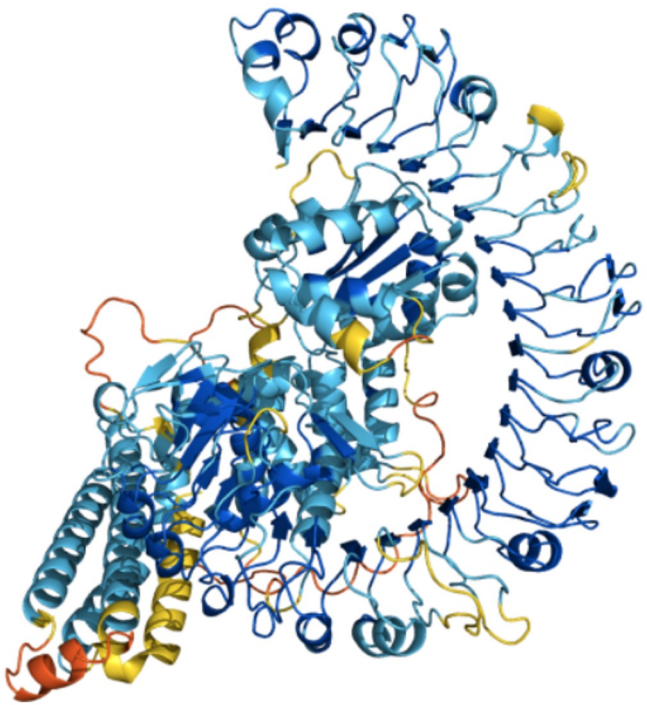
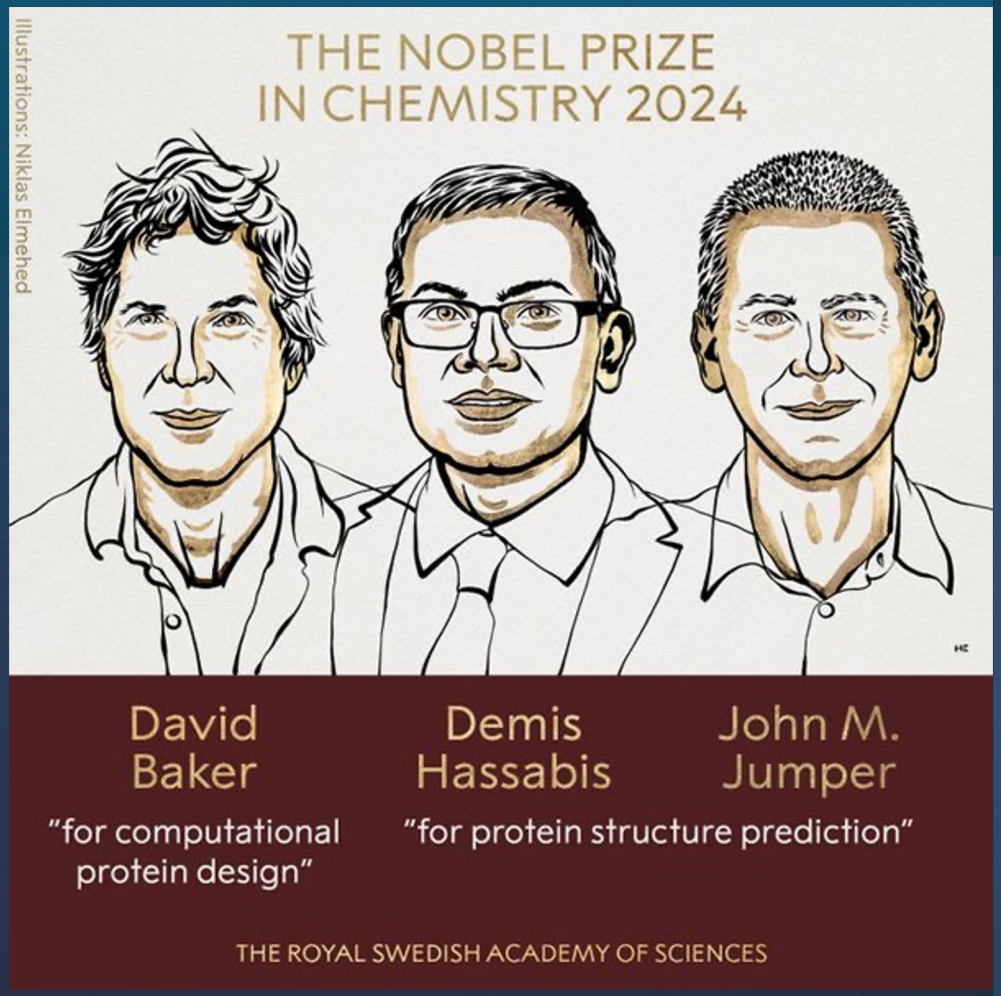


Protein structure, prediction, search, and analysis with AI workshop

(Day 3, November 21, 2024; 1:30 PM - 5:00 PM)

Hosted by the Protein Function and Phenotype Prediction Working Group


Illustrations: Niklas Elmehed
THE NOBEL PRIZE
IN CHEMISTRY 2024
David
Baker
"for computational
protein design"
Demis
Hassabis
John M.
Jumper
"for protein structure prediction"
THE ROYAL SWEDISH ACADEMY OF SCIENCES

AlphaFold 2&3

Google DeepMind

- **Optimal Use Cases**: Ideal for high-accuracy protein structure prediction, especially for proteins with extensive homologous sequences; requires substantial computational resources.
- **SCINet Accessibility**: Databases and singularity files are accessible on Ceres and Atlas for streamlined usage. Full GPU node.
- **Key Features**: Advanced deep learning architecture, high prediction accuracy, reliance on MSAs and templates, attention mechanisms for long-range interactions, and broad applicability across complex proteins. Now supports a variety of biomolecules, including proteins, DNA, RNA, and ligands.

mit-ll/**OmegaFold**

Meta

OmegaFold & ESMFold

- **Optimal Use Cases**: OmegaFold and ESMFold are ideal for rapid, resource-efficient structure prediction, especially for proteins with limited evolutionary data or no homologs, such as orphan genes or fast-evolving sequences.
- **SCINet Accessibility**: OmegaFold is available as modules on Ceres OmegaFold and ESMFold will need the installation of code and packages for Atlas. ESMFold requires Full GPU node. Omega fold uses either a Full GPU node or CPU node.
- **Key Features**: Both models use single sequence input without MSAs or templates, enabling quick predictions. OmegaFold typically provides higher accuracy, while ESMFold prioritizes speed, making them effective for large-scale analyses and novel protein research where approximate models are acceptable.

FoldSeek

Martin Steinegger Lab - Seoul National University

- **Optimal Use Cases**: Ideal for large-scale structural comparisons in protein biology, FoldSeek enables rapid, high-throughput screening of protein structures, aiding in functional annotation and evolutionary studies where sequence alignment is insufficient.
- **SCINet Accessibility**: Installation required on Ceres or Atlas. CPU node.
- **Key Features**: Ultra-fast structural alignment with sensitivity for low sequence identity, scalable to large datasets, and efficient in resource use. FoldSeek integrates with structural databases, supports multiple output formats, and offers a command-line interface for easy workflow integration.

ESM-Variants

Vasilis Ntranos Lab - University of California

- **Optimal Use Cases:** ESM-Variant is ideal for predicting the functional impact of missense mutations, especially in proteins without evolutionary or structural data, such as novel proteins or engineered variants. Useful for disease-associated mutation studies and large-scale variant analysis.
- **SCINet Accessibility:** Requires installation of code and Python packages on Ceres or Atlas. CPU node, Full GPU node, or GPU-mig node.
- **Key Features**: pLM-based predictions, single sequence input, and quantitative scoring for amino acid substitutions. Offers high-throughput capability, requires no structural data, and is easy to integrate into bioinformatics pipelines.

ESM-Variants

RFDiffusion

INSTITUTE FOR Protein Design
UNIVERSITY of WASHINGTON

David Baker'sLab - University of Washington

- **Optimal Use Cases:** RFDiffusion is designed for de novo protein binder creation, targeting specific protein regions like active sites or regulatory interfaces, making it ideal for precision binding studies and cases with no natural binders.
- **SCINet Accessibility:** Installation required on Ceres or Atlas. Full GPU node.
- **Key Features:** Employs diffusion probabilistic models to design novel binders with specified hot spots, customizable constraints, and high-throughput capability. Integrates structural data to enhance binder accuracy and is open-source for flexibility and experimental validation.

Acknowledgements

USDA Agricultural Research Service


Hye-Seon Kim


Stephen Harding


Carson Andorf


Olivia Haley

maizeGDB
Maize Genetics and Genomics Database

Corn Insects and Crop Genetics Research Unit

SCINet Initiative

National Center for Agricultural Utilization Research Unit

Oak Ridge Institute For Science and Education

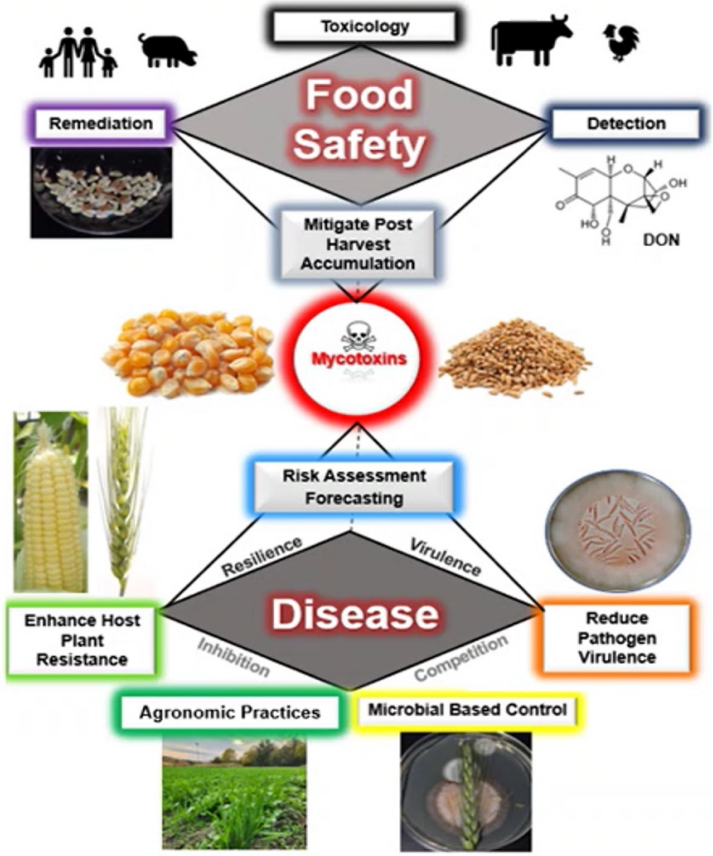
Mycotoxin Prevention & Applied Microbiology

Mission

Deliver scientific solutions to mycotoxin contamination

Mycotoxin Prevention & Applied Microbiology

Mycotoxins in Corn and Wheat



- Mycotoxins are toxic metabolite produced by fungi that contaminate crops.
- If consumed, they cause serious health effects and death of animals.
- ~$1 billion in annual yield losses caused by *Fusarium.*

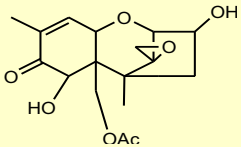
Goal

Identify proteins and metabolites that govern disease and mycotoxin contamination

Mycotoxin Prevention & Applied Microbiology

Fusarium is a plant pathogen on most crops

- **Fusarium Head blight of wheat and barley**
- **Corn ear rot**
- **Trichothecenes (Known as "Vomitoxin")**



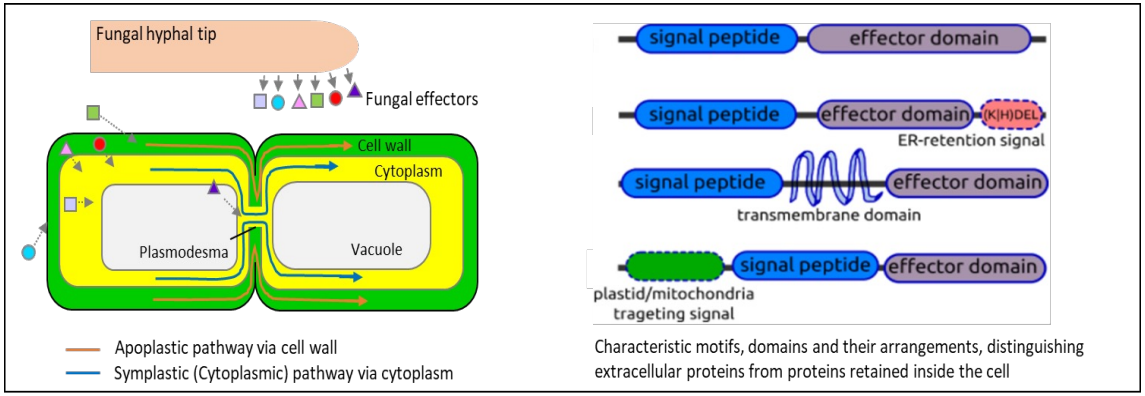
Toxicity: Inhibit Protein Synthesis
Organisms affected: humans, swine, poultry
Symptoms: nausea, vomiting, death

➔ Identify fungal genes that promote virulence and toxin production and investigate molecular mechanisms involved in plant-pathogen interactions.

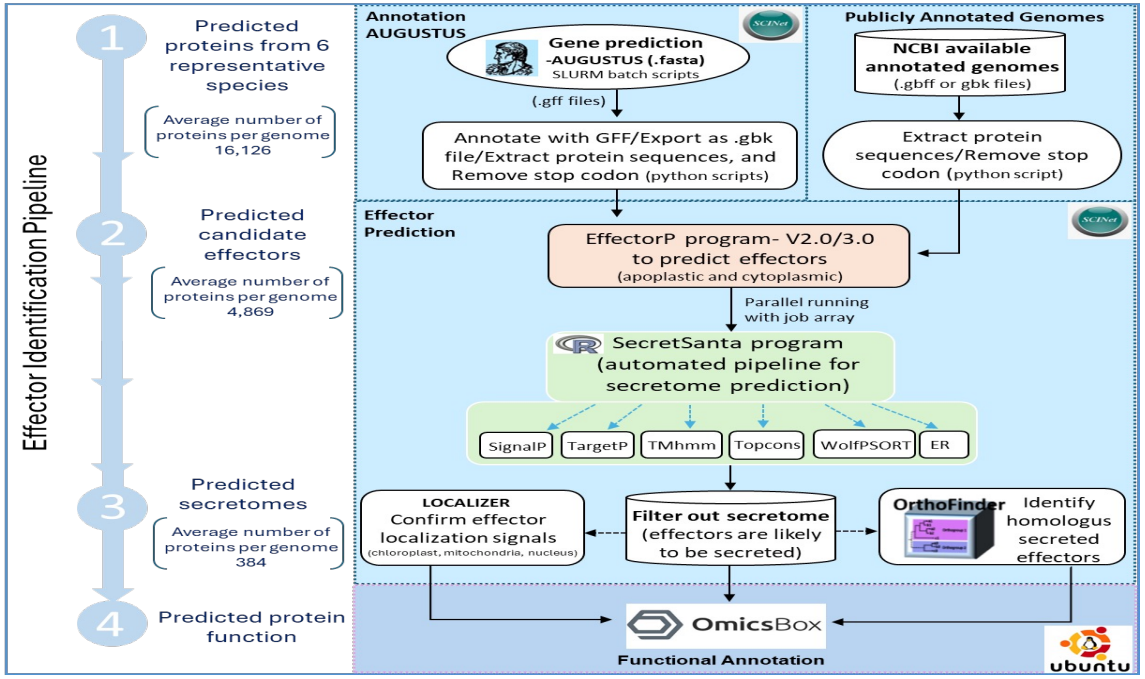
Approach

Develop pipeline to identify effectors of *Fusarium* that can be targeted to control disease

Mycotoxin Prevention & Applied Microbiology



Kim *et. al.* 2024 BMC Microbiology



maizeGDB
Maize Genetics and Genomics Database

Problem/Solution

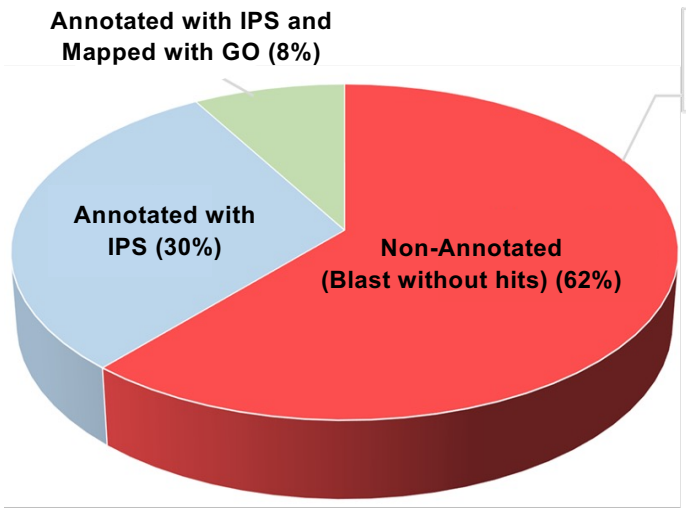
~62 % of putative effectors are hypothetical proteins of unknown function

Mycotoxin Prevention & Applied Microbiology

Identification of protein families of effectors

InterProScan (IPS) identifies protein families →

- Virulence factor
- Pectin lyase
- Peptidase
- Hydrolase family



62% of putative effectors are hypothetical proteins of unknown function.

Artificial intelligence (AI) program AlphaFold predicts 3D structure hypothetical proteins

Several AI/ML tools that uses language models to predict protein structures

AlphaFold

ESMFold

OmegaFold

maizeGDB
Maize Genetics and Genomics Database

WorkFlow & Applications of modeled protein structure

Mycotoxin Prevention & Applied Microbiology

Protein Structure Prediction

DEEPMIND ALPHAFOLD
(Jumper et. al. 2021 Nature)

ESMFold
Meta

mit-ll/**OmegaFold**

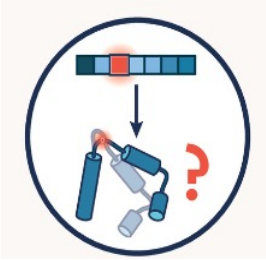
Protein Structure Search

FoldSeek

Comparisons 3D structures/ functional annotation

Variant Effect Predictions

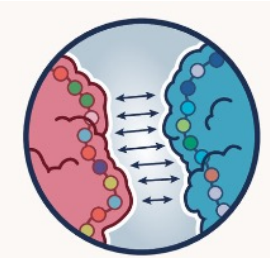
EMS-Variant/PanEffect



Estimating the effect of genetic variations on the protein structure/function

Host-Pathogen Interactions

RFDiffusion



Modeling interfaces of proteins that create protein-protein interactions

maizeGDB
MAIZE GENETICS AND GENOMICS DATABASE

Web-based resources tools & Databases

Mycotoxin Prevention & Applied Microbiology

PanEffect: a pan-genome visualization tool for variant effects in maize FREE

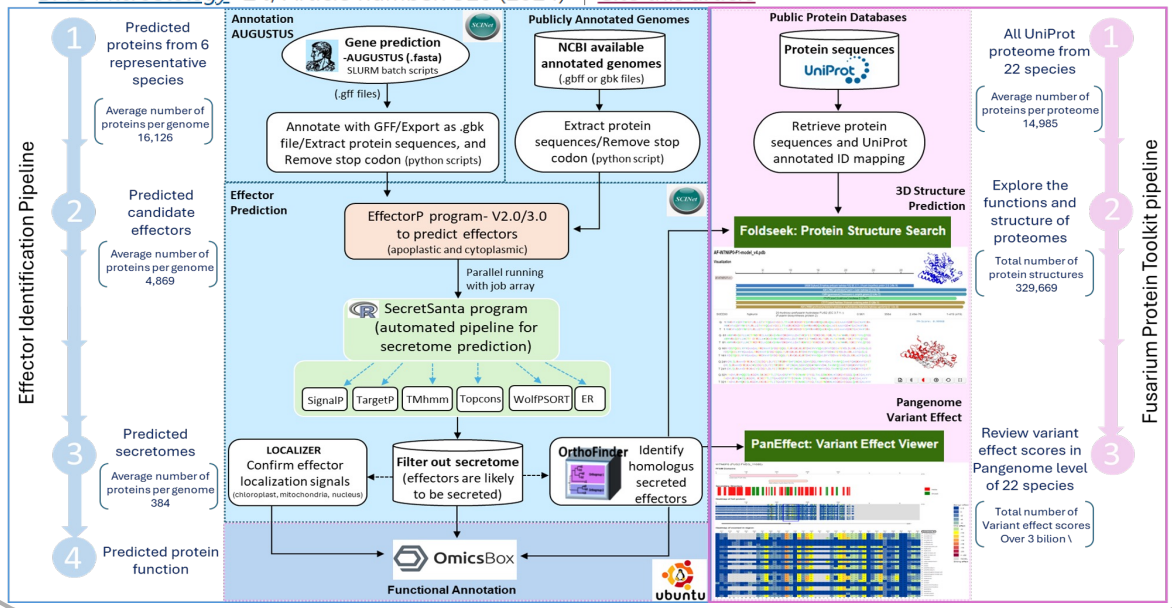
Carson M Andorf ✉, Olivia C Haley, Rita K Hayford, John L Portwood, II, Stephen Harding, Shatabdi Sen, Ethalinda K Cannon, Jack M Gardiner, Hye-Seon Kim, Margaret R Woodhouse
https://doi.org/10.1093/bioinformatics/btae073

Bioinformatics, Volume 40, Issue 2, February 2024, btae073,

Fusarium Protein Toolkit: a web-based resource for structural and variant analysis of *Fusarium* species

Hye-Seon Kim, Olivia C. Haley, John L. Portwood II, Stephen Harding, Robert H. Proctor, Margaret R. Woodhouse, Taner Z. Sen & Carson M. Andorf ✉
https://doi.org/10.1186/s12866-024-03480-5

BMC Microbiology **24**, Article number: 326 (2024) | Cite this article



maizeGDB
MAIZE GENETICS AND GENOMICS DATABASE

Web-based resources tools & Databases

Mycotoxin Prevention & Applied Microbiology

Application of RFdiffusion to predict interspecies protein-protein interactions between fungal pathogens and cereal crops

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

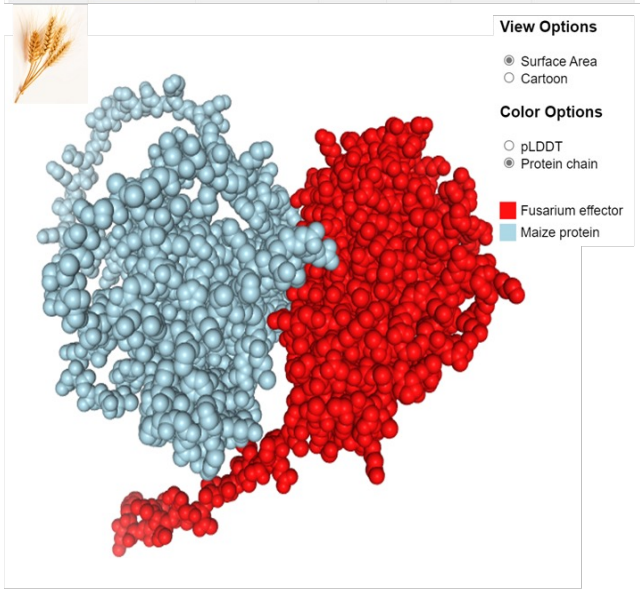
Olivia C. Haley, Stephen Harding, Taner Z. Sen, Margaret R. Woodhouse, Hye-Seon Kim, Carson Andorf
https://doi.org/10.1101/2024.09.17.613523

Fusarium-Wheat Interaction Database

https://sandbox.maizegdb.org/multimer/table_wheat.html

Predicted host/pathogen protein-protein using Fusarium effectors

EFFECTOR	PROTEIN	TYPE	IDENTITY	E-VALUE	BIT SCORE	ACTION
I1RHE5	A0A3B6KG28	Fusarium / Wheat	0.27	0.123	105	View Structure
I1RFB7	A0A3B6RN88	Fusarium / Wheat	0.176	0.186	105	View Structure
I1RFB7	W5I2D5	Fusarium / Wheat	0.176	0.186	105	View Structure
I1S2L3	A0A7H4LIP4	Fusarium / Wheat	0.272	0.266	100	View Structure



Fusarium-Maize Interaction Database

https://sandbox.maizegdb.org/multimer/

Predicted host/pathogen protein-protein using Fusarium effectors

EFFECTOR (UNIPROT)	EFFECTOR (ID)	MAIZE PROTEIN	TYPE	IDENTITY	E-VALUE	BIT SCORE	ACTION
Q876W5	N/A	Zm00001eb155870_P002	Fusarium -> Maize	0.218	1.36E-01	109	View structures
I1S5J8	FGSG_12119	Zm00001eb347240_P002	Fusarium -> Maize	0.125	0.1267	104	View structures
I1S2L3	FGSG_11011	Zm00001eb119510_P002	Fusarium -> Maize	0.236	0.1886	101	View structures