
plantdeepsea Documentation

Jan 19, 2023

OVERVIEW

1	Home Page	3
2	Model Selection	5
3	Submit Task	7
4	Query Results	11
5	Interpretation of Results	13
5.1	Brief Description	13
5.2	Main Functions	16
5.3	About the result files	17
6	Blast	19
7	Case Studies	23
7.1	Case1: The DEP1 gene of rice	23
7.2	Case2: The UPA2 gene of maize	23
8	Resources	25
8.1	Brief Description	25
8.2	Reference Genome Information (Download from Google Drive)	25
8.3	Reference Genome Information (Download from Baidu Cloud Disk Drive)	26
8.4	Trained model	26
8.5	Training data	26
8.6	Training config files	27
8.7	Model structure	27
8.8	A guide to applying the model locally	27
9	Statistics	29
10	Citation	45

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

HOME PAGE

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



PlantDeepSEA

Predicting Regulatory Effect of Genomic Variants

[Home](#)
[Results](#)
[BLAST](#)
[Statistics](#)
[Tutorial](#)

PlantDeepSEA is a webserver based on deep learning models of chromatin accessibility for multiple plant species. It can predict the impact of genomic variants on chromatin accessibility in multiple tissues. Therefore, it can be used to prioritize genomic variants and discover high-impact cis-regulatory sites within a sequence.

Select models

Trained models



Arabidopsis

Col-0

A small flowering plant that is widely used as a model organism in plant biology.



Rice

Zhenshan 97
Minghui 63

A vital staple crop and a model organism for monocotyledons.



Brachypodium

Bd21

A model plant for temperate grasses and herbaceous energy crops.



Foxtail millet

Yuguihao

An important grain crop and forage.



Sorghum

JDXBR

A biofuel crop and potential cellulosic feedstock.



Maize

B73

A staple food in many parts of the world.

Query your results

example task IDs:
2566_16148611193615174 or
2021_16084538926592038.

What's New

2021-04-02

We have updated the presentation of the results page.

2021-03-02

The rice model supports coordinates and intervals of multiple reference genomes.

2021-01-19

'Sequence Profiler' can accept input of a custom sequence for analysis.

2020-11-18



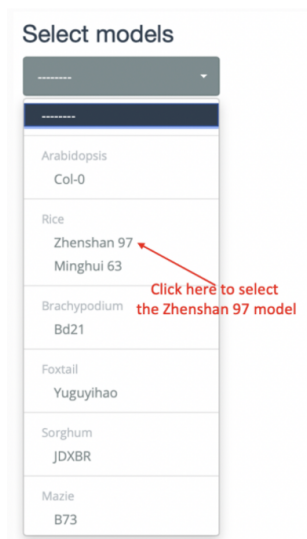
The online service of the Plant Multi-species Deep Learning Model is now available!

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.

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).



Trained models



Arabidopsis

Col-0

A small flowering plant that is widely used as a model organism in plant biology.



Rice

Zhenshan 97

Minghui 63

A vital staple crop and a model organism for monocotyledons.

Click here to select the Zhenshan 97 model

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

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

ZS97 Model

Here is the name of the model you have clicked

Notice:

This model was constructed based on ATAC-seq data from multiple 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

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.

ZS97 Model

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 multiple 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

Click here to select reference genome

Click here to submit VCF file

Click here to submit BED file

Click here to select chrome and position queries

Click here to enter the FASTA sequence

Click here to download the sample file

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:

Search Your Result

2803_16175216338581753

Submit

Your TaskID : 2803_16175216338581753

0 sub-mission completed.

Still running. Please wait ~

●●●●●●●●●●

Tips :

- 1) Your prediction results will be stored on our server for 7 days only, results beyond this time will not be retrieved.
- 2) Make sure you submit the correct task ID. After your submitting, this page will automatically refresh in every 20 seconds until your result shows up.
- 3) The task will only return results after it is run, so please wait for a while after submitting.
- 4) example task IDs: 2566_16148611193615174 or 2021_16084538926592038 .

For BED input:

Search Your Result

2809_16175219867692175

Submit

The following prediction requests were received.

Show 10 entries
Search:

Chromosome	Start	End
chr09	8998909	8998947
chr10	8461182	8461732
chr11	18836169	18836959
chr12	2794339	2794477
chr12	22076166	22076893

Showing 1 to 5 of 5 entries
Previous 1 Next

Your TaskID : 2809_16175219867692175

0 sub-mission completed.

Still running. Please wait ~

●●●●●●●●●●

Tips :

- 1) Your prediction results will be stored on our server for 7 days only, results beyond this time will not be retrieved.
- 2) Make sure you submit the correct task ID. After your submitting, this page will automatically refresh in every 20 seconds until your result shows up.
- 3) The task will only return results after it is run, so please wait for a while after submitting.
- 4) example task IDs: 2566_16148611193615174 or 2021_16084538926592038 .

For VCF input:

Search Your Result

The following prediction requests were received.

Show
10
▼
entries

Search:

Chromosome ▲	Position ▾	Variation ID ▾	Ref ▾	Alt ▾
chr09	17599142	vg0916410213	C	A
chr09	17599229	vg0916410299	C	G
chr09	17599532	vg0916410603	G	C
chr09	17599577	vg0916410648	G	T
chr09	17599855	vg0916410925	T	G
chr09	17599855	vg0916410925	T	G
chr09	17599991	vg0916411061	C	G
chr09	17599998	vg0916411068	G	T

Showing 1 to 8 of 8 entries

Previous
1
Next

Your TaskID : 2808_161752196048525
0 sub-mission completed.
Still running. Please wait ~
● ● ● ● ● ● ● ●

Tips :

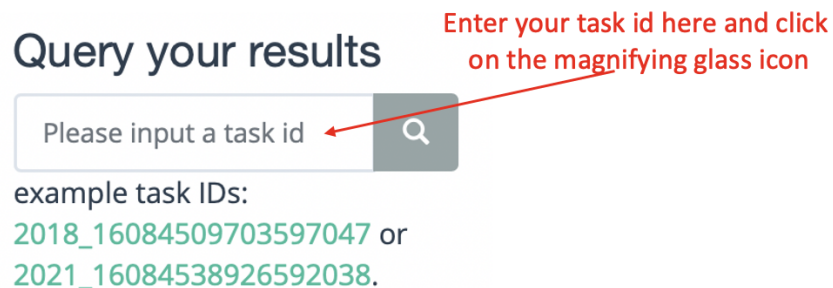
- 1) Your prediction results will be stored on our server for 7 days only, results beyond this time will not be retrieved.
- 2) Make sure you submit the correct task ID. After your submitting, this page will automatically refresh in every 20 seconds until your result shows up.
- 3) The task will only return results after it is run, so please wait for a while after submitting.
- 4) example task IDs: 2566_16148611193615174 or 2021_16084538926592038 .

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.

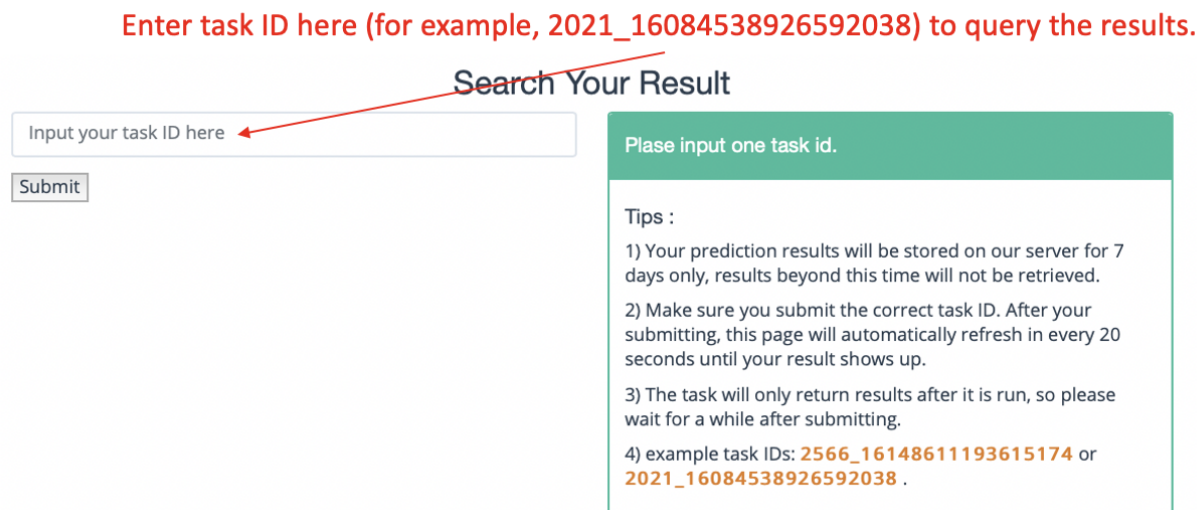
QUERY RESULTS

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.

Your TaskID : 2812_16175230132291112

0 sub-mission completed.

Still running. Please wait ~



Tips :

- 1) Your prediction results will be stored on our server for 7 days only, results beyond this time will not be retrieved.
- 2) Make sure you submit the correct task ID. After your submitting, this page will automatically refresh in every 20 seconds until your result shows up.
- 3) The task will only return results after it is run, so please wait for a while after submitting.
- 4) example task IDs: 2566_16148611193615174 or 2021_16084538926592038 .

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

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).

Search Your Result

Task ID : 2566_16148611193615174

Mission Info : The MH63 model was used to obtain the input information according to the chromosome number and position.

Position :

Chromosome : chr09

Start : 17599129

End : 17599329

Download the predicted results:

[chr09_17599129_17599329_16148611193615174.zip](#)

Click here to download the result file package

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.

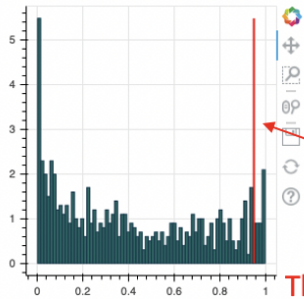
Results:

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).

ACACAGCTAAATGAATCGCATATGCAGTGAATGAGACTCTGAGAGCCAGTTAAGGTAGGAGTAGCATCAAATCACTAGTCTATACACGGGACGTCGTCGCAGTGCAATTTTATACGATCTTCATGTGTGCATCAGTGCATGCTGTTTGAGAAGGGAAGTTGGCCGCCTGCCCGTGCAGCTCAACTGAACGCTGGCTGCCTGCATGGCCAGTTCTGTGACTACAGTTACTACGCATCTACCCGGGGTAGCAGCGCGCGCTATCAGCTCTACCCGGAGTAGGCAGTAGCAGCAGTGGTGGTAGTAGTAGAAAGCGCGCAGTGGAAGGGATAGAGGAGTGCCACATTAATAACTCGTTTCAAATCGGACGGCGTGGCTACTCTAGTACACCATTTCACGGGCGGTTCCACTCCACTCCACCGTAACCTGCGCGCGTGGGGCGACCCCATCACGTGCCACGGTACGGGGCGTCGCTGACCGTCTCTGGCCACCCGAACCTGCTCCCCCAATCCAAACCCGCGCTGCACGCTCGCTGGCACTGGCCATTTTCGATCCATCGCGGTCGCGGCGCGGTATCACGACTGACGCGCGGGCGGCCCGGCACGCGAGACCGATGTAGACGTGCACCGGGGAAGCTGTCCGCTCGCGGTTGCCGCTGCGCGGGGGAGGGGAGGGCGTCGAGCCGTACGTCGTCGTCGATACCTCGGTGCGCTCCCTGTCAAAGCCGGCCATCGTGCCGGCTGCTCAATTTATCCCTTGCTGTTTCATTTCTGTACGTAGTCCGCGCTCGGGATGCGGCCATAGCCATATCGCCATATATCTCGGCGAGGCACCGTCACGCTCGCTCGGCAACTGTACGTGCCGTCTCAGGACGCGGCAGTCAATGCGGAGTTGGTACAACTTGCATGAAAGGAGTGTTACTGTGTACCACCACTGCCCTGTCTTTGCACACATCAGTCTTCCCGGAGTTAGC

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.

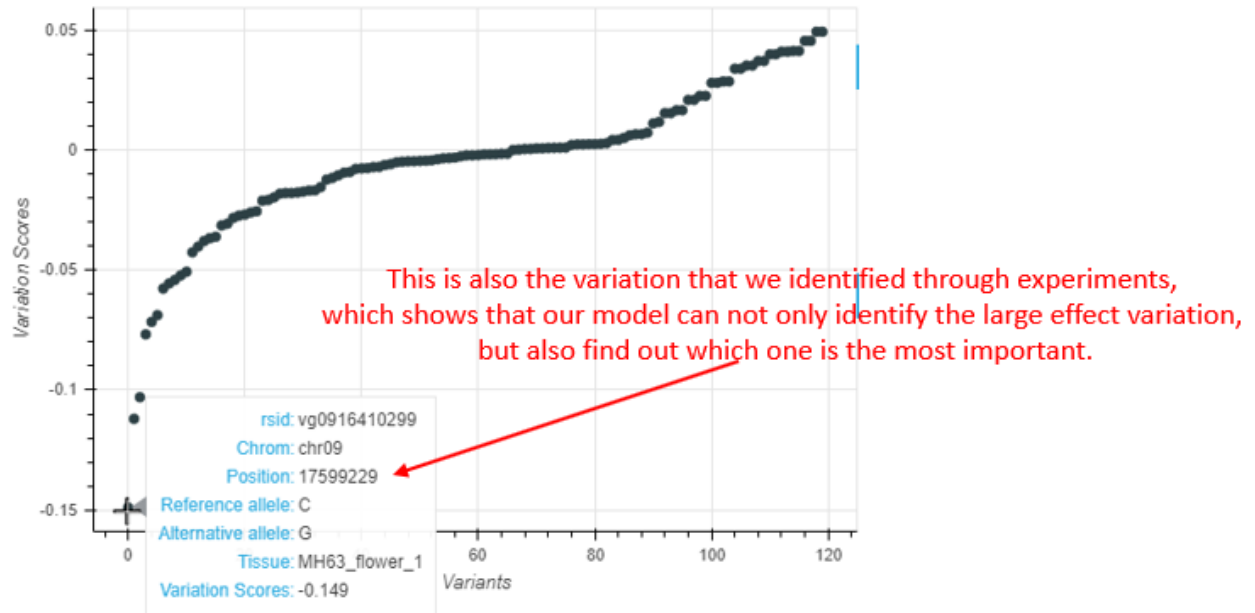
We have proved that an important motif, which can be identified by the model.



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

Results:

[display.html](#)



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 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	GCGACCCCATCAGTGCC	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 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 prediction 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 prediction 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 possible motif or important mutation in a 1-kb DNA sequence from reference genome file. You can generate a prediction request through 2 ways:

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

After receiving the chromosome 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 chromosome 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

Got task ID : 2021_16084538926592038

Mission Info : MH63_model_vcf_task

[Click here to download the compressed file of the results.](#)

The VCF File You Submit :

dep1_MH63_ref.vcf

Download the predicted results:

vcf_16084531203644767.zip

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

The contents of the compressed package are generally as follows:

File Name	Note
cismml.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 <i>logit(mut)</i> and <i>logit(ref)</i> predictions
ism_diffs.tsv	$P_{mut} - P_{ref}$, the difference between <i>mut</i> and <i>ref</i> predictions
ism_abs_diffs.tsv	$ P_{mut} - P_{ref} $, the absolute difference between <i>mut</i> and <i>ref</i> 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	$ P_{mut} - P_{ref} $, 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](#)

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).

PlantDeepSEA-Multi-Genome-BLAST

Input your sequence below:

GCGGAGGAGGCGGAAGCGGCGGC

Input query sequence →

Select database → Database: ZS97 ▼

E-value: 0.001 ▼

Word size: 11

Sensitivity: NORMAL ▼

Setting parameters →

Search

Copyright © 2020 National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University

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 chromosome and 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

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: GCGGAGGAGGCCGAAGCGGCGGC

M: |||

S: GCGGAGGAGGCCGAAGCGGCGGC

Click here to jump to the model prediction interface,
and the corresponding location information will be filled in automatically.

20

Chapter 6. Blast

ZS97 Model

Variant Effector

(1) By Upload VCF file

Upload one file:

选择文件
未选择文件

Upload

Sequence Profiler

(2) By Upload BED file

Upload one file:

选择文件
未选择文件

Upload

(3) By select a Chromosome 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:

The model corresponds to the choice of database

This model was constructed based on ATAC-seq data from multiple 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 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.

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) .

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/1TvqpN94yE6gg7VHVRNt6kA?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/1AqbUPws9AyTjnBv3vUMmQ?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.

STATISTICS

You can slide the table to see more information.

Species	Samples	Raw reads number	Mapping rate	Marked 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.	Ara-bidop-thaliana	151660948	0.953	0.2564	971300813	1.972950966	195	seq, stem cells, rep3	139540	0.02725399	5399	0.024930943	0.9438	0.1738
A.	Ara-bidop-thaliana	99733622	0.951	0.2035	49818576	1.498751687	27	seq, stem cells, rep2	240981	0.04706666	666	0.043669971	0.9790	0.7398
A.	Ara-bidop-thaliana	94079852	0.949	0.1017	36654398	1.319758499	61	seq, root non-hair, rep1	344379	0.06726452	340	0.064300287	0.9876	0.7826
A.	Ara-bidop-thaliana	47952800	0.9871	0.0603	28801395	1.476572272	50	seq, root tips, rep1	344701	0.06732631	406	0.065400497	0.9857	0.8505

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Marked 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.	Ara-bidop-thaliana stem_cell_rep1	846878	0.9949	0.2180	929652	462.56	325325	ATAC-seq, stem cells, rep1	356863	0.0696	989468	0.0690	10.9562	0.6636
A.	Ara-bidop-thaliana 7days_leaf_rep1	452255	0.9949	0.304	134615	588.32	25571	ATAC-seq, leaf, 7 days, rep1	393580	0.0768	710937	0.0710	10.9217	0.8547
A.	Ara-bidop-thaliana root_tip_rep2	121576	0.9858	0.0849	718948	585.75	220784	ATAC-seq, root tips, rep2	395322	0.0772	115881	0.0759	10.9143	0.8283
A.	Ara-bidop-thaliana root_hair_cell_rep1	109162	0.993	0.1039	628473	590.22	633856	ATAC-seq, root hair, rep1	414566	0.0809	699218	0.0819	10.9897	0.8597
A.	Ara-bidop-thaliana mesophyll_cell_rep1	112729	0.9914	0.3276	653150	416.34	927465	ATAC-seq, meso-phyll cells, rep1	509807	0.0995	716079	0.0966	10.9399	0.8562
A.	Ara-bidop-thaliana mesophyll_cell_rep3	100129	0.9927	0.2577	789457	30.49	876302	ATAC-seq, meso-phyll cells, rep3	546616	0.1067	609375	0.1078	10.9401	0.8346

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Marked reads number	dup30 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.	Ara-bidop-thaliana	2429695	0.9936	0.0578	5514667	0.0582	2346	ATAC-seq, root non-hair, rep2	5579590	0.1089	763671	0.1079	0.7364	0.12593
A.	Ara-bidop-thaliana	999383	0.9932	0.2860	1236870	0.0416	473928	ATAC-seq, mesophyll cells, rep2	5445920	0.1063	656875	0.1088	0.4945	0.09685
A.	Ara-bidop-thaliana	2917154	0.9938	0.0642	4381730	0.0385	264734	ATAC-seq, root hair, rep2	5709570	0.1115	1503906	0.1217	0.9482	0.48079
A.	Ara-bidop-thaliana	1493796	0.9917	0.3	3034571	0.054	33398	ATAC-seq, leaf, 7 days, rep2	5858680	0.1144	2782750	0.1128	0.9039	0.4588
B.	Bdis-tachyon	1454302	0.98715	0.0551	5170227	0.04744	296244	ATAC-seq, root, rep1	40093	0.0078	308640	0.00660	0.2551	0.025137
B.	Bdis-tachyon	1029164	0.99754	0.1146	7932578	0.14566	17113	ATAC-seq, young leaf, rep2	74266	0.0145	507281	0.01158	0.9512	0.13772
B.	Bdis-tachyon	1087020	0.9604	0.1215	8569899	0.07817	23358	ATAC-seq, young leaf, rep1	112185	0.0219	1180281	0.01885	0.9486	0.06820

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Marked 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.	Bdis-tachyon_flower_2	243299529589	0.5249	840396714.747946148	1.4747946148	1629020.031816796876.025390.9398	ATAC seq, flower, rep2	1629020.031816796876.025390.9398	0.031816796876.025390.9398	1629020.031816796876.025390.9398	0.031816796876.025390.9398	0.5461	0.5461
B.	Bdis-tachyon_root_2	45701165064	0.0649	173550649749663702	1.49749663702	1802460.035204296876.028690.9357586285	ATAC seq, root, rep2	1802460.035204296876.028690.9357586285	0.035204296876.028690.9357586285	1802460.035204296876.028690.9357586285	0.035204296876.028690.9357586285	0.5461	0.5461
B.	Bdis-tachyon_flower_1	395727269536	0.5504	13233688.728432655	1.728432655	2156630.042125679680.534348.96234082589	ATAC seq, flower, rep1	2156630.042125679680.534348.96234082589	0.042125679680.534348.96234082589	2156630.042125679680.534348.96234082589	0.042125679680.534348.96234082589	0.5461	0.5461
B.	Bdis-tachyon_flag_leaf_2	639324969606	0.3272	37316076.099082939	1.099082939	4394990.085839678430.572130.926720158794	ATAC seq, flag leaf, rep2	4394990.085839678430.572130.926720158794	0.085839678430.572130.926720158794	4394990.085839678430.572130.926720158794	0.085839678430.572130.926720158794	0.5461	0.5461
B.	Bdis-tachyon_panicle_1	116923882716	0.5741	395569215.731589336	1.731589336	4085960.079808906250.074350.45829786807	ATAC seq, young panicle, rep1	4085960.079808906250.074350.45829786807	0.079808906250.074350.45829786807	4085960.079808906250.074350.45829786807	0.079808906250.074350.45829786807	0.5461	0.5461
B.	Bdis-tachyon_flag_leaf_1	940677029695	0.3964	498210673.450757838	1.450757838	4526620.088410517876.075228.98706896854	ATAC seq, flag leaf, rep1	4526620.088410517876.075228.98706896854	0.088410517876.075228.98706896854	4526620.088410517876.075228.98706896854	0.088410517876.075228.98706896854	0.5461	0.5461
O.	ZS97_panicola	576204869742	0.3419	28421790.377087787	1.377087787	349130.0068189945310.508460.941660666666	ATAC seq, young panicle, rep1	349130.0068189945310.508460.941660666666	0.0068189945310.508460.941660666666	349130.0068189945310.508460.941660666666	0.0068189945310.508460.941660666666	0.5461	0.5461

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	30 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_panicl	85892512	0.151	0.123	584615	1757024	643603	ATAC-seq, stamen & pistil, rep1	49527	0.00967	13121	0.00821	0.9746	0.6851
O. sativa	MH63_panicl	11325206	0.2364	0.0853	761475	952217	926572	ATAC-seq, stamen & pistil, rep2	68509	0.01338	86540	0.01219	0.9839	0.6562
O. sativa	ZS97_panicl	66002916	0.3547	0.0873	291508	640718	667179	ATAC-seq, young panicle, rep5	140295	0.02740	48571	0.01609	0.9545	0.6349
O. sativa	ZS97_panicl	57332544	0.2823	0.0902	288070	153072	267392	ATAC-seq, young panicle, rep4	148634	0.02903	58781	0.01792	0.9558	0.6535
O. sativa	MH63_flag	64745602	0.2291	0.0861	306585	828341	29063	ATAC-seq, flag leaf, rep2	119482	0.02333	39281	0.01986	0.9508	0.6532
O. sativa	ZS97_young	67443808	0.5716	0.155	459014	878236	26667	ATAC-seq, young leaf, rep1	234188	0.04573	84237	0.02811	0.9453	0.5835

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	30 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_panicl	24324634	0.9875	0.3067	110478	18.9542	281875	ATAC-seq, young panicle, rep2	120640	0.0235	6283	0.0291	0.9748	0.5829
O. sativa	MH63_panicl	33309582	0.9863	0.3309	147324	16.5901	154837	ATAC-seq, young panicle, rep2	167176	0.0326	5435	0.0292	0.9756	0.6312
O. sativa	MH63_young	70109302	0.9834	0.4114	280297	14.2124	52855	ATAC-seq, young leaf, rep2	169796	0.0331	6428	0.0299	0.9663	0.5989
O. sativa	MH63_panicl	58832751	0.9843	0.4198	244094	18.8360	6717	ATAC-seq, young panicle, rep1	204168	0.0398	7556	0.0340	0.9637	0.6246
O. sativa	MH63_panicl	54162448	0.9863	0.3469	221575	17.7240	62842	ATAC-seq, young panicle, rep3	201314	0.0393	5913	0.0346	0.9683	0.4827
O. sativa	MH63_panicl	46253072	0.9874	0.2876	254832	14.9454	73604	ATAC-seq, young panicle, rep5	236853	0.0462	6685	0.0391	0.9583	0.6203

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Marked 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_panicula	5115396	0.9752	0.3268	2609708	1.752674025	ATAC-seq, young panicle, rep4	234237	0.04574984	40625	0.39330896	0.5507246	
O. sativa	MH63_young leaf	13893810	0.9864	0.4966	4473454	1.012965991	ATAC-seq, young leaf, rep1	229037	0.04473593	390625	0.40070747	0.689635	
O. sativa	MH63_root	6828499	0.987	0.4127	3256702	1.212471764	ATAC-seq, root, rep2	253463	0.04950439	21805	0.42632064	0.3648	
O. sativa	ZS97_young leaf	3823559	0.9843	0.3616	1178703	1.222785795	ATAC-seq, young leaf, rep2	356413	0.06961129	1740625	0.43302967	0.3675	
O. sativa	ZS97_panicula	3643606	0.9844	0.3379	1528250	1.704685220	ATAC-seq, young panicle, rep3	182557	0.03565896	40625	0.44570964	0.3629	
O. sativa	MH63_flag leaf	1126342	0.987	0.273	685454	1.039285130	ATAC-seq, flag leaf, rep1	292602	0.05714821	28126	0.50107958	0.206896	
O. sativa	ZS97_root	1669914	0.9781	0.5518	540006	1.533979269	ATAC-seq, root, rep1	407748	0.07963828	1250	0.52310949	0.598326	
O. sativa	ZS97_root	7464394	0.985	0.4001	344762	1.763224959	ATAC-seq, root, rep2	426907	0.08338027	4305	0.56890958	0.402768	

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Marked 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	14339198	0.867	0.2213	48732306	1.194293847	ATAC-seq, lemma & palea, rep1	330188	0.0644894737	894737	0.0570407498	0.7825	0.1635
O. sativa	MH63_lemma	85447604	0.9856	0.1906	56875872	0.9280929800375	ATAC-seq, lemma & palea, rep2	336402	0.0657085615626	85615626	0.0576509428689625	0.783	0.1635
O. sativa	MH63_root	23371175	0.769	0.5957	67402065	1.614483825	ATAC-seq, root, rep1	343090	0.067008755626	0.0589304956896631	0.0589304956896631	0.783	0.1635
O. sativa	ZS97_flag leaf	163670488	0.661	0.6583	39551243	1.215777310	ATAC-seq, flag leaf, rep2	270937	0.0529173828105	0.06780849388975838	0.06780849388975838	0.783	0.1635
O. sativa	ZS97_flag leaf	192513110	0.825	0.305	46956084	1.773185516	ATAC-seq, flag leaf, rep1	324916	0.063460195250	0.0797205641005917	0.0797205641005917	0.783	0.1635
O. sativa	ZS97_stamen & pistil	160666018	0.832	0.2821	61616632	1.8360810778	ATAC-seq, stamen & pistil, rep1	449357	0.0877630690605	0.084204785601021	0.084204785601021	0.783	0.1635

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	30 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_18860636	10860636	0.9858	0.2658	612624	13.1776	603872	ATAC-seq, stamens & pistil, rep2	462941	0.0904	131840	0.0932	0.9882	0.7844
O. sativa	ZS97_18844720	18844720	0.9865	0.1542	267580	10.0203	73452	ATAC-seq, lemma & palea, rep2	494763	0.0966	333984	0.0517	0.9563	0.6211
O. sativa	ZS97_18844720	18844720	0.9861	0.1391	261934	18.6618	85427	ATAC-seq, lemma & palea, rep1	601667	0.1175	128859	0.0542	0.9594	0.6366
S. bi-color	Sorghum_panicula_bottom	82882198	0.9871	0.5721	530417	40.1261	20294	ATAC-seq, bottom part of panicle, rep2	14006	0.0021	888675	0.0023	0.9769	0.5856
S. bi-color	Sorghum_panicula_top	17071098	0.9857	0.2408	162396	11.7467	50825	ATAC-seq, upper part of panicle, rep1	14920	0.0023	3105	0.0025	0.9764	0.5897

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	30 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.	Sorghum bicolor	63638	0.88728	0.4883	108270	10.8986	47897	ATAC-seq, root, rep1	37103	0.00579	46337	0.00579	0.7948	0.54359
S.	Sorghum bicolor	69446	0.86961	0.1574	114443	5.8668	3616149	ATAC-seq, lemma & palea, rep2	44558	0.00696	62875	0.00736	0.7784	0.6284
S.	Sorghum bicolor	85958	0.9875	0.1569	529513	13.0346	92824	ATAC-seq, young leaf, rep2	47084	0.00735	56875	0.00817	0.5856	0.48348
S.	Sorghum bicolor	138873	0.9647	0.15199	287659	13.3507	60239	ATAC-seq, young leaf, rep1	46878	0.00732	45875	0.00818	0.9096	0.5896
S.	Sorghum bicolor	84595	0.9831	0.1542	384766	7.4692	580589	ATAC-seq, bottom part of panicle, rep1	51301	0.00801	167812	0.00841	0.4615	0.46453
S.	Sorghum bicolor	39628	0.8290	0.3257	193204	8.8587	697830	ATAC-seq, upper part of panicle, rep2	54562	0.00852	59125	0.00863	0.8276	0.58861

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	30 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.	Sorghum bicolor	201182913	0.1957	0.1605	133533	456717	43324	ATAC-seq, stamen & pistil, rep1	64437	0.01006828	68125	0.01021697	0.972158	0.4123
S.	Sorghum bicolor	245343679	0.1966	0.1495	160133	7732019	460069	ATAC-seq, stamen & pistil, rep2	65513	0.01023638	638625	0.01049078	0.971858	0.77437
S.	Sorghum bicolor	303349282	0.1917	0.4356	523342	6424003	85991	ATAC-seq, flag leaf, rep1	68154	0.01064886	625	0.01103097	0.972301	0.694
S.	Sorghum bicolor	444504692	0.1921	0.412	601128	698226	69930	ATAC-seq, flag leaf, rep2	72738	0.01136926	625	0.01159397	0.976794	0.8591
S.	Sorghum bicolor	95665980	0.1952	0.1691	654426	9099071	105032	ATAC-seq, lemma & palea, rep1	75782	0.01184993	75	0.01239437	0.974530	0.7692
S.	Sorghum bicolor	397322692	0.1927	0.7644	605162	9839337	9326	ATAC-seq, root, rep2	97185	0.01518515	625	0.01547075	0.976163	0.8604

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	30 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.	Sital-ica_panicl_1	177674200883	0.2021	1262273889160802264	ATAC seq, young panicle, repl	48566	0.009485546876	0.005240.72008547688						
S.	Sital-ica_young_leaf_2c_1	700671389694	0.4212	296593202609069107	ATAC seq, young leaf, repl	65424	0.0127782125	0.007478.9724786257						
S.	Sital-ica_flower_1	131086539924	0.23	918095211961026573	ATAC seq, flower, repl	91008	0.01777518	0.010616.9491794872						
S.	Sital-ica_root_2c_1	257357508962	0.4955	7709618.4828 30919	ATAC seq, root, repl	102140	0.01949318750	0.011318.9886703586						
S.	Sital-ica_young_leaf_2c_2	57397792982	0.5489	196386697398941305	ATAC seq, young leaf, rep2	72681	0.01135636250	0.011969.9506400658						
S.	Sital-ica_flower_2	13835702992	0.2077	992705723444490325	ATAC seq, flower, rep2	121054	0.023643359376	0.014790.94912965837						
S.	Sital-ica_flag_leaf_1	14199660915	0.2786	859244790569965882	ATAC seq, flag leaf, repl	162685	0.0317741140605	0.020532.93256016257						

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Marked 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.	Sitalica_panicle_2	150142668	0.2297	100733619	619785594	699	ATAC-seq, young panicle, rep2	1925860	0.0376174	1433126	0.0247709	0.970685	0.4791
S.	Sitalica_flag_leaf_2	100706229	0.2437	636930610	2937368	992	ATAC-seq, flag leaf, rep2	2240320	0.0437582	5626	0.0285709	0.939820	0.638
26.	Zmays	967245089	0.3522	40851048	561478	382	ATAC-seq, upper part of ~1cm ear, rep1	39169	0.0062518	56250	0.0066109	0.969280	0.5392
26.	Zmays	965134809	0.3728	38466971	1.24558	0019	ATAC-seq, bottom part of ~1cm ear, rep2	53425	0.0083469	56250	0.0087109	0.942800	0.526
26.	Zmays	926365609	0.3325	40570461	0.986678	983	ATAC-seq, bottom part of ~1cm ear, rep1	61812	0.0096581	58125	0.0103039	0.960778	0.526

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	30 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.	Zmays_1011098009_32	1011098009	0.3889	0.3889	387844	10.2486	5514	ATAC-seq, upper part of ~1cm ear, rep2	87961	0.01374	19062	0.01378	0.9521	0.645
26.	Zmays_1063301027_22	1063301027	0.3883	0.3883	575815	14.0491	83836	ATAC-seq, young leaf, rep3	101406	0.01980	258593	0.01737	0.9813	0.8076
26.	Zmays_1584206243_2	1584206243	0.1117	0.1117	244422	19.7941	1615869	ATAC-seq, root, rep2	111460	0.02176	953125	0.02087	0.9832	0.6206
26.	Zmays_1593407597_5	1593407597	0.3186	0.3186	804690	11.9520	80430	ATAC-seq, root, rep3	158819	0.03101	1958593	0.03084	0.9811	0.5137
26.	Zmays_1613334252_2	1613334252	0.1359	0.1359	415877	16.6173	90227	ATAC-seq, young leaf, rep1	242864	0.04743	41675	0.04148	0.9886	0.5324
26.	Zmays_2266728050_24	2266728050	0.1413	0.1413	424312	15.0827	729812	ATAC-seq, root, rep1	226605	0.04425	628906	0.04175	0.9835	0.2437
26.	Zmays_1039760848_24	1039760848	0.2306	0.2306	306215	15.4010	74194	ATAC-seq, young leaf, rep2	234570	0.04581	122531	0.04182	0.9846	0.1507

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Marked 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.	Zmays_10590410023	10590410023	0.2676	538499	373660	624274	ATAC-seq, young tass-sel spikelet, repl	250943	0.049016234	680542	0.042025862	0.68285	0.285
26.	Zmays_94375200961	94375200961	0.3106	4241708	586916	36382	ATAC-seq, bottom part of ~1cm tass-sel, repl	243257	0.0475163728	1054290	0.0400860789	0.90360789	0.789
26.	Zmays_11260799975	11260799975	0.289	5519133	528452	5693	ATAC-seq, young tass-sel spikelet, rep2	264358	0.0516322	1876044	0.0441098534	0.8278	0.78
26.	Zmays_91087542144	91087542144	0.3956	3609411	504667	7335	ATAC-seq, ~5mm ear, repl	261472	0.0510683	0	0.0459946	0.912068362	0.862
26.	Zmays_86378300955	86378300955	0.3613	3535194	3707942	1634	ATAC-seq, bottom part of ~1cm tass-sel, rep2	254837	0.04977685	15605464	0.0396034825	0.4825	0.4825

continues on next page

Table 1 – continued from previous page

Species	Samples	Raw reads number	Mapping rate	Markdup rate	30 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.	Zmays_0601sh4024	960144024	0.325	0.030	42612007	1.053287925	ATAC-seq, ~5mm ear, rep2	276548	0.05401328	1250.048693426	0.0579	0.0579		
26.	Zmays_189483074963	189483074963	0.413	0.030	715798470775	68782	ATAC-seq, upper part of ~1cm tass-el, rep1	286296	0.05591751875	0.05098329241075379	0.0579	0.0579		
26.	Zmays_1007504951	1007504951	0.4233	0.030	398283974888620508	68782	ATAC-seq, flag leaf, rep2	386108	0.07541107147	50.0682902528600489	0.0579	0.0579		
26.	Zmays_052493619969	052493619969	0.3709	0.030	435707825500591086	68782	ATAC-seq, flag leaf, rep1	389894	0.0761510251876	0.06903286237934435	0.0579	0.0579		

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](https://doi.org/10.1093/nar/gkab383)