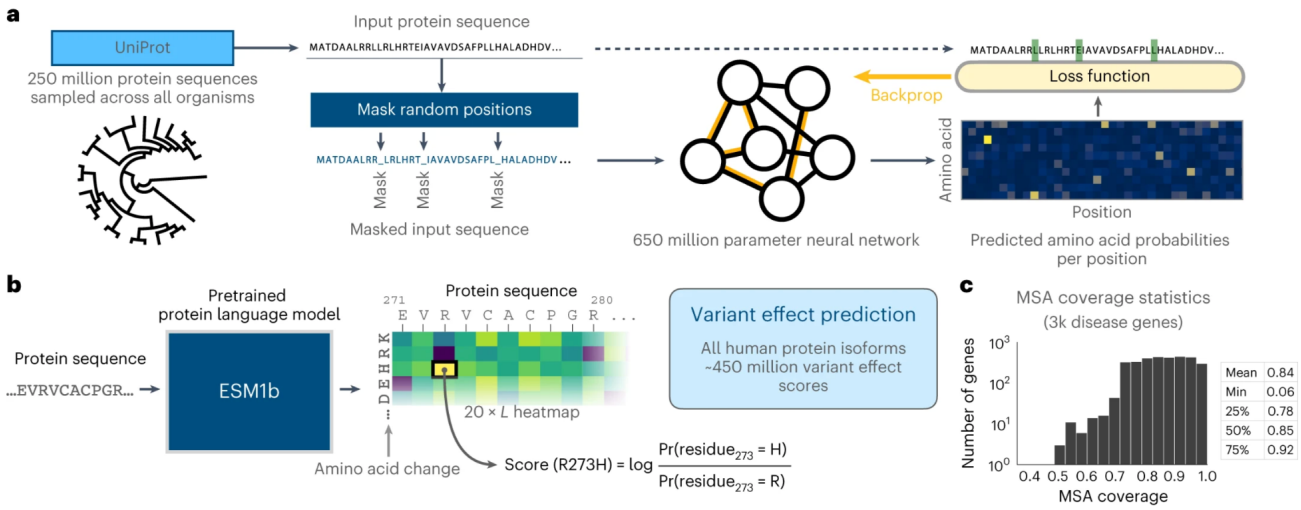
ESM-variant (Protein Variant Effect Prediction)

Task #1: Use ESM-variants to predict protein structures.



ESM-variant workflow, (a) ESM1b is a 650-million-parameter protein language model trained on 250 million protein sequences across all organisms. The model was trained via the masked language modeling task, where random residues are masked from input sequences and the model has to predict the correct amino acid at each position/. From Figure 1 of "Genome-wide prediction of disease variant effects with a deep protein language model." (Brandes, et al. 2023).

Build local files

```
cp -r /90daydata/shared/protein_structure_workshop/ESMVariant .
cd ESMFOLD
```

Run ESM-Variants

```
sbatch esm_variants.sh ./fasta/maize5.fasta ./variants/maize5_variants.tsv
```

```
#!/bin/bash -l
#SBATCH --account=scinet_workshop1              # Account name
#SBATCH --partition=gpu-a100-mig7               # Partition A100 MIG
#SBATCH --job-name=esm-variant                  # Job name
#SBATCH --output=./log/ESMVariant.%J.out        # Standard output
#SBATCH --error=./log/ESMVariant.%J.err         # Standard error
#SBATCH -t 04:00:00                             # Time limit for the job
#SBATCH --mem=32GB                              # CPU memory allocation
#SBATCH --ntasks=2                              # Number of tasks
#SBATCH --gres=gpu:a100_1g.10gb:1               # Request 1 A100 MIG GPU
#SBATCH --reservation=forum-gpu                 # Reservation

# Load necessary modules
module load miniconda3                          # Load Miniconda
module load cuda                                # Load CUDA for GPU support
module load python/3.12.5                       # Load Python

# Activate the ESMVariant environment
source activate /90daydata/shared/protein_structure_conda/esmvariant_env
export TORCH_HOME=/90daydata/shared/protein_structure_conda/.cache/

# Run the Python script for scoring missense mutations, passing input FASTA
and output CSV as arguments
python esm_score_missense_mutations.py --input-fasta-file $1 --output-csv-
file $2

# Deactivate the Conda environment after the script completes
conda deactivate

date                                            # Print the date and time
```

Check the status of slurm job:

```
squeue -u $USER
```

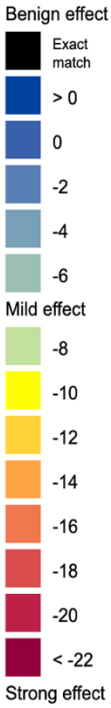
View log files:

```
cd log
ls -ltrh
tail ESMVariant.<JOBID>.out
tail ESMVariant.<JOBID>.err
cd ..
```

ESMVariant.<JOBID>.err output:
100%|| 5/5 [00:05<00:00, 1.18s/it]

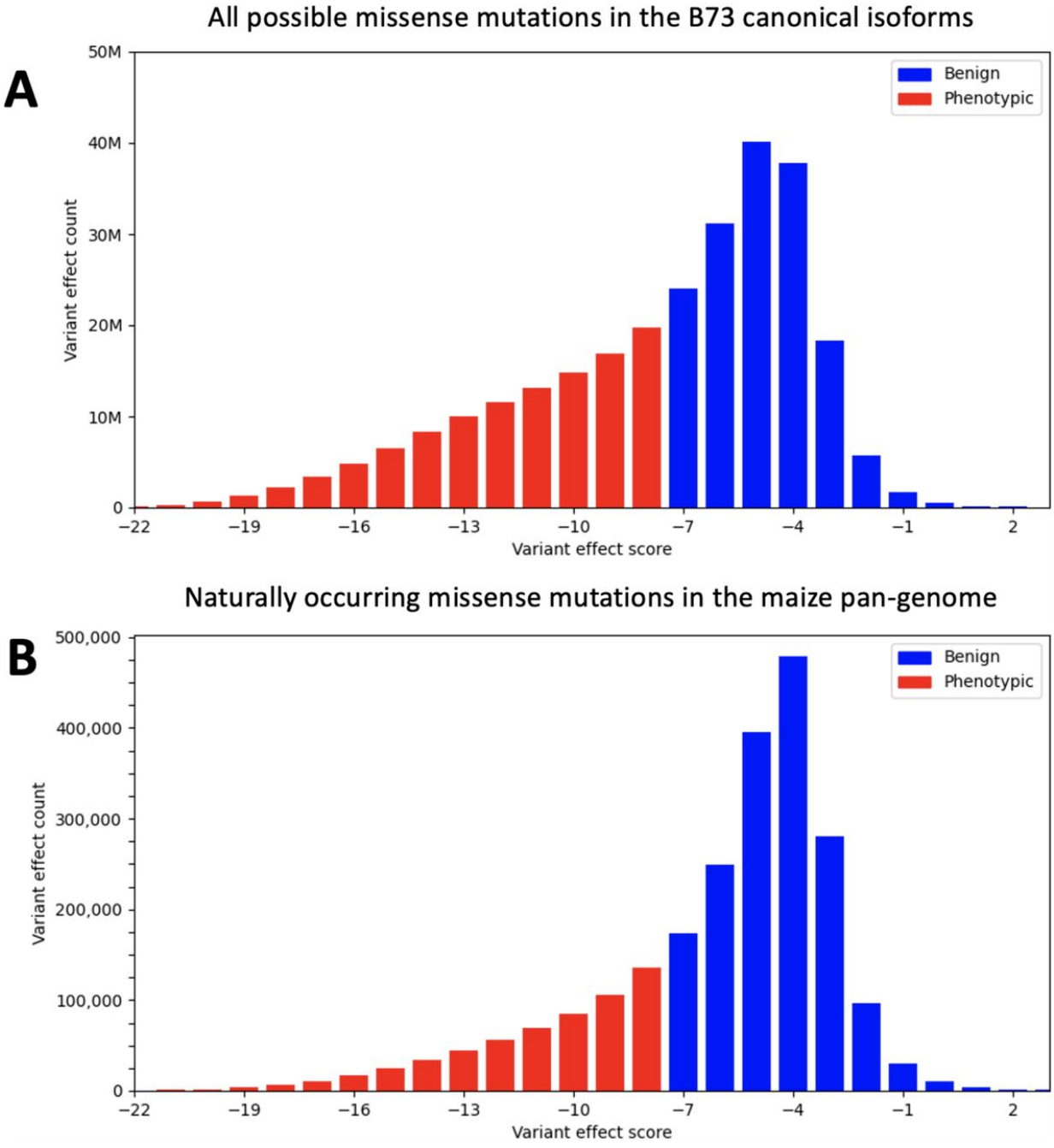
View variant output file:

```
head ./variants/maize5_variants.tsv
```




```
seq_id,mut_name,esm_score
Zea_mays_sugary1,M1K,-12.793287
Zea_mays_sugary1,M1R,-12.045446
Zea_mays_sugary1,M1H,-14.241626
Zea_mays_sugary1,M1E,-12.277544
Zea_mays_sugary1,M1D,-13.52236
Zea_mays_sugary1,M1N,-14.032594
Zea_mays_sugary1,M1Q,-12.64518
Zea_mays_sugary1,M1T,-13.319958
Zea_mays_sugary1,M1S,-12.58724
```

Variant effect scores in Maize:



Task #2: Explore the Maize and Fusarium PanEffect instances

- https://www.maizegdb.org/effect/maize/
- https://fusarium.maizegdb.org/

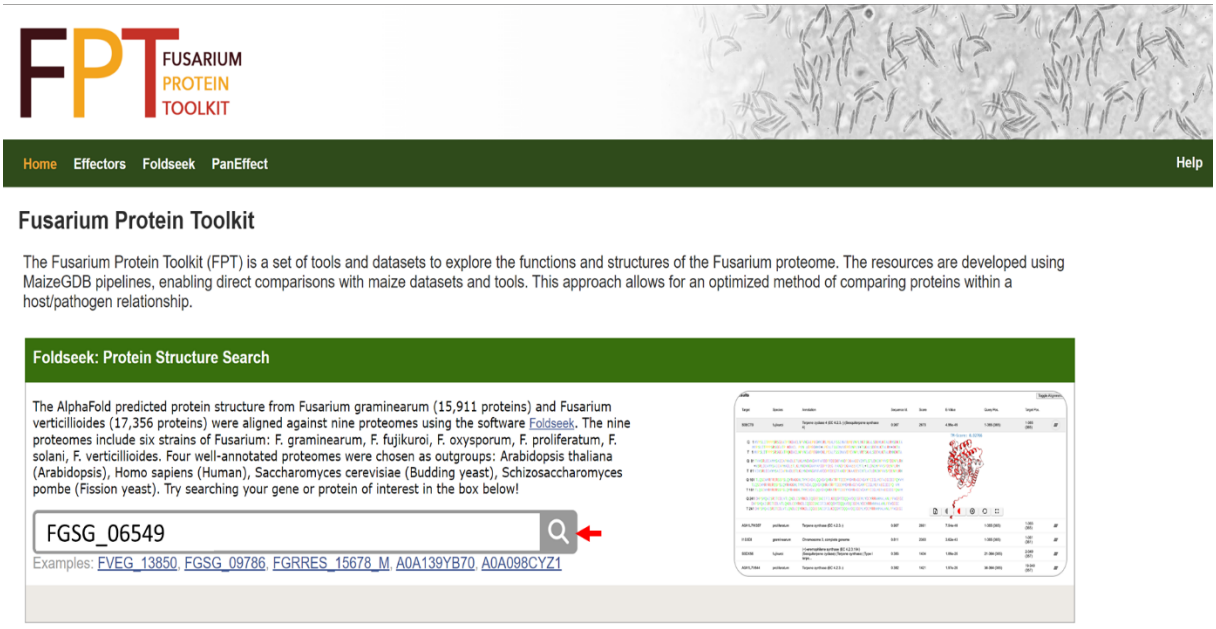
Maize examples:

- Zm00001eb174590 (sugary1)
- Zm00001eb374090 (shrunken1)
- Zm00001eb271860 (yellow endosperm1)
- Zm00001eb313510 (glossy1)
- Zm00001eb378140 (waxy1)

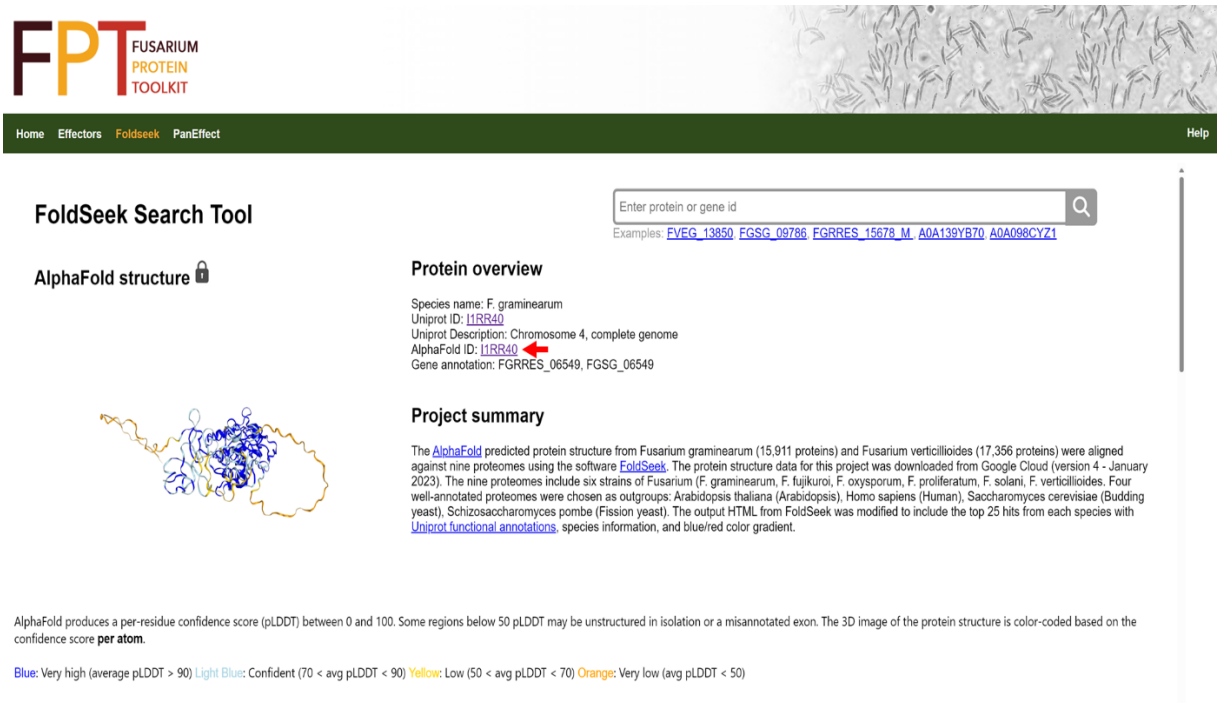
Fusarium examples:

- I1RR40 (FGSG_06549)
- I1RIB2 (FGSG_03543)
- I1RKZ1 (FGSG_04563)
- I1S303 (FGSG_11164)

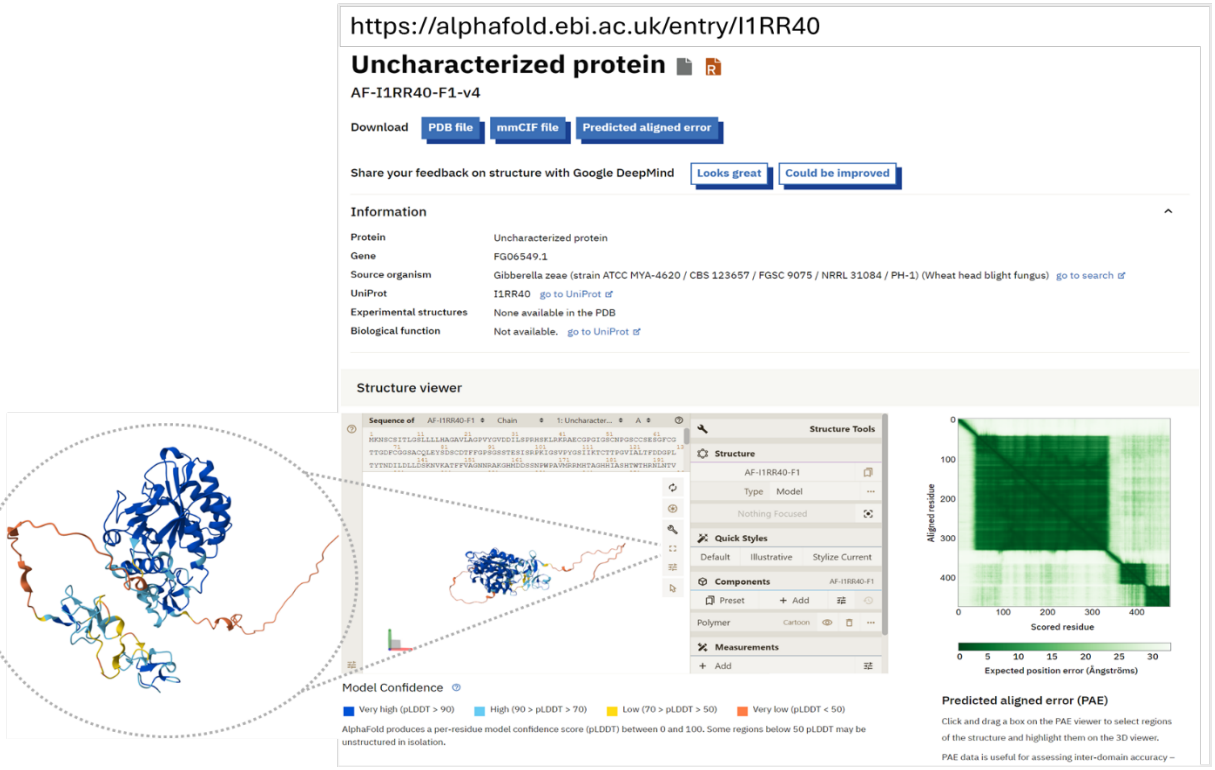
Fusarium PanEffect instances (https://fusarium.maizegdb.org/)



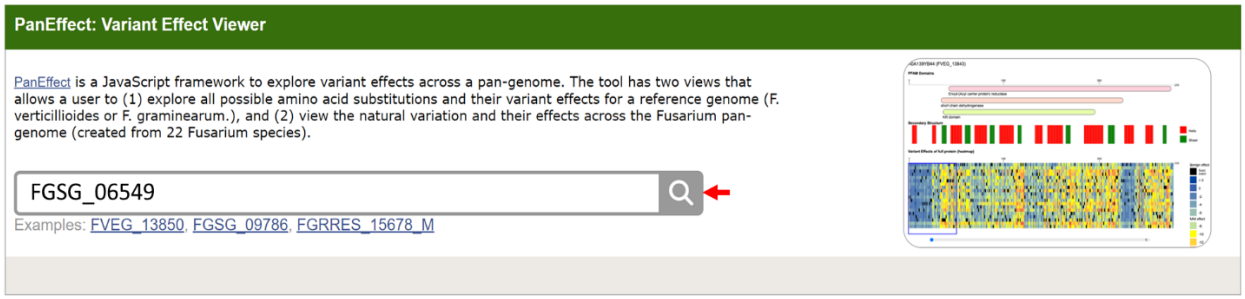
⇨ You can search a protein of interest by typing locus_tag prefix of FGSG for *F. graminearum*.



⇨ After clicking the search button, it will get to this page. The detail information about AlphaFold search also can be viewed by clicking the link of the AlphaFold ID.

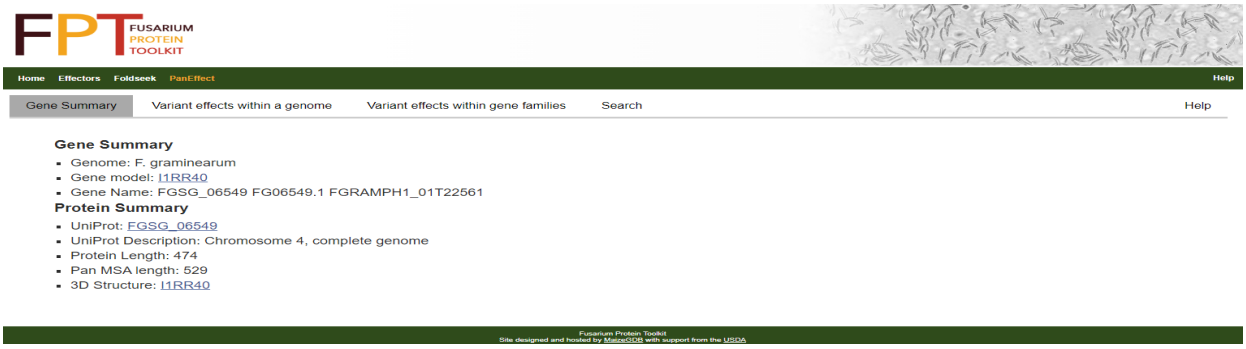


What are variant effects? What kinds of mutations in the genes or proteins can cause strong phenotypic results? Understanding structural variations in Fusarium proteins is also crucial for deciphering evolution, their pathogenicity, and virulence phenotypes. It will enable us to explore missense variation in amino acid sequences.



⇨ You can search a protein of interest by typing locus_tag prefix of FGSG for *F. graminearum*.

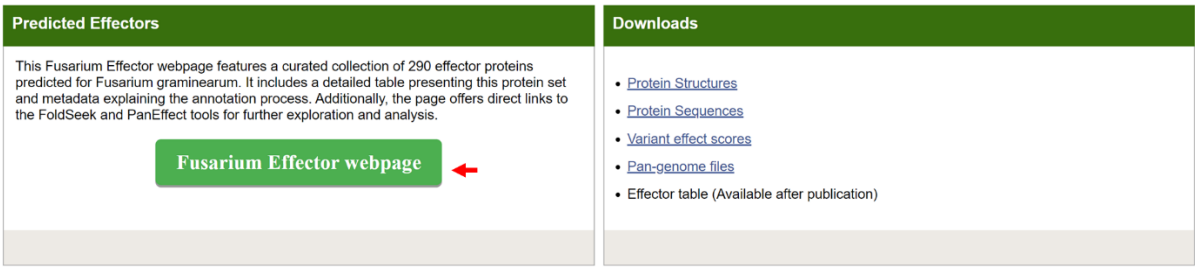
1. Gene Summary view



2. Variant effects within a genome view

3. Variant effects within gene families

⇨ Here is drop down menu to select either species or gene name for reviewing the variant effects in 22 Fusarium proteome.



⇨ You can search an effector protein of interest by typing locus_tag (e.g. FGSG_06549)

Protein Name	UniProt	Apoplastic	Cytoplasmic	Localizer Chloroplast	Localizer Mitochondria	Localizer Nucleus	Description	Gene Ontology (GO)	Enzyme Codes	Links
FGSG_06549	I1RR40	0.663					related to chitin binding	F:GO:0016810: hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds P:GO:0045493: xylan catabolic process F:GO:0008061: chitin binding F:GO:0016798: hydrolase activity, acting on glycosyl bonds		FoldSeek PanEffect AlphaFold

EMBL-EBI AlphaFold Protein Structures

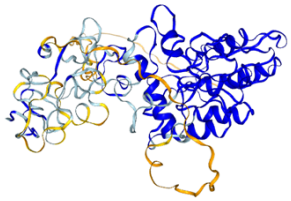
AlphaFold is an artificial intelligence system by DeepMind (reference), and was used to create predictions based on the Fusarium graminearum (15,911 proteins) and Fusarium verticillioides (17,356 proteins) sequences. Enter a gene or protein below to view the predicted AlphaFold structure.

The protein structure is color-coded based on per-residue confidence scores (pLDDT):

- Very high (avg pLDDT > 90)
- Confident (70 < avg pLDDT < 90)
- Low (50 < avg pLDDT < 70)
- Very low (avg pLDDT < 50)

FGSG_06549

Examples: FVEG_13850, FGSG_09786, FGRRES_15678_M, A0A139YB70, A0A098CYZ1



Protein overview

Uniprot ID: I1RR40

AlphaFold ID: I1RR40

Species: F. graminearum

Uniprot Description: Chromosome 4, complete genome

Gene annotation: FGRRES_06549

MetaAI ESMFold Protein Structures

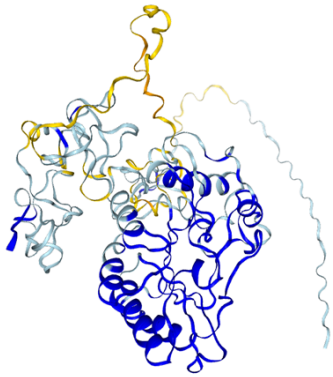
ESMFold is an artificial intelligence system by MetaAI (reference), and was used to create predictions of Fusarium graminearum (15,911 proteins) and Fusarium verticillioides (17,356 proteins) sequences. Enter a gene or protein below to view the predicted ESMFold structure.

The protein structure is color-coded based on per-residue confidence scores (pLDDT):

- Very high (avg pLDDT > 90)
- Confident (70 < avg pLDDT < 90)
- Low (50 < avg pLDDT < 70)
- Very low (avg pLDDT < 50)

FGSG_06549

Examples: FVEG_13850, FGSG_09786, FGRRES_15678_M, A0A139YB70, A0A098CYZ1



Protein overview

Uniprot ID: I1RR40

Protein: NA

Species: F. graminearum

Uniprot Description: Chromosome 4, complete genome

Gene annotation: FGRRES_06549