方法：

将54条已知血清型的HMTp210和345条未知血清型的HMTp210氨基酸序列用mafft进行序列比对，再用Trimal对比对结果切割对齐。之后去掉所得序列中去掉大于全长15%的gap，得到了399条拥有267个氨基酸的HMTp210序列。将已知血清型的54条HMTp210序列以比例7：3随机分为37条和17条，分别作为训练数据集和测试数据集。以37条HMTp210氨基酸序列的每个氨基酸的疏水性作为特征集，一共267个特征，训练机器训练模型。首先使用python包“sklearn”训练Support Vector Machine模型，该模型使用以下参数进行训练：kernel='linear',C=200,probability=True。除此之外还在默认设置下训练了Random Forest，Gradient Boosted Regression Tree ，Naïve Bayes，Logistic Regression，Multilayer Perceptron，Decision Tree模型。用 267个特征训练的支持向量机模型具有最高的准确性和整体性能，因此被选为预测 HMTp210血清型的首选模型。根据3中预测血清型的概率为每个预测分配置信度分数。具有最高概率的类表示预测的血清型。因此，第一个和第二个预测类之间的概率差异越大，预测就越可靠。为了解释这种分数差异，以下面方案划分三个置信度水平：如果第一个类别的概率是第二个预测类别的概率的3倍以上，则将其视为高置信度预测。如果差值小于随机概率的3倍但大于概率的1.5倍，则将其视为中等置信度预测。如果两者都不是，则它是一个低置信度预测。

The amino acid sequences of 54 HMTp210 with known serotypes and 345 HMTp210 with unknown serotypes were aligned using MAFFT. The aligned results were then trimmed by Trimal. After removing sequences with gaps exceeding 15% of the full length, 399 HMTp210 sequences, each containing 267 amino acids, were obtained. The 54 HMTp210 sequences with known serotypes were randomly divided into 37 and 17 sequences at a ratio of 7:3, serving as the training dataset and test dataset, respectively. The hydrophobicity of each amino acid in the 37 HMTp210 amino acid sequences, totaling 267 features, was used as the feature set to train a machine learning model. Initially, the Support Vector Machine (SVM) model was trained using the Python package "sklearn" with the following parameters: kernel='linear', C=200, probability=True. Additionally, Random Forest, Gradient Boosted Regression Tree, Naïve Bayes, Logistic Regression, Multilayer Perceptron, and Decision Tree models were trained using their default settings. The SVM model trained with 267 features exhibited the highest accuracy and overall performance, thus being selected as the preferred model for predicting HMTp210 serotypes. Confidence scores were assigned to each prediction based on the predicted serotype probabilities among the three serotypes. The serotype with the highest probability represented the predicted serotype. Therefore, the larger the difference in probability between the first and second predicted serotypes, the more reliable the prediction. To interpret this score difference, the following scheme was implemented to classify the confidence levels: if the probability of the first serotype was more than three times that of the second predicted serotype, it was considered a high-confidence prediction. If the difference was less than three times the random probability but more than 1.5 times, it was considered a medium-confidence prediction. If neither condition was met, it was classified as a low-confidence prediction.

结果与讨论：

HMTp210的region2 的基因同源性较低，是血清型差异分析的重点，但仅依据region2绘制出的进化树不足以用来预测血清型，因为region2序列中可能存在与决定血清型无关的序列，也有可能region2中只有结构上外露的氨基酸决定血清型，不外露的基因的进化没有明显规律。为了测试从这项整体分析中收集的进化特征是否可用于更好地预测未知HMTp210的血清型，我们采用了一个 ML 框架，该框架从HMTp210氨基酸序列的疏水性中学习，以预测研究不足的序列的特异性。用一 54个HMTp210血清型已知序列与345个未知序列的HMTp210进行序列比对，将公共核心内的每个对齐位置的氨基酸的疏水性作为特征集，然后，这些属性被用作训练多个 ML 模型的特征。在使用的 7 种方法中，支持向量机（SVM）模型基于使用267个贡献特征实现了最佳预测性能（准确率~98%），得到的准确度如表（图3-10B）。以测试数据集用SVM预测出的预测结果作混淆矩阵。为了进一步验证该模型，用54个HMTp210的验证集测试了它的性能（图3-10C）。用得到的支持向量机模型预测了345个未知血清型的HMTp210的血清型（图3-10D）。

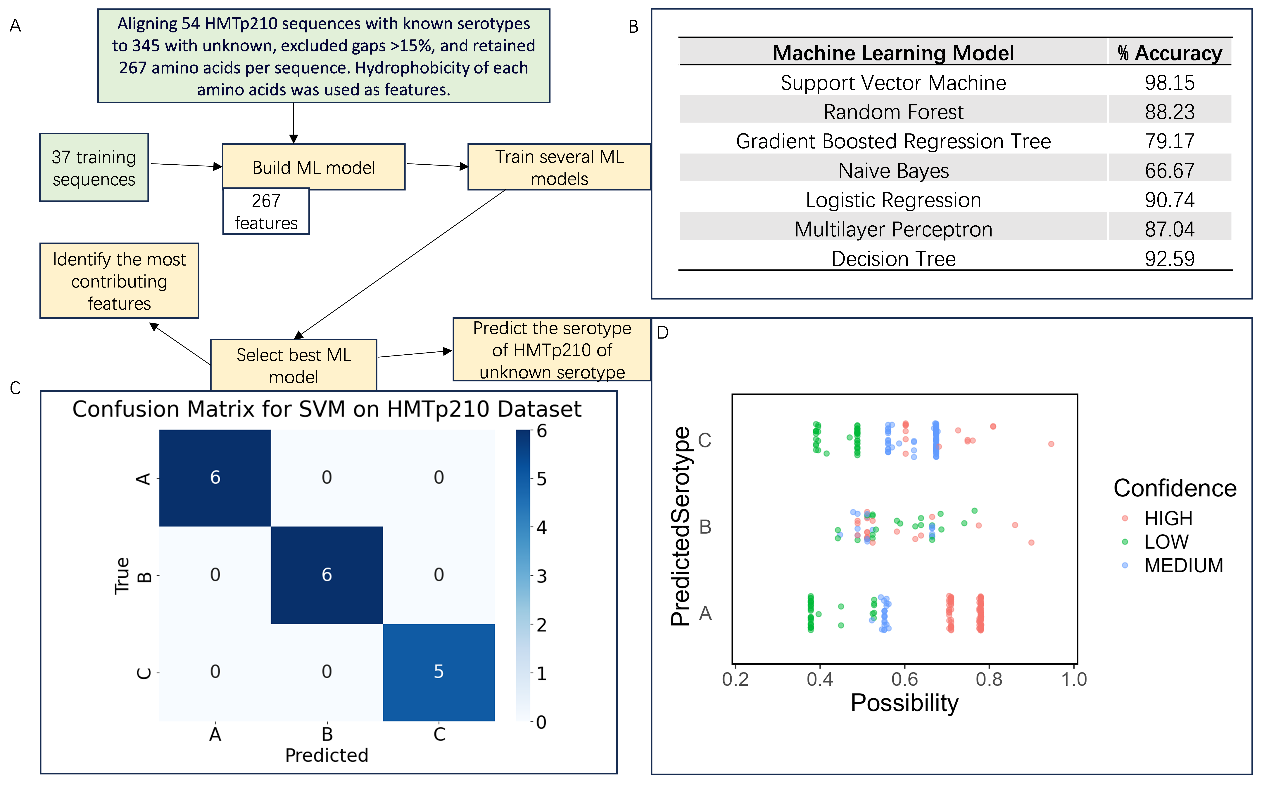
The gene homology in region 2 of HMTp210 is relatively low, making it a focus for serotype difference analysis. However, constructing an evolutionary tree solely based on region 2 is insufficient for predicting serotypes, as there may be sequences in region 2 that are unrelated to serotype determination, or it may be that only the structurally exposed amino acids in region 2 determine the serotype, while the evolution of non-exposed genes does not follow a clear pattern. To test whether evolutionary features collected from this comprehensive analysis can be used to better predict the serotypes of unknown HMTp210, we employed an ML framework that learns from the hydrophobicity of HMTp210 amino acid sequences to predict the specificity of understudied sequences. Sequence alignment was performed between 54 HMTp210 sequences with known serotypes and 345 HMTp210 sequences with unknown serotypes. The hydrophobicity of amino acids at each aligned position within the common core was used as the feature set. These properties were then used as features to train multiple ML models. Among the seven methods used, the Support Vector Machine (SVM) model achieved the best predictive performance (accuracy ~98%) based on 267 contributing features, with the obtained accuracy presented in Table (Fig 2(b)). The prediction results of the SVM on the test dataset were used to construct a confusion matrix. To further validate the model, its performance was tested using a validation set of 54 HMTp210 sequences (Fig 2(c)). The obtained SVM model was used to predict the serotypes of the 345 HMTp210 sequences with unknown serotypes (Fig 2(d)). 

Figure.2 用机器学习模型预测HMTp210的血清型。(a)是数据处理流程，(b)是7种模型预测HMTp210血清型结果的准确性，(c)是根据使用效果最佳的支持向量机模型预测测试数据集得到的数据制作的混淆矩阵，(d)是用该支持向量机模型预测345种未知血清型的HMTp210得到的结果制作的散点图，其中横坐标是为对应血清型的可能性，纵坐标是三种血清型，以决策函数值的绝对值为置信度，红色代表对应的置信度高，绿色代表置信度中等，蓝色代表对应的置信度低。