**Nghiên cứu áp dụng các công cụ học máy trong dự đoán chủng tộc sử dụng dữ liệu STR**

Tác giả: Dương Công Trình, Phạm Xuân Cường, Phùng Trí Đức

C06, Bộ Công an;

Email: [Trinh.DC241087M@sis.hust.edu.vn](mailto:Trinh.DC241087M@sis.hust.edu.vn); [Cuong.PX241090M@sis.hust.edu.vn](mailto:Cuong.PX241090M@sis.hust.edu.vn); [duc.pt241153m@sis.hust.edu.vn](mailto:duc.pt241153m@sis.hust.edu.vn).

*Tóm tắt.*

**Mục tiêu nghiên cứu:** Nghiên cứu khả năng sử dụng dữ liệu đoạn lặp ngắn STR (*Short Tandem Repeat*) trong việc dự đoán chủng tộc. Đầu tiên tiến hành khảo sát tập dữ liệu đa chủng tộc sử dụng các phương pháp xác định cấu trúc và khoảng cách di truyền để nhìn rõ sự khác biệt về chủng tộc. Sau đó, tiến hành chia tập dữ liệu thành tập huấn luyện và tập đánh giá sử dụng các phương pháp học máy khác nhau trong việc dự đoán chủng tộc trên cả dữ liệu thực tế và dữ liệu mô phỏng.

**Đối tượng và phương pháp nghiên cứu:** Nghiên cứu tiến hành so sánh hiệu năng dự đoán dân tộc của các mô hình học máy Bernoulli, Multi Layer Perceptron, Support Vector Machine, Decision Tree, Random Forest, XGBoost và AdaBoost. Tập dữ liệu sử dụng là tập dữ liệu ADN định dạng STR của 1036 mẫu thuộc 4 nhóm chủng tộc từ nguồn dữ liệu chia sẻ bởi Viện Tiêu chuẩn và Công nghệ Quốc gia Hoa Kỳ (NIST - National Institute of Standards and Technology) kết hợp với 408 mẫu dữ liệu người Việt. Ngoài ra đánh giá mở rộng trên tập dữ liệu gồm 4004 mẫu mô phỏng 4 chủng tộc trên dựa trên tần suất alen của tập dữ liệu NIST kết hợp với tập dữ liệu mô phỏng 1143 mẫu người Việt.

**Kết quả:** Kết quả trên tập dữ liệu thực tế cho thấy mô hình Bernoulli phát triển dựa trên Naive Bayes có hiệu năng dự đoán tốt nhất với F1-score đạt 0.68, độ chính xác 0.76 và độ nhạy 0.68. Kết quả trên tập mô phỏng không tốt đều giữa các mô hình cho thấy việc tạo dữ liệu mô phỏng chỉ dựa trên tần số không phản ánh được đúng bản chất dữ liệu ADN của các chủng tộc khác nhau. Mã nguồn và dữ liệu được chia sẻ theo link <https://github.com/DuongCongTrinh98/vn_population_genetics>.

**Kết luận:** Nghiên cứu này cho thấy việc sử dụng dữ liệu STR có tiềm năng trong việc dự đoán chủng tộc của người bất kỳ. Mô hình Bernoulli có hiệu quả phù hợp với dữ liệu đa chiều và sử dụng tính toán xác suất Bayes trong dự đoán mối quan hệ giữa các đặc trưng.

Từ khoá/Keywords: STR, Dự đoán chủng tộc, khoảng cách di truyền.

# Đặt vấn đề

Sự phát triển của nghiên cứu hệ gen trên thế giới đã mở ra những nghiên cứu mới về nhận diện và phân loại cá thể trong hoặc giữa các quần thể thông qua việc ứng dụng các mô hình thống kê, mô hình học máy lên dữ liệu di truyền, trong đó có dữ liệu trình tự lặp lại ngắn -Short Tandem Repeat (STR). So với loại dữ liệu phổ biến như đa hình đơn nucleotide (SNP), dữ liệu STR có tính đa hình tại một vị trí cao hơn nhiều (một locus STR có thể nhận đến 50 giá trị). Điều này giúp tăng khả năng phân biệt giữa các quần tại mỗi locus STR. Cùng với sự phát triển trong công nghệ giải trình tự cũng như đọc biến thể, hiện nay đã có các bộ kit giúp đọc đến hàng chục STR cũng như SNP cho hàng chục mẫu cùng lúc [(Li *et al.*, 2021)](https://paperpile.com/c/pWKXIn/o1lN), các ngân hàng dữ liệu STR cũng được hình thành, và nghiên cứu về chủng tộc cũng được đẩy mạnh. Các mô hình học máy tiên tiến cũng đã đạt được hiệu năng và độ chính xác rất cao trên tập dữ liệu quần thể lớn như người Châu Âu, Châu Mỹ, Châu Phi và Châu Á [(‘XGBoost as a reliable machine learning tool for predicting ancestry using autosomal STR profiles - Proof of method’, 2025)](https://paperpile.com/c/pWKXIn/kfWc) [(Devesse *et al.*, 2023)](https://paperpile.com/c/pWKXIn/77Pa). Tuy nhiên hiện chưa có nghiên cứu nào dành riêng cho quần thể người Việt.

# Đối tượng và phương pháp nghiên cứu

## 2.1 Đối tượng nghiên cứu

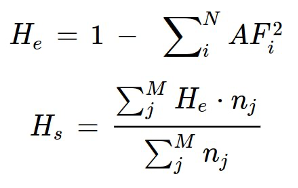
Dữ liệu ADN định dạng STR của các quần thể lớn trên thế giới và dữ liệu STR của người Việt. Từ nguồn chia sẻ công khai của Viên Tiêu chuẩn và Công nghệ quốc gia Hoa Kỳ (NIST), chúng tôi thu được 1036 dữ liệu STR trên các chủng tộc Mỹ gốc Phi, Mỹ Da trắng, Mỹ Latin và Châu Á [(Riman, no date)](https://paperpile.com/c/pWKXIn/ITuQ). Chúng tôi kết hợp với tập dữ liệu gồm 400 mẫu người Việt để cân bằng về tỉ lệ.

Với dữ liệu mô phỏng, chúng tôi dựa trên tần suất alen của tập mẫu thật sự ở trên để xây dựng tập dữ liệu mô phỏng

## 2.2 Phương pháp nghiên cứu

*Xác định khoảng cách di truyền dựa trên hệ khác biệt di truyền Fst (Fixation rate)*

Khoảng cách di truyền giữa các quần thể có thể được đo lường thông qua chỉ số cố định Fst, một công cụ quan trọng trong di truyền học quần thể. Chỉ số Fst phản ánh mức độ phân tách di truyền giữa các quần thể, với giá trị từ 0 đến 1, trong đó 0 biểu thị rằng các quần thể hoàn toàn giống nhau về di truyền, trong khi 1 cho thấy chúng hoàn toàn khác biệt. Fst được tính toán dựa trên tần suất alen của các gen trong quần thể góp phần đánh giá các yếu tố như di cư của quần thể, chọn lọc tự nhiên trong sự hình thành các quần thể. Việc sử dụng chỉ số Fst giúp xác định các mức độ khác nhau về sự khác biệt di truyền, từ đó bổ sung ý nghĩa bên cạnh những lịch sử tiến hóa và các mối quan hê của các quần thể [(Peter, 2016)](https://paperpile.com/c/pWKXIn/G3Zl).



*Sử dụng thuật toán UPGMA để tính toán khoảng cách sự di truyền giữa các dân tộc*

Thuật toán UPGMA xây dựng một cây có gốc (dendrogram) phản ánh cấu trúc có trong ma trận tương đồng từng cặp (hoặc ma trận không tương đồng ) [(Mastana and Singh, 2002)](https://paperpile.com/c/pWKXIn/4sgX). Ở mỗi bước, hai cụm gần nhất được kết hợp thành một cụm cấp cao hơn. Khoảng cách giữa bất kỳ hai cụm nào A và B, mỗi kích thước:



*Phân tích thành phần chính PCA để phân nhóm các quần thể di truyền*

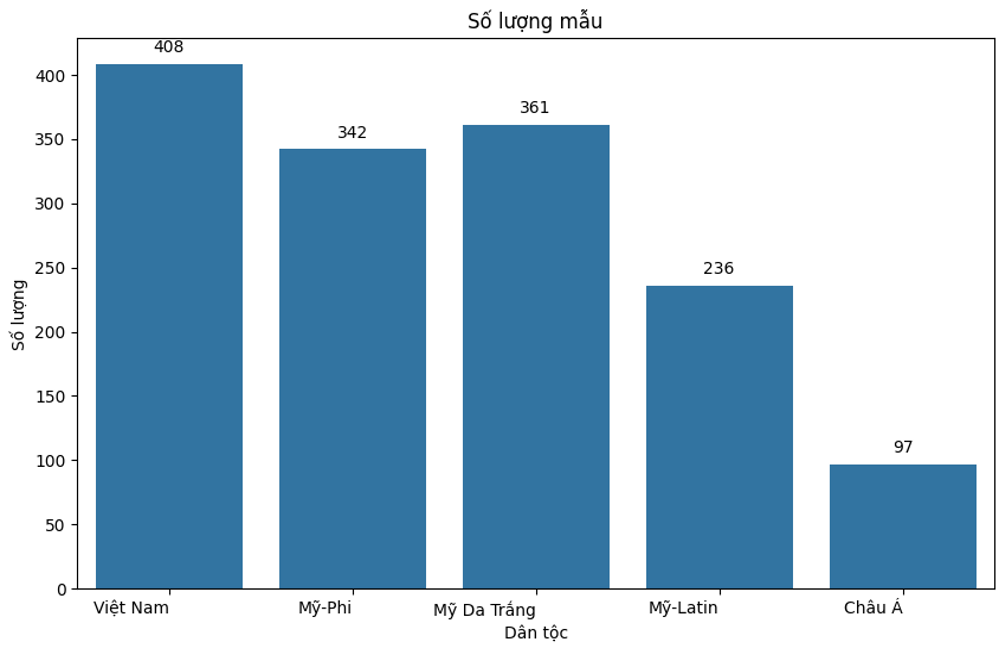
Phân tích thành phần chính (PCA) là một phương pháp hiệu quả trong việc phân nhóm các quần thể di truyền và quan sát mức độ tương quan giữa chúng. Bằng cách giảm số chiều của dữ liệu di truyền, PCA giúp làm nổi bật các yếu tố chính ảnh hưởng đến sự biến đổi di truyền trong các quần thể khác nhau. Khi áp dụng PCA, các cá thể được biểu diễn trong không gian hai hoặc ba chiều, từ đó dễ dàng nhận diện các nhóm cá thể có sự tương đồng cao về di truyền. Khoảng cách/ sự chồng chéo giữa các quần thể được thể hiện qua khoảng cách giữa các điểm trong không gian PCA; nếu các điểm nằm gần nhau, điều này cho thấy chúng có nhiều đặc điểm di truyền chung, trong khi các điểm xa nhau chỉ ra sự khác biệt di truyền rõ rệt. Nhờ vào PCA, ta có thể phần nào nắm bắt được về cấu trúc di truyền và mối quan hệ tiến hóa giữa các quần thể. [(Lu and Xu, 2013)](https://paperpile.com/c/pWKXIn/Jng5)

*Huấn luyện mô hình dự đoán*

Tiến hành chia dữ liệu thành hai phần, với tỷ lệ 0.7 cho việc huấn luyện và 0.3 cho việc kiểm tra hiệu quả, mô hình sẽ được huấn luyện để nhận diện các đặc trưng phân biệt giữa các nhóm dân tộc khác nhau. Để đánh giá hiệu suất của mô hình, chúng tôi dùng các chỉ số như F1-score, Recall và Precision. Chúng tôi sử dụng các mô hình học máy để huấn luyện và phân loại bao gồm: Naive Bayes, Support Vector Machine (SVM), Decision Tree, các mô hình Ensemble như Random Forest, XGBoost, AdaBoost và mô hình mạng nơ-ron Multi-layer Perceptron. [(Pedregosa *et al.*, 2012)](https://paperpile.com/c/pWKXIn/BDZx)

# Kết quả

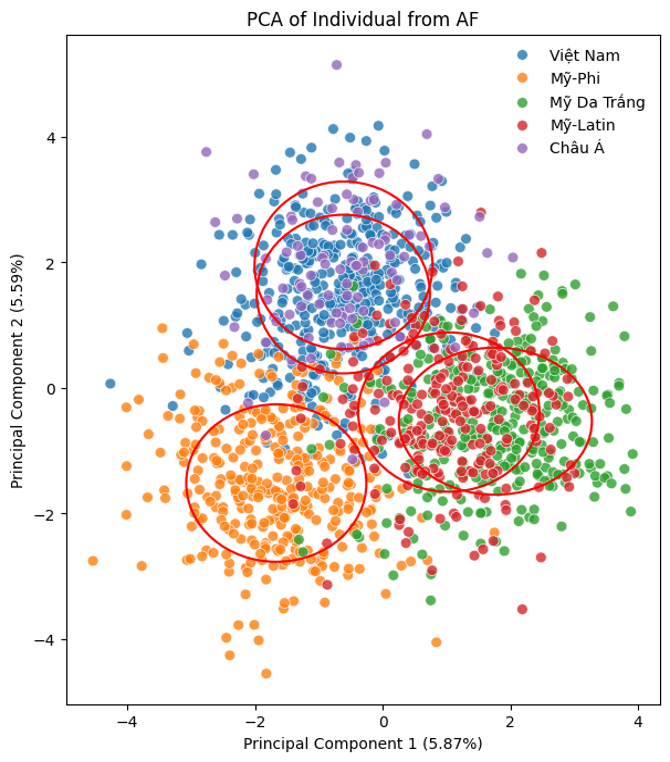
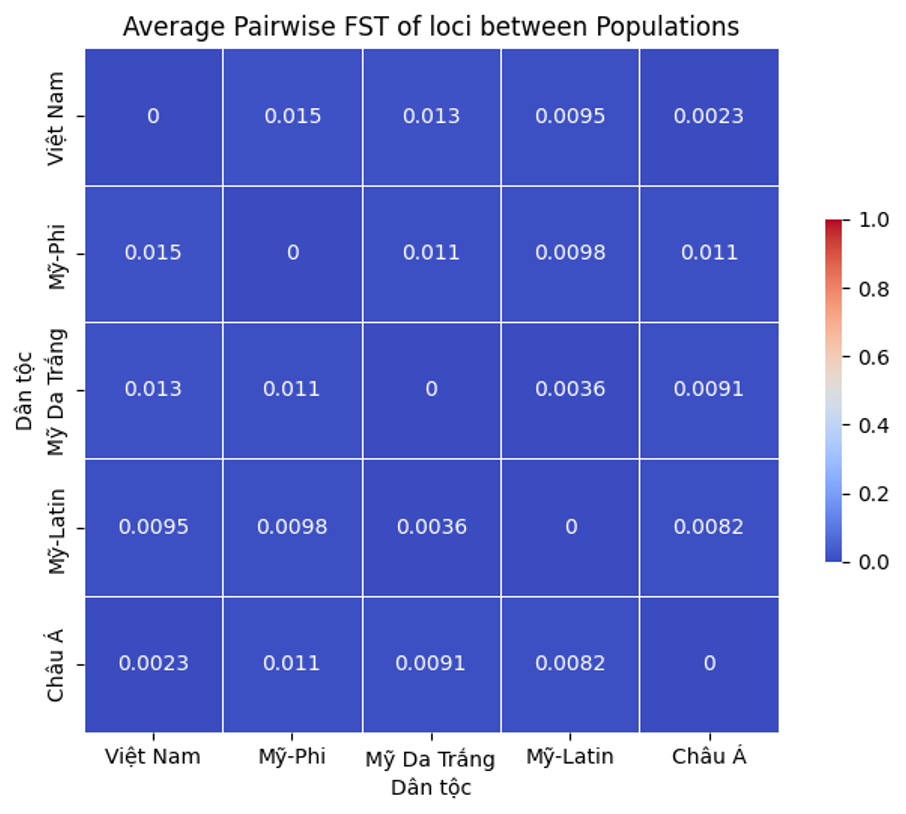
## 3.1 Cơ cấu dữ liệu



*Hình 1. Cơ cấu tập dữ liệu*

Tập dữ liệu bao gồm 1036 mẫu với 30 STR chia thành 4 chủng tộc: Mỹ-Phi, Mỹ Da trắng, Mỹ Latin và Châu Á [(Romani *et al.*, 2014)](https://paperpile.com/c/pWKXIn/qtaQ). Tập dữ liệu người Việt gồm 24 điểm STR. Khi tiến hành kết hợp với 408 mẫu người Việt có 23 STR chung. Tập dữ liệu cuối có 1444 mẫu.

## 3.2 Khoảng cách di truyền và kết quả PCA



*Hình 1. Kết quả tính khoảng cách di truyền và phân tích thành phần chính*

Kết quả phân tích quần thể dùng Fst và UPGMA cho thấy Việt Nam gần với châu Á nhất. Các dân tộc khác xếp theo khoảng cách di truyền là châu Mỹ Latin, Mỹ Da trắng và cuối cùng là châu Phi. Kết quả này đồng thuận với nghiên cứu trước đây trên 148 mẫu với 27 loci STR [(Pham *et al.*, 2021)](https://paperpile.com/c/pWKXIn/yJ8k)

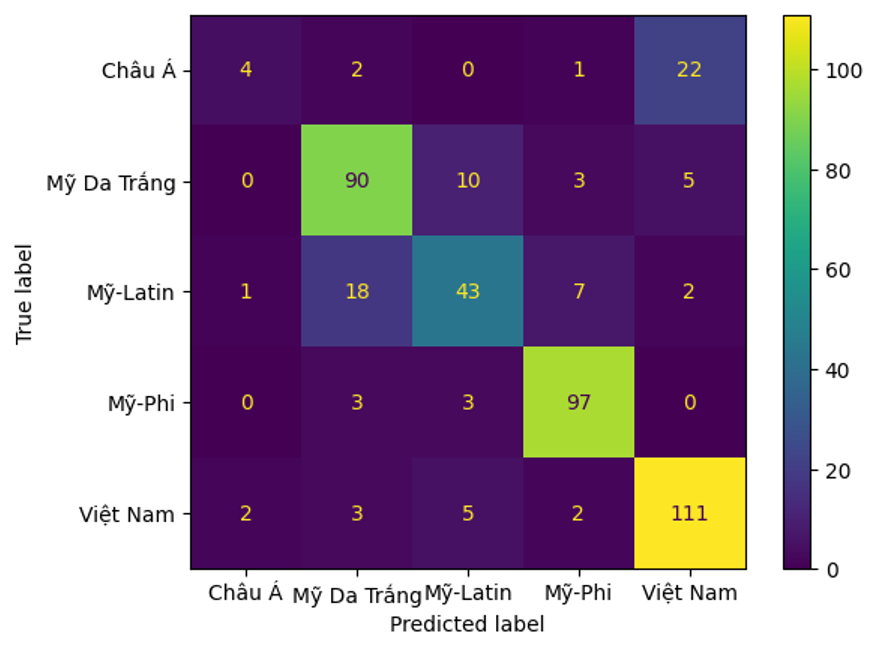
Kết quả phân tích thành phần chính cho thấy các cụm chủng tộc phân cụm rõ ràng, có sự tương đồng giữa nhóm Mỹ Latin với Mỹ da trắng và Việt nam cùng châu Á.

## 3.3 Kết quả huấn luyện mô hình

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Mô hình | Loại mô hình | Tập dữ liệu thật | | | Tập dữ liệu mô phỏng | | |
| Precision | Recall | F1-score | Precision | Recall | F1-score |
| **Bernoulli** | **Naive Bayes** | **0.76** | **0.68** | **0.68** | 0.40 | 0.41 | 0.37 |
| Multi Layer Perceptron | Neural Network | 0.63 | 0.63 | 0.63 | 0.41 | 0.41 | 0.41 |
| Support Vector Machine | - | 0.59 | 0.61 | 0.59 | 0.42 | 0.42 | 0.42 |
| Decision Tree | Tree | 0.46 | 0.45 | 0.46 | 0.46 | 0.45 | 0.46 |
| Random Forest | Ensemble Model | 0.55 | 0.54 | 0.51 | 0.42 | 0.41 | 0.41 |
| XGBoost | Ensemble Model | 0.66 | 0.60 | 0.60 | 0.41 | 0.41 | 0.41 |
| AdaBoost | Ensemble Model | 0.67 | 0.60 | 0.60 | 0.41 | 0.41 | 0.41 |

*Bảng 1. So sánh hiệu năng của các mô hình trên tập dữ liệu thật và dữ liệu mô phỏng*

Mô hình Naive Bayes tốt nhất vì mô hình này phù hợp nhất với dữ liệu đa chiều vì Naive Bayes không yêu cầu phải xây dựng mô hình phức tạp để mô phỏng mối quan hệ giữa các đặc trưng, mà chỉ cần tính toán xác suất và đếm tần suất. Trên dữ liệu mô phỏng kết quả của các mô hình đều không tốt.



*Hình 3. Ma trận nhầm lẫn của mô hình tốt nhất là mô hình Bernoulli*

Trên dữ liệu tập đánh giá, dữ liệu Mỹ Phi dự đoán có tỉ lệ đúng cao nhất với 97 mẫu dự đoán đúng là Mỹ Phi và với chỉ 13 mẫu lệch sang các chủng tộc khác. Dữ liệu Việt Nam có tỉ lệ dự đoán đúng là 111 mẫu với số lượng dự đoán sai gồm 22 mẫu nhầm sang nhóm Châu Á và 7 mẫu rơi vào các nhóm chủng tộc khác. Nhóm chủng tộc Mỹ Trắng và Mỹ Latin có tỉ lệ dự đoán lẫn lộn giữa hai nhóm khá cao có thể được lý giải từ kết quả khoảng cách di truyền hai nhóm này khá gần nhau.

# Bàn luận

Dữ liệu ADN có tính liên kết do đó nên việc mô phỏng chỉ dựa trên tần xuất từ tập gốc không phản ánh được sự đa dạng dữ liệu dẫn đến hiệu quả dự đoán mô hình trên tập mô phòng không tốt. Tuy nhiên, nghiên cứu vẫn còn hạn chế về số lượng mẫu nên cần thiết có các nghiên cứu tương tự với số mẫu lớn hơn. Theo kết quả dự đoán trên tập dữ liệu thật mô hình dựa trên Bayes đạt hiệu năng cao nhất cũng phù hợp với việc tính toán dựa trên xác suất của Bayes.

# Kết luận

Nghiên cứu này cho thấy việc sử dụng dữ liệu STR có tiềm năng trong việc dự đoán chủng tộc của người bất kỳ. Mô hình Bernoulli có hiệu quả phù hợp với dữ liệu đa chiều và sử dụng tính toán xác suất Bayes trong dự đoán mối quan hệ giữa các đặc trưng.

# Tên tác giả, địa chỉ, số điện thoại (liên hệ chính)

Dương Công Trình,

Địa chỉ: C06, Bộ Công an

Số điện thoại: 0818.033.666

# Danh mục tài liệu tham khảo

[Devesse, L. *et al.* (2023) ‘Biogeographical Ancestry Estimation from Autosomal Short Tandem Repeats in the Sequencing Era’, *Forensic Genomics* [Preprint]. Available at: https://doi.org/](http://paperpile.com/b/pWKXIn/77Pa)[10.1089/forensic.2023.0013](http://dx.doi.org/10.1089/forensic.2023.0013)[.](http://paperpile.com/b/pWKXIn/77Pa)

[Li, R. *et al.* (2021) ‘Developmental validation of the MGIEasy Signature Identification Library Prep Kit, an all-in-one multiplex system for forensic applications’, *International journal of legal medicine*, 135(3). Available at: https://doi.org/](http://paperpile.com/b/pWKXIn/o1lN)[10.1007/s00414-021-02507-0](http://dx.doi.org/10.1007/s00414-021-02507-0)[.](http://paperpile.com/b/pWKXIn/o1lN)

[Lu, D. and Xu, S. (2013) ‘Principal component analysis reveals the 1000 Genomes Project does not sufficiently cover the human genetic diversity in Asia’, *Frontiers in Genetics*, 4, p. 127.](http://paperpile.com/b/pWKXIn/Jng5)

[Mastana, S. and Singh, P.P. (2002) ‘Population genetic study of the STR loci (HUMCSF1PO, HUMTPOX, HUMTHO1, HUMLPL, HUMF13A01, HUMF13B, HSFESFPS and HUMVWA) in North Indians’, *Annals of human biology*, 29(6). Available at: https://doi.org/](http://paperpile.com/b/pWKXIn/4sgX)[10.1080/03014460210157402](http://dx.doi.org/10.1080/03014460210157402)[.](http://paperpile.com/b/pWKXIn/4sgX)

[Pedregosa, F. *et al.* (2012) ‘Scikit-learn: Machine Learning in Python’. Available at:](http://paperpile.com/b/pWKXIn/BDZx) <http://arxiv.org/abs/1201.0490> [(Accessed: 10 January 2025).](http://paperpile.com/b/pWKXIn/BDZx)

[Peter, B.M. (2016) ‘Admixture, Population Structure, and F-Statistics’, *Genetics*, 202(4), p. 1485.](http://paperpile.com/b/pWKXIn/G3Zl)

[Pham, N.S. *et al.* (2021) ‘The First Autosomal STR Population Data of Kinh Ethinic Group in Vietnam by Using Massively Parallel Sequencing’, *Russian Journal of Genetics*, 57(8), pp. 985–988.](http://paperpile.com/b/pWKXIn/yJ8k)

[Riman, S. (no date) *NIST Population Data*. Available at:](http://paperpile.com/b/pWKXIn/ITuQ) <https://strbase-archive.nist.gov/NISTpop.htm> [(Accessed: 10 January 2025).](http://paperpile.com/b/pWKXIn/ITuQ)

[Romani, C. *et al.* (2014) ‘Allele frequencies for 26 STR loci in a population of Tuscany (Central Italy)’, *International journal of legal medicine*, 128(3). Available at: https://doi.org/](http://paperpile.com/b/pWKXIn/qtaQ)[10.1007/s00414-013-0942-4](http://dx.doi.org/10.1007/s00414-013-0942-4)[.](http://paperpile.com/b/pWKXIn/qtaQ)

[‘XGBoost as a reliable machine learning tool for predicting ancestry using autosomal STR profiles - Proof of method’ (2025) *Forensic Science International: Genetics*, 76, p. 103183.](http://paperpile.com/b/pWKXIn/kfWc)