

AutoGP 帮助文档 V2

-  [快速开始](#)
-  [详细指南](#)
-  [补充说明](#)
-  [常见问题](#)
-  [联系我们](#)

简介

为有效推进基因组选择（GS）领域的持续发展，我们 开发了一个集成数据存储、数据前处理、数据分析、基因组选择、GS 论坛为一体的平台——AutoGP。

AutoGP 的设计目标是提供一个高度集成的环境，减少用户在编程方面的工作量，并提供一个直观、易用的操作界面，让用户能够轻松执行一系列复杂任务。具体而言，AutoGP 具备以下特点：

- **多模式的数据库管理平台：**支持多种模式的数据库管理，便于用户对数据进行高效、灵活的处理。
- **简单便捷的数据前处理工具：**对基因数据提取高质量 SNP 数据，通过视频能够实现表型便捷提取。
- **便捷直观的数据分析工具：**提供多种便捷且直观的数据分析工具，帮助用户快速了解和处理数据。
- **多种模型选择的基因组预测方法：**整合了多种传统统计方法、机器学习（ML）和深度学习（DL）技术，为基因组选择提供了强大且灵活的工具。

本文档提供两个版本的操作指南：

1. [快速开始](#)：适用于对 GS 领域有较为丰富经验且对 AutoGP 有一定了解的用户。您可以通过此指南快速完成个人需求任务。
2. [详细指南](#)：适用于对平台操作步骤不熟悉的用户。此指南明确各种输入数据格式要求，以帮助您顺利使用平台。

快速开始

-  [1.登录](#)
-  [2.数据管理](#)
-  [3.数据前处理](#)
-  [4.数据分析](#)
-  [5.基因组选择](#)

1.登录

请通过[导航栏右侧](#)进入，进行个人账户的登录或注册。

请使用个人用户名和密码登录平台。

2.数据管理

本平台为用户提供三种类型的数据库：

1. **个人数据库：**[用户对该数据库中的数据拥有全部权限](#)。此类数据为账号私有数据，仅能由账号本人进行上传、下载和删除。

2. **共享数据库：**[用户可以上传和下载共享数据](#)。此类数据由育种研究者自发上传，以促进 GS 领域的发展，并允许其他用户进行相关研究。经过平台认证后，部分共享数据将被升级为优质数据。

3. **优质数据库：**[用户仅能下载此类数据](#)。优质数据库中的数据经过平台严格筛选和确认，确保其为高品质数据。

以个人数据库为例，用户可以上传需要的基因 `vcf` 文件、表型 `csv` 文件和待预测子代的基因型序列数据 `txt` 文件。共享数据库、优质数据库操作同上，仅有文件上传、删除等权限差异。

此外，对于额外提供的“**表型数据库**”，为“表型提取”功能模块服务，起到对作物视频数据、点云数据的存储作用。

3.数据前处理

高质量 SNP 提取

需上传文件：

- 基因 vcf 文件（全基因组或含大规模的 SNP 数据）

在完成上述内容填写后，点击“提交”按钮提交任务。当“立即下载”按钮亮起时，表示通过特定基因调控网络对应的高质量 SNP 数据已完成提取。点击“立即下载”，即可下载对应 vcf 文件。

表型提取

请通过导航栏进入表型提取界面。

数据上传与选择：

- **上传视频：**点击“上传视频”按钮，用户可以通过扫描二维码，使用微信小程序上传视频数据。
- **视频选择：**数据上传后将被添加到网站数据库中，用户可以通过下拉框选择视频。选择完成后，点击提交，后台将开始进行三维重建。重建过程通常需要 45 分钟以上。完成后，用户可以下滑页面查看三维模型并进行表型提取，并可以鼠标拖动查看。

自动表型提取：

- 点击“自动提取”按钮后，将展示基于深度学习算法的参考表型。

交互式表型提取：

- 点击“交互提取”按钮后，将进入交互表型提取界面。
- 首先鼠标选取“标志物”的长度，并在左上角输入真实长度，从而形成参考长度，便于尺度变换。
- 提供两种交互方式：测量和计数。用户可以点击进行长度测量或获取计数点的数量。

4.数据分析

GWAS 分析

需上传文件：

- 基因 vcf 文件
- 目标 csv 文件

在完成上述内容填写后，点击“提交”按钮提交任务。当“立即下载”按钮亮起时，表示已完成相关 GWAS 分析。

群体划分

需上传文件：

- 基因 vcf 文件

在完成上述内容填写后，点击“提交”按钮提交任务。当“立即下载”按钮亮起时，表示已完成相关群体分析。

表型数据分析

需上传文件：

- 目标 csv 文件

在完成上述内容填写后，点击“提交”按钮提交任务。当“立即下载”按钮亮起时，表示已完成相关 csv 文件的描述性分析。

5.基因组选择

本平台为用户提供四种不同任务的算法功能，分别是模型训练、表型预测、训练预测一体化和选择最优亲本。

1. **模型训练**：该功能允许用户利用一批作物的基因型数据（VCF 格式文件）和对应的表型性状数据，通过机器学习（ML）或深度学习（DL）方法训练出具有一定预测能力的性状预测模型。
2. **表型预测**：该功能允许用户利用“模型训练”输出的模型文件，对待预测材料的基因型数据进行性状预测。
3. **训练预测一体化**：该功能整合了“模型训练”和“表型预测”两大算法功能，

提供了更加便捷的操作界面。

4. **选择最优亲本**：该功能允许用户通过已训练好的模型文件，对已知基因型的种子与一批纯合数据进行杂交，预测其中哪种组合的子代最符合期望。

接下来，依次介绍每个算法模型所需的类型文件要求：

模型训练

需上传文件：

- 基因 vcf 文件
- 单表型 csv 文件

需选择模型：

- 10 种 ML/DL 方法随机选择一种

选填内容：

备注信息（注意：不要出现空格，仅用于简易标注）

在完成上述内容填写后，点击“开始训练”按钮提交任务。当“立即下载”按钮亮起时，表示模型训练已完成。点击“立即下载”，即可下载对应的模型权重文件。

时间参考：

使用 1000 份材料的 5000 个 SNP 进行模型训练，时间约为 1 分钟。

模型训练（含环境信息）

需上传文件：

- 基因 vcf 文件
- 环境 csv 文件
- 各环境对应的单表型 csv 文件

选填内容：

备注信息（注意：不要出现空格，仅用于简易标注）

在完成上述内容填写后，点击“开始训练”按钮提交任务。当“立即下载”按钮亮起时，表示模型训练已完成。点击“立即下载”，即可下载对应的模型权重文件。

表型预测

需上传文件：

- 基因 vcf 文件
- “模型训练”产出的模型权重文件

选填内容：

如果您上传的模型文件为 DNNGP 模型，那么您需要额外上传一个 PCA 模型权重文件。

在完成上述内容填写后，点击“开始预测”按钮提交任务。当“立即下载”按钮亮起时，表示模型训练已完成。点击“立即下载”，即可下载对应文档。

时间参考：

使用 6000 份，时间约为半分钟。

表型预测（含环境信息）

需上传文件：

- 基因 vcf 文件
- 环境 csv 文件
- “模型训练（含环境信息）”产出的模型权重文件

选填内容：

如果您上传的模型文件为 DNNGP 模型，那么您需要额外上传一个 PCA 模型权重文件。

在完成上述内容填写后，点击“开始预测”按钮提交任务。当“立即下载”按钮亮起时，表示模型训练已完成。点击“立即下载”，即可下载对应文档。

训练预测一体化

需上传文件：

- 基因 vcf 文件
- vcf 文件对应的单表型 csv 文件
- 待预测材料的基因 vcf 文件

需选择模型：

- 10 种 ML/DL 方法随机选择一种

选填内容：

备注信息（注意：不要出现空格，仅用于简易标注）

在完成上述内容填写后，点击“开始训练”按钮提交任务。当“立即下载”按钮亮起时，表示模型训练已完成。点击“立即下载”，即可下载对应文档。

时间参考：

使用 1000 份材料的 5000 个 SNP 进行模型训练+6000 份材料做预测，时间约为 2 分钟。

训练预测一体化（含环境信息）

需上传文件：

- 基因 vcf 文件
- 环境 csv 文件
- vcf 文件对应的单表型 csv 文件
- 待预测材料的基因 vcf 文件

选填内容：

备注信息（注意：不要出现空格，仅用于简易标注）

在完成上述内容填写后，点击“开始训练”按钮提交任务。当“立即下载”按钮亮起时，表示模型训练已完成。点击“立即下载”，即可下载对应文档。

选择最优亲本

需上传文件：

- 基因 vcf 文件
- “模型训练”产出的模型权重文件
- 待评估材料的基因序列 txt 文件
- 待期望表型值（Max/Min/任意数值）

选填内容：

如果您上传的模型文件为 DNNGP 模型，那么您需要额外上传一个 PCA 模型权重文件。

在完成上述内容填写后，点击“开始预测”按钮提交任务。当“立即下载”按钮亮起时，表示模型训练已完成。点击“立即下载”，即可下载对应文档。

详细指南

 [1.登录](#)

 [2.数据管理](#)

 [3.数据前处理](#)

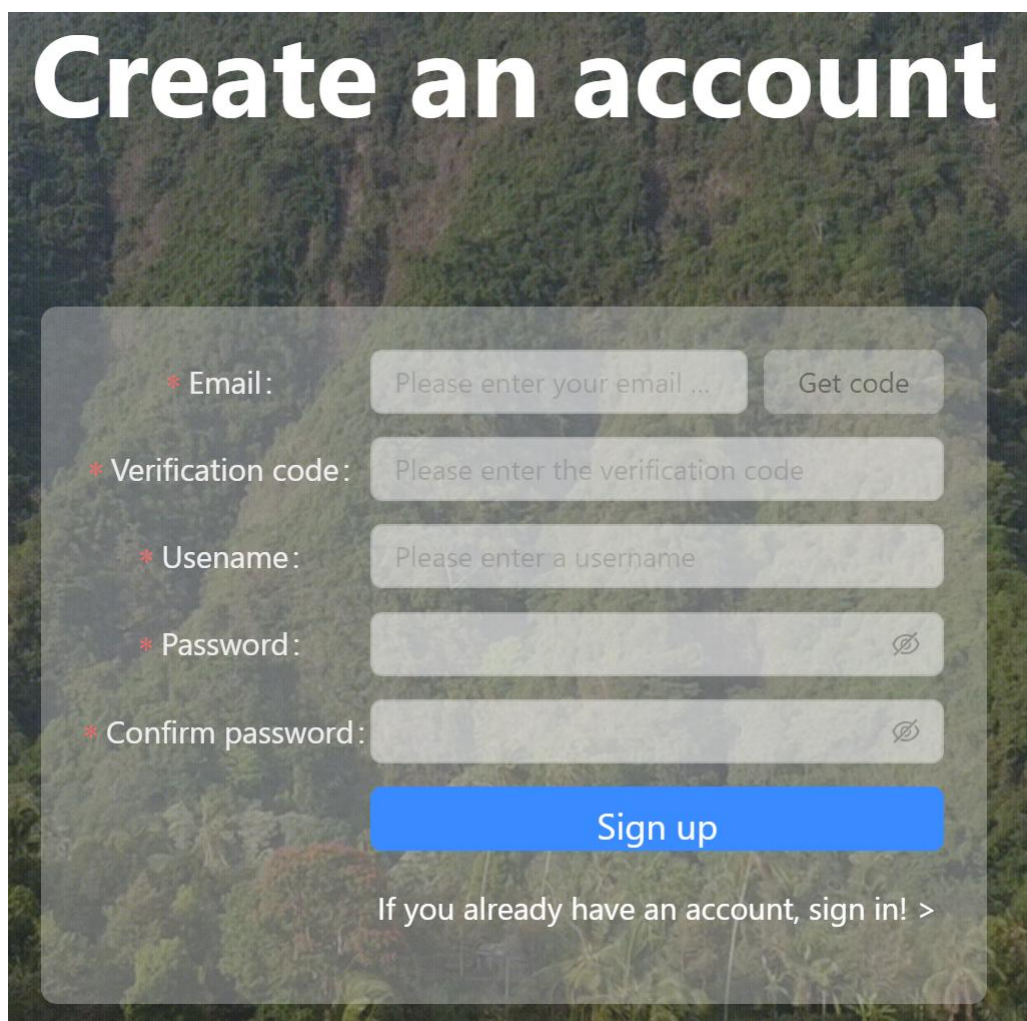
 [4.数据分析](#)

 [5.基因组选择](#)

详细指南将从[示例演示](#)进行展开

1.登录

对于新账户，用户可以通过登录页面上提供的两个账号进行体验登录，也可以通过[通过个人邮箱注册个人专属账号](#)。




The image shows a 'Create an account' form overlaid on a background of a dense green forest. The form is a semi-transparent white rectangle with rounded corners. It contains five input fields, each with a red asterisk icon to its left. The first field is for 'Email' with a placeholder 'Please enter your email ...' and a 'Get code' button to its right. The second field is for 'Verification code' with a placeholder 'Please enter the verification code'. The third field is for 'Username' with a placeholder 'Please enter a username'. The fourth field is for 'Password' with a placeholder 'Please enter a password' and a small circular icon with a diagonal line to its right. The fifth field is for 'Confirm password' with a placeholder 'Please enter a password' and a small circular icon with a diagonal line to its right. Below the input fields is a large blue button with the text 'Sign up'. At the bottom of the form, there is a link that says 'If you already have an account, sign in! >'.


Create an account

* Email:

* Verification code:

* Username:

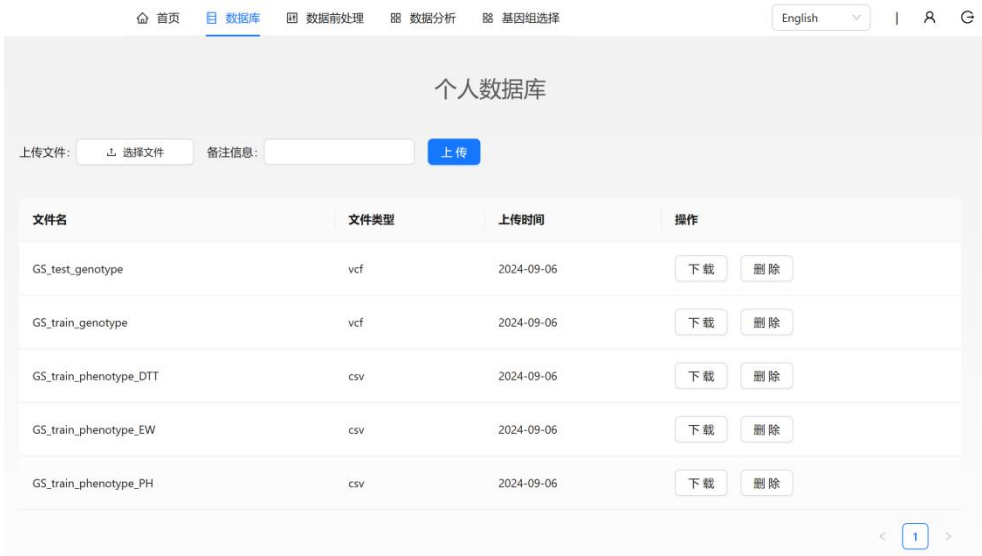
* Password: 

* Confirm password: 

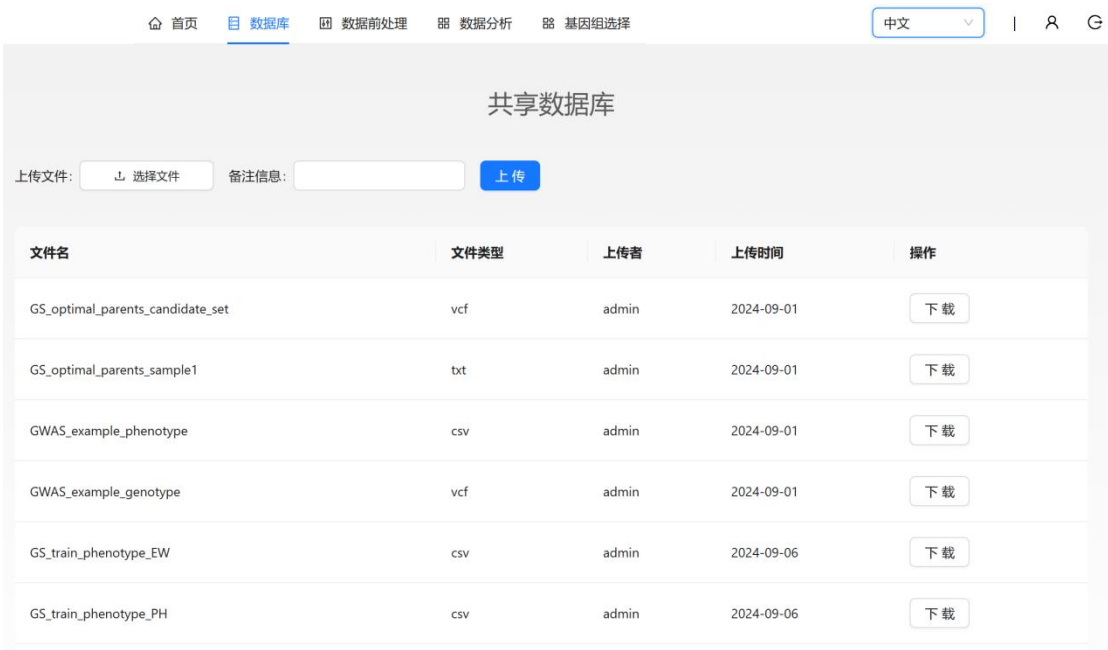
[If you already have an account, sign in! >](#)

2.数据管理

个人数据库：用户本人自己管理，具有上传、下载、删除权限。



共享数据库：用户可以上传个人数据到共享平台，并允许其他用户使用。也可以通过该数据库使用他人数据。



精选数据库：该数据库用平台自行收集整理管理，用户均能使用。

优质数据库			
文件名	文件类型	上传时间	操作
CUBIC_related_genotype	vcf	2024-09-27	下载

3.数据前处理

高质量 SNP 提取

平台提供示例数据：

➤ High-confidence_original_data.vcf

用户通过提取原始（未通过调控网络筛选 SNP）的基因 VCF 文件，点击提交，就可以获得“DTT 对应的基因调控网络”的基因 VCF 文件。

请上传待分析基因型数据：

⬇ VCF 格式

High-confidence_original_data.vcf

提交

下载结果

表型提取

视频拍摄要求

1. 拍摄需放置标志物与目标玉米一旁，固定高度的方形物体即可，不宜太小，如插入地中的标志牌，应达到玉米高度的 1/3，如果苗子小而标志牌大，可以剪短标志牌以便于拍摄。

2. 注意拍摄时目标玉米不要太晃，注意风或者碰撞，晃动的玉米会影响最终的重建效果。

3. 视频尽量拍摄 1280*720 即 720p 高清的视频格式。

拍摄技巧

（1）技巧一：首先进行远景拍摄，围绕物体 360° 环绕扫描。接着进行近景拍摄，同样环绕 360°。拍摄时速度应适中，总时长约为 25 秒，避免过快以防止运动模糊。如果物体较大，如玉米，可以适当放慢速度。

(2) 技巧二：从上到下进行扫描拍摄。确保上下移动时速度适中，并从大约六个不同方向进行上下扫描。

我们以 Corndata1 作为实例演示表型提取流程。

数据上传与选择：

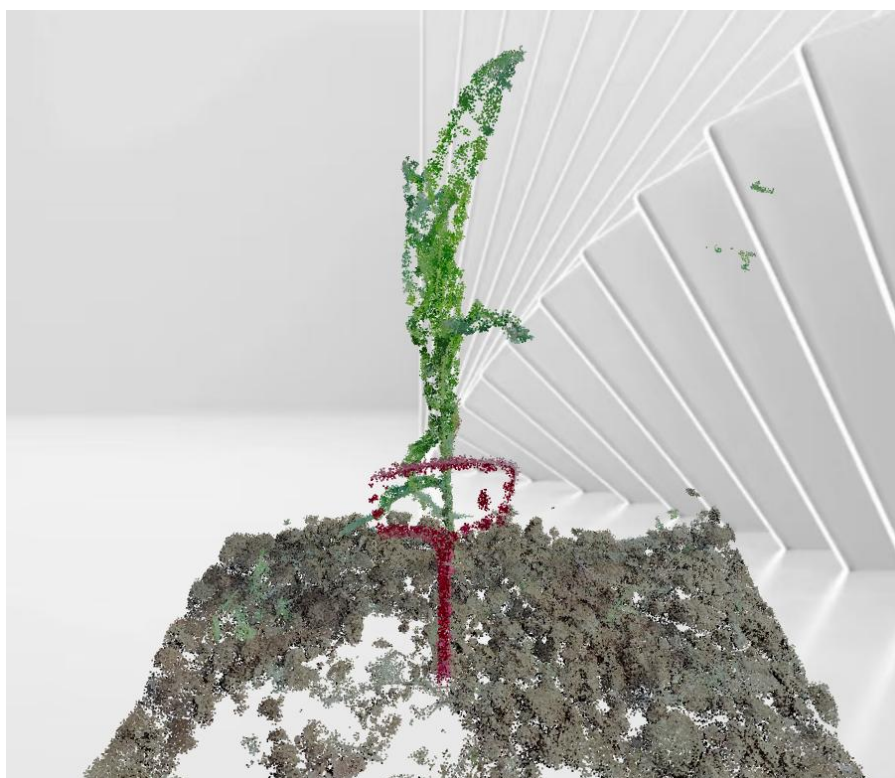
- **上传视频：**点击“上传视频”按钮，通过扫描二维码，上传 Corndata1 视频数据，在小程序填写上传者姓名、区域、品种等信息。



- **视频选择：**数据上传后将被添加到网站数据库中，点击下拉框，选择 Corndata1 数据，视频自动开始三维重建，等待约 45min 即可重建完成。



三维重建完成后，下拉网页，即可看到 Corndata1 的三维模型。



自动表型提取：

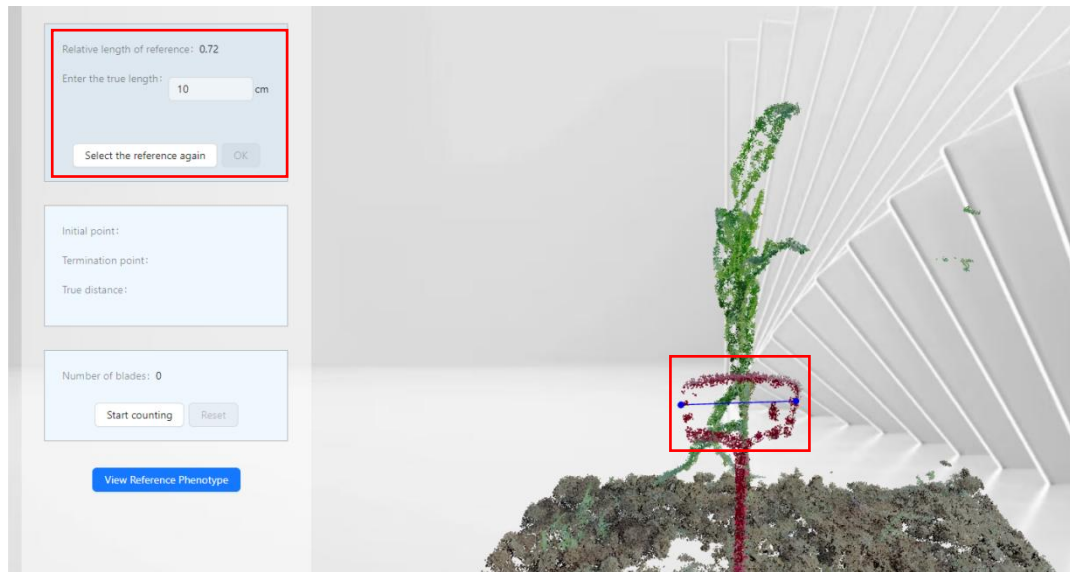
- 点击“自动提取”按钮后，将在网页左端展示参考表型（该功能正在升级维护中）。

自动表型提取稳健性仍需改进，用户可以选择使用交互式表型提取进行进一步分析。

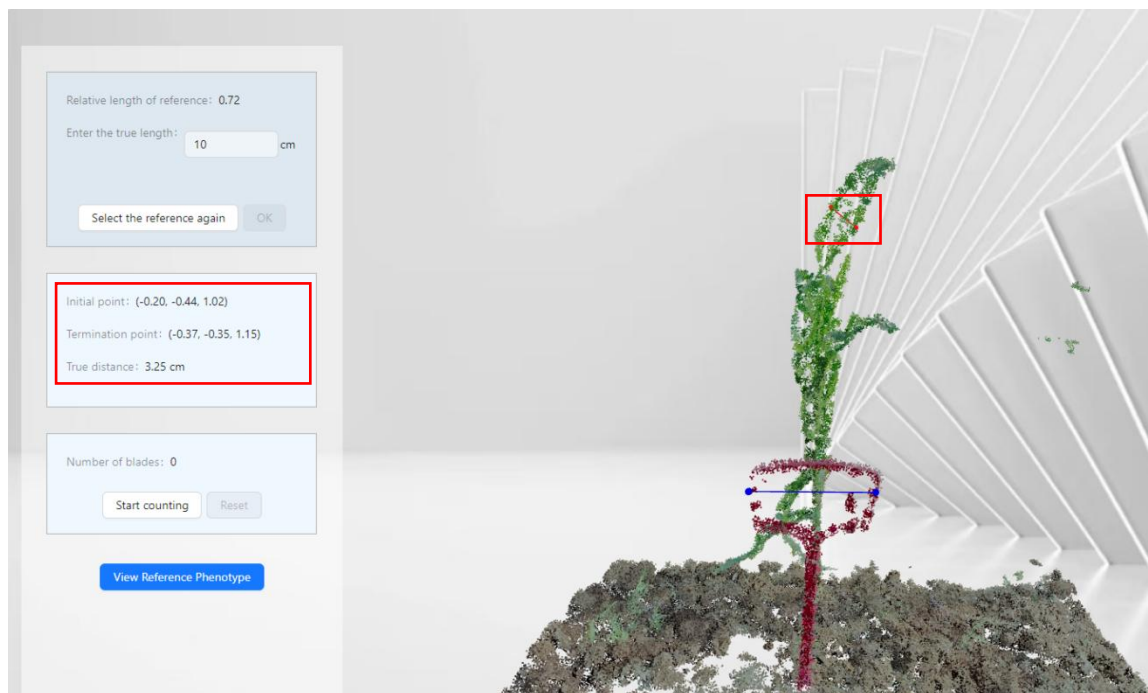
交互式表型提取：

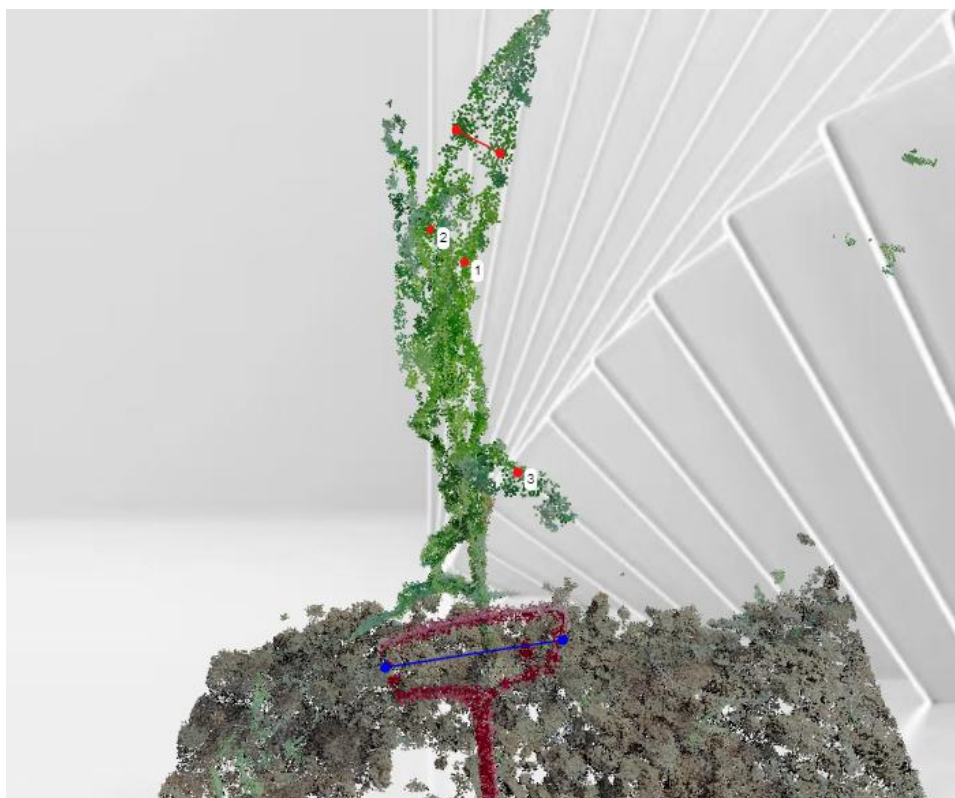
- 页面左端点击“交互提取”按钮后，用户可以通过鼠标选取三维点云中的关键点进行交互。
- 首先鼠标点击 Corndata1 “标志物”的两侧，获取标志物的相对长度，

并在左上角输入真实长度 10cm，从点击 OK，即可生成转换比例。



- 随后，我们可以交互点击叶宽，株高等表型，如下图，红色线段所示，还可以点击“开始计数”按钮，点击叶片进行叶片计数，计数模式下，点数将显示在植物点云上。





4.数据分析

GWAS 分析

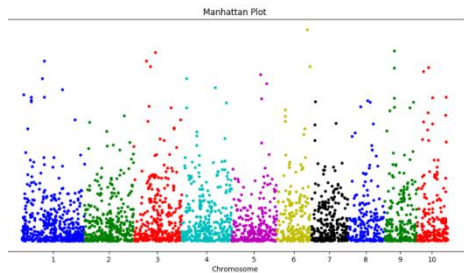
平台提供示例数据：

- GWAS_example_genotype.vcf
- GWAS_example_genotype.csv

用户通过提取基因的 VCF 文件和 CSV 文件，点击提交，可获得关于个 SNP 的 P 值及对应的曼哈顿图。【下载可获得对应曼哈顿图和各 SNP 的 P 值 csv 文件】

* 请上传待分析基因型数据: GWAS_example_genotype.vcf

* 请上传待分析单表型数据: GWAS_example_phenotype.csv



群体划分

平台提供示例数据:

➤ Population_division_example.vcf

用户通过提取基因的 VCF 文件和“期望群体划分规模数”，点击提交，可获得对应的群体划分分布图，以及获得对应的分群详情的 CSV 文件。【下载可获得对应群体划分分布图和分群详情的 CSV 文件】

群体划分

* 上传待训练基因型数据: Population_division_example.vcf

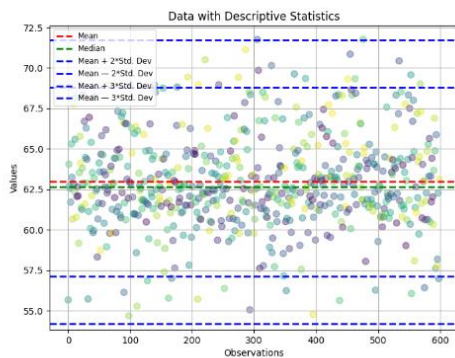
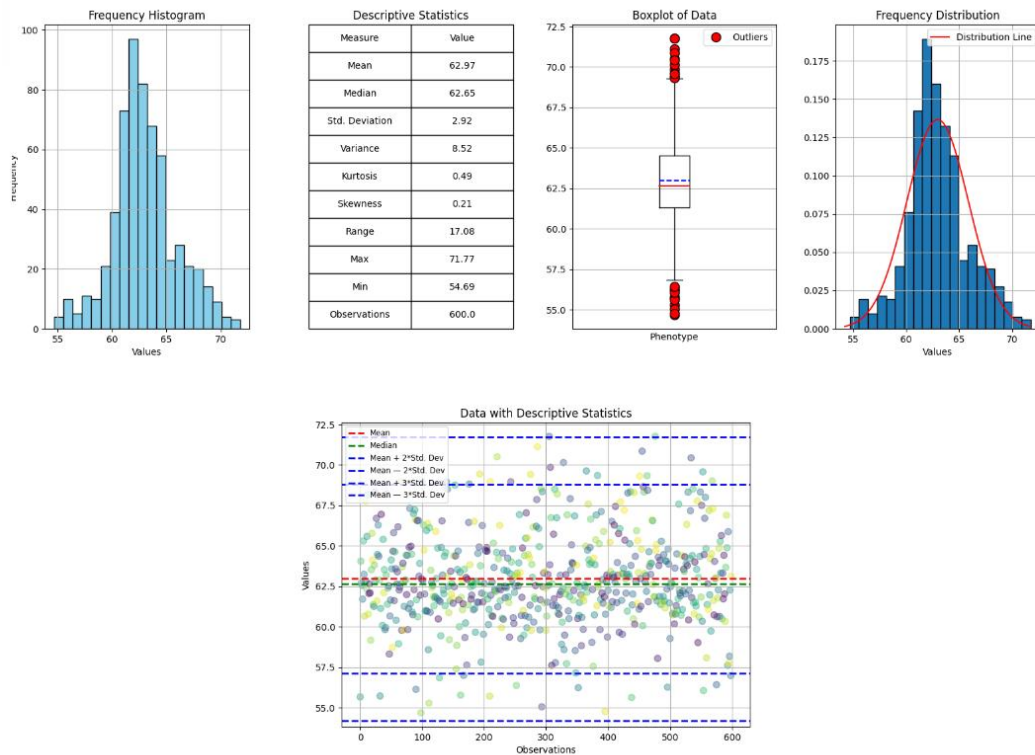
请输入期望群体个数:

表型数据分析

平台提供示例数据:

➤ Data_analysis_example.csv 【任意一个 GS 的 csv 文件都行】

用户通过 CSV 文件，点击提交，可获得对应的描述性分析图，从而直观上查看数据基本情况。



5.基因组选择

模型训练

平台提供示例数据：【按照顺序】

- GS_train_genotype.vcf
- GS_train_phenotype_PH.csv

用户通过提交文件（并完成其他部署），点击提交，可获得对应模型文件。

【注：输出为.pth.zip，不需要进行解压操作，可直接作为权重文件用于后续输入操作】

* 上传待训练基因型数据:

* 上传待训练单表型数据:

模型选择: XGBOOST ▼

备注信息:

XGBOOST

GBDT

MLP

SVM

RandomFor...

DNNGP

DLGWAS

DeepGS

模型训练（含环境信息）

平台提供示例数据：【按照顺序】

- mul_genotype_train.vcf
- mul_environment_train.csv 【与“模型训练”的差异点】
- mul_phenotype_train.csv 【含环境信息，所以可以填入几个环境对应的表型】

用户通过提交文件（并完成其他部署），点击提交，可获得对应模型文件。

* 上传待训练基因型数据:

* 上传待训练环境数据:

* 上传待训练各环境表型数据:

备注信息:

表型预测

平台提供示例数据：【按照顺序】

- GS_test_genotype.vcf
- GS_test_model.pth.zip

【使用平台输出权重文件，并且避免出现文件名有（1）（2）情况】

添加环境数据

* 上传待预测基因型数据:

vcf 格式 GS_test_genotype.vcf

* 上传模型权重

zip 格式 GS_test_model_PH_XGBOOST_1_R_0.89.pth.zip

开始预测

下载CSV文件

用户通过提交文件（并完成其他部署），点击提交，可获得对应预测的 csv 文件。

【注：输出的 csv 文件可以在“表型数据分析”进行可视化查看】

id	predict
MG_115_X_MG_1528	238.24701
MG_991_X_MG_1524	257.93466
MG_162_X_MG_1540	274.8416
MG_204_X_MG_1520	270.39908
MG_68_X_MG_1545	262.42966
MG_621_X_MG_1532	241.12639
F349_X_MG_1535	244.07361
MG_204_X_MG_1536	242.68143
MG_923_X_MG_1541	361.16336
MG_1303_X_MG_1527	263.64697
MG_556_X_MG_1535	239.821
MG_447_X_MG_1518	270.0877

表型预测（含环境信息）

平台提供示例数据：【按照顺序】

- mul_test_genotype.vcf
- mul_environment_test.csv 【与“表型预测”的差异点】
- mul_test_model.pth.zip 【模型来源必须来自于“模型训练（含环境信息）”】

用户通过提交文件（并完成其他部署），点击提交，可获得对应结果。

ID	BJ	HeB	JL	LN	HN
MG_115_X_MG_1528	17.143845	17.152086	17.18957	17.165375	17.111414
MG_991_X_MG_1524	17.140156	17.147812	17.185732	17.160892	17.109158
MG_162_X_MG_1540	17.136444	17.144053	17.18276	17.157333	17.105593
MG_204_X_MG_1520	17.13905	17.146381	17.184267	17.159435	17.10772
MG_68_X_MG_1545	17.136644	17.144276	17.183714	17.158031	17.106506
MG_621_X_MG_1532	17.13686	17.144392	17.180227	17.156788	17.107155
F349_X_MG_1535	17.14204	17.149588	17.18623	17.162268	17.111845
MG_204_X_MG_1536	17.138123	17.145449	17.180069	17.157507	17.109976
MG_923_X_MG_1541	17.138506	17.145596	17.180561	17.15708	17.107817
MG_1303_X_MG_1527	17.136435	17.14428	17.183271	17.157352	17.105667
MG_556_X_MG_1535	17.14391	17.152077	17.189358	17.165434	17.112888
MG_447_X_MG_1518	17.138042	17.145672	17.182762	17.158417	17.10751
MG_631_X_MG_1539	17.143354	17.15051	17.186666	17.162378	17.112217
MG_1236_X_MG_1526	17.13969	17.148052	17.192438	17.163284	17.108932
MG_1242_X_MG_1544	17.137096	17.14543	17.187044	17.159653	17.105282

训练预测一体化

平台提供示例数据：【按照顺序】

- GS_train_genotype.vcf
- GS_train_phenotype_PH.csv
- GS_test_genotype.vcf

用户通过提交文件（并完成其他部署），点击提交，可获得对应权重文件和预测表型文件。

添加环境数据

* 上传待训练基因型数据:

vcf 格式
GS_train_genotype.vcf

* 上传待训练单表型数据:

csv 格式
GS_train_phenotype_PH.csv

* 上传待预测基因型数据:

vcf 格式
GS_test_genotype.vcf

模型选择:
XGBOOST

备注信息:

开始训练和预测

下载模型文件

下载CSV文件

训练预测一体化（含环境信息）

平台提供示例数据：【按照顺序】

- mul_genotype_train.vcf
- mul_environment_train.csv 【与“训练预测一体化”的差异点】
- mul_phenotype_train.csv

➤ mul_genotype_test.vcf

用户通过提交文件（并完成其他部署），点击提交，可获得对应结果。

选择最优亲本

平台提供示例数据：【按照顺序】

➤ GS_optimal_parents_candidate.vcf

➤ GS_optimal_parents_sample1.txt

➤ GS_test_model.pth.zip

➤ 预测相关数值：55

用户通过提交文件（并完成其他部署），点击提交，可获得对应结果。

* 上传杂交亲本基因型数据:

📁 vcf 格式

GS_optimal_parents_candidate_set.vcf

* 待杂交材料的基因序列txt文件:

📁 txt 格式

GS_optimal_parents_sample1.txt

* 上传模型权重:

📁 zip 格式

GS_test_model_PH_XGBOOST_1_R_0.89.pth.zip

* 期望表型值大小:

222

提交

立即下载

	id	predict
74	MG_1021_X_MG_1521	272.217712
723	MG_1021_X_MG_1518	272.764648
546	MG_675_X_MG_1518	273.256104
366	MG_552_X_MG_1527	274.746185
680	MG_911_X_MG_1528	275.095947

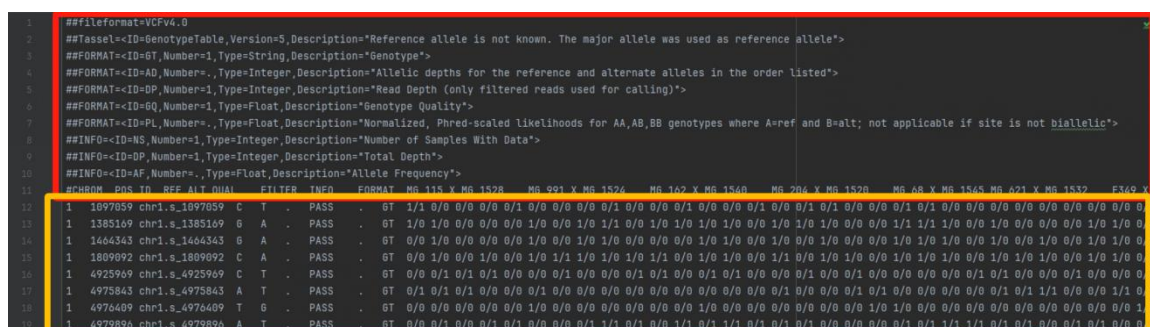
补充说明

1.数据格式要求

作为补充，下面严格要求了对于 AutoGP 平台关于 vcf 文件、csv 文件、txt 文件的格式要求：

VCF 文件

vcf 文件对应：行表示为各 SNP，列表示为各样本。



The image shows a screenshot of a VCF file. The first 11 lines are header information, including file format, version, and various annotations. The data section starts at line 12, with columns for chromosome, position, reference allele, alternate allele, filter, and genotype. The genotype column contains values like '0/0', '0/1', and '1/1' for different samples.

vcf 文件前 11 行为常规 vcf 格式的备注信息，从第 12 行开始为待读取信息，其中前 9 列数据为关于 SNP 数据的备注信息，第 10 列开始每一列对应每个材料在各 SNP 上的详细情况，其中“0/0”“1/0”“0/1”“1/1”分别表示等位基因未突变、单突变、双突变。

CSV 文件

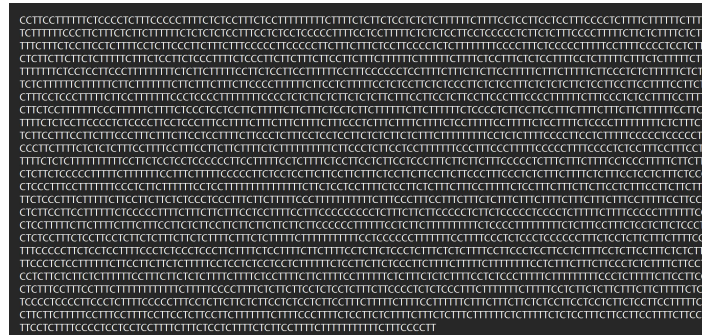
csv 文件对应：行表示为各样本，列表示为表型。

csv 文件，第一列的每行表示为样本名（这个样本顺序应该和 VCF 文件中第 10 行中样本顺序保持一致），第二列的每行表示为样本对应的表型值。

注意：目前本平台提供的单表型的预测，因此用户仅提供上述两列格式的 csv 格式文件。

Lineid	DTT
MG_115_X_MG_1528	55
MG_991_X_MG_1524	64
MG_162_X_MG_1540	63
MG_204_X_MG_1520	64
MG_68_X_MG_1545	63
MG_621_X_MG_1532	65
F349_X_MG_1535	63
MG_204_X_MG_1536	60
MG_923_X_MG_1541	68
MG_1303_X_MG_1527	63
MG_556_X_MG_1535	63
MG_447_X_MG_1518	66
MG_631_X_MG_1539	
MG_1236_X_MG_1526	60
MG_1242_X_MG_1544	63
MG_891_X_MG_1533	62
MG_552_X_MG_1546	62
MG_161_X_MG_1536	59
MG_522_X_MG_1520	63
MG_220_X_MG_1525	66
MG_447_X_MG_1523	64

TXT 文件



txt 文件，表示为待预测单样本的对应的 SNP 拼接而成的基因序列。例如，vcf 中涉及 20000 个 SNP，则单样本对应的 txt 文件就表示为 20000 个 SNP 拼接而成的长度为 20000 的基因序列。

2.算法设计细节

为了更加详细给 GS 研究者了解各算法架构设计，本平台下面详细讲述个算法架构设计细节：

SVM

在本平台中，如果选择了支持向量机 (SVM) 作为模型配置，系统将自动使用 SVR(支持向量回归)进行训练和预测。本文并没做太多操作，仅仅使用 `model = SVR(verbose=0)` `model.fit(train_x, train_y)`进行相关训练。

XGBOOST

在本平台中，如果选择了 XGBoost 作为模型配置，系统将使用 XGBRegressor 进行训练，并通过网格搜索进行参数调优。首先，定义模型的初始参数，其中学习率 (learning_rate) 设置为 0.05，初始的决策树数量 (n_estimators) 为 500，最大深度 (max_depth) 为 5，最小子节点权重 (min_child_weight) 为 1,随机种子 (seed) 为 0,子样本比例 (subsample) 为 0.8,列采样比例 (colsample_bytree) 为 0.8，树的节点分裂所需的最小损失函数下降值 (gamma) 为 0，L1 正则化系数 (reg_alpha) 为 0，L2 正则化系数 (reg_lambda) 为 1。同时，设定网格搜索的参数范围，候选的决策树数量 (n_estimators) 为 500、600 和 700。然后，使用这些参数初始化一个 XGBRegressor 模型对象。接着，通过 GridSearchCV 对模型进行参数调优，评

估指标为 R^2 ，并采用 5 折交叉验证。最后，使用提供的训练数据对调优后的模型进行训练，训练过程中输出详细信息。通过以上步骤，XGBoost 模型完成了初始化、参数调优和训练，可以用于后续的数据预测任务。

GBDT

在本平台中，如果选择了梯度提升决策树 (GBDT) 作为模型配置，系统将使用 GradientBoostingRegressor 进行训练。首先，系统会初始化一个 GradientBoostingRegressor 模型对象，其中随机种子 (random_state) 设置为 123，以确保结果的可重复性，并设置 verbose 为 0，以关闭训练过程中的详细信息输出。然后，系统会使用提供的训练数据 (train_x 和 train_y) 对模型进行训练。通过这些步骤，GBDT 模型完成了初始化和训练，可以用于后续的数据预测任务。

MLP

在本平台中，如果选择了多层感知器 (MLP) 作为模型配置，系统将使用 MLPRegressor 进行训练。首先，系统会初始化一个 MLPRegressor 模型对象，其中隐藏层的结构 (hidden_layer_sizes) 设置为两层，第一层有 3060 个神经元，第二层有 64 个神经元，激活函数 (activation) 选择 ReLU，优化算法 (solver) 选择 Adam，L2 正则化项的参数 (alpha) 设置为 0.01，最大迭代次数 (max_iter) 设置为 200。通过这些步骤，MLP 模型完成了初始化，可以用于后续的数据训练和预测任务。

RF

在本平台中，如果选择了随机森林 (RandomForest) 作为模型配置，系统将使用 RandomForestRegressor 进行训练。首先，系统会初始化一个 RandomForestRegressor 模型对象，其中决策树的数量 (n_estimators) 设置为 20。然后，系统会使用提供的训练数据 (train_x 和 train_y) 对模型进行训练。通过这些步骤，随机森林模型完成了初始化和训练，可以用于后续的数据预测任务。本平台的设计确保了模型配置的灵活性和易用性，使得研究者可以专注于研究本身，而无需过多关心底层实现的复杂性。

DeepGS、DLGWAS、DNNGP、SoyDNGP

本平台对于 DL 模型均仿照源论文架构重新进行复现搭建，将其设计成确保

其架构能够对于不同大小矩阵能够重复使用。

下面是对于其他参数的配置情况：

训练时的批处理大小由 `train_batch` 参数设置，默认值为 64；验证时的批处理大小由 `valid_batch` 参数设置，默认值也为 64。训练的总轮数通过 `epochs` 参数指定，默认值为 28。为了确保结果的可重复性，可以使用 `seed` 参数设置随机种子，默认值为 6。学习率由 `lr` 参数控制，默认值为 0.001，而权重衰减系数用于 L2 正则化，通过 `weight_decay` 参数设置，默认值为 $1e-5$ 。

在模型训练中，使用了平滑 L1 损失函数来计算损失。优化器选用了 Adam，其学习率和权重衰减系数分别由 `lr` 和 `weight_decay` 参数指定。为了调整学习率，采用了阶梯式学习率调度器，每 10 个 `epoch` 学习率会降低到原来的 10%。通过这些配置和设置，模型的训练过程可以根据需要进行灵活调整，从而实现更好的性能和结果。

常见问题

若对 AutoGP 有任何建议或疑问，请通过平台提供的反馈途径或电子邮件与我们联系。

当前为 V2 版本，希望这个版本符合您的要求。

联系我们

如果您对 AutoGP 有任何建议或疑问，请通过以下方式与我们联系：

- 电子邮箱：wuh_hzau@163.com
- 地址：湖北省武汉市华中农业大学信息学院
- 电话：

在联系我们之前，您可以访问我们的常见问题（FAQ）页面，查看是否已有相关问题的解答。