of the ear, then combines rows of pixels from sequential frames to create a flat projection of the entire surface of the ear. This method creates a digital record of each ear, which can then be quantified to track kernel markers. In our pollen transcriptome study, we hand-counted over 200,000 kernels using images generated by this method. While this was faster than counting seeds on the physical ear, there was still room for significant improvements. To increase throughput, I created an automated computer vision system using deep learning. A deep learning approach requires training data. To satisfy this requirement, we generated 300 annotated images labeled with bounding boxes, identifying every kernel and marker identity. I used a transfer learning approach to retrain the top layers of a previously trained neural network. My final trained model was able to identify kernel positions and marker types with a high degree of accuracy, creating a truly high-throughput phenotyping system that categorized over 300,000 kernels. This method has helped to analyze many additional ears from various projects that are not included in this dissertation.

Taken together, our results point to a framework where high-throughput methods like transcriptomics and large-scale phenotyping provide context for individual studies of gene

function, which can in turn provide more inputs for larger functional models. For example, expression levels and transmission rates for *nop* genes can fit into the larger context of our pollen transcriptome study (Figure 1A). While *nop1* and *nop2* are on the lower end of the expression levels tested in the larger study (which encompassed a randomized set of genes, all in the Top 20% by FPKM), they have stronger transmission defects than any other genes in the pollen transcriptome study. Based on their expression values, the *nop* genes would have fallen into the "Vegetative Cell" category (VC genes). Descriptions of *nop* tube length phenotypes support the idea that their relatively high expression in mature pollen, and lower expression in the sperm cell, provide some correlative information pointing toward functional importance in the vegetative cell. In addition, their strong transmission defect suggests transmission rates associated with loss of function in VC genes can vary widely, likely pointing to a wide range of functions and fitness costs for these genes. However, unlike genes in the transcriptome study, *nop* genes were chosen based on their transmission defects. Thus, these results represent two approaches: in the larger transcriptome study, results are generalizable across all highly expressed genes in representative categories, while in the *nop* study, results are specific to *nop1* and *nop2.*

Combining quantitative phenotypes like transmission rates and tube lengths with high-throughput sequencing data in meaningful ways will likely be a fruitful area of future research (Figure 1B). For example, if a mutant has a transmission defect but no tube length defect, the integration of these phenotyping data could guide the researcher towards defects at germination or fertilization. In addition, specific phenotypes could be linked to specific sequences, such as regulatory elements or functional domains. The major challenge in this approach would be how to most effectively integrate these heterogeneous data sources to answer the biological questions at hand. Before any such system could be put into place, improvements in phenotyping methods would need to be made. The time required for pollen tube measurements is a major throughput bottleneck, but could likely be automated with a computer vision approach. In addition, linking *wx1* markers to large numbers of genes would be a time-consuming process in the field. The generation of a *Ds-GFP* insertion library with GFP expression in pollen as well as seeds would expedite this process immensely.

Our use of phenotyping methods at varying scales of complexity has laid a framework for future studies, both of small sets of genes and entire systems. I have shown the *nop* genes are influential in pollen tube growth, given insights into pollen developmental transcriptomes and gene functions, and created a new high-throughput phenotyping system.



(A) Percent marker transmission for male crosses of all mutant alleles described in this work. On the x-axis, gene expression level is shown for each allele, based on their expression in their representative category (see Chapter 3). *nop* genes show the strongest transmission defect of all genes studied here. Because *nop* genes were chosen for study based on their transmission defect, this result is not surprising.

(B) Combining phenotypic data from multiple sources, such as kernel marker transmission (left), pollen tube lengths (middle), and sequence-based data (right), can potentially drive future discovery. However, generating this data on a large scale and integrating these varied datasets will likely be major challenges.

Bibliography

Abadi, M. et al. (2016a). TensorFlow: A System for Large-Scale Machine Learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265–283.

Abadi, M. et al. (2016b). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv [cs.DC].

Adams, J., Qiu, Y., Xu, Y., and Schnable, J.C. (2020). Plant segmentation by supervised machine learning methods. Plant phenome j. **3**: 6980.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389–3402.

Anderson, S.N., Stitzer, M.C., Zhou, P., Ross-Ibarra, J., Hirsch, C.D., and Springer, N.M. (2019). Dynamic Patterns of Transcript Abundance of Transposable Element Families in Maize. G3 **9**: 3673–3682.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. Mol. Syst. Biol. **12**: 878.

Arthur, K.M., Vejlupkova, Z., Meeley, R.B., and Fowler, J.E. (2003). Maize ROP2 GTPase provides a competitive advantage to the male gametophyte. Genetics **165**: 2137–2151.

Arvidsson, S., Pérez-Rodríguez, P., and Mueller-Roeber, B. (2011). A growth phenotyping pipeline for Arabidopsis thaliana integrating image analysis and rosette area modeling for robust quantification of genotype effects. New Phytol. **191**: 895–907.

Awlia, M., Nigro, A., Fajkus, J., Schmoeckel, S.M., Negrão, S., Santelia, D., Trtílek, M., Tester, M., Julkowska, M.M., and Panzarová, K. (2016). High-Throughput Non-destructive Phenotyping of Traits that Contribute to Salinity Tolerance in Arabidopsis thaliana. Front. Plant Sci. **7**: 1414.

Bai, F., Daliberti, M., Bagadion, A., Xu, M., Li, Y., Baier, J., Tseung, C.-W., Evans, M.M.S., and Settles, A.M. (2016). Parent-of-Origin-Effect rough endosperm Mutants in Maize. Genetics **204**: 221–231.

Bai, L., Singh, M., Pitt, L., Sweeney, M., and Brutnell, T.P. (2007). Generating novel allelic variation through Activator insertional mutagenesis in maize. Genetics **175**: 981–992.

Baron, K.N., Schroeder, D.F., and Stasolla, C. (2014). GEm-Related 5 (GER5), an ABA and stress-responsive GRAM domain protein regulating seed development and inflorescence architecture. Plant Sci. **223**: 153–166.

Becker, J.D., Boavida, L.C., Carneiro, J., Haury, M., and Feijó, J.A. (2003). Transcriptional profiling of Arabidopsis tissues reveals the unique characteristics of the pollen transcriptome. Plant Physiol. **133**: 713–725.

Begcy, K., Nosenko, T., Zhou, L.-Z., Fragner, L., Weckwerth, W., and Dresselhaus, T.

(2019). Male Sterility in Maize after Transient Heat Stress during the Tetrad Stage of Pollen Development. Plant Physiol.

Berger, P., Schaffitzel, C., Berger, I., Ban, N., and Suter, U. (2003). Membrane association of myotubularin-related protein 2 is mediated by a pleckstrin homology-GRAM domain and a coiled-coil dimerization module. Proc. Natl. Acad. Sci. U. S. A. **100**: 12177–12182.

Berry, D.B., Guan, Q., Hose, J., Haroon, S., Gebbia, M., Heisler, L.E., Nislow, C., Giaever, G., and Gasch, A.P. (2011). Multiple means to the same end: the genetic basis of acquired stress resistance in yeast. PLoS Genet. **7**: e1002353.

Besprozvannaya, M., Dickson, E., Li, H., Ginburg, K.S., Bers, D.M., Auwerx, J., and Nunnari, J. (2018). GRAM domain proteins specialize functionally distinct ER-PM contact sites in human cells. Elife **7**.

Borges, F., Parent, J.-S., van Ex, F., Wolff, P., Martínez, G., Köhler, C., and Martienssen, R.A. (2018). Transposon-derived small RNAs triggered by miR845 mediate genome dosage response in Arabidopsis. Nat. Genet. **50**: 186–192.

Brink, R.A. and MacGillivray, J.H. (1924). Segregation for the Waxy Character in Maize Pollen and Differential Development of the Male Gametophyte. Am. J. Bot. **11**: 465–469.

Bushnell, B. (2014). BBTools software package. URL http://sourceforge.net/projects/bbmap.

Calarco, J.P., Borges, F., Donoghue, M.T.A., Van Ex, F., Jullien, P.E., Lopes, T., Gardner, R., Berger, F., Feijó, J.A., Becker, J.D., and Martienssen, R.A. (2012). Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. Cell **151**: 194–205.

Camacho, L. and Malhó, R. (2003). Endo/exocytosis in the pollen tube apex is differentially regulated by Ca2+ and GTPases. J. Exp. Bot. **54**: 83–92.

Chaivivatrakul, S., Tang, L., Dailey, M.N., and Nakarmi, A.D. (2014). Automatic morphological trait characterization for corn plants via 3D holographic reconstruction. Comput. Electron. Agric. **109**: 109–123.

Cheng, C.-Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., and Town, C.D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. **89**: 789–804.

Chen, J., Strieder, N., Krohn, N.G., Cyprys, P., Sprunck, S., Engelmann, J.C., and Dresselhaus, T. (2017). Zygotic Genome Activation Occurs Shortly after Fertilization in Maize. Plant Cell **29**: 2106–2125.

Chettoor, A.M., Givan, S.A., Cole, R.A., Coker, C.T., Unger-Wallace, E., Vejlupkova, Z., Vollbrecht, E., Fowler, J.E., and Evans, M.M. (2014). Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. Genome Biol. **15**: 414.

Ching, T. et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface **15**.

Chollet, F. and Others (2015). Keras.

Choudhury, P., Srivastava, S., Li, Z., Ko, K., Albaqumi, M., Narayan, K., Coetzee, W.A., Lemmon, M.A., and Skolnik, E.Y. (2006). Specificity of the myotubularin family of phosphatidylinositol-3-phosphatase is determined by the PH/GRAM domain. J. Biol. Chem. **281**: 31762–31769.

Choudhury, S.D., Stoerger, V., Samal, A., Schnable, J.C., Liang, Z., and Yu, J.-G. (2016). Automated vegetative stage phenotyping analysis of maize plants using visible light images. In KDD workshop on data science for food, energy and water, San Francisco, California, USA (researchgate.net).

Clark, R.T., Famoso, A.N., Zhao, K., Shaff, J.E., Craft, E.J., Bustamante, C.D., McCouch, S.R., Aneshansley, D.J., and Kochian, L.V. (2013). High-throughput two-dimensional root system phenotyping platform facilitates genetic analysis of root growth and development. Plant Cell Environ. **36**: 454–466.

Cole, R.A., Synek, L., Zarsky, V., and Fowler, J.E. (2005). SEC8, a subunit of the putative Arabidopsis exocyst complex, facilitates pollen germination and competitive pollen tube growth. Plant Physiol. **138**: 2005–2018.

Collins, G.N. (1909). A new type of Indian corn from China (US Government Printing Office).

Conrad, L.J. and Brutnell, T.P. (2005). Ac-immobilized, a stable source of Activator transposase that mediates sporophytic and gametophytic excision of Dissociation elements in maize. Genetics **171**: 1999–2012.

Coussens, L., Parker, P.J., Rhee, L., Yang-Feng, T.L., Chen, E., Waterfield, M.D., Francke, U., and Ullrich, A. (1986). Multiple, distinct forms of bovine and human protein kinase C suggest diversity in cellular signaling pathways. Science **233**: 859–866.

Demerec, M. (1924). A Case of Pollen Dimorphism in Maize. Am. J. Bot. **11**: 461–464.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**: 15–21.

Dobos, O., Horvath, P., Nagy, F., Danka, T., and Viczián, A. (2019). A Deep Learning-Based Approach for High-Throughput Hypocotyl Phenotyping. Plant Physiol. **181**: 1415–1424.

Dobrescu, A., Giuffrida, M.V., and Tsaftaris, S.A. (2020). Doing More With Less: A Multitask Deep Learning Approach in Plant Phenotyping. Front. Plant Sci. **11**: 141.

Doerks, T., Strauss, M., Brendel, M., and Bork, P. (2000). GRAM, a novel domain in glucosyltransferases, myotubularins and other putative membrane-associated proteins.

Trends Biochem. Sci. **25**: 483–485.

Dooner, H.K., Wang, Q., Huang, J.T., Li, Y., He, L., Xiong, W., and Du, C. (2019). Spontaneous mutations in maize pollen are frequent in some lines and arise mainly from retrotranspositions and deletions. Proc. Natl. Acad. Sci. U. S. A.

Douarre, C., Schielein, R., Frindel, C., Gerth, S., and Rousseau, D. (2018). Transfer Learning from Synthetic Data Applied to Soil–Root Segmentation in X-Ray Tomography Images. Journal of Imaging **4**: 65.

Dresselhaus, T., Sprunck, S., and Wessel, G.M. (2016). Fertilization Mechanisms in Flowering Plants. Curr. Biol. **26**: R125–39.

Engel, M.L., Chaboud, A., Dumas, C., and McCormick, S. (2003). Sperm cells of Zea mays have a complex complement of mRNAs. Plant J. **34**: 697–707.

Engel, M.L., Holmes-Davis, R., and McCormick, S. (2005). Green sperm. Identification of male gamete promoters in Arabidopsis. Plant Physiol. **138**: 2124–2133.

Erbar, C. (2003). Pollen Tube Transmitting Tissue: Place of Competition of Male Gametophytes. Int. J. Plant Sci. **164**: S265–S277.

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

Feldmann, M.J., Hardigan, M.A., Famula, R.A., López, C.M., Tabb, A., Cole, G.S., and Knapp, S.J. (2020). Multi-dimensional machine learning approaches for fruit shape phenotyping in strawberry. Gigascience **9**.

Feldmann, M., Tabb, A., and Knapp, S.J. (2019). Cost-effective, high-throughput 3D reconstruction method for fruit phenotyping. Computer Vision Problems in Plant Phenotyping (CVPPP): 1.

Gao, H., Smith, J., Yang, M., Jones, S., Djukanovic, V., Nicholson, M.G., West, A., Bidney, D., Falco, S.C., Jantz, D., and Lyznik, L.A. (2010). Heritable targeted mutagenesis in maize using a designed endonuclease. Plant J. **61**: 176–187.

Gatta, A.T., Sauerwein, A.C., Zhuravleva, A., Levine, T.P., and Matthews, S. (2018). Structural insights into a StART-like domain in Lam4 and its interaction with sterol ligands. Biochem. Biophys. Res. Commun. **495**: 2270–2274.

Giaever, G. et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature **418**: 387–391.

Gibbon, B.C., Kovar, D.R., and Staiger, C.J. (1999). Latrunculin B has different effects on pollen germination and tube growth. Plant Cell **11**: 2349–2363.

Gibbs, J.A., Burgess, A.J., Pound, M.P., Pridmore, T.P., and Murchie, E.H. (2019). Recovering Wind-Induced Plant Motion in Dense Field Environments via Deep Learning

and Multiple Object Tracking. Plant Physiol. **181**: 28–42.

Gilles, L.M. et al. (2017). Loss of pollen-specific phospholipase NOT LIKE DAD triggers gynogenesis in maize. EMBO J. **36**: 707–717.

González Montoro, A. and Ungermann, C. (2015). StARTing to understand membrane contact sites. Trends Cell Biol. **25**: 497–498.

Guo, J. and Yang, Z. (2020). Exocytosis and endocytosis: coordinating and fine-tuning the polar tip growth domain in pollen tubes. J. Exp. Bot. **71**: 2428–2438.

Gu, Y., Fu, Y., Dowd, P., Li, S., Vernoud, V., Gilroy, S., and Yang, Z. (2005). A Rho family GTPase controls actin dynamics and tip growth via two counteracting downstream pathways in pollen tubes. J. Cell Biol. **169**: 127–138.

Hafidh, S., Fíla, J., and Honys, D. (2016). Male gametophyte development and function in angiosperms: a general concept. Plant Reprod. **29**: 31–51.

Hammond, G.R.V. and Balla, T. (2015). Polyphosphoinositide binding domains: Key to inositol lipid biology. Biochim. Biophys. Acta **1851**: 746–758.

Haralick, R.M., Sternberg, S.R., and Zhuang, X. (1987). Image analysis using mathematical morphology. IEEE Trans. Pattern Anal. Mach. Intell. **9**: 532–550.

Heilmann, I. and Ischebeck, T. (2016). Male functions and malfunctions: the impact of phosphoinositides on pollen development and pollen tube growth. Plant Reprod. **29**: 3–20.

Helmann, T.C., Deutschbauer, A.M., and Lindow, S.E. (2019). Genome-wide identification of Pseudomonas syringae genes required for fitness during colonization of the leaf surface and apoplast. Proc. Natl. Acad. Sci. U. S. A.

Hepler, P.K. and Winship, L.J. (2015). The pollen tube clear zone: clues to the mechanism of polarized growth. J. Integr. Plant Biol. **57**: 79–92.

He, S., Vickers, M., Zhang, J., and Feng, X. (2019). Natural depletion of histone H1 in sex cells causes DNA demethylation, heterochromatin decondensation and transposon activation. Elife **8**.

Higashiyama, T. and Takeuchi, H. (2015). The mechanism and key molecules involved in pollen tube guidance. Annu. Rev. Plant Biol. **66**: 393–413.

Higashiyama, T., Yabe, S., Sasaki, N., Nishimura, Y., Miyagishima S, Kuroiwa, H., and Kuroiwa, T. (2001). Pollen tube attraction by the synergid cell. Science **293**: 1480–1483.

Honys, D. and Twell, D. (2003). Comparative analysis of the Arabidopsis pollen transcriptome. Plant Physiol. **132**: 640–652.

Hormaza, J.I. and Herrero, M. (1996). Dynamics of pollen tube growth under different

competition regimes. Sex. Plant Reprod. **9**: 153–160.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., and Murphy, K. (2016). Speed/accuracy trade-offs for modern convolutional object detectors. arXiv [cs.CV].

Huang, J.T., Wang, Q., Park, W., Feng, Y., Kumar, D., Meeley, R., and Dooner, H.K. (2017). Competitive Ability of Maize Pollen Grains Requires Paralogous Serine Threonine Protein Kinases STK1 and STK2. Genetics **207**: 1361–1370.

Hüther, P., Schandry, N., Jandrasits, K., Bezrukov, I., and Becker, C. (2020). araDEEPopsis: From images to phenotypic traits using deep transfer learning. bioRxiv: 2020.04.01.018192.

Jiang, H., Xu, Z., Aluru, M.R., and Dong, L. (2014). Plant chip for high-throughput phenotyping of Arabidopsis. Lab Chip **14**: 1281–1293.

Jiang, N., Floro, E., Bray, A.L., Laws, B., Duncan, K.E., and Topp, C.N. (2019). Three-Dimensional Time-Lapse Analysis Reveals Multiscale Relationships in Maize Root Systems with Contrasting Architectures. Plant Cell **31**: 1708–1722.

Jiang, S.Y., Cai, M., and Ramachandran, S. (2005). The Oryza sativa no pollen (Osnop) gene plays a role in male gametophyte development and most likely encodes a C2-GRAM domain-containing protein. Plant Mol. Biol. **57**: 835–853.

Jiang, S.-Y., Ramamoorthy, R., and Ramachandran, S. (2008). Comparative transcriptional profiling and evolutionary analysis of the GRAM domain family in eukaryotes. Dev. Biol. **314**: 418–432.

Jiao, Y. et al. (2017). Improved maize reference genome with single-molecule technologies. Nature **546**: 524–527.

Johnson, M.A., Harper, J.F., and Palanivelu, R. (2019). A Fruitful Journey: Pollen Tube Navigation from Germination to Fertilization. Annu. Rev. Plant Biol. **70**: 809–837.

Jones-Rhoades, M.W., Borevitz, J.O., and Preuss, D. (2007). Genome-wide expression profiling of the Arabidopsis female gametophyte identifies families of small, secreted proteins. PLoS Genet. **3**: 1848–1861.

Junker, A., Muraya, M.M., Weigelt-Fischer, K., Arana-Ceballos, F., Klukas, C., Melchinger, A.E., Meyer, R.C., Riewe, D., and Altmann, T. (2014). Optimizing experimental procedures for quantitative evaluation of crop plant performance in high throughput phenotyping systems. Front. Plant Sci. **5**: 770.

Kelliher, T. et al. (2017). MATRILINEAL, a sperm-specific phospholipase, triggers maize haploid induction. Nature **542**: 105–109.

Kelliher, T. et al. (2019). One-step genome editing of elite crop germplasm during haploid

induction. Nat. Biotechnol. **37**: 287–292.

Khafif, M., Balagué, C., Huard-Chauveau, C., and Roby, D. (2017). An essential role for the VASt domain of the Arabidopsis VAD1 protein in the regulation of defense and cell death in response to pathogens. PLoS One **12**: e0179782.

Khafif, M., Cottret, L., Balagué, C., and Raffaele, S. (2014). Identification and phylogenetic analyses of VASt, an uncharacterized protein domain associated with lipid-binding domains in Eukaryotes. BMC Bioinformatics **15**: 222.

Kim, C.Y. et al. (2003). Rice C2-domain proteins are induced and translocated to the plasma membrane in response to a fungal elicitor. Biochemistry **42**: 11625–11633.

Krishnakumar, V. et al. (2015). Araport: the Arabidopsis information portal. Nucleic Acids Res. **43**: D1003–9.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. **305**: 567–580.

Lam, S.K., Siu, C.L., Hillmer, S., Jang, S., An, G., Robinson, D.G., and Jiang, L. (2007). Rice SCAMP1 defines clathrin-coated, trans-golgi-located tubular-vesicular structures as an early endosome in tobacco BY-2 cells. Plant Cell **19**: 296–319.

Lausser, A., Kliwer, I., Srilunchang, K.-O., and Dresselhaus, T. (2010). Sporophytic control of pollen tube growth and guidance in maize. J. Exp. Bot. **61**: 673–682.

Letunic, I. and Bork, P. (2018). 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. **46**: D493–D496.

Liang, X., Wang, K., Huang, C., Zhang, X., Yan, J., and Yang, W. (2016). A high-throughput maize kernel traits scorer based on line-scan imaging. Measurement **90**: 453–460.

Liao, B.-Y. and Weng, M.-P. (2015). Unraveling the association between mRNA expressions and mutant phenotypes in a genome-wide assessment of mice. Proc. Natl. Acad. Sci. U. S. A. **112**: 4707–4712.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**: 923–930.

Li, H., Lin, Y., Heath, R.M., Zhu, M.X., and Yang, Z. (1999). Control of pollen tube tip growth by a Rop GTPase-dependent pathway that leads to tip-localized calcium influx. Plant Cell **11**: 1731–1742.

Liu, C. et al. (2017). A 4-bp Insertion at ZmPLA1 Encoding a Putative Phospholipase A Generates Haploid Induction in Maize. Mol. Plant **10**: 520–522.

Liu, L., Li, C., Liang, Z., and Yu, H. (2018). Characterization of Multiple C2 Domain and

Transmembrane Region Proteins in Arabidopsis. Plant Physiol. **176**: 2119–2132.

Li, Y., Kabbage, M., Liu, W., and Dickman, M.B. (2016). Aspartyl Protease-Mediated Cleavage of BAG6 Is Necessary for Autophagy and Fungal Resistance in Plants. Plant Cell **28**: 233–247.

Li, Y., Segal, G., Wang, Q., and Dooner, H.K. (2013). Gene Tagging with Engineered Ds Elements in Maize. In Plant Transposable Elements: Methods and Protocols, T. Peterson, ed (Humana Press: Totowa, NJ), pp. 83–99.

Lloyd, S. (1982). Least squares quantization in PCM. IEEE Trans. Inf. Theory **28**: 129–137.

Lorrain, S., Lin, B., Auriac, M.C., Kroj, T., Saindrenan, P., Nicole, M., Balagué, C., and Roby, D. (2004). Vascular associated death1, a novel GRAM domain-containing protein, is a regulator of cell death and defense responses in vascular tissues. Plant Cell **16**: 2217–2232.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**: 550.

Lunardon, A., Forestan, C., Farinati, S., Axtell, M.J., and Varotto, S. (2016). Genome-Wide Characterization of Maize Small RNA Loci and Their Regulation in the required to maintain repression6-1 (rmr6-1) Mutant and Long-Term Abiotic Stresses. Plant Physiol. **170**: 1535–1548.

Lyons, E. and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences: How to usefully compare plant genomes. Plant J. **53**: 661–673.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., and Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. **47**: W636–W641.

Mahlein, A.-K. (2016). Plant Disease Detection by Imaging Sensors - Parallels and Specific Demands for Precision Agriculture and Plant Phenotyping. Plant Dis. **100**: 241–251.

Makanza, R., Zaman-Allah, M., Cairns, J.E., Eyre, J., Burgueño, J., Pacheco, Á., Diepenbrock, C., Magorokosho, C., Tarekegne, A., Olsen, M., and Prasanna, B.M. (2018). High-throughput method for ear phenotyping and kernel weight estimation in maize using ear digital imaging. Plant Methods **14**: 49.

Martínez, G., Panda, K., Köhler, C., and Slotkin, R.K. (2016). Silencing in sperm cells is directed by RNA movement from the surrounding nurse cell. Nat Plants **2**: 16030.

Martínez, G. and Slotkin, R.K. (2012). Developmental relaxation of transposable element silencing in plants: functional or byproduct? Curr. Opin. Plant Biol. **15**: 496–502.

Martinez, G., Wolff, P., Wang, Z., Moreno-Romero, J., Santos-González, J., Conze, L.L., DeFraia, C., Slotkin, R.K., and Köhler, C. (2018). Paternal easiRNAs regulate parental

genome dosage in Arabidopsis. Nat. Genet. **50**: 193–198.

Mauri, N., Fernández-Marcos, M., Costas, C., Desvoyes, B., Pichel, A., Caro, E., and Gutierrez, C. (2016). GEM, a member of the GRAM domain family of proteins, is part of the ABA signaling pathway. Sci. Rep. **6**: 22660.

McCarty, D.R., Latshaw, S., Wu, S., Suzuki, M., Hunter, C.T., Avigne, W.T., and Koch, K.E. (2013). Mu-seq: sequence-based mapping and identification of transposon induced mutations. PLoS One **8**: e77172.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. U. S. A. **36**: 344–355.

McCormick, S. (1993). Male Gametophyte Development. Plant Cell **5**: 1265–1275.

Miller, N.D., Haase, N.J., Lee, J., Kaeppler, S.M., de Leon, N., and Spalding, E.P. (2017). A robust, high-throughput method for computing maize ear, cob, and kernel attributes automatically from images. Plant J. **89**: 169–178.

Mitchell, A.L. et al. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. **47**: D351–D360.

Mizukami, A.G. et al. (2016). The AMOR Arabinogalactan Sugar Chain Induces Pollen-Tube Competency to Respond to Ovular Guidance. Curr. Biol. **26**: 1091–1097.

Mohanty, S.P., Hughes, D.P., and Salathé, M. (2016). Using Deep Learning for Image-Based Plant Disease Detection. Front. Plant Sci. **7**: 1419.

Mori, T., Igawa, T., Tamiya, G., Miyagishima, S.-Y., and Berger, F. (2014). Gamete attachment requires GEX2 for successful fertilization in Arabidopsis. Curr. Biol. **24**: 170–175.

Moscatelli, A. and Idilli, A.I. (2009). Pollen tube growth: a delicate equilibrium between secretory and endocytic pathways. J. Integr. Plant Biol. **51**: 727–739.

Moscatelli, A., Idilli, A.I., Rodighiero, S., and Caccianiga, M. (2012). Inhibition of actin polymerisation by low concentration Latrunculin B affects endocytosis and alters exocytosis in shank and tip of tobacco pollen tubes: Actin-dependent endo/exocytosis in pollen tubes. Plant Biol. **14**: 770–782.

Murley, A., Sarsam, R.D., Toulmay, A., Yamada, J., Prinz, W.A., and Nunnari, J. (2015). Ltc1 is an ER-localized sterol transporter and a component of ER-mitochondria and ER-vacuole contacts. J. Cell Biol. **209**: 539–548.

Muschietti, J.P. and Wengier, D.L. (2018). How many receptor-like kinases are required to operate a pollen tube. Curr. Opin. Plant Biol. **41**: 73–82.

Nalefski, E.A. and Falke, J.J. (1996). The C2 domain calcium-binding motif: structural and

functional diversity. Protein Sci. **5**: 2375–2390.

Nelms, B. and Walbot, V. (2019). Defining the developmental program leading to meiosis in maize. Science **364**: 52–56.

Neuffer, M.G., Coe, E.H., and Wessler, S.R. (1997). Mutants of maize (Cold Spring Harbor Laboratory Press).

Otsu, N. (1979). A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**: 62–66.

Palanivelu, R. and Preuss, D. (2006). Distinct short-range ovule signals attract or repel Arabidopsis thaliana pollen tubes in vitro. BMC Plant Biol. **6**: 7.

Panda, K., Ji, L., Neumann, D.A., Daron, J., Schmitz, R.J., and Slotkin, R.K. (2016). Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. Genome Biol. **17**: 170.

Passot, S., Moreno-Ortega, B., Moukouanga, D., Balsera, C., Guyomarc'h, S., Lucas, M., Lobet, G., Laplaze, L., Muller, B., and Guédon, Y. (2018). A New Phenotyping Pipeline Reveals Three Types of Lateral Roots and a Random Branching Pattern in Two Cereals. Plant Physiol. **177**: 896–910.

Paszke, A. et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, eds (Curran Associates, Inc.), pp. 8026–8037.

Phillips, A.R. and Evans, M.M.S. (2011). Analysis of stunter1, a maize mutant with reduced gametophyte size and maternal effects on seed development. Genetics **187**: 1085–1097.

Polder, G., Blok, P.M., de Villiers, H.A.C., van der Wolf, J.M., and Kamp, J. (2019). Potato Virus Y Detection in Seed Potatoes Using Deep Learning on Hyperspectral Images. Front. Plant Sci. **10**: 209.

Portwood, J.L., 2nd et al. (2019). MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic Acids Res. **47**: D1146–D1154.

Pound, M.P., Atkinson, J.A., Townsend, A.J., Wilson, M.H., Griffiths, M., Jackson, A.S., Bulat, A., Tzimiropoulos, G., Wells, D.M., Murchie, E.H., Pridmore, T.P., and French, A.P. (2017). Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. Gigascience **6**: 1–10.

Price, M.N., Deutschbauer, A.M., Skerker, J.M., Wetmore, K.M., Ruths, T., Mar, J.S., Kuehl, J.V., Shao, W., and Arkin, A.P. (2013). Indirect and suboptimal control of gene expression is widespread in bacteria. Mol. Syst. Biol. **9**: 660.

Rawat, W. and Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification:

A Comprehensive Review. Neural Comput. **29**: 2352–2449.

Redmon, J. and Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv [cs.CV].

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Advances in Neural Information Processing Systems 28, C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, eds (Curran Associates, Inc.), pp. 91–99.

Rodriguez, L. et al. (2014). C2-domain abscisic acid-related proteins mediate the interaction of PYR/PYL/RCAR abscisic acid receptors with the plasma membrane and regulate abscisic acid sensitivity in Arabidopsis. Plant Cell **26**: 4802–4820.

Running, M.P., Clark, S.E., and Meyerowitz, E.M. (1995). Chapter 15 Confocal Microscopy of the Shoot Apex. In Methods in Cell Biology, D.W. Galbraith, H.J. Bohnert, and D.P. Bourque, eds (Academic Press), pp. 217–229.

Rutter, M.T., Murren, C.J., Callahan, H.S., Bisner, A.M., Leebens-Mack, J., Wolyniak, M.J., and Strand, A.E. (2019). Distributed phenomics with the unPAK project reveals the effects of mutations. Plant J.

Schindelin, J. et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods **9**: 676–682.

Schnable, J.C. (2019). Genes and gene models, an important distinction. New Phytol.

Schoft, V.K., Chumak, N., Mosiolek, M., Slusarz, L., Komnenovic, V., Brownfield, L., Twell, D., Kakutani, T., and Tamaru, H. (2009). Induction of RNA-directed DNA methylation upon decondensation of constitutive heterochromatin. EMBO Rep. **10**: 1015–1021.

Schreiber, D.N. and Dresselhaus, T. (2003). In vitro pollen germination and transient transformation of Zea mays and other plant species. Plant Mol. Biol. Rep. **21**: 319–319.

Settles, A.M. et al. (2007). Sequence-indexed mutations in maize using the UniformMu transposon-tagging population. BMC Genomics **8**: 116.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdzić, M., Becker, J.D., Feijó, J.A., and Martienssen, R.A. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell **136**: 461–472.

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

Steffen, J.G., Kang, I.-H., Macfarlane, J., and Drews, G.N. (2007). Identification of genes expressed in the Arabidopsis female gametophyte: Female gametophyte-expressed genes. Plant J. **51**: 281–292.

Steinhorst, L. and Kudla, J. (2013). Calcium - a central regulator of pollen germination and

tube growth. Biochim. Biophys. Acta **1833**: 1573–1581.

Stone, L.M., Seaton, K.A., Kuo, J., and McComb, J.A. (2004). Fast pollen tube growth in Conospermum species. Ann. Bot. **93**: 369–378.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv [cs.CV].

Taghavi Namin, S., Esmaeilzadeh, M., Najafi, M., Brown, T.B., and Borevitz, J.O. (2018). Deep phenotyping: deep learning for temporal phenotype/genotype classification. Plant Methods **14**: 66.

Tardieu, F., Cabrera-Bosquet, L., Pridmore, T., and Bennett, M. (2017). Plant Phenomics, From Sensors to Knowledge. Curr. Biol. **27**: R770–R783.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., and Su, Z. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. **45**: W122–W129.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28**: 511–515.

Tse, Y.C., Mo, B., Hillmer, S., Zhao, M., Lo, S.W., Robinson, D.G., and Jiang, L. (2004). Identification of multivesicular bodies as prevacuolar compartments in Nicotiana tabacum BY-2 cells. Plant Cell **16**: 672–693.

Tsujita, K., Itoh, T., Ijuin, T., Yamamoto, A., Shisheva, A., Laporte, J., and Takenawa, T. (2004). Myotubularin regulates the function of the late endosome through the gram domain-phosphatidylinositol 3,5-bisphosphate interaction. J. Biol. Chem. **279**: 13817–13824.

Ubbens, J., Cieslak, M., Prusinkiewicz, P., Parkin, I., Ebersbach, J., and Stavness, I. (2020). Latent Space Phenotyping: Automatic Image-Based Phenotyping for Treatment Studies. Plant Phenomics **2020**: 5801869.

Ubbens, J.R. and Stavness, I. (2017). Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks. Front. Plant Sci. **8**: 1190.

Uzal, L.C., Grinblat, G.L., Namías, R., Larese, M.G., Bianchi, J.S., Morandi, E.N., and Granitto, P.M. (2018). Seed-per-pod estimation for plant breeding using deep learning. Comput. Electron. Agric. **150**: 196–204.

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., Coppens, F., and Vandepoele, K. (2018). PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. **46**: D1190–D1196.

Vasseur, F., Bresson, J., Wang, G., Schwab, R., and Weigel, D. (2018). Image-based

methods for phenotyping growth dynamics and fitness components in Arabidopsis thaliana. Plant Methods **14**: 63.

Vejlupkova, Z. and Fowler, J.E. (2003). Maize DNA preps for undergraduate students: a robust method for PCR genotyping. Maize Genetics Cooperation Newsletter **77**: 24–25.

Vermeer, J.E.M., van Leeuwen, W., Tobeña-Santamaria, R., Laxalt, A.M., Jones, D.R., Divecha, N., Gadella, T.W.J., Jr, and Munnik, T. (2006). Visualization of PtdIns3 P dynamics in living plant cells : PtdIns3 P dynamics in plant cells. Plant J. **47**: 687–700.

Vidali, L., McKenna, S.T., and Hepler, P.K. (2001). Actin polymerization is essential for pollen tube growth. Mol. Biol. Cell **12**: 2534–2545.

Vollbrecht, E. et al. (2010). Genome-wide distribution of transposed Dissociation elements in maize. Plant Cell **22**: 1667–1685.

Vollbrecht, E. and Hake, S. (1995). Deficiency analysis of female gametogenesis in maize. Dev. Genet. **16**: 44–63.

Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., and Briggs, S.P. (2016). Integration of omic networks in a developmental atlas of maize. Science **353**: 814–818.

Wang, G., Jiang, H., Del Toro de León, G., Martinez, G., and Köhler, C. (2018). Sequestration of a Transposon-Derived siRNA by a Target Mimic Imprinted Gene Induces Postzygotic Reproductive Isolation in Arabidopsis. Dev. Cell **46**: 696–705.e4.

Wang, G., Sun, Y., and Wang, J. (2017). Automatic Image-Based Plant Disease Severity Estimation Using Deep Learning. Comput. Intell. Neurosci. **2017**: 2917536.

Wang, H. et al. (2015). Arabidopsis Synaptotagmin 2 Participates in Pollen Germination and Tube Growth and Is Delivered to Plasma Membrane via Conventional Secretion. Mol. Plant **8**: 1737–1750.

Wang, H., Zhuang, X., Cai, Y., Cheung, A.Y., and Jiang, L. (2013). Apical F-actin-regulated exocytic targeting of NtPPME1 is essential for construction and rigidity of the pollen tube cell wall. Plant J. **76**: 367–379.

Wang, T., Rostamza, M., Song, Z., Wang, L., McNickle, G., Iyer-Pascuzzi, A.S., Qiu, Z., and Jin, J. (2019). SegRoot: A high throughput segmentation method for root image analysis. Comput. Electron. Agric. **162**: 845–854.

Warman, C. and Fowler, J.E. (2019). Custom built scanner and simple image processing pipeline enables low-cost, high-throughput phenotyping of maize ears. bioRxiv: 780650.

Warman, C., Panda, K., Vejlupkova, Z., Hokin, S., Unger-Wallace, E., Cole, R.A., Chettoor, A.M., Jiang, D., Vollbrecht, E., Evans, M.M.S., Slotkin, R.K., and Fowler, J.E. (2020). High expression in maize pollen correlates with genetic contributions to pollen fitness as well as with coordinated transcription from neighboring transposable elements. PLoS

Genet. **16**: e1008462.

Wen, W., Guo, X., Lu, X., Wang, Y., and Yu, Z. (2019). Multi-scale 3D Data Acquisition of Maize. In Computer and Computing Technologies in Agriculture XI (Springer International Publishing), pp. 108–115.

Williams, J.H. and Reese, J.B. (2019). Evolution of development of pollen performance. Curr. Top. Dev. Biol. **131**: 299–336.

Wilson-Sánchez, D., Rubio-Díaz, S., Muñoz-Viana, R., Pérez-Pérez, J.M., Jover-Gil, S., Ponce, M.R., and Micol, J.L. (2014). Leaf phenomics: a systematic reverse genetic screen for Arabidopsis leaf mutants. Plant J. **79**: 878–891.

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**: e00052.

Wolfgruber, T.K. et al. (2009). Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons. PLoS Genet. **5**: e1000743.

Xu, Z. and Dooner, H.K. (2006). The maize aberrant pollen transmission 1 gene is a SABRE/KIP homolog required for pollen tube growth. Genetics **172**: 1251–1261.

Yamada, K., Shimada, T., Nishimura, M., and Hara-Nishimura, I. (2005). A VPE family supporting various vacuolar functions in plants. Physiol. Plant. **123**: 369–375.

Yang, H., Li, Y., and Hua, J. (2006). The C2 domain protein BAP1 negatively regulates defense responses in Arabidopsis. Plant J. **48**: 238–248.

Yang, W.-C., Shi, D.-Q., and Chen, Y.-H. (2010). Female gametophyte development in flowering plants. Annu. Rev. Plant Biol. **61**: 89–108.

Yang, W.-Q., Lai, Y., Li, M.-N., Xu, W.-Y., and Xue, Y.-B. (2008). A novel C2-domain phospholipid-binding protein, OsPBP1, is required for pollen fertility in rice. Mol. Plant **1**: 770–785.

Yazdanbakhsh, N. and Fisahn, J. (2009). High throughput phenotyping of root growth dynamics, lateral root formation, root architecture and root hair development enabled by PlaRoM. Funct. Plant Biol. **36**: 938–946.

Yokotani, N., Ichikawa, T., Kondou, Y., Maeda, S., Iwabuchi, M., Mori, M., Hirochika, H., Matsui, M., and Oda, K. (2009). Overexpression of a rice gene encoding a small C2 domain protein OsSMCP1 increases tolerance to abiotic and biotic stresses in transgenic Arabidopsis. Plant Mol. Biol. **71**: 391–402.

Zhai, J., Zhang, H., Arikit, S., Huang, K., Nan, G.-L., Walbot, V., and Meyers, B.C. (2015). Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. Proc. Natl. Acad. Sci. U. S. A. **112**: 3146–3151.

Zhang, H., Qu, X., Bao, C., Khurana, P., Wang, Q., Xie, Y., Zheng, Y., Chen, N., Blanchoin, L., Staiger, C.J., and Huang, S. (2010). Arabidopsis VILLIN5, an actin filament bundling and severing protein, is necessary for normal pollen tube growth. Plant Cell **22**: 2749–2767.

Zhang, X. et al. (2017). High-Throughput Phenotyping and QTL Mapping Reveals the Genetic Architecture of Maize Plant Growth. Plant Physiol. **173**: 1554–1564.

Zhang, X., Hause, R.J., Jr, and Borevitz, J.O. (2012). Natural Genetic Variation for Growth and Development Revealed by High-Throughput Phenotyping in Arabidopsis thaliana. G3 **2**: 29–34.

Zhong, Y. et al. (2019). Mutation of ZmDMP enhances haploid induction in maize. Nat Plants **5**: 575–580.

Zhou, L.-Z. and Dresselhaus, T. (2019). Chapter Seventeen - Friend or foe: Signaling mechanisms during double fertilization in flowering seed plants. In Current Topics in Developmental Biology, U. Grossniklaus, ed (Academic Press), pp. 453–496.

Zhou, L.-Z., Juranić, M., and Dresselhaus, T. (2017). Germline Development and Fertilization Mechanisms in Maize. Mol. Plant **10**: 389–401.

Zhu, Y., Aoun, M., Krijn, M., and Vanschoren, J. (2018). Data Augmentation using Conditional Generative Adversarial Networks for Leaf Counting in Arabidopsis Plants. BMVC.

Ziamtsov, I. and Navlakha, S. (2019). Machine Learning Approaches to Improve Three Basic Plant Phenotyping Tasks Using Three-Dimensional Point Clouds. Plant Physiol. **181**: 1425–1440.