plantdeepsea Documentation

—

Jan 19, 2023

OVERVIEW

Welcome! This is the documentation for PlantDeepSEA, a deep learning web service based on Selene Python SDK and Django. The web sever address is located here. PlantDeepSEA is freely accessible for all users.

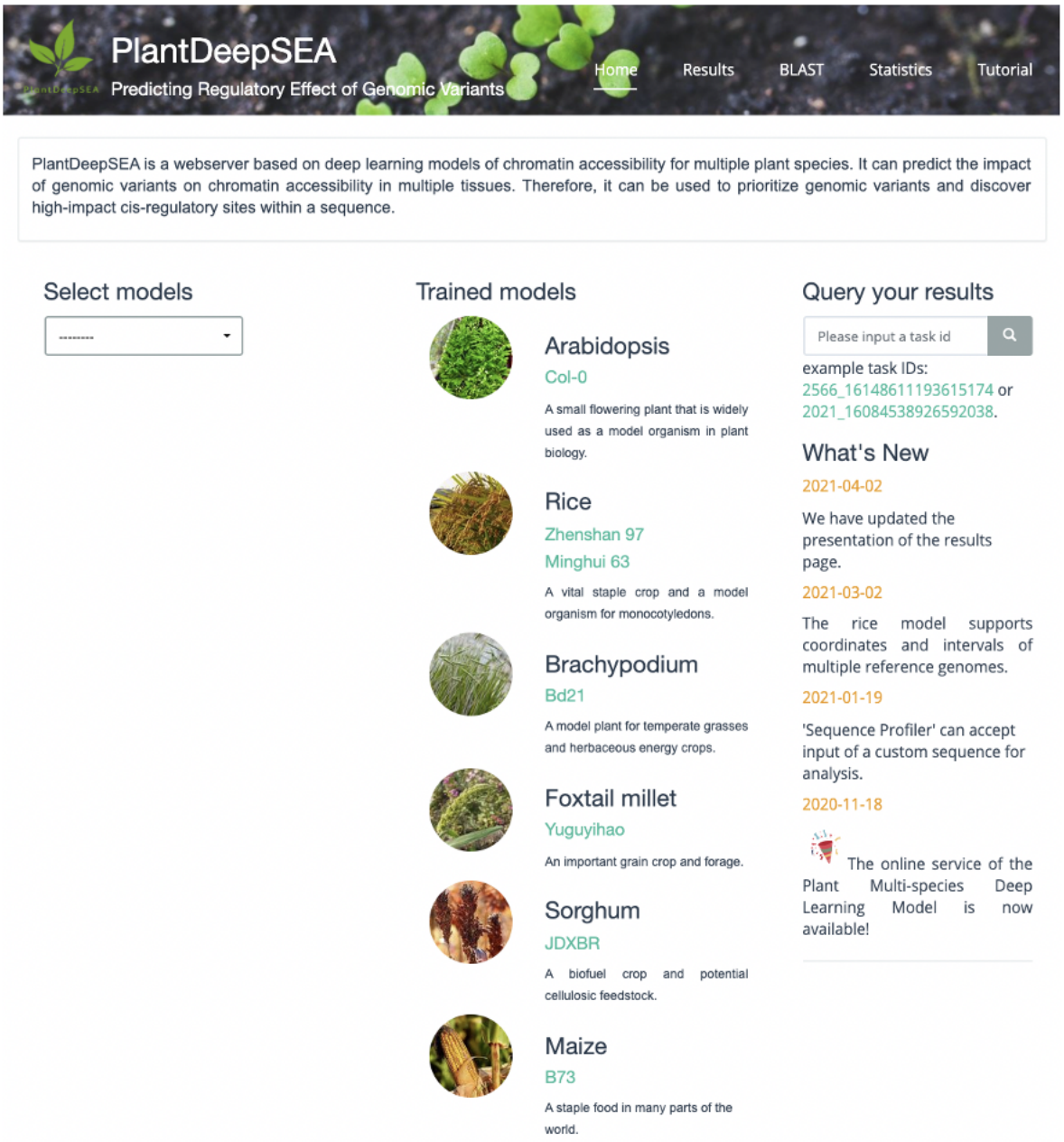
Citation : Hu Zhao#, Zhuo Tu#, Yinmeng Liu, Zhanxiang Zong, Jiacheng Li, Hao Liu, Feng Xiong, Jinling Zhan, Xuehai Hu, and Weibo Xie* (2021). PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. *Nucleic Acids Research*, doi: 10.1093/nar/gkab383

CHAPTER

ONE

HOME PAGE

The appearance of the home page is shown in the figure below.



Here, you can select the model, go to result, statistics, tutorial, blast Interface.

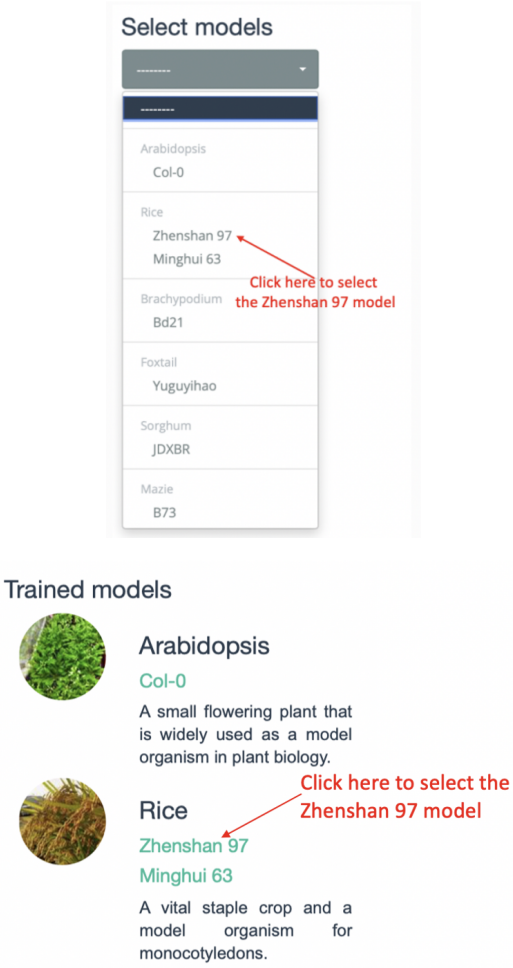
You can click the home button to return to the home page.

CHAPTER

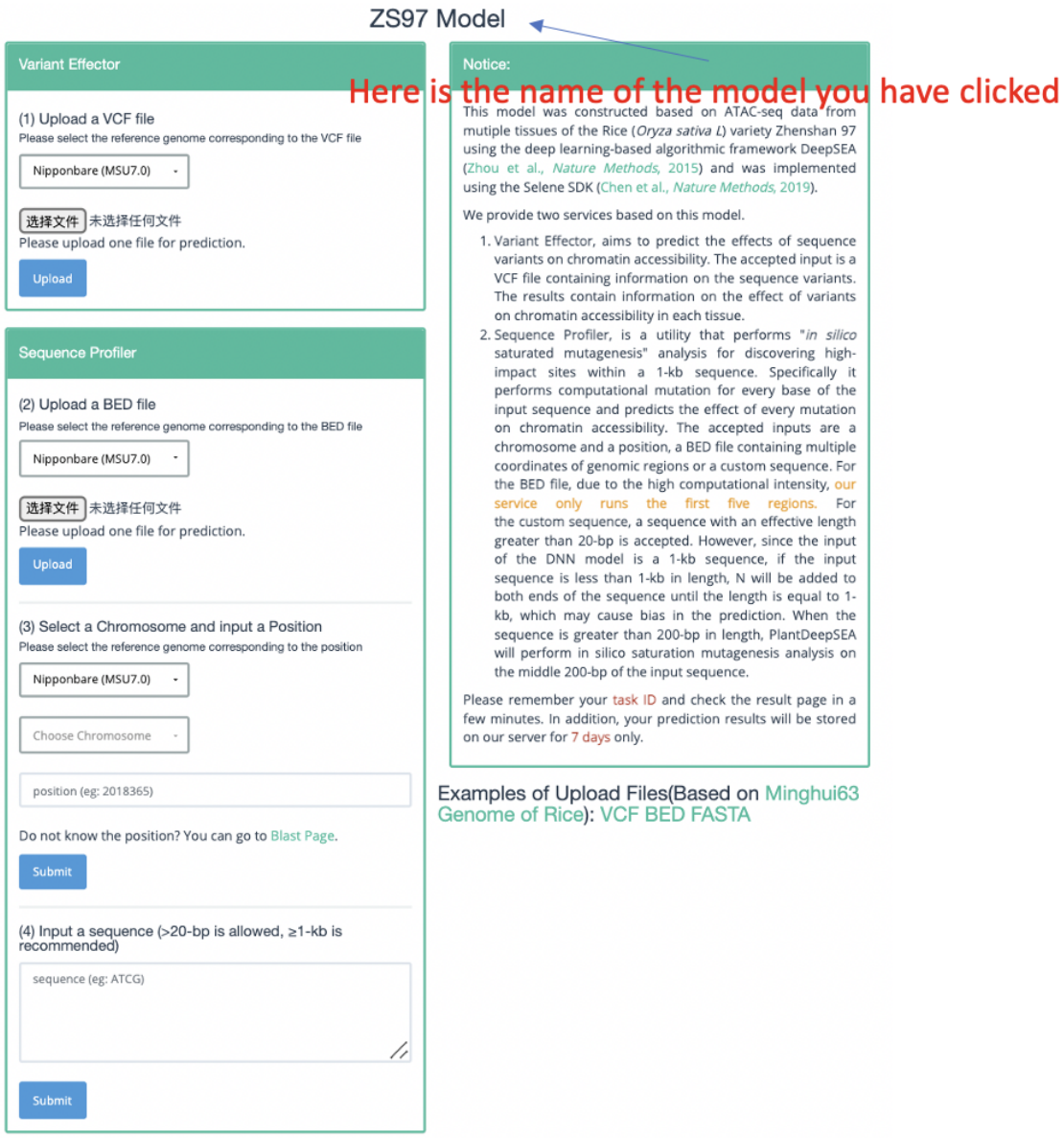
TWO

MODEL SELECTION

You can click the drop-down box on the left side of the home page or the species reference genome name in the middle column to select the appropriate model for prediction (as shown in the figure below).



Then, jump to the prediction interface, where you will see the model name you have chosen (as shown in the following figure).


ZS97 Model
Here is the name of the model you have clicked
Variant Effector
(1) Upload a VCF file
Please select the reference genome corresponding to the VCF file
Nipponbare (MSU7.0)
选择文件 未选择任何文件
Please upload one file for prediction.
Upload
Sequence Profiler
(2) Upload a BED file
Please select the reference genome corresponding to the BED file
Nipponbare (MSU7.0)
选择文件 未选择任何文件
Please upload one file for prediction.
Upload
(3) Select a Chromosome and input a Position
Please select the reference genome corresponding to the position
Nipponbare (MSU7.0)
Choose Chromosome
position (eg: 2018365)
Do not know the position? You can go to Blast Page.
Submit
(4) Input a sequence (>20-bp is allowed, ≥1-kb is recommended)
sequence (eg: ATCG)
Submit
Notice:
This model was constructed based on ATAC-seq data from mutiple tissues of the Rice (Oryza sativa L) variety Zhenshan 97 using the deep learning-based algorithmic framework DeepSEA (Zhou et al., Nature Methods, 2015) and was implemented using the Selene SDK (Chen et al., Nature Methods, 2019).
We provide two services based on this model.
1. Variant Effector, aims to predict the effects of sequence variants on chromatin accessibility. The accepted input is a VCF file containing information on the sequence variants. The results contain information on the effect of variants on chromatin accessibility in each tissue.
2. Sequence Profiler, is a utility that performs "in silico saturated mutagenesis" analysis for discovering high-impact sites within a 1-kb sequence. Specifically it performs computational mutation for every base of the input sequence and predicts the effect of every mutation on chromatin accessibility. The accepted inputs are a chromosome and a position, a BED file containing multiple coordinates of genomic regions or a custom sequence. For the BED file, due to the high computational intensity, our service only runs the first five regions. For the custom sequence, a sequence with an effective length greater than 20-bp is accepted. However, since the input of the DNN model is a 1-kb sequence, if the input sequence is less than 1-kb in length, N will be added to both ends of the sequence until the length is equal to 1-kb, which may cause bias in the prediction. When the sequence is greater than 200-bp in length, PlantDeepSEA will perform in silico saturation mutagenesis analysis on the middle 200-bp of the input sequence.
Please remember your task ID and check the result page in a few minutes. In addition, your prediction results will be stored on our server for 7 days only.
Examples of Upload Files(Based on Minghui63 Genome of Rice): VCF BED FASTA

CHAPTER

THREE

SUBMIT TASK

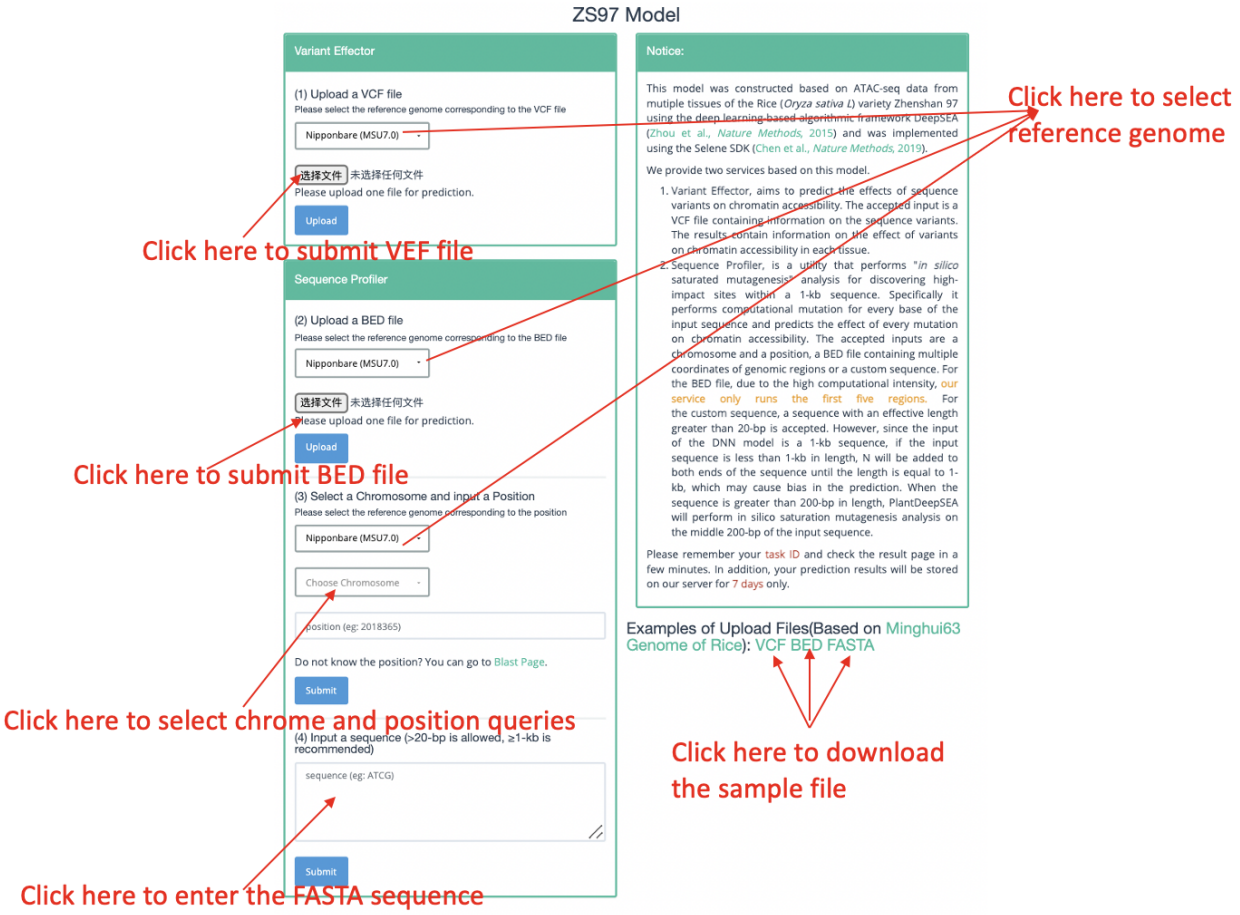
There are three ways to submit a task.

For Variant Effector, you can upload a VCF for query. The uploaded file must be a .vcf file ending. The format of the file can be viewed by downloading the sample file in the lower right corner.

For Sequence Profiler, you can:

(1) Upload a BED for query. The uploaded file must be a BED file ending. The format of the file can be viewed by downloading the sample file in the lower right corner. For BED files, we only predict the first 5 lines of the file due to limited GPU resources.

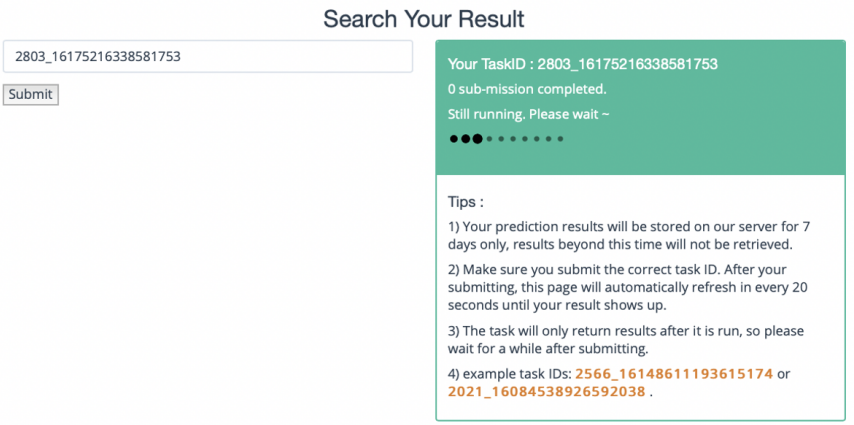
(2) Using chromosome and position queries. The results will output the effect on the open region of chromatin before and after the 200 bp sequence mutation around the position.



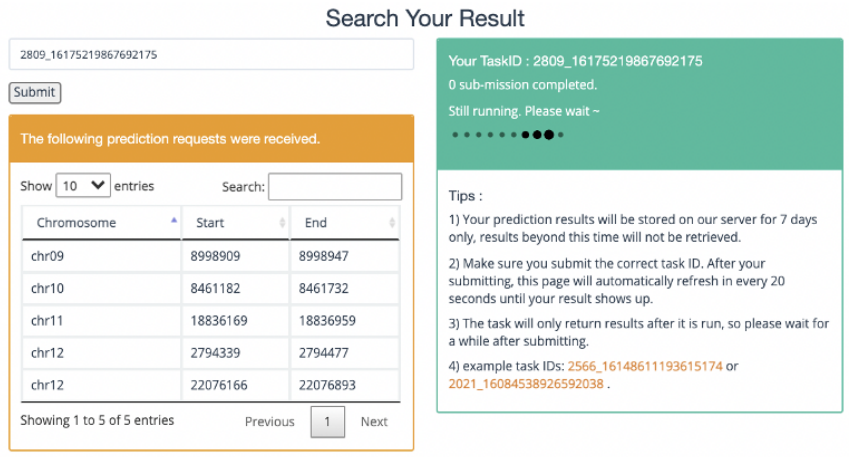
Note:Only rice model supports coordinates and intervals of multiple reference genomes.

After the task is submitted, it will return to Task ID and run in the background queue. After the task is submitted successfully, it will jump to the waiting interface as shown in the figure below.

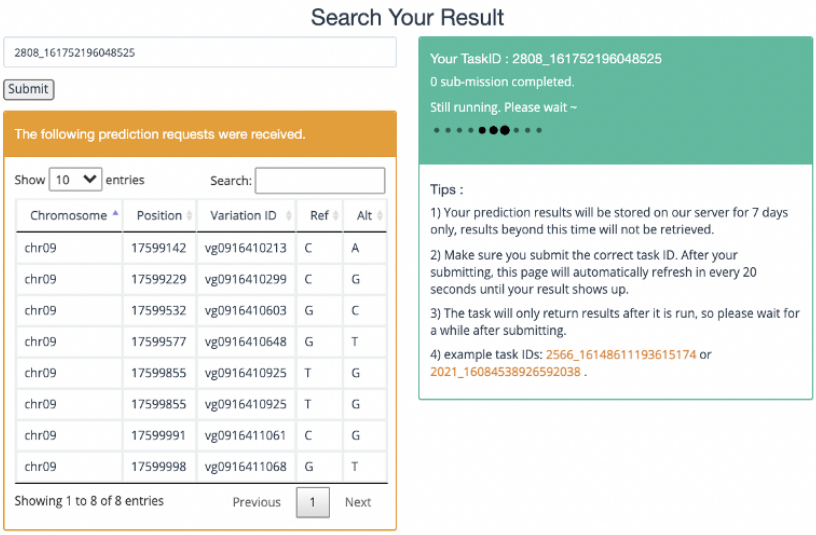
For sequence or position input:



For BED input:



For VCF input:



Remember your task ID and query the results after a while (a few minutes), and your prediction result will store on our server for only 7 days.

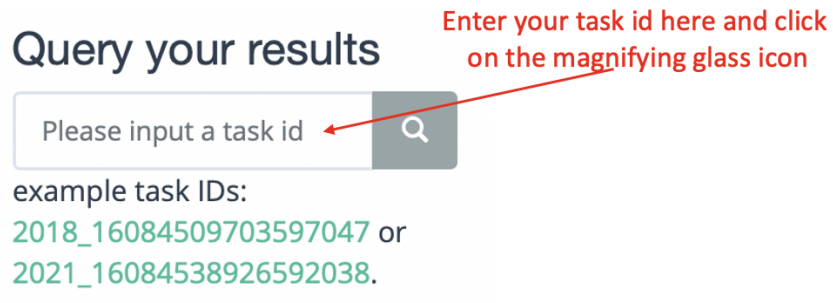
CHAPTER

FOUR

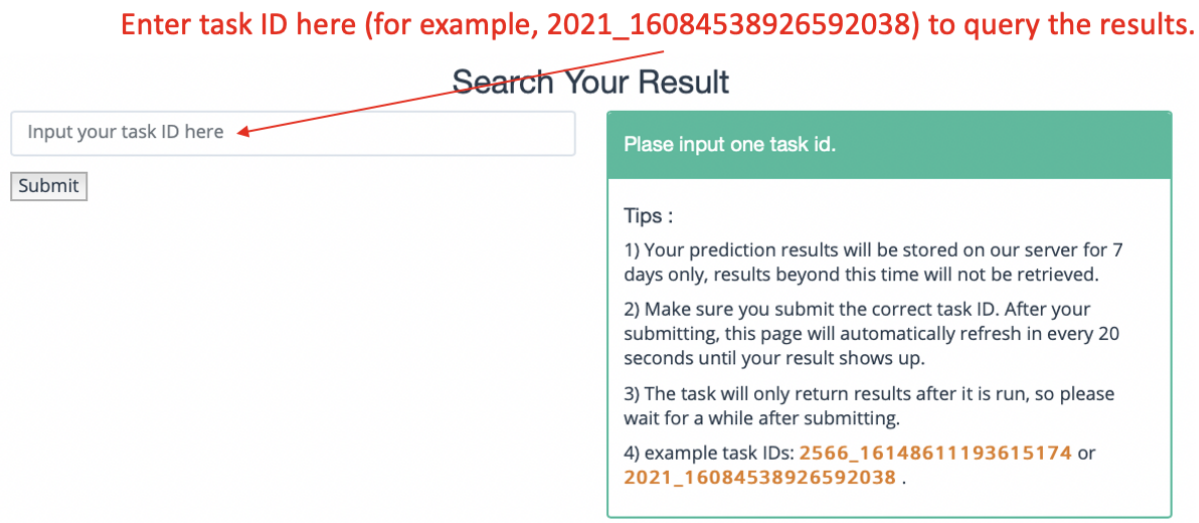
QUERY RESULTS

There are two ways to query the results of submitted tasks.

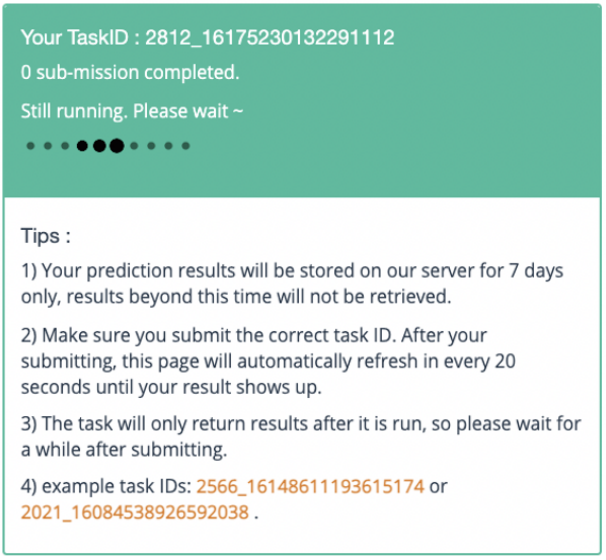
You can enter your task ID in the query your results column on the right side of the home page and click on the magnifying glass icon (as shown in the figure below).



You can also click the results button on the home page to enter the result query page, and then enter the task ID in the interface and submit.



After you submit the query, the information as shown in the following figure will be displayed on the right side of the interface.



The interface will refresh every 10 seconds until the result is returned.

CHAPTER

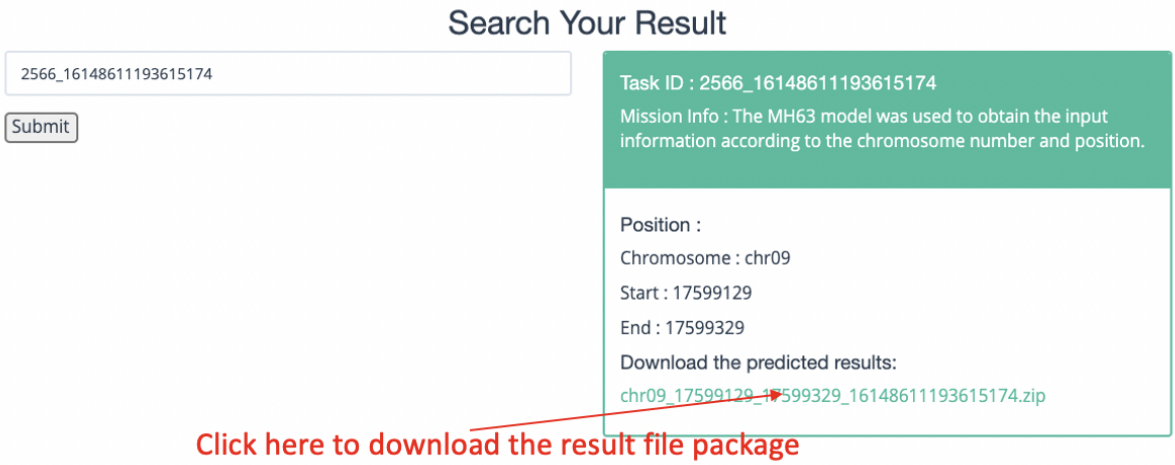
FIVE

INTERPRETATION OF RESULTS

After the task is completed, enter the task ID query to obtain the model prediction results.

5.1 Brief Description

The left side of the upper part of the result interface is the task ID query column, and the right side shows the download link of the forecast result compression package and task information (as shown in the figure below,take chromosome number and position input as an example).



The lower part of the result interface is a visual display of the prediction results generated. For different task types, the returned results are slightly different, see the **Main Functions** part of this section for details. Here are two examples (see the figure below) for reference.

Enter the chromosome number and position, and return the following result. The visualization result is generated by bokeh.

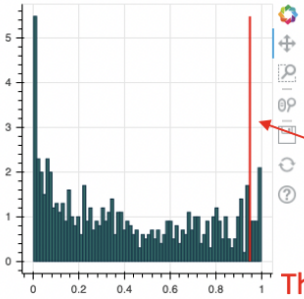
Results:

You can download the predicted results on this page or select one of the sample you would like to display:

ATAC-seq, flag leaf, rep1

Click here to select the sample to display

The currently selected model was trained using the data "ATAC-seq, flag leaf, rep1". The distribution of the prediction scores for 744 sequences labeled as open chromatin regions (OCRs) in the test set is shown below. The prediction score (0.95) of the submitted sequence "chr09:17599129-17599329" is at the 93.7% quantile of this distribution, as shown by the red line.



The more right the red line is, the higher the confidence is.

Sequence used for prediction

The bases shown in the '*in silico* saturation mutagenesis map' section are highlighted (with line width: 100 bp).

ACACAGCTAAAATGAATCGCATATGCAGTGAATGAGACTCTGAGAGCCCAGTTTAAGGTAGGAGTAGCATCAAATTCACTAGTCCTATACACCGGCACGT
CGTCGCAGTGCAATTTTATACGATCTTCATGTGTGCATCAGTGCATGCTGTTTGCAGAAGGGAAGTTGGCCGCCTGCCCGTGCTGAGCTCAACTGAACGC
TGGCTGCCTGCATGGCCAGTTCGTGACTACAGTTACTACGCATACTACCCGGGGTAGCAGCGGCGCGCTATCAGCTCTACCCGGAGTAGGCAGTAGCAGC
AGTGGTGGTAGTAGTAGAAAGCGCGGCAGTGGAAGGGATAGAGGAGTGCCCACATTAATAACTCGTTTCAAATCGGACGGCGTGGCTACTCTAGTACACC
CATTTCAACGGGCCGTTCCACTCCACTCCACCGTAACTGCGCGCGTGGGGCGACCCCATCACGTGCCACGGTACGGGGGGCGTCGCTGACCGTCTCTGGCC
CACCGCAACCCTGCTCCCCCAATCCAAACCCCGCGCTGCACGTCTCGCTGGCACTGGCCATTTCGATCCATCGCGGTCGCGGCCGCGGTATCACGACTGA
CGCGCGGGGCCGGCCCGGCACGCGAGACCGATGTAGACGTGTCACCGGGGAAGCTTGTCCGCTCGCGGTTGCCGCTGCGCGGGGGGGAGGGGAGGGCGTC
GAGCCGTACGTCGTCGTCGCATACCTCGGTCGCGTCCCTGTCAAAGCCGGCCATCGCTGCCGGCTGCTCAATTTATTCCCTTGCTGTTTCATTTCGTACG
TAGTCCGCGCTCGGGATGCGGCCATAGCCATATCGCCATATATCTCGGGCAGGCACCGTCACGCTCGCTCGGCAACTGTACGTGCCGTCTCAGGACGCGG
CAGTCAATGCGGAGTTGGTACAAACTTGCGATGAAAGGAGTGGTACTGTGGTACCACCACTGCCCTGTCTTTGCACACATCAGTCTTCCCGGAGTTTAGC

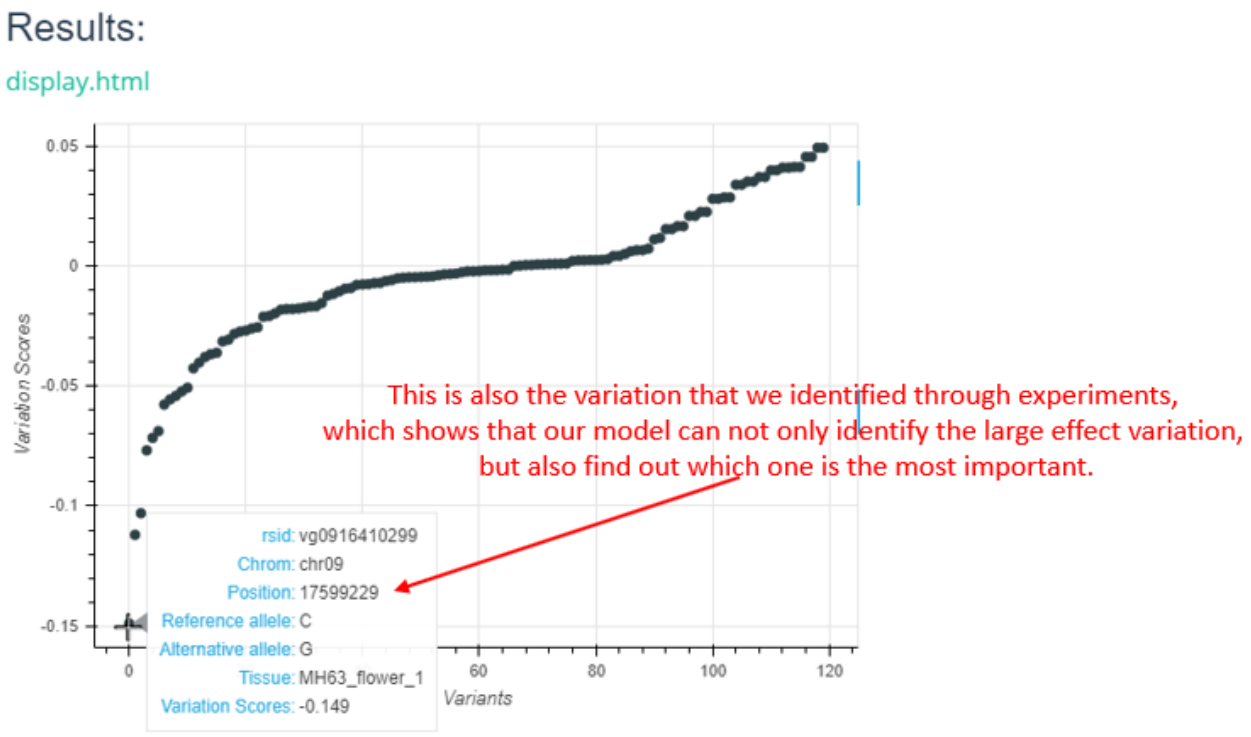
in silico saturated mutagenesis map

In the heatmap, each row represents mutations in the corresponding nucleotide, each column represents a position in the sequence. The color of the heatmap presents the predicted change in probability that an allele belongs to open chromatin. The loss score refers to the maximum decrease in probability that an allele belongs to open chromatin compared to the reference nucleotide in all mutations at each site. And the gain score refers to the maximum increase.

Tip: You can drag the graph to view information about other positions.

We have proved that an important motif, which can be identified by the model.

For VCF format input, the interactive scatter plot is returned as follows.



In addition, at the bottom of the results interface, you may also get a fimo scan result, if fimo does scan the result(as shown below).

Regulatory elements identified by FIMO

Detailed Fimo results and the parameters used can be viewed at fimo.html.

For the motifs identified by FIMO (with the *P*-value < 1e-4), we calculated the average loss and gain values in the *in silico* saturated mutagenesis map corresponding to the motif regions, which are listed in the table below as 'Loss value' and 'Gain value', respectively.

Show 10 entries Search:

Motif ID	Name	Start	End	Strand	p-value	q-value	Matched Sequence	Loss value	Gain value
MP00534	AT5G38860	449	466	+	2.13e-06	5.81e-04	GCGACCCCATCACGTGCC	1.13e-02	3.45e-03
MP00594	AT5G67000	573	587	+	1.04e-05	3.07e-03	CGGTCGCGGCCGCGG	5.76e-03	5.85e-03
MA1261.1	AT5G67000	573	587	+	1.04e-05	3.07e-03	CGGTCGCGGCCGCGG	5.76e-03	5.85e-03
MP00078	AT3G22170	437	448	-	1.16e-05	3.63e-03	CCCACGCGCGCA	9.72e-03	1.16e-03
MA0557.1	FHY3	437	448	-	1.16e-05	3.63e-03	CCCACGCGCGCA	9.72e-03	1.16e-03
MA1011.1	PHYPADRAFT_72483	457	466	+	1.22e-05	2.30e-03	ATCACGTGCC	6.18e-03	2.99e-03
MA1011.1	PHYPADRAFT_72483	457	466	-	1.22e-05	2.30e-03	GGCACGTGAT	6.18e-03	2.99e-03
MP00102	AT4G14410	457	466	+	1.41e-05	2.15e-03	ATCACGTGCC	6.18e-03	2.99e-03
MP00102	AT4G14410	457	466	-	1.41e-05	2.15e-03	GGCACGTGAT	6.18e-03	2.99e-03
MA0960.1	BHLH104	457	466	+	1.41e-05	2.15e-03	ATCACGTGCC	6.18e-03	2.99e-03

Showing 1 to 10 of 119 entries Previous 1 2 3 4 5 ... 12 Next

5.2 Main Functions

1.Variant Effector

We predicted changes in chromatin features in a 1-kb region around the mutation (upstream 500 bp and downstream 500 bp). You are supposed to submit a vcf file to generate a background prediciton task, note that there is no limit for the num of lines of your VCF file by now. For each mutation in your VCF file, we predict the scores of changes (may be positive or negative) in every sample of this model, then display all results via scatterplot.

Input: A VCF file in any length.

Output: After your submission, our system will return a task id, which is essential, go to the result page to find out your prediciton result via the task id. Input your task id to search your result stored on our server (contains the vcf file you submitted and all scores file and result scatterplot in html format)

2.Sequence Profiler

Sequence Profiler is used to analyse posible motif or important mutation in a 1-kb DNA sequence from reference genome file. You can generate a prediciton request through 2 ways:

(1)By select a chromomsome and input a position:

After receiveing the chromomsome and position information, we predict the mutation of the middle 200-bp sequence in the range of the middle 1-kb region (for each position of the 200-bp, our server simulate the difference mutation effect of the reference allele and the other 3 alter allele, thus return a 200*4 matrix afterwards, and display the matrix in the html result page through a interactive bokeh heatmap).

Input: Select a chromomsome and input a position

Output: The web page returns a heatmap, a line plot and fimo result, the heatmap shows the 200*4 matrix and line plot shows continuous positive or negative scores of a possible motif (named gain line or loss line respectively). Fimo result shows possible motif get in meme database of this DNA sequence.

(2)By upload a BED file:

Unlike predicting VCF file in Variant Effector, we only calculate the top 5 lines of your BED file due to our limited GPU resource. For each line in the top 5 lines of your BED file, the server does the same work described in (1).

Input: A BED file in any length.

Output: Prediction result of the top 5 lines of your BED file, web result description same as (1).

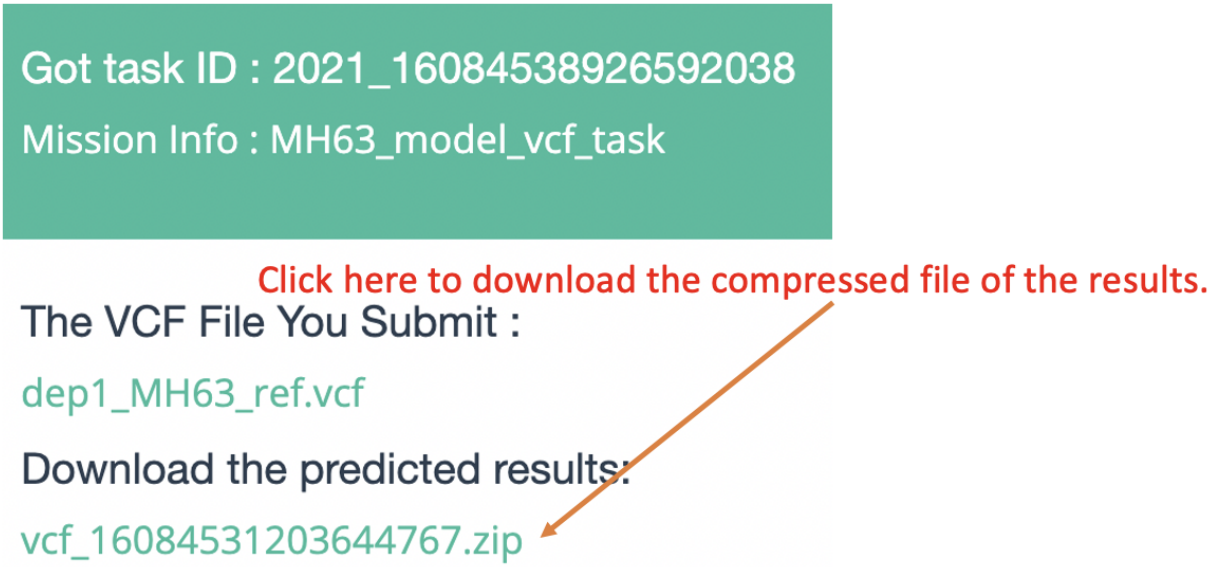
(3)By inputting the FASTA sequence:

Input any length of FASTA sequence you are interested in, and the model will intercept the middle 1000bp (if the input length is greater than 1000bp) or fill 1000bp at both ends (if the input length is less than 1000bp) for prediction.

Input: Any length of FASTA sequence you are interested in.

Output: The web page returns a heatmap, a line plot and fimo result, the heatmap shows the 200*4 matrix and line plot shows continuous positive or negative scores of a possible motif (named gain line or loss line respectively). Fimo result shows possible motif get in meme database of this DNA sequence.

5.3 About the result files



1.The task of predicting sequence (For example,ID2847_16177604444503944)

The contents of the compressed package are generally as follows:

File Name	Note
cisml.xml	Motif information (CisML format) matched by 200 bp segment scanned by Fimo
fimo.gff	Motif information (GFF format) matched by 200 bp segment scanned by Fimo
fimo.html	Motif information (HTML format) matched by 200 bp segment scanned by Fimo
fimo.tsv	Motif information (TSV format) matched by 200 bp segment scanned by Fimo
fimo.xml	Motif information (in XML format) matched by 200 bp segment scanned by Fimo
fimo_cut.tsv	Optimized(by PlantDeepSEA) version of fimo.tsv
do_fimo.bed	Record the position information of 200 bp section
do_fimo.fasta	The FASTA file of the sequence
ism_logits.tsv	$\log(P_{mut} / (1-P_{mut})) - \log(P_{ref} / (1-P_{ref}))$, the difference between *logit(mut)* and *logit(ref)* predictions
ism_diffs.tsv	$P_{mut} - P_{ref}$, the difference between *mut `and `ref* predictions
ism_abs_diffs.tsv	$\lvert P_{mut} - P_{ref} \rvert$, the absolute difference between *mut* and *ref* predictions
ism_predictions.tsv	After Selene SDK prediction, the input sequence was scored three times per mutation

The mut in the formula in the table refers to mutagenesis.

The interpretation of various formats of Fimo results is detailed here.

ism_predictions.tsv is the score of each mutation of the input 1000bp sequence three times after the prediction of Selene SDK. It records the absolute value of the difference, the true value of the difference, the Logits value and the true score of each mutation sequence.

Each PDF file in the compressed package is the drawing of each organization.

2.VCF variation annotation task (For example,ID2870_16178001127678075)

The contents of the compressed package are generally as follows:

File Name	Note
dep1_MH63_ref.vcf	VCF files submitted by users
dep1_MH63_ref_abs_diffs.tsv	$\lvert P_{mut} - P_{ref} \rvert$, Absolute value of difference before and after mutation of each locus
dep1_MH63_ref.alt_predictions.tsv	Scoring value of each locus variation
dep1_MH63_ref_diffs.tsv	$P_{mut} - P_{ref}$, value of difference before and after mutation of each locus
dep1_MH63_ref_logits.tsv	Logits value before mutation of each locus
dep1_MH63_ref.ref_predictions.tsv	Score before variation of each locus

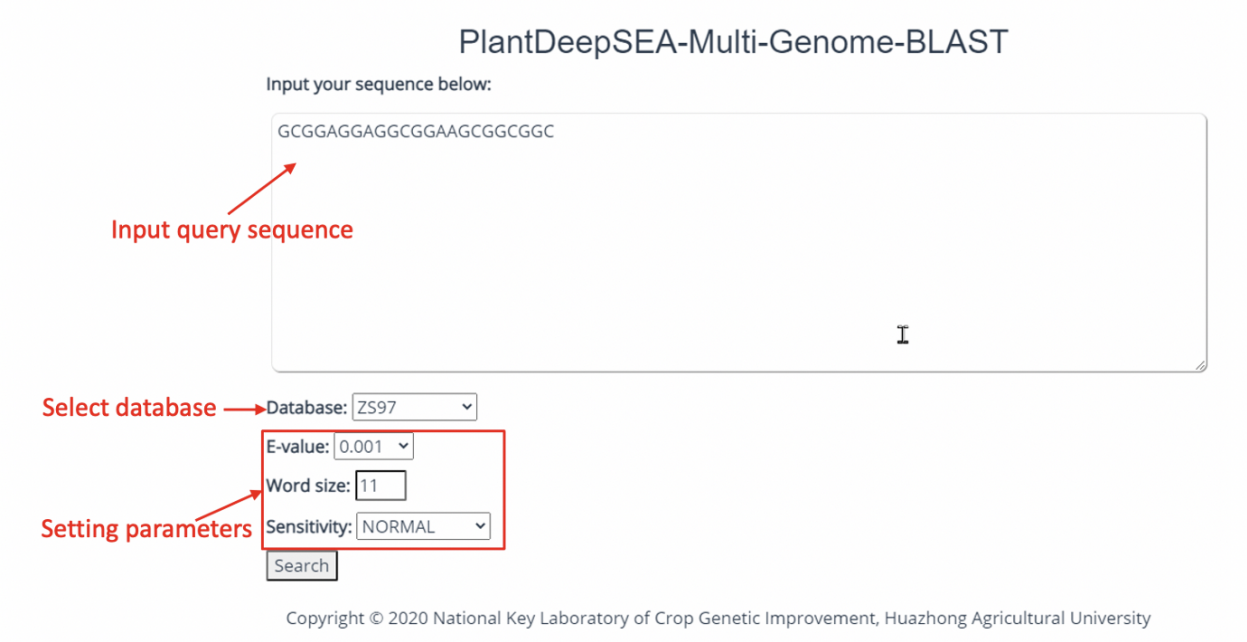
The above TSV files are the scoring files predicted by selene SDK

CHAPTER

SIX

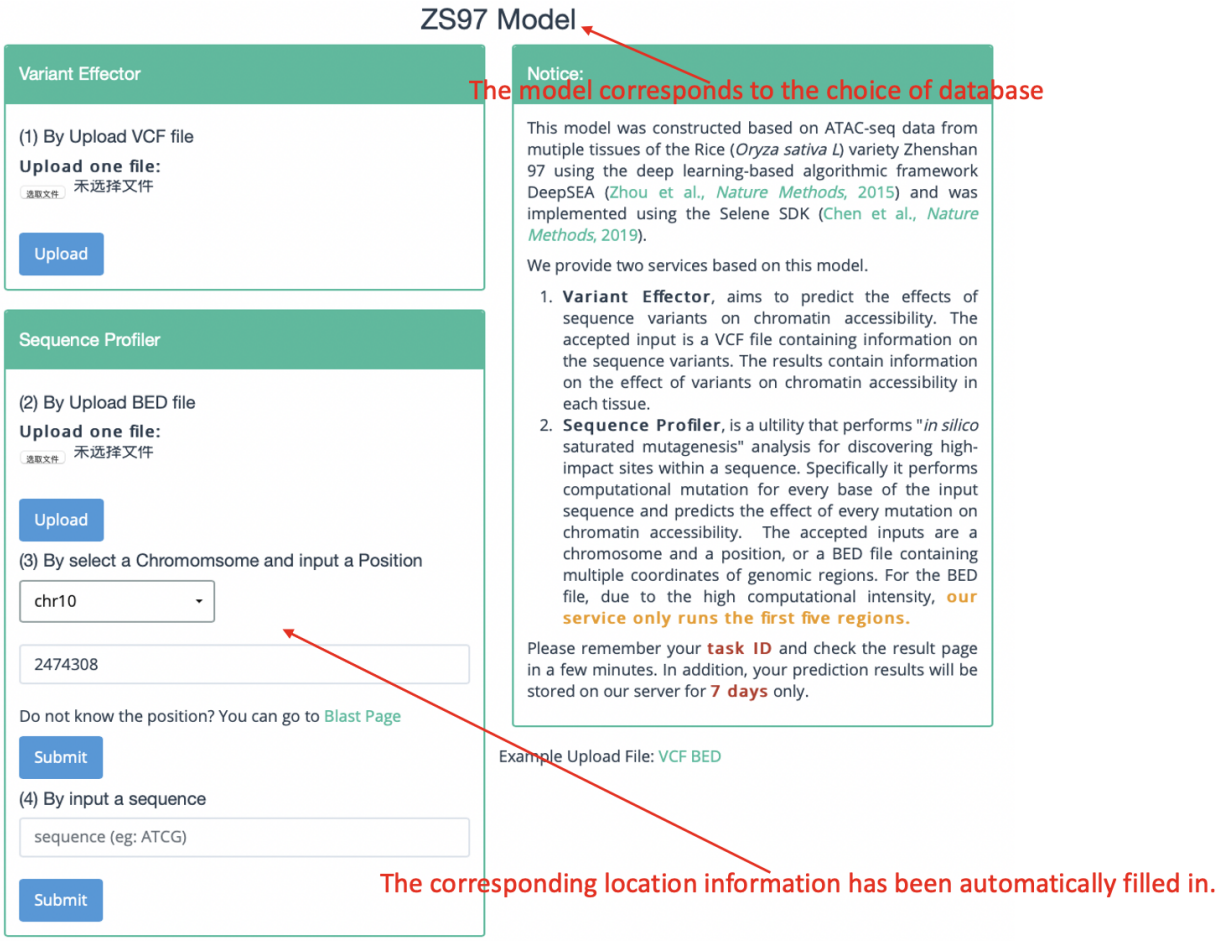
BLAST

Click the blast button in the main interface to enter the blast (for conservative sequence analysis) page. On this page, you can enter the sequence of interest and select the appropriate database and parameters to query (as shown in the figure below).



In the returned blast result interface, click the corresponding result to jump to the model (same as database) prediction interface, and automatically fill the corresponding position to (3) by select a chromsomeand input a position under the sequence profiler. The user can directly click submit to submit the task for prediction.


PlantDeepSEA-Multi-Genome-BLAST-Result
Got Zhenshan 97 BLASTN Result !
BLASTN 2.9.0+
Description
contig query length e-value score ident %
1 chr10 seq1 25619316 0.000155868 46.0 100.0
2 chr08 seq1 30215435 0.000155868 46.0 100.0
3 chr01 seq1 44482313 0.000155868 46.0 100.0
Click here to jump to the model prediction interface,
and the corresponding location information will be filled in automatically.
Alignments
chr10 length = 25619316
hsp 1
length 23
e-value 0.000155868
score 46.0
identities 23
positives 23
bits 42.8
query start 1
query end 23
subject start 2474308
subject end 2474286
Q: GCGGAGGAGGCGGAAGCGGCGGC
M: |||||||||||||||||||||||
S: GCGGAGGAGGCGGAAGCGGCGGC


ZS97 Model
The model corresponds to the choice of database
Variant Effector
(1) By Upload VCF file
Upload one file:
选取文件 未选择文件
Upload
Sequence Profiler
(2) By Upload BED file
Upload one file:
选取文件 未选择文件
Upload
(3) By select a Chromomsome and input a Position
chr10
2474308
Do not know the position? You can go to Blast Page
Submit
(4) By input a sequence
sequence (eg: ATCG)
Submit
Notice:
This model was constructed based on ATAC-seq data from mutiple tissues of the Rice (Oryza sativa L) variety Zhenshan 97 using the deep learning-based algorithmic framework DeepSEA (Zhou et al., Nature Methods, 2015) and was implemented using the Selene SDK (Chen et al., Nature Methods, 2019).
We provide two services based on this model.
1. Variant Effector, aims to predict the effects of sequence variants on chromatin accessibility. The accepted input is a VCF file containing information on the sequence variants. The results contain information on the effect of variants on chromatin accessibility in each tissue.
2. Sequence Profiler, is a ultility that performs "in silico saturated mutagenesis" analysis for discovering high-impact sites within a sequence. Specifically it performs computational mutation for every base of the input sequence and predicts the effect of every mutation on chromatin accessibility. The accepted inputs are a chromosome and a position, or a BED file containing multiple coordinates of genomic regions. For the BED file, due to the high computational intensity, our service only runs the first five regions.
Please remember your task ID and check the result page in a few minutes. In addition, your prediction results will be stored on our server for 7 days only.
Example Upload File: VCF BED
The corresponding location information has been automatically filled in.

CHAPTER

SEVEN

CASE STUDIES

7.1 Case1: The DEP1 gene of rice

A recent study(Fu , Xu etal. ,2019) have shown that nine NCVs in the DEP1 promoter region can regulate the gene expression and leaf-trait variation .

(1) Mapping these nine variants to RiceVarMap database and constructing the VCF file .

(2) Using PlantDeepSEA to make predictions for this VCF file .

(3) One SNP, named vg0916410299 in RiceVarMap, has the greatest effect score (example result ID:2870_16178001127678075) .

(4) Entering the genomic coordinates(Chromosome:chr09 ; position:17599229) of vg0916410299 into 'Sequence Profiler' for prediction. The sequence TGGCCC, which overlaps with vg0916410299, has the extreme effect scores. FIMO results also indicate that this sequence overlaps with a binding motif of the TCP transcription factors.(example result ID: 2847_16177604444503944) .

7.2 Case2: The UPA2 gene of maize

Two validated 240bp regions of the (UPA2) haplotype(Tian , Wang etal. ,2019) were analyzed with "Sequence Profiler" .

The in silico saturated mutagenesis map showed the haplotype of CIMMYT 8759 (with AGTGTG) has intensive high effect scores in the C2C2 motif region compared to the haplotype of W22 (with AGTG–) (example result ID: 2867_16177642225564935 and 2868_1617764237554985) .

CHAPTER

EIGHT

RESOURCES

8.1 Brief Description

Click on the "Resources" link in the main menu to view training data statistics, quality control information, and model evaluation results. In addition, we provide links to download the reference genomes, the trained models and files for training models.

Please note: *We are still analyzing the ATAC-seq data we generated from multiple tissues of grass species. We have provided lists of OCRs obtained from these ATAC-seq data so that researchers can use it to reproduce the results of PlantDeepSEA. However, please do not use these lists for genome-scale analyses, until the publication of our article on data analysis (by May 2022 at the latest). Or contact us for permission. Users who download data from this site are considered to have accepted this statement.*

8.2 Reference Genome Information (Download from Google Drive)

variety	Species	Version	Reference link
Arabidopsis	Arabidopsis thaliana	TAIR10.1	Download link
Rice	Oryza sativa	MSU v7.0	Download link
Rice	Oryza sativa	MH63 RS2	Download link
Rice	Oryza sativa	ZS97 RS2	Download link
Brachypodium	Brachypodium distachyon	Bd21-3 v1.1	Download link
Foxtail millet	Setaria italica	v2.0	Download link
Sorghum	Sorghum bicolor	v3.1.1	Download link
Mazie	Zea mays	AGPv4	Download link

8.3 Reference Genome Information (Download from Baidu Cloud Disk Drive)

variety	Species	Version	Reference link
Arabidopsis	Arabidopsis thaliana	TAIR10.1	https://pan.baidu.com/s/1TvqpN94yE6gq7VHVRNt6kA?pwd=iv43 iv43
Rice	Oryza sativa	MSU v7.0	: https://pan.baidu.com/s/1SYpyzJLd6X_3sTVoTbyxbw : tdgi
Rice	Oryza sativa	MH63 RS2	https://pan.baidu.com/s/1uXqWbSdipnNHfckH6xOZoA?pwd=7fed 7fed
Rice	Oryza sativa	ZS97 RS2	https://pan.baidu.com/s/1Xm8GoRQ2wCt1aSjU7DEqUQ?pwd=bhtw bhtw
Brachypodium	Brachypodium distachyon	Bd21-3 v1.1	: https://pan.baidu.com/s/1hiUYOE9g3KwtDucUkRkPTw : r4ka
Foxtail millet	Setaria italica	v2.0	https://pan.baidu.com/s/1qjwnowHMEf_psOLpTgEo0g?pwd=4m10 4m10
Sorghum	Sorghum bicolor	v3.1.1	https://pan.baidu.com/s/1AqbUPws9AyTtjnBv3vUMmQ?pwd=3j0q 3j0q
Mazie	Zea mays	AGPv4	https://pan.baidu.com/s/1Ppexb6mzC_FtfvIzfYb4vw?pwd=1xf8 1xf8

8.4 Trained model

variety	Reference link
Arabidopsis	Download link
Zhenshan 97	Download link
Minghui 63	Download link
Brachypodium	Download link
Foxtail millet	Download link
Sorghum	Download link
Mazie	Download link

8.5 Training data

variety	Reference link
Arabidopsis	Download link
Zhenshan 97	Download link
Minghui 63	Download link
Brachypodium	Download link
Foxtail millet	Download link
Sorghum	Download link
Mazie	Download link

8.6 Training config files

variety	Reference link
Arabidopsis	Download link
Zhenshan 97	Download link
Minghui 63	Download link
Brachypodium	Download link
Foxtail millet	Download link
Sorghum	Download link
Mazie	Download link

8.7 Model structure

model name	Reference link
DeeperDeepSEA	Download link

This model structure is Selene's DeeperDeepSEA structure .

8.8 A guide to applying the model locally

You can click here to see how the models are trained and how to apply the trained models provided in PlantDeepSEA locally.

CHAPTER

NINE

STATISTICS

You can slide the table to see more information.

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
A. thaliana	Arabidopsis_stem_cell_rep3	[illegible]	[illegible]	0.2564[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, stem cells, rep3	139540	0.0272[illegible]	[illegible]	0.0249[illegible]	[illegible]	[illegible]
A. thaliana	Arabidopsis_stem_cell_rep2	9973[illegible]	[illegible]	0.2035[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, stem cells, rep2	240981	0.0470[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
A. thaliana	Arabidopsis_root_non_hair_cell_rep1	9407985[illegible]	[illegible]	0.1017[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, root non-hair, rep1	344379	0.0672[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
A. thaliana	Arabidopsis_root_tip_rep1	479528[illegible]	[illegible]	0.0603[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, root tips, rep1	344701	0.0673[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
A.	Arabidopsis_thaliana_stem_cell_rep1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, stem cells, rep1	356863	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
A.	Arabidopsis_thaliana_7days_leaf_rep1	[illegible]	[illegible]	0.304	[illegible]	[illegible]	25571	ATAC-seq, leaf, 7 days, rep1	393580	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
A.	Arabidopsis_thaliana_root_tip_rep2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, root tips, rep2	395322	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
A.	Arabidopsis_thaliana_root_hair_cell_rep1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, root hair, rep1	414566	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
A.	Arabidopsis_thaliana_mesophyll_cell_rep1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, mesophyll cells, rep1	509807	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
A.	Arabidopsis_thaliana_mesophyll_cell_rep3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, mesophyll cells, rep3	546616	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
A. thaliana	Arabidopsis_root_non_hair_cell_rep2	242969529936		0.0578	5514	7	2	ATAC-seq, root non-hair, rep2	557959	0.1089	6	1	0.7	0.7
A. thaliana	Arabidopsis_mesophyll_cell_rep2	99938389932		0.2860	1	4	2	ATAC-seq, mesophyll cells, rep2	544592	0.1063	6	1	0.1	0.1
A. thaliana	Arabidopsis_root_hair_cell_rep2	291715489938		0.0642	3	3	2	ATAC-seq, root hair, rep2	570957	0.1115	1	0	0.5	0.4
A. thaliana	Arabidopsis_7days_leaf_rep2	149379647171		0.3	3034575	18.54	33398	ATAC-seq, leaf, 7 days, rep2	585868	0.1144	2	1	0.7	0.7
B. distachyon	Bdistachyon_root_1	145430284715		0.0551	5170227	7.474429	6244	ATAC-seq, root, rep1	40093	0.0078	3	0	0.9	0.9
B. distachyon	Bdistachyon_young_leaf_2	102916469754		0.1146	7932578	8.145661	7113	ATAC-seq, young leaf, rep2	74266	0.0145	0	0	0.9	0.9
B. distachyon	Bdistachyon_young_leaf_1	108702149604		0.1215	8569899	9.781782	3358	ATAC-seq, young leaf, rep1	112185	0.0219	1	0	0.9	0.9

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
B. distachyon	Bdistachyon_flower_2	24329950	0.9589	0.5249	8403967	14.7479	46148	ATAC-seq, flower, rep2	162902	0.0318	16796.875	0.02539	0.6598	0.5461
B. distachyon	Bdistachyon_root_2	45701156	0.5064	0.0649	17355036	4.9749	63702	ATAC-seq, root, rep2	180246	0.0352	20429.6875	0.02869	0.7537	0.5864
B. distachyon	Bdistachyon_flower_1	39572726	0.9536	0.5504	13233688	8.7284	42655	ATAC-seq, flower, rep1	215663	0.0421	21607.6875	0.03434	0.9602	0.4085
B. distachyon	Bdistachyon_flag_leaf_2	63932496	0.9606	0.3272	37316076	6.0990	52939	ATAC-seq, flag leaf, rep2	439499	0.0858	39671.84375	0.07213	0.9387	0.5874
B. distachyon	Bdistachyon_panicle_1	116923382	0.9716	0.5741	39556925	5.7315	59336	ATAC-seq, young panicle, rep1	408596	0.0798	39006.25	0.07435	0.9458	0.7068
B. distachyon	Bdistachyon_flag_leaf_1	94067702	0.9695	0.3964	49821067	3.4507	57838	ATAC-seq, flag leaf, rep1	452662	0.0884	41054.6875	0.07522	0.9878	0.9065
O. sativa	ZS97_panicle_1	57620486	0.9742	0.3419	28421790	4.3770	67787	ATAC-seq, young panicle, rep1	34913	0.0068	18945.3125	0.00846	0.9416	0.6666

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
O. sativa	MH63_flower_rep1	[illegible]	[illegible]	0.151	[illegible]	[illegible]	[illegible]	ATAC-seq, stamen & pistil, rep1	49527	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
O. sativa	MH63_flower_rep2	[illegible]	[illegible]	0.2364	[illegible]	[illegible]	[illegible]	ATAC-seq, stamen & pistil, rep2	68509	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
O. sativa	ZS97_panicle_rep5	[illegible]	[illegible]	0.3547	[illegible]	[illegible]	[illegible]	ATAC-seq, young panicle, rep5	140295	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
O. sativa	ZS97_panicle_rep4	[illegible]	[illegible]	0.2823	[illegible]	[illegible]	[illegible]	ATAC-seq, young panicle, rep4	148634	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
O. sativa	MH63_flag_leaf_rep2	[illegible]	[illegible]	0.2291	[illegible]	[illegible]	[illegible]	ATAC-seq, flag leaf, rep2	119482	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
O. sativa	ZS97_young_leaf_rep1	[illegible]	[illegible]	0.5716	[illegible]	[illegible]	[illegible]	ATAC-seq, young leaf, rep1	234188	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
O. sativa	ZS97_panicle_rep2	2432162349875	0.3067	11047821	13.954281875			ATAC-seq, young panicle, rep2	120640	0.02356283		0.0291967	0.7743	0.5827
O. sativa	MH63_panicle_rep2	3330958299863	0.3309	14732470	15.5901584837			ATAC-seq, young panicle, rep2	167176	0.03265435625		0.0292967	0.8786	0.6312
O. sativa	MH63_young_leaf_rep2	7010930298824	0.4114	28029716	14.21245272855			ATAC-seq, young leaf, rep2	169796	0.03316425	81250	0.0299703	0.8962	0.7059
O. sativa	MH63_panicle_rep1	5883275699843	0.4198	24409468	18.83604646717			ATAC-seq, young panicle, rep1	204168	0.03987556	56625	0.0340783	0.8963	0.7162
O. sativa	MH63_panicle_rep3	5116214899863	0.3469	22157517	17.72406242842			ATAC-seq, young panicle, rep3	201314	0.03931915	40625	0.0346843	0.8706	0.6827
O. sativa	MH63_panicle_rep5	4623530799874	0.2876	25483264	14.945473604			ATAC-seq, young panicle, rep5	236853	0.04626885	156250	0.0391299	0.8958	0.7203

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
O. sativa	MH63_panicle_rep4	55158396752	0.3268	26097080	10.75267	74025	ATAC-seq, young panicle, rep4	234237	0.04574984	40602	0.03933	0.8963	0.7246
O. sativa	MH63_young_leaf_rep1	13893810864	0.4966	44734548	13.01296	5991	ATAC-seq, young leaf, rep1	229037	0.04473978	39062	0.04007	0.9476	0.8655
O. sativa	MH63_root_rep2	68284902987	0.4127	32567024	14.21247	71764	ATAC-seq, root, rep2	253463	0.04950449	42187	0.04263	0.9563	0.8408
O. sativa	ZS97_young_leaf_rep2	38335049243	0.3616	11787053	17.22273	85795	ATAC-seq, young leaf, rep2	356413	0.06961129	60625	0.04330	0.9675	0.8753
O. sativa	ZS97_panicle_rep3	36435609844	0.3379	15282506	16.70468	85220	ATAC-seq, young panicle, rep3	182557	0.03565896	30406	0.04457	0.9546	0.8259
O. sativa	MH63_flag_leaf_rep1	136342029487	0.273	68545456	9.610392	45130	ATAC-seq, flag leaf, rep1	292602	0.05714842	48812	0.05010	0.9586	0.8896
O. sativa	ZS97_root_rep1	166991452781	0.5518	54000696	15.53397	59269	ATAC-seq, root, rep1	407748	0.07963852	67912	0.05231	0.9495	0.8206
O. sativa	ZS97_root_rep2	74642946985	0.4001	34476273	15.76327	24959	ATAC-seq, root, rep2	426907	0.08338027	71437	0.05689	0.9504	0.8268

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
O. sativa	MH63_lemma_1	[illegible]	[illegible]	0.2213	[illegible]	[illegible]	[illegible]	ATAC-seq, lemma & palea, rep1	330188	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
O. sativa	MH63_lemma_2	[illegible]	[illegible]	0.1906	[illegible]	[illegible]	[illegible]	ATAC-seq, lemma & palea, rep2	336402	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
O. sativa	MH63_root_1	[illegible]	[illegible]	0.5957	[illegible]	[illegible]	[illegible]	ATAC-seq, root, rep1	343090	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
O. sativa	ZS97_flag_leaf_2	[illegible]	[illegible]	0.6583	[illegible]	[illegible]	[illegible]	ATAC-seq, flag leaf, rep2	270937	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
O. sativa	ZS97_flag_leaf_1	[illegible]	[illegible]	0.305	[illegible]	[illegible]	[illegible]	ATAC-seq, flag leaf, rep1	324916	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
O. sativa	ZS97_flower_1	[illegible]	[illegible]	0.2821	[illegible]	[illegible]	[illegible]	ATAC-seq, stamen & pistil, rep1	449357	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
O. sativa	ZS97_flower_rep2	98606280	69858	0.2658	61262472	13.17765	503872	ATAC-seq, stamen & pistil, rep2	462941	0.09041	218406	0.04264	0.91932	0.70344
O. sativa	ZS97_lemma_rep2	38347720	49865	0.1542	26758092	12.02037	333452	ATAC-seq, lemma & palea, rep2	494763	0.09663	233484	0.04561	0.9563	0.6211
O. sativa	ZS97_lemma_rep1	37055650	09861	0.1391	26193428	11.66188	345427	ATAC-seq, lemma & palea, rep1	601667	0.11751	288593	0.05636	0.94039	0.63366
S. bicolor	Sorghum_panicle_bottom_rep2	18288230	94987	0.5751	5304178	4.012619	220294	ATAC-seq, bottom part of panicle, rep2	14006	0.00218	8867	0.00138	0.72696	0.53846
S. bicolor	Sorghum_panicle_upper_rep1	37071098	85267	0.2408	16239688	11.74675	330825	ATAC-seq, upper part of panicle, rep1	14920	0.00233	7025	0.00109	0.94764	0.55897

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
S. bicolor	[illegible]	[illegible]	[illegible]	0.4883	[illegible]	[illegible]	[illegible]	ATAC-seq, root, rep1	37103	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
S. bicolor	[illegible]	[illegible]	[illegible]	0.1574	[illegible]	[illegible]	[illegible]	ATAC-seq, lemma & palea, rep2	44558	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
S. bicolor	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, young leaf, rep2	47084	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
S. bicolor	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, young leaf, rep1	46878	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
S. bicolor	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, bottom part of panicle, rep1	51301	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
S. bicolor	[illegible]	[illegible]	[illegible]	0.3257	[illegible]	[illegible]	[illegible]	ATAC-seq, upper part of panicle, rep2	54562	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
S. bicolor	[illegible]	[illegible]	[illegible]	0.1605	[illegible]	[illegible]	[illegible]	ATAC-seq, stamen & pistil, rep1	64437	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
S. bicolor	[illegible]	[illegible]	[illegible]	0.1495	[illegible]	[illegible]	[illegible]	ATAC-seq, stamen & pistil, rep2	65513	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
S. bicolor	[illegible]	[illegible]	[illegible]	0.4356	[illegible]	[illegible]	[illegible]	ATAC-seq, flag leaf, rep1	68154	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
S. bicolor	[illegible]	[illegible]	[illegible]	0.412	[illegible]	[illegible]	[illegible]	ATAC-seq, flag leaf, rep2	72738	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
S. bicolor	[illegible]	[illegible]	[illegible]	0.1691	[illegible]	[illegible]	[illegible]	ATAC-seq, lemma & palea, rep1	75782	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
S. bicolor	[illegible]	[illegible]	[illegible]	0.7644	[illegible]	[illegible]	[illegible]	ATAC-seq, root, rep2	97185	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
S. italica	Sitalica_panicle_1	1776742009883	0.2021	1262277638	916080	2264	ATAC-seq, young panicle, rep1	48566	0.009485574687	5	0.005240	1.7	4	8
S. italica	Sitalica_young_leaf_2c_1	700671389694	0.4212	296593820	609069	9107	ATAC-seq, young leaf, rep1	65424	0.012778245	25	0.007478	8.9	7	4
S. italica	Sitalica_flower_1	1310865369924	0.23	91809577	119610	26573	ATAC-seq, flower, rep1	91008	0.017775	18	0.010616	6.9	8	9
S. italica	Sitalica_root_2c_1	257357508962	0.4955	77096187	.4828	30919	ATAC-seq, root, rep1	102140	0.019949	9218	750.011318	8.9	8	6
S. italica	Sitalica_young_leaf_2c_2	573977902982	0.5489	196386697	398	41305	ATAC-seq, young leaf, rep2	72681	0.011359	6	250.011969	9.9	6	4
S. italica	Sitalica_flower_2	1383570202492	0.2077	99270572	3444	70325	ATAC-seq, flower, rep2	121054	0.023644	9	0.014790	2.9	9	6
S. italica	Sitalica_flag_leaf_1	1419966009915	0.2786	85924476	056	65882	ATAC-seq, flag leaf, rep1	162685	0.031774	406	250.020530	3.9	5	6

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
S. italica	Sitalica_panicle_2	[illegible]	[illegible]	0.2297	[illegible]	[illegible]	[illegible]	ATAC-seq, young panicle, rep2	192586	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
S. italica	Sitalica_flag_leaf_2	[illegible]	[illegible]	0.2437	[illegible]	[illegible]	[illegible]	ATAC-seq, flag leaf, rep2	224032	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
26. mays	[illegible]	[illegible]	[illegible]	0.3522	[illegible]	[illegible]	[illegible]	ATAC-seq, upper part of ~1cm ear, rep1	39169	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
26. mays	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, bottom part of ~1cm ear, rep2	53425	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
26. mays	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ATAC-seq, bottom part of ~1cm ear, rep1	61812	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
26. mays	Zmays_[illegible]	[illegible]	[illegible]	0.3889	38784430	10.248675514		ATAC-seq, upper part of ~1cm ear, rep2	87961	0.013743906250	0.013784	[illegible]	[illegible]	[illegible]
26. mays	Zmays_young_[illegible]	[illegible]	[illegible]	0.3883	57581514	14.049183836		ATAC-seq, young leaf, rep3	101406	0.019803588593750	0.017376	[illegible]	[illegible]	[illegible]
26. mays	Zmays_root_[illegible]	[illegible]	[illegible]	0.1117	24442269	4.794161615869		ATAC-seq, root, rep2	111460	0.021769531250	0.020878	[illegible]	[illegible]	[illegible]
26. mays	Zmays_root_[illegible]	[illegible]	[illegible]	0.3186	80469075	11.952080430		ATAC-seq, root, rep3	158819	0.031014938593750	0.030846	[illegible]	[illegible]	[illegible]
26. mays	Zmays_young_[illegible]	[illegible]	[illegible]	0.3596	41587718	6.617397227		ATAC-seq, young leaf, rep1	242864	0.047434863675	0.041487	[illegible]	[illegible]	[illegible]
26. mays	Zmays_root_[illegible]	[illegible]	[illegible]	0.1413	42431275	5.082777729812		ATAC-seq, root, rep1	226605	0.044258789062500	0.041756	[illegible]	[illegible]	[illegible]
26. mays	Zmays_young_[illegible]	[illegible]	[illegible]	0.3064	30621551	5.401074194		ATAC-seq, young leaf, rep2	234570	0.045814453125000	0.041823	[illegible]	[illegible]	[illegible]

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
26. mays	Zmays_flower_1	[illegible]	[illegible]	0.2676	[illegible]	[illegible]	[illegible]	ATAC-seq, young tassel spikelet, rep1	250943	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
26. mays	Zmays_tassel_bottom_1	[illegible]	[illegible]	0.3106	[illegible]	[illegible]	[illegible]	ATAC-seq, bottom part of ~1cm tassel, rep1	243257	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
26. mays	Zmays_flower_2	[illegible]	[illegible]	0.289	[illegible]	[illegible]	[illegible]	ATAC-seq, young tassel spikelet, rep2	264358	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
26. mays	Zmays_ear_small_1	[illegible]	[illegible]	0.3956	[illegible]	[illegible]	[illegible]	ATAC-seq, ~5mm ear, rep1	261472	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
26. mays	Zmays_tassel_bottom_2	[illegible]	[illegible]	0.3613	[illegible]	[illegible]	[illegible]	ATAC-seq, bottom part of ~1cm tassel, rep2	254837	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	Q30 reads number	TSS Enrichment	Peak number	Sample name	Number of sequences labeled as OCR in training set	Ratio of sequences labeled as OCR in training set	Number of sequences labeled as OCR in test set	Ratio of sequences labeled as OCR in test set	AUC	AUPRC
26. mays	Zmays_[illegible]	[illegible]	[illegible]	0.325	[illegible]	[illegible]	87925	ATAC-seq, ~5mm ear, rep2	276548	0.0540	[illegible]	0.0486	[illegible]	[illegible]
26. mays	Zmays_[illegible]	[illegible]	[illegible]	0.413	[illegible]	[illegible]	68782	ATAC-seq, upper part of ~1cm tassel, rep1	286296	0.0559	[illegible]	0.0509	[illegible]	[illegible]
26. mays	Zmays_[illegible]	[illegible]	[illegible]	0.4233	[illegible]	[illegible]	20508	ATAC-seq, flag leaf, rep2	386108	0.0754	[illegible]	0.0682	[illegible]	[illegible]
26. mays	Zmays_[illegible]	[illegible]	[illegible]	0.3709	[illegible]	[illegible]	91086	ATAC-seq, flag leaf, rep1	389894	0.0761	[illegible]	0.0690	[illegible]	[illegible]

CHAPTER

TEN

CITATION

If you use this web sever for your research, please cite our paper:

Hu Zhao#, Zhuo Tu#, Yinmeng Liu, Zhanxiang Zong, Jiacheng Li, Hao Liu, Feng Xiong, Jinling Zhan, Xuehai Hu, and Weibo Xie* (2021). PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. *Nucleic Acids Research*, doi: 10.1093/nar/gkab383