creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

의학박사 학위논문

아시아인의 민족 예측을 위한 상염색체 단일 염기 다형성 활용 연구

2019 년 2 월

서울대학교 대학원

의학과 법의학 전공

이 지 현

요약(국문초록)

서론: 범죄현장에서 획득된 생물학적 유래물로부터 대상의 출신 민족을 추정하는 것은 그 사건과 관련된 대상의 범위를 줄여 효율적으로 범죄를 해결하는 데에 유용하게 이용될 수 있다. 법과학 영역에서 개인 식별을 위해 사용하는 STR(Short Tandem Repeat) 마커는 사람의 표현형과는 관련이 적어 얻을 수 있는 정보가 제한적이고, 각 나라들마다 엄격하게 데이터베이스를 관리하고 있어 다른 나라의 자료를 활용하기에 무척 어려움이 따른다. 국제결혼 등이 증가하는 상황에서 Y 염색체와 미토콘드리아 DNA를 이용한 부계 혹은 모계의 민족 예측으로는 유용한 정보를 얻는데 정확성과 실용성 측면에서 아쉬움이 있을 것으로 예상되는 상황이다. 이에 따라 국제적으로 민족 식별과 관련한 상염색체 단일염기다형성 마커에 대한 많은 연구가 진행되고 있어, 이 해당 마커를 이용한 실험 모델을 아시아인을 대상으로 하여 구축하여 보고, 아시아인을 세부적으로 예측 할 수 있는 요인들에 대해 탐색해 보기로 하였다.

방법: 아시아인 750명을 대상으로 상염색체 단일 염기 다형성 분석을 진행하였다. 단일 염기 다형성 분석에는 대규모 병렬 시퀀싱(Massively Parallel Sequencing; MPS) 기기를 사용하는

상용화된 패널을 이용하였다. 이 패널은 민족 간에 유의한 차이가 있다고 알려져 있는 단일 염기 다형성 마커 165개가 포함되어 있는 민족 추정을 위해 개발된 패널이다. 획득된 실험 결과를 바탕으로 통계적인 방법을 적용하여 예측 모델을 개발하였다.

결과: 민족 예측 모델 구축을 위해 아시아인의 민족 혹은 출신지를 동북아시아/동남아시아/서남아시아로 중 한 지역으로 예측하는 데 통계적으로 유의한 9개의 SNP 마커를 선정하였다. 동북/동남아시아 예측모델의 오분류율은 0.277, 동북/서남아시아 예측 모델의 오분류율은 0.015, 동남/서남아시아 예측모델의 오분류율은 0.027이 도출되었다.

결론: 본 연구에서는 타 연구자들의 연구에서 동/남아시아인으로 구분되었던 아시아인을 동북/동남/서남아시아로 세분화하여 세 지역을 두 민족씩 나누어 예측하는 모델을 구축하였다. 이와 같은 연구 결과를 바탕으로 본 연구에 포함되지 않은 민족들의 데이터가 반영된다면 실제 영역에서 보다 활용성이 높은 민족 예측 모델이 구축될 것을 기대한다.

주요어: 법과학, 대규모 병렬 시퀀싱, 민족 예측 모델, 단일 염기 다형성, 아시아인

학 번: 2013 - 23510

목 차

초록.....i

목차..... iii

표 및 그림 목록.....iv

서론.....1

실험재료 및 방법 16

결과.....25

고찰.....56

참고문헌 66

초록 (영문) 78

List of Tables

Table 1. Details of populations analyzed.....33

Table 2. The likelihood ratio of 14 individuals in Northeast Asian based on 165 SNPs.....34

Table 3. The admixture prediction, likelihood ratio results of 42 Vietnamese data, based on 165 SNPs.....35

Table 4. The admixture prediction and likelihood ratio results of unclassified Southwest Asian.....38

Table 5. The results of applying variable selection method (significance level = 0.20, maximum depth = 30) and decision tree (significance level = 0.20,

maximum depth = 30).....41

Table 6. The optimal critical variable obtained by fitting the combination of variables to the prediction model.....45

Table 7. The results of the selected variables.....48

Table 8. Evaluation of predicted models.....49

List of Figures

Figure 1. Read depth per marker based on analyses of 750 samples (mean and standard deviation).....50

Figure 2. Example of HID SNP Genotyper Report from Han Chinese.....51

Figure 3. STRUCTURE analysis results based on 165 SNPs.....53

Figure 4. Classification Rules.....54

서 론

1. 법의학 분야에서 민족 추정

1.1 법의유전학 영역에서의 DNA 활용

법과학 영역 중에서도 법의유전학 영역에서 도입된 유전자 지문(DNA fingerprint)의 개념은 증거물을 통한 과학수사의 해석에 있어서 수치적으로 결과를 해석할 수 있는 시작점이 되었다는 점에서 그 의의가 있다(Butler JM, 2011). 유전자 검사는 통계에 기반하여 이전부터 친자 확인 등에서 많이 활용되어 왔다. 특히 개인을 특정하는 개인 식별 영역에서 DNA 프로파일(DNA profile)을 이용하면서 통계의 이해와 활용이 더욱 중요해 지고 있다.

법의유전학적 접근에서 유전자 검사는 개인식별에 중점을 두고 진행되어 왔지만, 범죄 유형이 급변하는 실정에서 개인 식별의 목적으로 사용되는 기존 유전자 검사로는 충분한 정보를 얻을 수 없다. 이에 최근에는 추가적인 수사정보를 획득하기 위한 방향으로 연구가 진행되고 있다. 사건 현장에 유류된 증거물의 유전자형과 동일한 유전자형을 데이터베이스에서 확인할 수 없는 경우 DNA를 이용하여 나이, 민족(출신지), 주요 외형이나 습성 등을 추정할 수 있는 정보를 얻을 수

있는 경우 자칫 장기화 되거나 미제로 남을 수 있는 사건의 수사 단계에 매우 유용할 수 있다. 이에 유전자 정보를 활용한 새로운 접근이 요구되고 있는 상황이다.

1.2 유전정보를 이용한 민족 추정

민족 혹은 출신지 확인은 목적에 따라 대륙간 구분 혹은 특정 지역의 구분과 배제 등으로 범위를 설정할 수 있다. 이러한 연구들은 주로 민족 간 이동이 큰 유럽을 중심으로 활발하게 진행되어 왔기 때문에 이러한 연구 결과들을 바로 아시아 지역 민족에 적용하기 에는 한계가 있다. 또한, 최근에는 아시아 지역의 사회 경제적인 수준이 높아지고 교류가 활발해 지면서 아시아 내에서도 민족을 구분할 필요성이 증가하고 있는 상황이다.

유전적으로 민족 간 차이가 있음을 뒷받침하는 자료에는 Y 염색체나 미토콘드리아 DNA가 널리 이용되어 왔지만, 이를 활용하는 민족 추정 방법은 부계와 모계를 따로 분석한 결과로 판단해야 하는 어려움이 있었다. 따라서 법과학적인 수사 목적으로 성염색체의 유전자 정보를 사용하여 외국인임을 판단할 때는 추가적인 정보를 수집하여 판단의 근거를 보완할 필요가 있었다. 이에 부계와 모계의 뿌리가 전혀 다른 국제 결혼 등이 증가함에 따라 상염색체를 이용하려는 움직임이 시작되었다. 이러한 움직임과 함께 실험 기술의 발달에 따라 조상 정보

표식자(Ancestry Informative Markers; AIMs)로 사용할 수 있는 다수의 상염색체 단일염기다형성(Single Nucleotide Polymorphism; SNP) 마커가 보고됨에 따라 생물학적 증거물로부터 얻은 유전적 정보를 이용하여 시료를 남긴 사람의 민족 계통을 추정하는 연구 뿐 아니라 이를 이용하여 아프리카, 아시아, 유럽 기원의 구분이 가능하게 되었다(Royal CD et al., 2010).

민족 간 유전적인 차이를 수사에 활용하기 위한 노력은 다양하게 진행되어 왔고, 이를 활용해 특정 돌연변이를 가진 사람이 어디에서 왔는지를 비교적 정확하게 구분해 낼 수 있다. 유전적 차이는 대상 민족들의 기반이 되는 자료를 모아 각 민족 간의 차이를 비교 분석하여 지도상에 각 민족이 어떻게 분포하는지 유전적 거리 차이 혹은 빈도 등으로 표시하게 된다(Pritchard JK et al., 2000).

아시아 인류는 복잡한 식민지 역사와, 다양한 원인으로부터 발생한 이주가 많은 것이 특징이다. 국가를 지리적 영역으로 나누기 위하여 유전자의 흐름을 활용하는 것에는 제한이 있지만 이전의 DNA 정보를 이용한 계통 학적 연구는 북아시아와 남부의 동아시아 인구간에 유전적 차이가 있음을 보여 주었다. 이는 대부분 동아시아로의 남부 및 북부의 이주 경로에 기인한 것으로 나타났다(Underhill PA et al., 2000, Quintana-Murci L et al., 1999, Zhong H et al., 2011, Zhang X et al., 2014)

1.3 최근 민족 추정 연구 동향

조상 정보 표식자(AIMs)의 개발은 민족 추정연구에서 큰 가치를 지니고 있다. 수백 개의 마커를 사용하는 연구를 통해 대륙 별 분포와 밀접하게 일치하는 대립 유전자의 빈도 차이를 기반으로 전 세계 인구 집단을 그룹화 할 수 있다고 제안하기 때문이다(Allocco DJ et al., 2007, Enoch MA et al., 2006, Li JZ et al., 2008, Rosenberg NA et al., 2012). 이와 관련해 수많은 SNP 마커에 대한 동아시아인과 동남아시아인의 대립 유전자 빈도에 대한 정보가 발표되었고, 아시아 지역 중에서는 북부 한족과 남부 한족의 대립 유전자 빈도에 대한 연구가 활발하게 진행되었다(Suo C et al., 2012, Chen J et al., 2009, Xu S et al., 2009, Xu S, 2012).

법의학 실험실 내에서 대규모 병렬 시퀀싱(Massively Parallel Sequencing; MPS) 혹은 차세대 염기서열 분석법(Next Generation Sequencing; NGS) 기술의 도입은 수천 개의 SNP를 동시에 분석하는 해결책이 되었다(Børsting C et al., 2015). 이와 관련하여 법과학 시료에 맞게 적은 양의 DNA로 민족 정보 표식자(AIMs)를 동시에 수십 개 이상 분석 가능하도록 제작된 패널을 구성하는 연구가 진행되고 있다. EUROFORGEN-NoE 컨소시엄이 대표적으로 31개의 SNP 마커를 한번에 증폭할 수 있는 패널을 제작하였지만(De la Puente M et al., 2016) 유럽인에 초점을 맞추고 있으며 대륙의 수준에서 민족을

구분한 연구로 세부적인 민족을 추정하기에는 어려움이 있다. 이 외에도 중국에서 27개의 마커로 이루어진 패널이 개발되었고(Wei YL et al., 2016), Pacifiplex라는 이름으로 아프리카/유럽/동아시아/오세아니아를 구분하는 29개 SNP multiplex panel이 제작(Santos C et al., 2016)되어 발표한 경우도 있다.

대륙의 구분을 세분화해서 시료의 민족 혹은 출신지를 Sub-Saharan 아프리카, 북아프리카, 유럽, 서남아시아, 남아시아, 북아시아, 동아시아, 동남아시아, 태평양지역, 미국으로 추정할 수 있는 74개의 SNP 패널도 발표(Li CX et al., 2016)되었다.

1.4 민족 추정을 위한 상업용 패널

민족을 예측하는데 필요한 민족 정보 표식자(AIMs) 중에서 165개의 SNP 마커를 포함하는 대규모 병렬 시퀀싱 용 패널이 개발되었다. Applied Biosystems™ Precision ID Ancestry Panel(Thermo Fisher Scientific, Waltham, USA)라는 이름으로 개발된 이 패널은 1 ng의 DNA에 최적화되어 있으며, 이 패널에는 민족 추정 연구를 대규모로 진행하고 있는 두 연구소에서 개발한 패널에 포함된 165개의 SNP 마커들이 포함되어 있다(Kidd KK et al., 2012, Kosoy R et al., 2009). 패널 제조사가 제공하는 분석 프로그램에는 65개의 민족 데이터(Kidd KK et al., 2012)가 참조 데이터로 포함되어있고, 이러한 분석

프로그램이 출신을 알 수 없는 개체를 적절한 인구 그룹에 할당하고 우도 비율과 같은 다양한 통계 값을 계산한다. 다시 말해, 기기와 연결된 HID SNP Genotyper Plugin(Thermo Fisher)이라는 소프트웨어에서 분석한 시료가 속할 가능성이 높은 지리적 위치가 지도에 표시되고 각 지역 군에 해당될 확률이 퍼센트(%)값으로 계산되어 출력된다. 또한, 소프트웨어에 포함되어 있는 참조 민족들의 유전자 자료들과의 무작위 매치 확률을 계산하여 관찰된 유전자형과 유사할 확률이 높은 민족 순으로 순위를 매겨준다. 이것은 참조 집단들 사이에서 관찰된 프로파일의 조상의 랭킹 우도 목록을 생성한다. 다만, 본 Plugin 소프트웨어에서 예측되는 민족은 유럽인, 오세아니아인, 동아시아인, 아프리카인, 남아시아인, 서남아시아 인으로 아시아인에 대한 세부적인 예측은 제공하지 않는다는 단점이 있고 참조 집단에 포함되지 않은 개체군의 조상을 정확하게 식별할 수 있는지 여부는 알려져 있지 않다(Themudo GE et al., 2016). 특히 동북아시아와 동남아시아를 동아시아로 합하여 결과를 제공해주는 단점이 있어, 아시아인에 직접적으로 적용하기에는 어려움이 있다.

2. 예측 모델 구축을 위한 통계기법

2.1 분류분석

민족 예측에서의 분류분석이란 본인의 민족이나 출신지 정보를 알고 있는 개인의 유전 데이터를 민족 별로 모아 민족 별로 집단을 만들어 분류하는 모형을 학습 시킨 후, 민족 정보가 없는 다른 개인의 유전정보를 학습된 집단 중 한 집단으로 분류하는 분석을 말한다. 분류분석에는 다양한 종류의 모형이 존재하고 각 모형 별로 각기 다른 분류함수를 사용하게 된다. 의사결정나무(decision tree) 모형, 로지스틱회귀(logistic regression) 모형 그리고 신경망(Neural Network)모형 등이 대표적으로 알려져 있다.

2.1.1 의사결정나무(Decision Tree) 모형

의사결정나무 모형은 가장 대표적인 분류모형으로 규칙에 따라 나무구조로 표현된다. 목표변수가 이산형일때 의사결정나무, 연속형일때 회귀나무로 불린다. 이 모형의 장점은 분류 기준이 되는 규칙이 명확하게 표현되어 결과를 쉽게 해석할 수 있다는 점이다(Quinlan JR, 1987).

의사결정나무는 크게 결절(노드, node), 잎(leaves),

가지(branch)로 구성되어 있으며 결절은 대상 변수에 대한 분류 기준을 만드는 변수에 해당한다. 잎은 최종 분류 된 그룹을 나타내고, 이러한 구성요소는 가지가 갈라지는 기준과 가지 뻗기가 멈추는 기준대로 나무구조가 된다. 추론 규칙(induction rule)과 오류를 적절하게 판단하여 가지치기(pruning)를 통해 모형을 형성한다(Kamiński B et al., 2018). 분할 기준에는 카이 제곱 통계를 사용하여 그룹의 분포가 독립적인지 여부를 테스트하여 나무구조를 형성하는 방법과 Gini 지수 또는 Entropy 지수와 같은 불순도(impurity) 지수를 사용하여 불순도 값을 감소시키는 방향으로 나무를 형성하는 방법이 있다.

2.1.2 랜덤포레스트(Random Forest) 모형

랜덤포레스트(Random Forest) 모형은 Breiman(Breiman L, 2001)에 의해 개발된 분류 방법으로 전통적인 의사결정나무 기법을 단일모형이 아닌 여러 개의 나무모형들로 확정시킨 기계학습 기법이다. 랜덤포레스트 모형을 구성하는 각각의 의사결정나무는 무작위로 선택된 학습 데이터와 입력변수들에 의해 형성되는데, 이 경우 각 개별 의사결정나무의 정밀도는 떨어질 수 있으나, 이들을 종합하여 예측을 수행하게 되는 전체 모형의 정확도와 안정성은 높아지게 된다.

랜덤포레스트 모형은 대수의 법칙에 의해 숲 또는 나무의 크기가 커질수록 일반화 오류의 특정 값으로 수렴하게 되어 과적합을 방지할 수

있으며, 각 단일 의사결정나무들을 학습시킬 때 전체 학습용 자료에서 무작위로 복원 추출된 데이터를 사용하고 있어 이상치(outlier)로부터 크게 영향을 받지 않는다(Barandiaran I, 1998). 무엇보다 가장 큰 장점은 분석자가 입력변수를 선정하는데 있어 자유로울 수 있다는 점이다. 발표된 한 연구(Brown I, 2012)에 따르면 계급불균형(imbalanced) 이항분류의 예측에 있어 가장 우수한 예측력을 보인 것으로 보고되고 있다.

2.1.3 그래디언트 부스팅(Gradient Boosting) 모형

그래디언트 부스팅은 Friedman(Friedman JH, 2001)에 의해 제안된 방법으로 단일 분류기를 사용하는 의사결정나무에 비해 여러 분류기들의 예측을 종합하여 분류의 정확도를 향상시키는 앙상블기법의 하나이다. 분석용 데이터 관측값의 가중치가 같은 상태에서 시작되어 형성된 분류기에 의해 오분류된 관측값은 다음 관측값에 높은 가중치를 주게 되지만, 정분류된 관측값은 낮은 가중치를 주는 과정을 반복함으로써 최종 분류기를 형성한다. 즉, 순차적으로 기본 분류기가 분류하기 어려운 자료들에 집중하도록 자료들의 분포를 변경하는 절차를 반복적으로 수행하며, 자료로부터 수정된 새로운 자료를 생성하게 되며 기본 모형에 가중치를 결합한 형태를 보인다. Gradient descent 알고리즘을 이용하여 기대손실을 최소화 시키는 방향으로 이루어지며,

Gradient descent 알고리즘의 아이디어는 Newton-Raphson 방법과 연관이 깊다(Friedman JH, 2002). 이 알고리즘은 반복이 계속됨에 따라 추가적인 함수는 항상 부스팅 기법이 적용되어 목표변수 값들에 대한 중요도 값이 차례로 부여되기 때문에 이전단계의 반복까지의 영향을 받는다. 이러한 순차적 반복 단계에 따라 목표 변수들의 값은 손실을 줄일 수 있게 된다. 이와 같은 과정을 통하여 분류 정확도가 높은 최적의 모형을 찾을 수 있다.

2.1.4 신경망(Neural Network) 모형

신경망(Neural Network) 모형은 비선형함수를 이용하여 집단을 분류하는 분류모형으로 인간의 뇌를 모방하여 만들어진 모형이다(Rosenblatt F, 1958). 인간의 뇌는 신경세포인 뉴런(neuron)들이 서로 연결되어 있다. 이를 신경망이라고 부르게 되는데, 이러한 신경세포들은 인간이 의사결정을 내릴 때 하나의 뉴런이 자극을 받으면 다른 뉴런들에게 그 자극을 전달되게 된다. 이러한 자극을 받으면 각 뉴런에서 자극에 대한 정보를 생성하게 되는데 이러한 정보들을 종합하여 인간이 어떻게 반응할지에 대한 최종적인 의사결정을 판단 내리게 된다.

신경망 모형은 이러한 뉴런들의 의사결정 방식을 모방하여 인간의 뇌와 유사하게 여러 가지의 결절(노드, node)들을 서로 연결하는

네트워크를 만들어 최종적인 분류를 하게 된다(Freund Y et al., 1999).

2.1.5 앙상블(Ensemble) 모형

앙상블(Ensemble) 모형은 동일 종류 또는 다른 종류의 여러 분류 모형의 분류 결과를 결합하여 최종 분류를 유도하는 모델이다(Opitz D et al., 1999). 이론적으로, 앙상블 모델을 낮은 오분류율을 갖는 것으로 알려져 있다.

3. 연구의 필요성

현대 과학기술에 발전에 따른 교통과 정보통신 기술의 발전, 국가 간 무역의 증가와 활발한 자본 이동 등의 결과로 각 나라 사이에 교류가 증가하고, 이러한 움직임은 아시아 내부에서 더욱 활발한 경향을 띠고 있다. 이와 함께 법과학 영역에서는 기존의 STR 프로파일을 이용한 개인식별의 차원을 넘어 유전자 정보를 활용해 새로운 수사정보를 추가적으로 확보하려는 적극적인 움직임이 있었고, 실제로 사건 해결에 도움을 주고자 하는 연구들이 진행되고 있다(Lee HY et al., 2016). 유전자 정보를 이용해 생김새 등 외형을 추정하는 연구와 함께 민족추정 연구 또한 이러한 흐름의 하나를 형성하고 있다(Conomos MP et al., 2015, Pritchard JK et al., 2000, Enoch MA et al., 2006). 이 외에도, 한국에 체류하는 외국인 수는 지속적으로 증가하고 있는 추세이며, 특히 장기간 국내에 체류하는 외국인이 증가하고 있는 상황에서 외국인 범죄의 절대 건수도 상승세에 있다.

각 나라마다 유전적 차이가 있음은 널리 알려져 있고, 관련 연구 또한 매우 많아 실제로 유전자 정보를 이용하여 시료의 주인이 어느 대륙에서 왔는지 구분하는 것은 그리 어렵지 않다. 하지만 이와 같은 연구들의 대부분이 유럽 등 서구를 중심으로 진행되었기 때문에, 아시아인을 대상으로 한 대규모 연구는 그리 쉽게 접할 수 있지 않다.

유전자 정보를 통해 추정된 민족 정보는 다양한 상황에서 활용될 가능성이 높으나 아직 이에 대한 연구 자료가 충분하지 않은 상황이므로, 활용을 위한 다양한 연구가 진행되어야 한다.

또한, DNA의 손상이 있어 충분한 양의 DNA를 얻지 못하는 법과학 시료의 특성으로 미루어 볼 때, 다른 생명과학 분야의 실험과 달리 원하는 결과를 충분히 얻지 못할 가능성이 있다. 따라서 대량의 마커를 한번에 분석해내는 타 분야의 실험과 달리 적은 수의 마커를 적은 양의 DNA를 이용해 분석하는 연구가 법과학 시료에는 더 적합할 수 있다.

단순히 민족 간 차이를 관찰하는 수준을 넘어 수사 실무에서 유용하게 활용되기 위해서는 정확한 예측모델을 확립하는 것이 중요하다. 유전자 감식이 진행되고 30년 이상 경과한 지금, 여러 민족을 대상으로 다양한 자료들이 쌓이고 있는 시점에서는 다양한 통계적인 접근을 활용해 새로운 시작으로 자료를 분석하여 볼 필요가 있다. 법과학 분야에서 실제적으로 예측 모델을 사용할 수 있으려면 소수 집단간에 얼마나 많은 유전적 차이가 존재하는지, 얼마나 많은 다른 민족 집단이 구별될 수 있는지, 그리고 오분류 비율과 같은 통계적 가치를 결정하는 것이 필요하다. 일부 상업 회사는 신청자의 다양한 조상 기원을 결정하는 이른바 대륙간 파티셔닝(intercontinental partitioning)을 기반으로 한 서비스를 제공하기도 한다(Eriksson N et al., 2010, Bryc K et al., 2015). 그러나 분류하고자 하는 민족의 범위가

대륙보다 좁은 경우에는 이러한 접근법을 적용하기 어려워진다.

이와 같은 흐름에 맞추어 아시아인을 대상으로 민족 간 차이가 크다고 이미 알려져 있는 SNP 마커를 분석하여 데이터를 수집하고, 실제적인 활용을 위하여 유의한 마커를 이용해 예측 결과를 도출해 내고자 하였다. 이때 다른 마커보다 영향을 많이 미치는 특정한 마커들이 있을 것으로 가정하였고, 이러한 특정 마커들을 선별해 내는 데에 통계적 방법을 이용한 변수 선택법을 이용하기로 하였다.

이것은 유전자 정보를 이용한 민족 예측 방법이 실제적인 활용으로 나아가기 위해서 필요한 연구이며, 유전자 정보를 이용하여 아시아를 세부적으로 구분하고 예측하는 모델의 기반을 만드는 것은 매우 시의 적절하다고 할 수 있다.

연구 목적

본 연구에서는 아시아인을 대상으로 단일염기다형성(SNP) 마커 정보를 이용하여 아시아인을 세부적으로 예측하는 모델을 개발함으로써 새로운 통계적 접근법의 유용성과 한계를 이해하고자 한다. 세부 목표는 다음과 같다.

1. 아시아 7개 민족의 SNP 유전자 정보 분석 및 통계적 결과 도출
2. 특정인의 민족을 동북아시아/동남아시아/서남아시아 중 한 지역으로 예측하기 위해 선별한 마커를 이용한 예측 모델 검증
3. 아시아인을 세부적으로 구분하기 위한 배경 지식과 보다 정확한 민족 예측 연구의 기초를 제공

실험 재료 및 방법

1. 시료

1.1 상업 시료의 수집

중국 남부의 한족(CHS) 99명, 중국 베이징 지역의 한족(CHB) 100명, 일본 도쿄 지역의 일본인(JPT) 101명, 베트남 호치민시의 Kinh 민족(KHV) 100명의 DNA 시료를 Coriell 연구소(Coriell Institute for Medical Research in Camden, New Jersey)에서 구매하였다(Table 1). 시료는 DNA 농축액의 형태로 분양되었다. 시료의 구매는 IRB에서 승인 이후 이루어졌다(C-1208-057-421).

1.2 이 외 아시아인 시료의 수집

법의학교실에서 운영 중인 시료은행에서 보관중인 아시아인 시료 중 2차 사용에 동의한 시료에 한해 DNA를 일부 분양 받았다. 본 시료 은행에서 보관 중인 시료는 모두 실험에 대해 적절한 설명을 시행하고 동의를 받은 후 채취하였으며, 이는 IRB에서 승인을 받았다(C-1501-078-641). 사용한 시료는 한국인(KOR) 100명 네팔인(NP) 100명, 인도 Deccan 지역에 거주하는 인도인(IN) 51명,

파키스탄의 Pathan 지역인(PT) 49명, 파키스탄의 Punjab 지역인(PJ) 50명이다(Table 1). 보관된 시료은행의 시료 사용에 대해 IRB에서 승인을 받았다(C-1404-068-572).

1.3 DNA 정량

수집한 DNA는 Quantifiler® Human DNA Quantification Kit (Thermo Fisher)를 이용하여 QuantStudio® 5 Real-Time PCR System(Thermo Fisher) 기기로 정량하였다.

2. 라이브러리 제작 및 시퀀싱

2.1 DNA 증폭

Applied Biosystems™ Precision ID Ancestry Panel과 Ion Precision ID Library Kit(Thermo Fisher)를 사용하여 DNA의 표적부분 증폭을 진행했다. 1 ng DNA에 5X Ion AmpliSeq™ HiFi Mix 4 ㎕, Precision ID Ancestry Panel 10 ㎕를 포함하는 총 20 ㎕의 반응액으로 PCR을 수행하였다. GeneAmp® PCR System 9700(Thermo Fisher)기기를 이용하여 반응 조건은 99℃에서 2분 간 효소를 활성화 시키는 과정을 거친 후, 99℃에서 15초, 60℃에서 4분을 한 주기로 21회 반복한 뒤 10℃에서 보관하였다. 이 후,

반응하지 않은 시발체 부분의 제거를 위해 2 ㎕ FuPa 시약을 직접 첨가해 50℃에서 10분, 55℃에서 10분, 60℃에서 20분을 배양하였다.

2.2 어댑터 결합 및 정제

앰플리콘에 대한 바코드 어댑터의 결합 및 증폭되지 않은 라이브러리의 정제를 위해 가이드라인에 따라 4 ㎕의 Switch Solution, 2 ㎕의 Precision ID IonCode™ Barcode Adapters, 2 ㎕의 DNA Ligase를 앰플리콘에 연속적으로 첨가해 22 ℃에서 30 분간, 72 ℃에서 10 분간, 10 ℃에서 1 시간 동안 배양하였다. 증폭되지 않은 라이브러리 제거를 위해 사용자 가이드에 따라 1.5x Agencourt AMPure® XP 시약(Beckman Coulter, FL, USA)을 사용하여 정제하였다.

2.3 정량

모든 과정을 거친 앰플리콘 2 ㎕에 순순한 물 198 ㎕을 더해 1:100 희석액을 준비한 후, Ion Library TaqMan™ Quantitation Kit(Thermo Fisher)를 이용하여 QuantStudio® 5 Real-Time PCR System 에서 정량하였다. 반응 조건은 50℃에서 2분, 95℃에서 20초 과정을 거친 후, 95℃에서 3초, 60℃에서 30초을 한 주기로 40회 반복하였다. 모든 시료에 대해 PCR 반응은 2회씩 반복 수행하였으며,

결과는 같은 제조사의 HID Real-Time PCR Analysis Software v1.1을 이용하여 분석하였다.

2.2 시퀀싱

모든 시료는 30pM로 희석되어 200명씩 하나의 칩에 로딩되도록 준비하였다. 사용된 칩은 Ion 540™ Chip Kit(Thermo Fisher)로 형광을 사용하지 않고 중합효소(polymerase) 기반의 결합을 전기적으로 감지한다. 200염기쌍(base pair)의 데이터를 10~15Gb 생산할 수 있다. 칩에 시료를 심기 위한 자동화 장비인 Ion Chef(Thermo Fisher) 기기를 이용하였고, 칩에 샘플이 심어지면 Ion S5xl System(Thermo Fisher)을 이용하여 시퀀싱을 수행하였다.

3. 데이터 분석

3.1 시퀀싱 데이터 분석

시퀀싱 데이터 중 signal processing, base-calling, barcode de-convolution 과정은 시퀀싱 기기인 Ion S5xl System에서 자동으로 연계되는 HID SNP Genotyper v4.3.2 Plugin(Thermo Fisher)으로 분석되었다. 이 Plugin은 우도비 계산을 위해 동아시아, 아시아,

오세아니아, 시베리아, 북미, 유럽, 아프리카 및 남미로 나뉘어진 총 65개 민족을 지역 표준으로 분류한 참조 데이터를 포함한다. 동아시아에는 10개의 민족(Japanese-HapMap, Japanese, Koreans, Han-HapMap, Cambodians Khmer, Ami, Hakka, Taiwanese Han, Malaysians, Atayale)이 포함되어 있고, 아시아에는 14개의 민족(Lao Loum, Hazara, Khanty, Kachari, Keralite, Pashtun, Mohanna, Thoti, Negroid Makrani, Komi-Zyrian, Kuwaiti, Palestinian, Druze, Jews Yemenite)이 포함되어 있다. 이 데이터를 사용하여 우도 값을 계산하고 높은 값에서 낮은 값 순으로 나열한다.

또한, Plugin은 유럽, 오세아니아, 동아시아, 아프리카, 남아시아, 미국 및 서남아시아 중에서 시료가 각 민족에 속할 확률을 산출한다.

각 시료 별 165개의 SNP 마커에 대한 유전자형 정보는 Microsoft Excel 프로그램을 사용하여 정리하였다.

3.2 인구 구조 분석

이형 접합체 균형은 하나의 뉴클레오타이드에 대한 판독 횟수를 다른 뉴클레오타이드에 대한 판독 횟수로 나눈 값으로 계산되었다. 연관 비평형(Linkage Disequilibrium; LD), 대립 유전자 빈도, 관찰된 이형 접합체 및 예상된 이형 접합체는 HAPLOVIEW 소프트웨어를 사용하여 산정(Barrett JC et al., 2005)하였으며, Sigmaplot

12.0(Systat Software, Erkrath, Germany)프로그램을 이용하여 수치를 나타내었다.

관련 대립 유전자 빈도를 이용해 인구 구조를 분석하였다. STRUCTURE v. 2.3.4 프로그램(Pritchard JK et al., 2000, Falush D et al., 2003)이 사용되었고, 추정 민족 클러스터(K)를 1에서 9까지 변경하면서 각 클러스터(K) 값에 대한 독립 실행을 수행하였다. 각 분석은 사후 확률의 정확한 추정을 달성하기 위해 40,000개의 burnin 단계와 40,000개의 Markov chain Monte Carlo(MCMC) 단계로 구성되었다. 또한 K의 최적 값을 선택해주는 프로그램인 Structure Harvester v0.6.94.(Earl DA et al., 2011)를 사용하였다.

4. 민족 예측 모델 구축

4.1 데이터 전처리

본 자료의 예측 대상에는 세 종류의 민족(동북아시아, 동남아시아, 서남아시아)이 존재한다. 효과적인 민족 예측을 위해서 두 민족만을 포함하는 자료들로 재 생성하였다. 즉, 1)동북-동남아시아 자료, 2)동북-서남아시아 자료 그리고 3)동남-서남아시아 자료로 생성하였다.

생성한 세 가지의 자료는 두 집단의 데이터 수가 다른 불균형

데이터(Imbalance data)이다. 이를 계급 불균형 자료라고 말하는데, 이를 균형 자료로 만들어 주기 위해 언더샘플링(Under-sampling) 기법을 사용하였다. 언더샘플링이란 비복원 추출 방법을 이용하여 다수인 집단의 비율을 낮추어주는 방법으로, 원 데이터만을 가지고 분석한다는 장점과 원 데이터의 정보가 손실된다는 단점이 존재한다. 본 연구에서 다루는 데이터는 유전자 데이터로 원 데이터만 사용하는 것이 적절하다고 판단하여 언더샘플링을 통해 계급불균형의 자료를 균형으로 맞추어 분석을 진행하였다. 각 자료마다 난수 초기값(seed)을 다르게 하여 언더샘플링을 5번 적용하였다. 이후, 분류 모형 구축을 위해 데이터 분할을 수행하였다. 훈련용, 검증용 데이터를 70대 30으로 분할하였다.

4.2 변수 선택 방법

변수 선택은 총 세가지 방법을 사용하였다. 첫 번째 방법은 명목형 변수 군집화이다. 변수들을 군집화 시키는 방법을 $n \geq k$ 라고 한다면 n개의 변수들을 k개의 변수들로 차원을 축소시키는 방법이다. 일반적으로 자료가 연속형 변수일 때 사용이 가능한 기법이지만 자료가 명목형인 경우 제한적인 방법을 통해 변수 군집화를 이루기도 한다. 본 분석에서는 명목형 자료에 대해 변수 군집화를 진행하였으나 군집분석의 결과가 해석하기 어려운 점과 묶인 군집변수들에 대한

유사도 거리가 상대적으로 낮게 나왔다고 판단이 되어 이 방법을 사용할 수 없다고 판단하였다.

두 번째 방법은 카이제곱 통계량을 이용한 변수선택법이다. 명목형(범주형) 자료 분석에서 주로 사용되는 카이제곱 통계량을 이용하여 목표변수에 대한 입력변수들의 동질성 및 독립성을 검정하여 변수를 선택하는 방법이다. 보통 이러한 방법을 변수 선택(Variable Selection)이란 용어로 부른다. 변수선택법에서는 매개변수(parameter)값으로 최대 깊이를 총 5종류(6, 12, 18, 24, 30)fh 변동시켜 진행하였으며, 카이제곱 값의 하한에 대한 유의수준을 총 4종류(5%, 10%, 15%, 20%)로 변동시켜가며 분석을 진행하였다. 따라서 각 언더샘플링 데이터 셋마다 총 20종류의 변수 선택법이 적용되었다.

세 번째 방법은 의사결정나무 분석을 통한 변수선택법이다. 이 방법은 SNP 데이터를 가지고 아시아 민족을 분류하는 의사결정나무 모형을 구축한 이후 의사결정나무 분석 결과에서 도출되는 변수 중요도를 이용하여 변수를 선택하는 방법이다. 의사결정나무 모형의 매개변수로 카이제곱 통계량을 사용하였으며, 유의수준은 0.2로 통일한 상태로 최대 깊이를 총 5종류(6, 12, 18, 24, 30)로 변동시켜 모형을 구축하였다. 따라서 각 언더샘플링 데이터 셋마다 총 5종류의 의사결정나무 모형이 적용되었다.

4.3 분석 다이어그램 생성

본 분석은 SAS® Enterpirse Miner™ 13.1을 사용하였다. 세 종류의 데이터 자료를 난수 초기 값을 다르게 하여 5종류의 언더샘플링 데이터 세트로 만든 이후, 각 데이터 세트에 대해 20가지의 변수선택법(Variable selection)과 카이제곱 통계값의 유의수준을 0.2로 동일하게 분석한 5가지의 의사결정나무 분석(Decision Tree)을 진행한 다이어그램을 생성하였다. 즉, 한 데이터 자료당 총 125가지의 분석 결과가 도출 되었으며, 전체 자료로 확장하면 총 375가지의 결과가 도출되었다.

결 과

1. 패널 성능 확인

Applied Biosystems™ Precision ID Ancestry Panel을 통해 총 750 명의 아시아인의 SNP 마커가 타이핑되었다. 1명당 165개의 마커가 타이핑되는데, 본 연구 결과에서는 123,750개의 유전자형 중 123,000개(99.4%)개의 유전자형이 입력 기준을 넘어 유전자형으로 나타났다. 6개 유전자형(0.005 %)은 시퀀싱 기기에서 읽히지 않았고(no coverage), 318개(0.26 %) 유전자형은 유전자 위치에 대한 커버리지 범위가 평균 범위의 2 표준 편차 값보다 작음을 의미하는 'NOC'로 불렸다. NOC가 있는 마커 중 rs12498138과 rs2196051은 750명 중 58명과 42명의 유전체형을 읽지 못한 것으로 나타났다. 165개의 SNP 마커 중 최소 커버리지는 193x(rs1407434)이고, 최대 커버리지는 5369x(rs2073821)로 평균 1,918x(Fig. 1)를 나타내었다.

총 3,973개(3.21%)의 유전자형은 DNA 사(strand)의 불균형으로 인한 평균 커버리지의 범위가 2,276이상인 커버리지의 양성 범위(percent of positive coverage; PPC) 현상이 나타났다.

750명 중 127명이 적어도 하나의 마커에서 유전자형이 나타나지 않았고, 최대 19개의 마커에서 유전자 형의 드롭아웃(drop-out)이 나타났다. 일부 상태가 좋은 샘플에서도 노콜(no-call) 유전자형이 관찰되었다. 이는 유사한 연구에서 사용 된 상위 90 %의 완전한 유전자형만을 나타내는 높은 엄격성을 가진 임계값(Nassir R et al., 2009)을 적용한 결과이다.

이형 접합체 균형(Heterozygote balance; Hb)값의 평균은 0.33에서 3.68의 범위를 나타냈다. rs1871534 마커는 모든 시료에서 GG 유전체형만을 나타내는 단일형(monomorphic)을 보였고, rs3827760 마커도 베트남의 한 명을 제외하면 GG 유전체를 보였다. 또한 165개 마커 중에서는 본 연구에 포함된 7개 민족에서 연관 비평형(LD)을 보이는 마커가 발견되지 않았다.

2. 아시아인의 유전자형 분포

패널 제조사에서 제공하는 HID SNP Genotyper v4.3.2 Plugin(Thermo Fisher)을 이용하여 750명의 아시아인 유전체 정보를 분석하였다. 이 소프트웨어에 포함되어 있는 참조 민족들의 유전자 자료들과의 무작위 매치 확률을 계산하여 관찰된 유전자형과 유사할 확률을 계산하고 우도비를 계산하여 목록을 생성한다. 결과를 보면,

우도비가 높은 지역을 지도에 나타내주고, 유럽, 오세아니아, 동아시아, 아프리카, 남아시아, 아메리카, 서남아시아 중 어디에 속하는지를 확률로 나타내 준다. 또, 참조 민족과의 우도비를 계산하여 값이 높은 순서대로 순위를 나타내는 자료를 같이 제공해준다(Fig. 2).

한국, 중국, 일본 출신의 동북아시아인 400명의 유전체 정보를 분석한 결과, 400명 중 14명을 제외하고는 동아시아 지역 출신일 확률이 제일 높게 예측되었다. 14명 중 13명은 모두 아시아지역으로 예측되었는데, 그 중에서 Lao Loum 지역인과 가장 높은 우도비를 지닌 것으로 나타났다. 다른 1명 역시 아시아지역으로 예측되었고, 캄보디아 지역의 Khmer민족과 가장 높은 우도비를 나타내었다(Table 2).

동남아시아 지역 중 유일하게 연구에 포함된 베트남인은 동아시아 출신으로 예측되었다. 100명 중 42명은 1명이 오세아니아의 Micronesians 민족과 가장 높은 우도비를 나타낸 것을 제외하고는, 대만(14/42), Hakka(13/42)와 같은 동아시아 민족과 높은 우도비를 나타내는 것으로 계산되었다(Table 3).

서남아시아(네팔, 인도, 파키스탄)인은 동북아시아와 동남아시아 지역보다 더 복잡한 결과를 나타내었다. 서남아시아인은 주로 남아시아인으로 예측되었지만 250명 중 6명은 유럽인, 3명은 동아시아인으로 예측되었다. 다만, 확률 퍼센트가 여러 지역에

균등하게 분포되어 50% 미만의 예측 확률을 나타내는 결과가 27명에게서 나타났고, 일부 시료는 예측한 지역과 가장 높은 우도비를 나타낸 지역이 일치 하지 않기도 하였다(Table 4). 예를 들면, 파키스탄 Punjab 지역인인 PJ-F-020시료의 경우 남아시아인일 확률이 65 %로 나타났지만, 우도비가 가장 높은 민족은 유럽의 Chuvash (Turkic ethnic group) 민족으로 나타났다.

3. 아시아인의 인구 구조

165개의 SNP 마커의 우도비를 이용하여 7개의 아시아 민족의 인구 구조도를 분석하였다(Fig. 3). 또한, 클러스터의 최적값을 확인하기 위한 프로그램을 사용한 결과, 클러스터 개수가 2인 K=2 값이 최적의 클러스터 개수로 확인되었다.

동북아시아인은 다중 클러스터에서 가장 복잡한 패턴을 보여주었고, 동남아시아로 분류되는 베트남 민족과 혼합된 형태를 보였는데, 특히 동북아시아 민족 중에서도 중국인과의 혼합된 모습이 두드러졌다. 그러나 전체 7 민족을 두 민족으로 나눌 때와(K=2) 세 민족(K=3), 네 민족으로 나눌 때(K=4) 구조도의 형태를 비교해 보면, 동북아시아와 베트남인이 구별되어 베트남인이 세 번째 클러스터를 형성하는 것을 알 수 있다. 또한 중국인 민족은 클러스터

개수가 3일 때(K=3), 동북아시아 지역에서 일본인, 한국인과 분리가 되는 것을 볼 수 있었다.

4. 아시아인 세부 예측을 위한 변수선택

동북아시아-동남아시아, 동북아시아-서남아시아, 동남아시아-서남아시아 자료에 대해 변수선택법(유의수준=0.20, 최대 깊이=30)과 의사결정나무(유의수준=0.20, 최대 깊이=30)을 적용하여 변수의 상대중요도 순으로 변수를 나열하였다(Table 5a, 5b and 5c).

각 자료에 대한 5 가지 언더샘플링 데이터에 대해 선택되는 변수들이 조금씩 달라지는 것을 볼 수 있다. 이는 언더샘플링을 진행할 때마다 추출되는 표본들이 달라지기 때문에 변수에 대한 카이제곱값들 역시 달라지기 때문이다. 따라서 각 자료의 5 가지 데이터의 결과를 합하면, 동북-동남아시아 자료는 변수선택법으로 총 56 개의 변수(17 개 변수 중복), 의사결정나무로 22 개의 변수(5 개 변수 중복)가 선택되었다. 동북-서남아시아 자료는 변수선택법으로 총 10 개의 변수(5 개 변수 중복), 의사결정나무로 7 개의 변수(5 개 변수 중복)가 선택되었고, 동남-서남아시아 자료에서는 변수선택법으로 총

13 개의 변수(3 개 변수 중복), 의사결정나무로 총 5 개 변수(2 개 변수 중복)가 선택되었다.

각 자료 별로 변수선택법과 의사결정나무의 변수중요도로부터 나온 모든 변수들(All), 변수선택법과 의사결정나무의 변수중요도로부터 나온 모든 변수들 중 두 번 이상 중복된 변수들(All_twice), 변수선택법으로부터 나온 모든 변수들(VS_total), 변수선택법으로부터 나온 모든 변수들 중 두 번 이상 중복된 변수들(VS_twice), 의사결정나무의 변수중요도로부터 나온 모든 변수들(DT_total), 의사결정나무의 변수중요도로부터 나온 모든 변수들 중 두 번 이상 중복된 변수들(DT_twice)의 총 6 가지 조합을 통해 각 조합으로 나온 입력변수들을 예측 모델에 적합 시킨 이후 각 민족을 분류하는 최적의 중요 변수들을 선택하고자 하였다. 6 개의 조합에 따른 변수 중에서 변수중요도가 0.2 이상인 변수 중 중복으로 선택된 변수를 최종적인 예측모델의 입력변수로 사용하였다(Table 6a, 6b and 6c, Table 7). 동북-동남아시아 자료에서는 rs3811801, rs174570, rs7997709, rs12629908, rs9319336, rs1083988 변수가 중복되어 총 6 개의 변수가 선택되었다. 동북-서남아시아 자료에서는 rs1426654, rs1229984, rs206690 의 3 개 변수가 중복되었고, 동남-서남아시아 자료에서는 동북-서남아시아

자료에서도 선택된 rs1426654 변수만 중복되었다. 따라서 총 9 개 변수가 선택되었다.

5. 아시아인 예측 모델

변수선택법에서 변수중요도가 0.2 이상이면서 공통적으로 선택된 9 개 변수를 입력변수로 예측모델을 구축하였다. 예측모델의 오분류율, ROC index, 민감도, 특이도를 보면(Table 8), 동북-서남아시아 예측모델이 오분류율 0.015, ROC 0.992, 민감도 0.986, 특이도 0.979 로 가장 우수한 모델인 것으로 나타났다. 그러나 오분류율이 0.277 로 제일 높은 동북-동남아시아 예측모델도 특이도가 0.984 에 달했다.

모델별로 각각의 분류 규칙을 살펴 보면(Fig. 4a, 4b and 4c), 동북-동남아시아 예측 모델에서는 rs174570 마커가 TT 유전자형을 나타내고, rs3811801 마커에서 GG 유전자형이면 83.33%의 확률로 동북아시아인으로 예측되었다. 같은 방법으로 rs174570 마커가 CC 또는 CT 유전자형을 나타내고, rs3811801 마커에서 AA 유전자형이면 90.91%의 확률로 동남아시아인으로 예측되었다. 동북-서남아시아 예측 모델에서는 rs1426654 마커에서 AA 혹은 AG

유전자형이면 서남아시아인일 확률이 97.86%로 계산되었다. 동남-서남아시아 예측모델에서는 rs1426654 마커가 GG 유전자형이면 100% 동남아시아인으로 예측되었다.

Table 1. Details of populations analyzed.

Geographical region	Population sample description	sample size (N)	Data Source
Northeast Asia	Han Chinese South [CHS]	99	Coriell Institute
	Han Chinese in Beijing, China [CHB]	100	Coriell Institute
	Japanese in Tokyo, Japan [JPT]	101	Coriell Institute
	Korean from Seoul, Korea[KOR]	100	Seoul National Univ.
Southeast Asia	Kinh in Ho Chi Minh City, Vietnam [KHV]	100	Coriell Institute
Southwest Asia	Nepalese[NP]	100	Seoul National Univ.
	Indian[IN] (Residing at Deccan region)	51	Seoul National Univ.
	Pakistani in Pathan, Pakistan[PT]	49	Seoul National Univ.
	Punjabi in Lahore, Pakistan[PJ]	50	UHS Lahore
		750	

Table 2. The likelihood ratio of 14 individuals in Northeast Asian based on 165 SNPs. Those were observed to have the highest values for Southeast Asia. Only the highest value is indicated in this table.

Sample	Admixture Prediction		Population Likelihoods		
	Population	%	Population	Geo Region	Highest Likelihood
[CHS]-013	East Asia	100	Lao Loum	Asia	6.02E-47
[CHS]-043	East Asia	100	Lao Loum	Asia	5.60E-46
[CHS]-067	East Asia	100	Lao Loum	Asia	4.20E-47
[CHS]-083	East Asia	100	Lao Loum	Asia	1.11E-48
[CHS]-093	East Asia	95	Lao Loum	Asia	2.81E-43
[CHS]-094	East Asia	95	Lao Loum	Asia	2.51E-48
[CHB]-004	East Asia	95	Lao Loum	Asia	4.06E-46
[CHB]-010	East Asia	75	Lao Loum	Asia	3.01E-49
[CHB]-020	East Asia	100	Lao Loum	Asia	1.02E-46
[CHB]-034	East Asia	75	Lao Loum	Asia	1.33E-54
[CHB]-046	East Asia	100	Lao Loum	Asia	3.89E-48
[CHB]-073	East Asia	70	Lao Loum	Asia	2.00E-45
[JPT]-070	East Asia	80	Cambodians, Khmer	EastAsia	1.07E-49
[KOA]-074	East Asia	95	Lao Loum	Asia	5.31E-49

Table 3. The admixture prediction, likelihood ratio results of 42 Vietnamese data, based on 165 SNPs.

Sample	Admixture Prediction		Population Likelihoods		
	Population	%	Population	Geo Region	Likelihood
[KHV]-001	East Asia	90	Taiwanese Han	EastAsia	1.33E-49
[KHV]-003	East Asia	95	Taiwanese Han	EastAsia	2.18E-42
[KHV]-017	East Asia	100	Japanese-HapMap	EastAsia	1.02E-45
[KHV]-019	East Asia	85	Hakka	EastAsia	1.19E-42
[KHV]-020	East Asia	90	Micronesians	Oceania	2.04E-46
[KHV]-021	East Asia	95	Hakka	EastAsia	9.44E-48
[KHV]-022	East Asia	100	Taiwanese Han	EastAsia	2.26E-49
[KHV]-025	East Asia	90	Japanese-HapMap	EastAsia	1.75E-45
[KHV]-026	East Asia	100	Hakka	EastAsia	1.85E-45
[KHV]-028	East Asia	100	Taiwanese Han	EastAsia	7.53E-46
[KHV]-029	East Asia	85	Taiwanese Han	EastAsia	1.23E-54
[KHV]-030	East Asia	80	Koreans	EastAsia	2.51E-51
[KHV]-034	East Asia	90	Hakka	EastAsia	1.15E-44
[KHV]-035	East Asia	95	Han-HapMap	EastAsia	5.35E-46
[KHV]-038	East Asia	95	Taiwanese Han	EastAsia	1.43E-44

[KHV]-039	East Asia	90	Hakka	EastAsia	2.16E-45
[KHV]-043	East Asia	90	Hakka	EastAsia	2.97E-49
[KHV]-044	East Asia	95	Hakka	EastAsia	4.54E-45
[KHV]-046	East Asia	95	Hakka	EastAsia	1.52E-43
[KHV]-048	East Asia	95	Koreans	EastAsia	5.99E-47
[KHV]-052	East Asia	100	Han-HapMap	EastAsia	2.16E-45
[KHV]-055	East Asia	100	Koreans	EastAsia	3.51E-50
[KHV]-057	East Asia	100	Taiwanese Han	EastAsia	2.45E-46
[KHV]-058	East Asia	95	Taiwanese Han	EastAsia	6.50E-49
[KHV]-059	East Asia	95	Hakka	EastAsia	2.18E-47
[KHV]-060	East Asia	80	Japanese-HapMap	EastAsia	1.15E-50
[KHV]-067	East Asia	90	Koreans	EastAsia	2.52E-41
[KHV]-072	East Asia	95	Taiwanese Han	EastAsia	1.62E-48
[KHV]-074	East Asia	100	Han-HapMap	EastAsia	2.57E-46
[KHV]-075	East Asia	95	Hakka	EastAsia	8.93E-45
[KHV]-076	East Asia	95	Hakka	EastAsia	5.19E-48
[KHV]-079	East Asia	100	Hakka	EastAsia	2.37E-46
[KHV]-081	East Asia	100	Hakka	EastAsia	5.16E-45
[KHV]-083	East Asia	100	Ami	EastAsia	9.72E-46
[KHV]-084	East Asia	95	Han-HapMap	EastAsia	2.07E-49

[KHV]-085	East Asia	85	Taiwanese Han	EastAsia	2.04E-49
[KHV]-090	East Asia	100	Taiwanese Han	EastAsia	2.03E-47
[KHV]-091	East Asia	100	Taiwanese Han	EastAsia	2.67E-44
[KHV]-092	East Asia	100	Japanese-HapMap	EastAsia	4.17E-48
[KHV]-094	East Asia	100	Han-HapMap	EastAsia	2.74E-48
[KHV]-097	East Asia	95	Taiwanese Han	EastAsia	3.69E-49
[KHV]-098	East Asia	95	Taiwanese Han	EastAsia	3.62E-44
[KHV]-100	East Asia	100	Ami	EastAsia	2.25E-43

Table 4. The admixture prediction and likelihood ratio results of unclassified Southwest Asian.

Sample	Admixture Prediction (%)							Population Likelihoods		
	Europe	Oceania	East Asia	Africa	South Asia	America	Southwest Asia	Population	Geo Region	Highest Likelihood
NP-001	50	10	0	0	30	0	10	Adygei	Europe	4.75E-47
NP-006	0	10	55	0	35	0	0	Cambodians,Khmer	EastAsia	7.40E-53
NP-015	0	15	50	0	35	0	0	Thoti	Asia	1.68E-52
NP-017	55	0	0	0	45	0	0	Adygei	Europe	1.08E-43
NP-024	30	0	25	0	40	5	0	Hazara	Asia	1.11E-53
NP-025	0	15	50	0	35	0	0	Hazara	Asia	3.00E-49
NP-026	0	15	45	0	35	5	0	Hazara	Asia	3.04E-53
NP-030	35	5	0	0	45	15	0	Pashtun	Asia	2.54E-57
NP-032	0	0	50	0	45	5	0	Hazara	Asia	3.73E-48
NP-047	0	0	30	0	45	0	25	Hazara	Asia	1.17E-48
NP-052	20	0	25	0	30	0	25	Keralite	Asia	5.20E-51
NP-056	10	10	0	0	40	0	40	Mohanna Asia	Asia	3.20E-50
NP-057	0	10	40	5	45	0	0	Hazara	Asia	9.28E-55

NP-060	15	0	0	0	40	10	35	Keralite	Asia	4.49E-51
NP-061	0	0	45	0	50	5	0	Hazara	Asia	3.27E-52
NP-064	0	10	40	5	45	0	0	Cambodians,Khmer	EastAsia	5.62E-56
NP-067	0	0	80	5	15	0	0	Cambodians,Khmer	EastAsia	6.77E-47
NP-069	0	0	40	0	35	0	25	Hazara	Asia	3.67E-47
NP-074	0	5	70	0	25	0	0	Yakut	Siberia	9.39E-47
NP-078	0	5	80	5	10	0	0	Japanese	EastAsia	2.36E-47
NP-080	30	5	25	0	40	0	0	Hazara	Asia	6.12E-52
NP-105	0	0	75	0	20	5	0	Han-HapMap	EastAsia	8.21E-50
IN-011	20	0	0	0	45	5	30	Kuwaiti	Asia	7.42E-49
PT-004	35	0	0	0	35	0	30	Pashtun	Asia	1.26E-44
PT-010	40	0	0	0	35	10	15	Pashtun	Asia	2.95E-49
PT-011	55	25	0	0	10	0	10	Mohanna	Asia	2.95E-49
PT-024	45	5	0	0	40	0	10	Pashtun	Asia	3.12E-46
PT-043	55	0	0	0	45	0	0	Chuvash	Europe	8.79E-42
PT-059	40	0	0	0	35	0	25	Jews, Ashkenazi	Europe	1.25E-44
PT-060	20	0	0	0	35	0	45	Pashtun	Asia	1.11E-49
PT-065	40	0	0	0	50	0	10	Adygei	Europe	1.18E-49
PT-068	40	0	0	0	35	0	25	Kuwaiti	Asia	1.18E-49
PT-088	15	0	0	0	35	5	45	Kuwaiti	Asia	7.61E-52

PT-104	65	0	0	0	25	0	10	Russians, Archangel'sk	Europe	1.53E-42
PT-154	0	0	10	0	5	0	85	Palestinian	Asia	2.87E-45
PT-155	35	0	0	0	45	0	20	Pashtun	Asia	2.36E-46
PT-158	50	0	0	0	45	5	0	Adygei	Europe	1.88E-49
PT-172	35	5	0	0	10	0	50	Druze	Asia	4.20E-46
PT-177	40	0	15	5	0	0	40	Pashtun	Asia	5.63E-52
PT-179	25	0	0	0	45	0	30	Kuwaiti	Asia	6.43E-52
PJ-M-011	20	0	0	5	0	20	55	Mohanna	Asia	9.73E-60
PJ-M-014	10	0	0	0	45	0	45	Pashtun	Asia	4.90E-47
PJ-M-016	30	0	0	0	45	0	25	Pashtun	Asia	1.57E-50
PJ-M-017	35	0	0	0	60	0	5	Jews, Sephardic	Europe	5.17E-46
PJ-F-005	55	5	0	0	35	5	0	Chuvash	Europe	6.63E-48
PJ-F-008	15	10	0	0	35	0	40	Keralite	Asia	5.94E-46
PJ-F-009	45	0	15	0	35	0	5	Adygei	Europe	1.85E-49
PJ-F-020	25	0	5	5	65	0	0	Chuvash	Europe	2.06E-52
PJ-F-021	0	25	0	0	25	0	50	Pashtun	Asia	7.03E-49
PJ-F-025	25	5	0	0	40	0	30	Pashtun	Asia	4.66E-47
PJ-F-100	30	0	15	0	0	5	50	Keralite	Asia	2.31E-53

Table 5. The results of applying variable selection method(significance level = 0.20, maximum depth = 30) and decision tree(significance level = 0.20, maximum depth = 30)

(a) Northeast−Southeast

Variable Selection (alpha=0.20, pass number=30)									
Undersampling-1		Undersampling-2		Undersampling-3		Undersampling-4		Undersampling-5	
Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance
rs174570	1.000	rs174570	1.000	rs174570	1.000	rs174570	1.000	rs174570	1.000
rs3811801	0.499	rs4984913	0.459	rs1513056	0.746	rs3811801	0.848	rs2702414	0.675
rs2030763	0.409	rs10513300	0.401	rs7226659	0.678	rs2532060	0.636	rs3811801	0.671
rs731257	0.405	rs9809104	0.400	rs214678	0.639	rs1760921	0.512	rs13400937	0.650
rs12629908	0.368	rs2504853	0.395	rs4670767	0.637	rs1513056	0.507	rs1229984	0.649
rs1760921	0.366	rs4908343	0.392	rs10496971	0.625	rs1572018	0.444	rs1950993	0.551
rs10513300	0.361	rs1950993	0.382	rs1500127	0.600	rs10839880	0.423	rs1760921	0.537
rs9809104	0.347	rs4463276	0.381	rs6422347	0.574	rs1229984	0.408	rs192655	0.522
rs917115	0.346	rs10236187	0.364	rs9291090	0.555	rs310644	0.391	rs7997709	0.469
rs192655	0.290	rs2702414	0.324	rs2330442	0.526	rs1079597	0.387	rs10496971	0.396

rs2070586	0.289	rs772262	0.310	rs2238151	0.485	rs4717865	0.354	rs2030763	0.380
rs4411548	0.288	rs2073821	0.305	rs772262	0.401	rs10497191	0.306	rs9319336	0.355
rs10236187	0.277	rs948028	0.302	rs4918664	0.391	rs10511828	0.289	rs1040045	0.329
rs1040404	0.259	rs4955316	0.302	rs1040404	0.335	rs10513300	0.289	rs10108270	0.269
rs12498138	0.259	rs4891825	0.294	rs2166624	0.225			rs10497191	0.179
rs10007810	0.211	rs11227699	0.281						
rs4798812	0.206	rs1325502	0.281						
		rs10108270	0.272						

Decision Tree (alpha=0.20, pass number=30)

Undersampling-1		Undersampling-2		Undersampling-3		Undersampling-4		Undersampling-5	
Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance
rs174570	1.000	rs174570	1.000	rs174570	1.000	rs174570	1.000	rs174570	1.000
rs4821004	0.441	rs4984913	0.459	rs7226659	0.678	rs3811801	0.848	rs12629908	0.636
rs9845457	0.383	rs10513300	0.401	rs214678	0.639	rs2532060	0.636	rs192655	0.609
rs7803075	0.355	rs3793791	0.388	rs4670767	0.637	rs1513056	0.507	rs3811801	0.551
		rs4471745	0.344	rs1513056	0.628	rs1572018	0.444	rs4471745	0.490
		rs10108270	0.341	rs1500127	0.600	rs10512572	0.335	rs7803075	0.470
		rs4908343	0.337	rs3811801	0.483			rs13400937	0.418

(b) Northeast−Southwest

Variable Selection (alpha=0.20, pass number=30)									
Undersampling-1		Undersampling-2		Undersampling-3		Undersampling-4		Undersampling-5	
Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance
rs1426654	1.000	rs1426654	1.000	rs1426654	1.000	rs1426654	1.000	rs1426654	1.000
rs3811801	0.285	rs3811801	0.254	rs3811801	0.285	rs3811801	0.285	rs3811801	0.285
rs1229984	0.258	rs3827760	0.239	rs1229984	0.258	rs1229984	0.258	rs1229984	0.258
rs1950993	0.247	rs1229984	0.235	rs192655	0.247	rs192655	0.247	rs192655	0.247
rs10236187	0.123	rs6422347	0.132	rs12498138	0.123	rs12498138	0.123	rs12498138	0.123
		rs1040404	0.105						

Decision Tree (alpha=0.20, pass number=30)									
Undersampling-1		Undersampling-2		Undersampling-3		Undersampling-4		Undersampling-5	
Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance
rs1426654	1.000	rs1426654	1.000	rs1426654	1.000	rs1426654	1.000	rs1426654	1.000
rs3811801	0.285	rs3827760	0.239	rs3811801	0.285	rs3811801	0.286	rs1229984	0.267
rs1229984	0.258	rs1229984	0.235	rs1229984	0.258	rs1229984	0.260	rs260690	0.239
rs1950993	0.247	rs260690	0.206	rs192655	0.247	rs3827760	0.222	rs3827760	0.211

(C) Southeast−Southwest

Variable Selection (alpha=0.20, pass number=30)									
Undersampling-1		Undersampling-2		Undersampling-3		Undersampling-4		Undersampling-5	
Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance
rs1426654	1.000	rs1426654	1.000	rs1426654	1.000	rs1426654	1.000	rs1426654	1.000
rs1800414	0.306	rs10839880	0.203	rs1040404	0.131	rs2835370	0.253	rs4984913	0.240
rs1229984	0.211	rs1513181	0.159	rs10007810	0.131	rs1229984	0.211	rs10007810	0.210
				rs3793791	0.129	rs10007810	0.171	rs10108270	0.135
				rs10236187	0.125			rs2627037	0.133

Decision Tree (alpha=0.20, pass number=30)									
Undersampling-1		Undersampling-2		Undersampling-3		Undersampling-4		Undersampling-5	
Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance
rs1426654	1.000	rs1426654	1.000	rs1426654	1.000	rs1426654	1.000	rs1426654	1.000
rs1800414	0.253					rs2835370	0.253	rs4984913	0.240
rs1229984	0.211					rs1229984	0.211		

Table 6. The optimal critical variable obtained by fitting the combination of variables to the prediction model. SNP markers with an importance value of 0.2 from duplicated markers are the selected variables.

(a) Northeast−Southeast

Combination	Misclassification rate (train)	Variable (Importance)
All	0.11	rs3811801 (1.00)
		rs174570 (0.96)
		rs3821004 (0.52)
		rs7997709 (0.50)
		rs12629908 (0.49)
		rs9319336 (0.37)
		rs10839880 (0.24)
All_twice	0.15	rs3811801 (1.00)
		rs174570 (0.96)
		rs10108270 (0.44)
		rs192655 (0.40)
		rs2702414 (0.35)
		rs10513300 (0.32)
VS_total	0.1	rs3811801 (1.00)
		rs174570 (0.96)
		rs7997709 (0.50)
		rs12629908 (0.49)
		rs192655 (0.48)
		rs10108270 (0.43)
		rs9319336 (0.37)
		rs10839880 (0.24)
VS_twice	0.13	rs3811801 (1.00)

		rs174570 (0.96)
		rs4821004 (0.52)
		rs12629908 (0.49)
DT_total	0.13	rs3811801 (1.00)
		rs174570 (0.96)
		rs4821004 (0.52)
		rs12629908 (0.49)
DT_twice	0.17	rs3811801 (1.00)
		rs174570 (0.94)
		rs1513056 (0.33)
		rs4471745 (0.19)

(b) Northeast-Southwest

Combination	Misclassification rate (train)	Variable (Importance)
All	0.017	rs1426653 (1.00)
		rs1229984 (0.26)
		rs260690 (0.21)
All_twice	0.017	rs1426653 (1.00)
		rs1229984 (0.26)
		rs260690 (0.21)
VS_total	0.017	rs1426653 (1.000)
		rs1229984 (0.26)
		rs3827760 (0.21)
VS_twice	0.02	rs1426653 (1.00)
		rs1229984 (0.26)
		rs192655 (0.21)
DT_total	0.017	rs1426653 (1.00)
		rs1229984 (0.26)
		rs260690 (0.21)
DT_twice	0.017	rs1426653 (1.00)
		rs1229984 (0.26)
		rs260690 (0.21)

(C) Southeast-Southwest

Combination	Misclassification rate (train)	Variable (Importance)
All	0.016	rs1426653 (1.000) rs10007810 (0.1773) rs1800414 (0.1633)
All_twice	0.024	rs1426653 (1.000)
VS_total	0.024	rs1426653 (1.000)
VS_twice	0.024	rs1426653 (1.000)
DT_total	0.016	rs1426653 (1.000) rs1229984 (0.1773) rs1800414 (0.1633)
DT_twice	0.024	rs1426653 (1.000)

Table 7. The results of the selected variables. Only markers with an importance value higher than 0.2 were selected. The nine markers in this table were finally selected to use in an ancestry prediction model.

Model		Variable (Importance)		Total
Northeast Southeast	–	rs3811801	(1.00)	rs10839880
		rs174570 rs7997709	(0.96)	rs1229984
		rs12629908 rs9319336	(0.50)	rs12629908
		rs10839880	(0.49)	rs1426654
			(0.37)	rs174570
			(0.24)	rs260690
Northeast Southwest	–	rs1426654	(1.00)	rs3811801
		rs1229984	(0.22)	rs7997709
		rs260690	(0.20)	rs9319336
Southeast Southwest	–	rs1426654	(1.00)	

Table 8. Evaluation of predicted models.

Model	Training Misclassification Probability	Training AOC	Training Sensitivity	Training Specificity
Northeast – Southeast	0.277	0.723	0.698	0.984
Northeast – Southwest	0.015	0.992	0.986	0.979
Southeast – Southwest	0.027	0.973	0.952	1.000

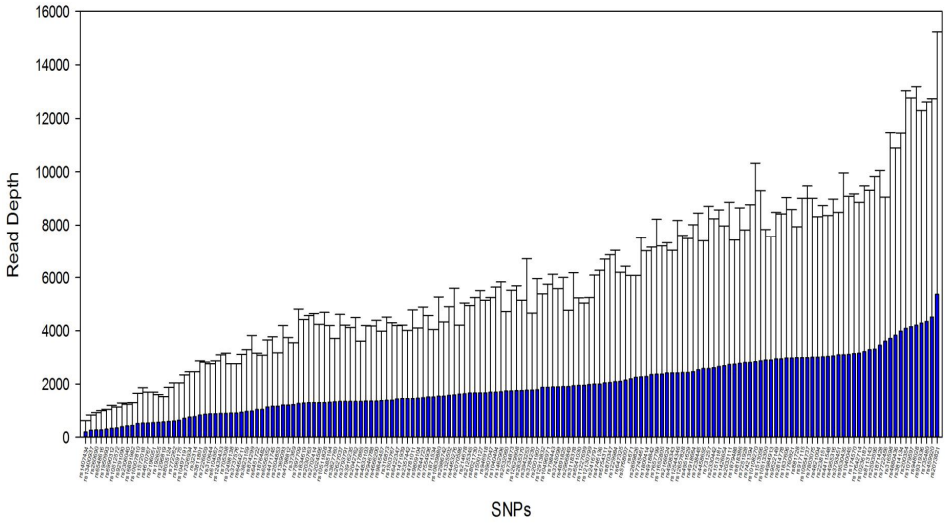


Figure 1. Read depth per marker based on analyses of 750 samples (mean and standard deviation).

Sample: Han-001	Admixture Prediction - Set of 151 AISNPs	Barcode: IonCode_0101

Population Name	Percentage
Europe	0.0
Oceania	0.0
East Asia	100.0
Africa	0.0
South Asia	0.0
America	0.0
Southwest Asia	0.0

Confidence: **HIGH.** Log-likelihood value: 104.06. The likelihood value of the sample is within the 95% confidence interval obtained from 10,000 random samples with the same admixture proportions as the test sample

Sample: Han-001	Population Likelihoods - Set of 151 AISNPs	Barcode: IonCode_0101

Population Name	Geo Region	Likelihood
Hakka	EastAsia	8.38E-45
Taiwanese Han	EastAsia	6.55E-46
Han-HapMap	EastAsia	2.47E-46
Lao Loum	Asia	1.78E-46
Koreans	EastAsia	3.64E-47
Cambodians, Khmer	EastAsia	8.84E-48
Japanese-HapMap	EastAsia	2.48E-48
Japanese	EastAsia	8.36E-50
Atayal	EastAsia	6.35E-52
Ami	EastAsia	3.54E-52
Micronesians	Oceania	4.83E-53
Malaysians	EastAsia	4.65E-55
Yakut	Siberia	6.22E-57
Hazara	Asia	1.81E-59
Samoans	Oceania	2.30E-60
Kachari	Asia	1.04E-62
Thoti	Asia	1.98E-65
Khanty	Asia	1.04E-68
Keralite	Asia	1.20E-71
Mohanna	Asia	4.44E-75

Maya, Yucatan	NorthAmerica	1.37E-75
Pashtun	Asia	7.08E-76
Chuvash	Europe	7.19E-77
Quechua	SouthAmerica	6.22E-80
Negroid Makrani	Asia	2.69E-83
Kuwaiti	Asia	5.99E-84
Melanesian, Nasioi	Oceania	3.33E-84
Ticuna	SouthAmerica	9.87E-87
Guihiba	SouthAmerica	3.59E-88
Palestinian	Asia	2.22E-88
Adygei	Europe	9.71E-90
Somali	Africa	1.96E-90
Russians	Europe	3.20E-91
Pima, Mexico	NorthAmerica	1.77E-91
Papuan New Guinean	Oceania	8.60E-94
Komi-Zyrian	Asia	8.50E-94
Jews, Ethiopian	Africa	2.52E-94
Finns	Europe	1.91E-94
Russians, Archangel'sk	Europe	2.73E-95
Hungarian	Europe	1.06E-99
Druze	Asia	8.47E-100
Greeks	Europe	4.09E-100
Jews, Ashkenazi	Europe	2.42E-100
Jews, Sephardic	Europe	4.57E-101
Danes	Europe	2.14E-104
Jews, Yemenite	Asia	1.60E-106
Europeans-HapMap	Europe	9.10E-107
African Americans	Africa	4.67E-107
European Americans	Europe	2.41E-107
Sardinian	Europe	7.28E-109
Irish	Europe	8.74E-110
Samaritans	Europe	6.92E-113
Karitiana	SouthAmerica	1.67E-113
Surui	SouthAmerica	1.36E-116
Sandawe	Africa	2.64E-121
Masai	Africa	1.81E-125
Lisongo	Africa	2.01E-129
Chagga	Africa	8.96E-130
Hausa	Africa	6.09E-144
Zaramo	Africa	2.15E-153
Ibo	Africa	3.68E-167
Mbuti	Africa	6.39E-170
Yoruba	Africa	1.24E-170
Yoruba-HapMap	Africa	3.92E-175
Biaka	Africa	7.28E-183

Figure 2. Example of HID SNP Genotyper Report from Han Chinese

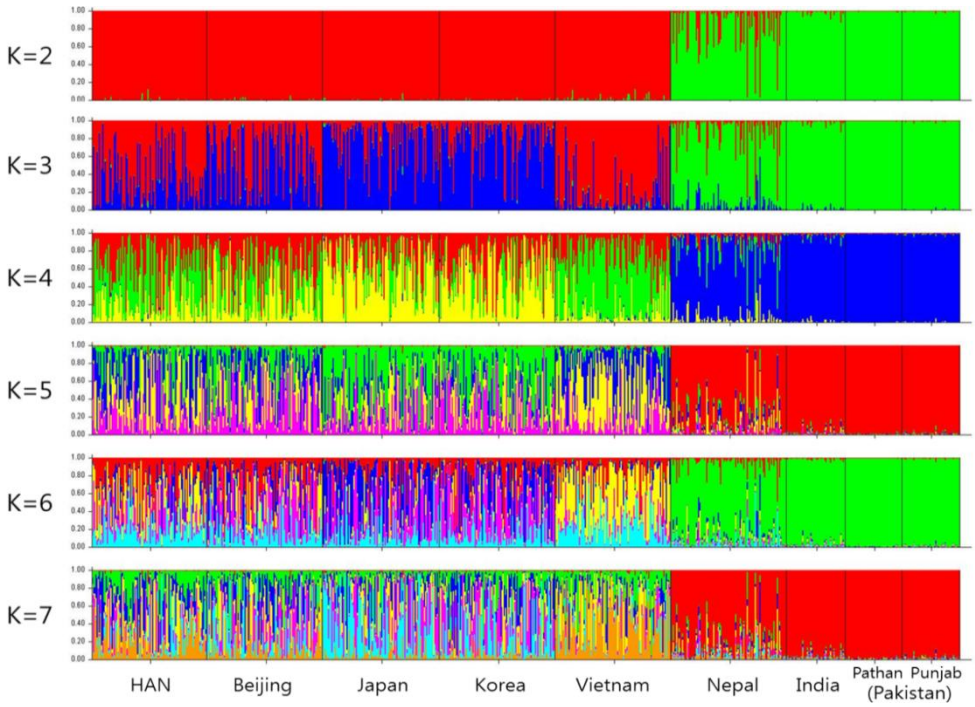
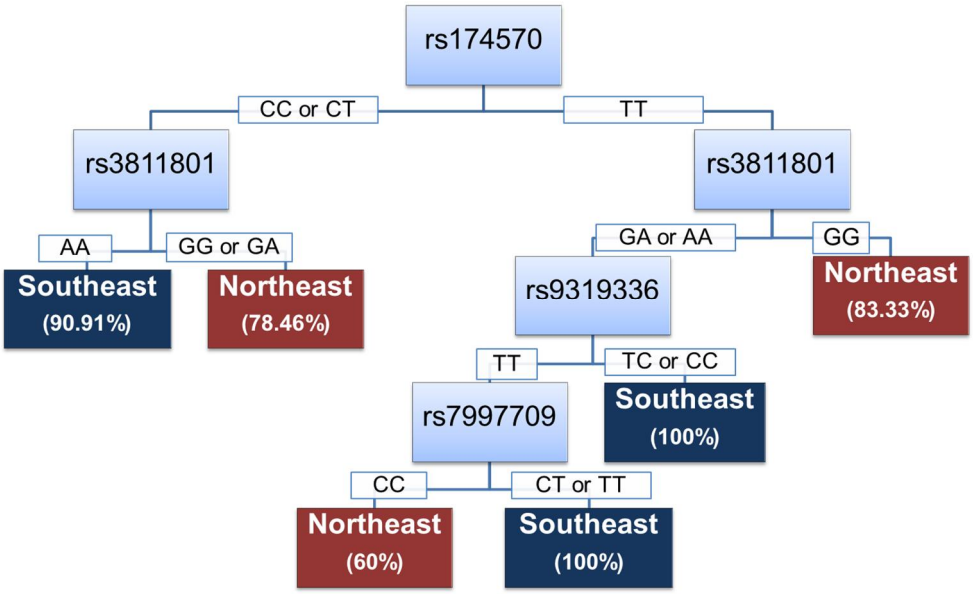
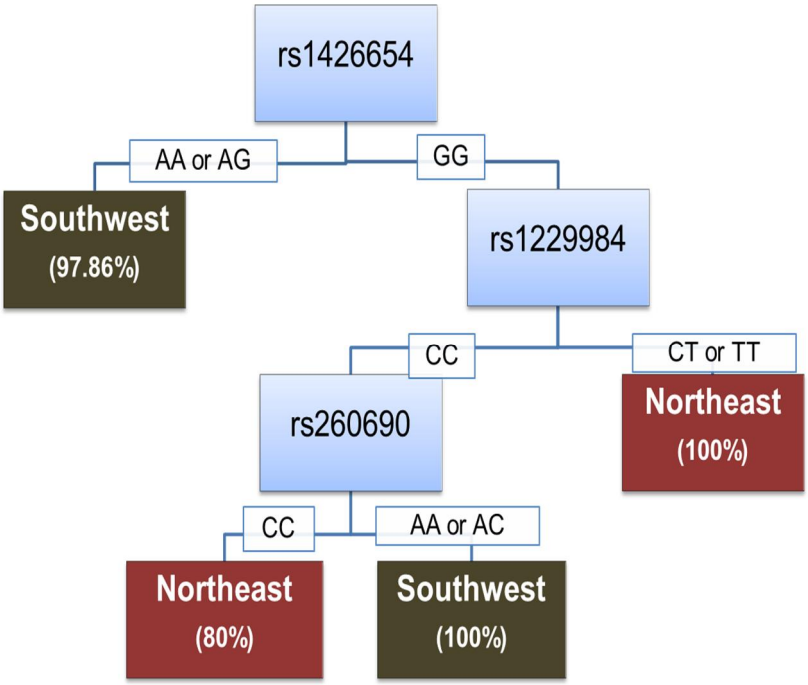


Figure 3. STRUCTURE analysis results based on 165 SNPs. The most likely number of clusters (K) is two.

(a) Northeast – Southeast



(b) Northeast – Southwest



(c) Southeast – Southwest

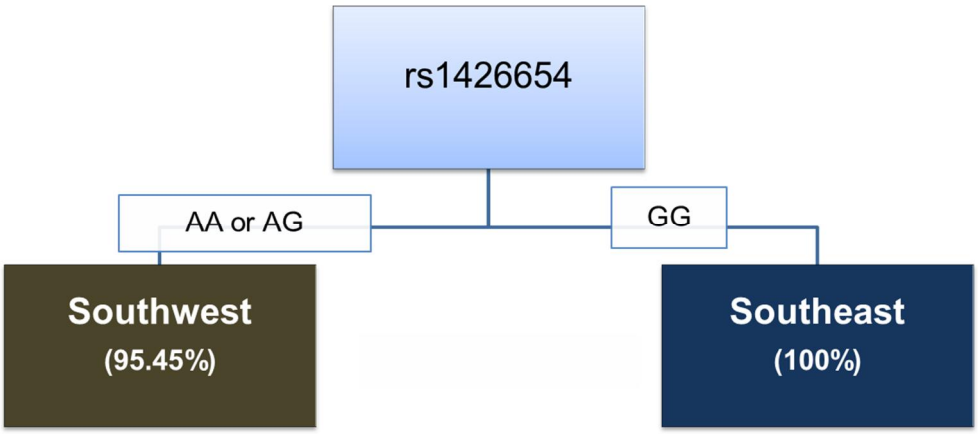


Figure 4. Classification Rules. This is an illustration that explains classification rules using decision trees. This probability is calculated as the probability among the samples included in this model.

고 찰

유전자를 이용한 민족의 구분은 기본적으로 구분하고자 하는 민족 간의 유전적 차이를 밝힘으로써 시작된다. 유전적 차이는 단순한 민족의 구분 이외에도 다양한 목적으로 진행될 수 있는데, 수사 목적에서는 민족 간 차이를 이용한 구분이라는 접근보다는 어느 민족 혹은 지역에 속할 것인지 예측하려는 접근이 더 유용할 것으로 예상된다. 이와 관련해 어느 민족으로 예측되는지 수치화된 결과가 주어진다면 수사의 범위를 좁히고 법적 증거물로 사용되는 데에 적지 않은 도움을 줄 것으로 기대된다. 물론 현재의 연구결과만으로는 효용성이 그리 높지 않을 수도 있겠지만, 대상 민족의 확충이나 성염색체 검사 결과 등의 추가를 통한 보완으로 결과의 정확성을 높일 수 있을 것으로 예상한다. 또한, 이 연구는 대규모 병렬 시퀀싱 실험이 일반화되어 빅데이터의 시대를 맞이한 이때, 다양한 민족 집단에 관한 데이터가 누적되기 시작한 것을 시기 적절하게 고려하고 있다. 따라서 본 연구에서는 기존에 법과학 분야에서 흔히 시도되지 않았던 예측 모델을 유전자 자료 분석에 사용하여 아시아인들을 세 개의 아시아 지역으로 분류하는 모형을 구축하였다. 이 과정에서 선행된 연구에 비해 몇 가지 통계적인 방법을 적용해 보았고, 통계적으로 유의한 마커를 선정하고자 하였다. 또한, 통계적 분류 규칙의 따른 예측

모식도를 구축해보고자 하였다.

SNP 마커들의 커버리지는 깊이(depth; x)로 나타나는데, 일반적으로 102x 이상이면 같은 시료를 다른 시퀀싱 시약과 기기를 이용해서 시퀀싱 했을 때 99.999937% 동일한 유전자형을 나타낸다고 알려져 있다(Ajay SS et al., 2011). 따라서 본 연구에서도 최소 커버리지가 102x를 이상을 나타내도록 한 칩에 포함되는 시료의 개수를 200개로 제한하여 실험을 설계하였다. 본 연구는 대규모 병렬 시퀀싱 패널을 사용한 초기 단계의 연구인 관계로 패널의 성능 등을 파악해 보고자 평균 깊이를 약 2000x가 될 수 있도록 한 칩에 포함되는 시료의 개수를 계산하였지만, SNP 마커의 정보만 얻는 실험인 경우에는 한 칩에 포함되는 시료의 개수를 조절할 수 있겠다.

일부 샘플에서 유전자형이 나타나지 않는 결과가 관찰되었는데, 동일한 DNA를 이용하여 상염색체 STR, Y-염색체 STR, 미토콘드리아 조절 부위(control region)를 타이핑 해본 결과, DNA의 상태에는 이상이 없는 것으로 나타났다. 동일한 패널을 이용한 다른 논문(Themudo GE et al., 2016)에서도 DNA의 손상과는 상관없이 일부 유전자 위치에서 유전자 형이 나타나지 않는 것으로 보아 패널의 시발체의 구조적인 문제인 것으로 생각된다. 다만 유전자형이 나타나지 않는 유전자 위치가 달라 본 연구에서 유전자형이 나타나지 않는 마커의 시발체 접합 부위에 아시아인의 특이적인 구조적 문제가

있는지는 좀 더 확인이 필요한 상황이다. 또한 본 연구에서는 연관 비평형이 발견되지 않았는데 이는 전체 시료의 크기가 750명 밖에 되지 않기 때문일 수도 있다. 따라서 표본의 수나 집단의 수가 증가할 경우에는 단일형 및 연관 비평형 여부를 다시 고려할 필요가 있다.

패널 제조사에서 제공하는 Plugin을 이용한 결과를 살펴보면, Lao Loum 지역인과 가장 높은 우도비를 지닌 동북아시아인 13명 중 12명은 중국인으로 확인되었다(Table 2). 라오스(Lao People's Democratic Republic)는 동남아시아의 메콩강(Great Mekong) 하위 지역에 위치한 내륙 국가로 지역에 따라 사회경제적 및 생태적 역학 특성이 크게 다르다. 특히 북부 지역에서는 산과 고지대가 많고, 중국 남부 지역과 비슷한 생태계가 발견된다고 한다(Sayasone S et al., 2011). 때문에 9 세기 또는 10 세기 이후 중국 남부에서 라오스 북부 지방으로 중국인의 이주가 일어났고(Stuart-Fox M, 2008), 이러한 이유로 199 명의 중국인 중 12 명이 Lao Loum 민족과 높은 우도비를 보인 것으로 예상된다.

베트남인 Plugin 결과를 보면(Table 3), 14명이 대만인과 우도비가 높은 것으로 계산되었다. 대만의 인구 분포는 매우 다양해, 원주민(1.5 %), Min Nan 과 Hakka 민족(91 %), 중국 본토인(7.5 %) 등으로 구성되어있다고 한다. Min Nan 과 Hakka 족은 17 세기 이래 중국 동남부에서 대만으로 이민 온 대만인(Lin M et al., 2000)으로,

대만은 지리적으로 동북아시아의 일부이지만 역사적으로 동남아시아와 많은 유사점이 있다. 따라서 동남아시아 민족을 정확하게 예측하기 위해서는 많은 수의 동남아시아 민족이 참조 데이터베이스에 포함되어야 하겠다. 마찬가지로 인도인도 Negrito, Australoid, Mongoloid, Caucasoid의 4가지 유형으로 분류될 수 있다고 알려져 있다(Majumder PP, 1998). 이러한 이유로 동남아시아와 서남아시아인을 예측하거나 예측 백분율을 추가 증거로 사용하는데 주의를 기울여야 한다.

STRUCTURE 프로그램을 이용한 7개 민족의 인구 구조도를 살펴보면(Fig. 3), 동북아시아 지역의 일본과 한국 지역은 베트남 민족과는 다른 기원을 반영한다는 것을 알 수 있고, 베트남 민족과 중국의 한족이 장기간 혼합되어 왔을 것이라는 예상을 할 수 있다. 또한, 같은 동북아시아 지역 내에서도 중국과 일본, 한국 지역이 구분될 가능성이 보이기도 한다. 따라서 민족 추정 용도의 SNP를 이용한 마커 패널이 재구성되거나 마커를 추가적으로 보완한다면 아시아인을 세부 집단으로 구분할 충분한 가능성이 있고, 더 나아가 지역적 구분이 아닌 나라 별로도 구분을 할 수 있음을 암시한다.

본 연구는 아시아 지역에 있는 7 개 민족만 대상으로 하여, 이 7개 민족 중에는 현재 Torrent Server HID SNP Genotyper 데이터베이스에 포함되어 있는 민족도 있고, 포함되어 있지 않은

민족도 있다. 이러한 소프트웨어는 이 민족 예측용 패널의 장점이라고 할 수 있다. HID-SNP Genotyper v.4.2 Plugin(Thermo Fisher)은 해당 개체군에서 SNP 마커의 유전자형 빈도를 고려하여 데이터베이스 내의 다양한 참조 집단에서 유전자형의 발생 확률을 계산한다. 결과적으로 이 Applied Biosystems™ Precision ID Ancestry Panel은 Plugin 결과를 사용하여 동북아시아 사람들을 동아시아인으로 비교적 정확하게 예측할 수 있다. 그러나 이러한 소프트웨어를 사용하여 동남 및 서남아시아인을 확인하려면 다양한 측면에서의 접근이 필요하다. 법과학 분야에서 개인의 가장 적절한 민족 혹은 출신지를 추론하는 것은 전 세계에 포괄적인 참조 그룹을 두는 것에 달려 있다. Plugin은 현재 소프트웨어가 제공하는 것 이외의 참조 인구를 포함할 수 없고, 데이터베이스내의 참조 인구의 정보 등을 수정할 수 없다. 이에 이러한 소프트웨어를 아시아인에게 바로 적용하기는 어려움이 있을 것으로 예상하여, 아시아인을 예측하는데 좀 더 유용한 모델을 만들어 보고자 하였다.

예측 모델 구축에는 의사결정나무(decision tree) 알고리즘을 사용하였다. 의사결정나무는 과대적합(overfitting)이 일어나기 쉽고 결과 또는 성능의 변동 폭이 크다는 점과 연속 변수가 분할 과정에 의해 암묵적으로 이산화되어 정보가 손실된다는 단점이 있다. 하지만, 다른 기계 학습(machine learning) 알고리즘과 비교할 때, 의사결정

나무는 분류의 규칙이 겉으로 드러나지 않는 블랙박스(black-box) 모델이 아니어서 분류 규칙으로 쉽게 알고리즘을 설명할 수 있다는 장점이 있다(Dreiseitl S et al., 2002). 본 연구에서는 분류 규칙을 효과적으로 설명하고, 실용성을 높이고자 화이트 박스(white-box) 모형을 예측 모형으로 사용하였지만 모델의 성능을 높이기 위해 인공 신경망(artificial neural networks)과 같은 블랙박스 모델에 대한 도입도 고려할 필요가 있다.

또한, 본 연구를 통해 획득한 SNP 유전자 정보를 기반으로 법과학 시료에도 적용 가능하면서도 가능한 한 적은 마커를 변수로 사용하는 민족 예측 모델을 구축하고자 하였다. 165개의 SNP 마커 변수 중에서 최종적으로 9개의 SNP 마커가 선택되었다. 이 중에서 rs1426654 마커는 동남아시아와 서남아시아를 구분하는 단 하나의 마커로 선정되었다. 이 마커는 피부 색소 침착과 연관성을 확인된 마커로 전체 표현형 변이의 27% 가량을 설명할 수 있는 것으로 알려져 있다. 이 유전자 변이는 남 아시아 인들 사이의 피부색 변이와 관련이 있다고 알려져 있다(Mallick CB et al., 2013).

선택된 마커에 대한 추가적인 정보를 얻고자, 다양한 대규모의 시퀀싱 프로젝트로부터 얻은 exome과 genome 데이터를 제공하는 Genome Aggregation Database(gnomAD)(Karczewski K et al., 2017)에서 9개의 마커를 검색해 보았다. 이 데이터베이스는

125,748개의 Exome 데이터와 15,708개의 genome 데이터를 보유하고 있으며 현재까지 가장 많은 인간 genome을 축적한 데이터베이스이다. 이 데이터베이스에는 아시아인이 동아시아인과 남아시아인으로 나누어져 분류되어 있지만, 아쉽게도 남아시아인의 genome 데이터는 대부분 포함되어 있지 않다. 동북-동남을 분류한 6개의 마커는 동아시아인에서 대립유전자 빈도(allele frequency)가 0.3140부터 0.6947로 나타났다. 이는 패널에 속한 SNP 마커들이 선정되었을 때, 대립유전자 빈도(MAF; minor allele frequency)를 고려해 선정하였기 때문인 것으로 생각된다. 이들의 남아시아인 대립유전자 빈도 정보는 포함되지 않았다. 동북-서남, 동남-서남 분류 시 공통으로 선택되었던 마커인 rs1426654의 대립유전자 빈도는 동아시아인에서 0.9931, 남아시아인에서 0.2251로 나타났다. 이러한 대립유전자 빈도의 차이가 이 마커를 중요도가 높은 마커로 선정한 것으로 보인다. 동북-서남 분류에서 선택된 rs1229987 마커 역시 마찬가지로 동아시아인에서 0.2623, 남아시아인에서 0.9496로 대립유전자 빈도의 차이를 보였다. 동북-서남 분류에서 선택된 rs260690의 동아시아인 대립유전자 빈도는 0.04955로 매우 낮은 수치를 나타냈다. 남아시아인의 대립유전자 정보는 포함되지 않았지만 동아시아인과 비교해서 상대적으로 높은 대립유전자 빈도값을 나타낼 것으로 예상한다.

예측 모델에 대한 평가 테이블을 살펴보면(Table 8), 동북-동남아시아의 오분류 비율이 상대적으로 높지만 이는 동남아시아인이 베트남만 포함되어 있고 그 수도 매우 작기 때문이라는 판단이다. 그러나 예측 모형의 훈련용 자료의 ROC가 모두 0.7이상으로 계산되었고, 세 모델 모두에서 특이성이 높기 때문에 이것으로 미루어 보아 새로운 분류 분석 접근법에 타당성이 있음을 확인하였다.

유전적 차이를 사용하여 국적 혹은 출신지를 예측하는 것은 그렇게 어렵지는 않지만 실용적 측면에서 어려움이 있어 왔다. 가장 큰 어려움은 민족이나 국적의 개념이 반드시 유전적 차이로 반영되지 않는다는 것이다. 이러한 관점에서 볼 때 대륙을 단순히 세 지역으로 나누는 것이 대단히 어려울 수 있다. 특히 아시아는 민족 별로 특정 지역에 집중되어 있는 민족도 있고, 역사적으로 여러 나라에 널리 분산되어 명맥이 이어져 있는 다양한 민족이 있다. 또한, 동일 국가 출신이라도 역사적인 이유로 국가 내 지역에 따라 구분되는 대륙적 기반이 다를 수 있다. 이러한 이유로 민족의 유전 데이터 수집 시 피험자의 출신지에 대해 자세히 확인할 필요가 있고, 피험자 모집 기준을 명확히 할 필요가 있다. 이와 별개로, 본 연구에서는 국가적인 기준에 따라 동북, 동남, 서남 아시아를 구분하였지만, 지리적으로 그리고 역사적인 민족의 이동 등을 고려하여 새로운 기준으로 아시아를 세부 지역으로 분류하는 연구도 진행하여 국가적인 기준과

지리적인 기준에 대한 비교와 이 둘을 통합한 새로운 세부 지역 분류에 대한 기준을 세우는 것도 필요할 것으로 생각한다. 다만 실제 연구에서는 국가를 대상으로 그룹을 구분하여 진행할 수밖에 없었다. 결국 유전적인 접근일 수밖에 없는 본 연구결과가 상황에 따라 제한적일 수도 있다. 연구가 진행됨에 따라 그룹 구분이 좀 더 세분화되고 다양한 분석 방법 개발을 통해 이러한 문제를 완화할 수 있을 것으로 기대한다. 모델을 구축함에 있어 여러 다양한 통계적인 접근 방법이 시도될 수 있을 것으로 보며, 본 연구 또한 최종적인 예측 모델의 완성이라기보다는 이런 다양한 시도 가운데 하나라고 보는 것이 적절하다.

한편 본 연구에 활용된 일부 민족의 경우에는 시료 혹은 정보가 충분하지 못하였을 가능성이 있다. 특히 동남아시아인의 시료 수가 다른 지역보다 적었는데 이는 동남아시아와 다른 아시아 지역을 분류하는데 정확도를 감소시키는 기능을 하였을 가능성이 있다. 추후 연구에서 동남아시아인들의 데이터를 더욱 많이 확보하고, 동남아 지역 외에도 다양한 아시아 지역의 유전자 데이터가 포함된다면 본 분류 모형의 성능이 더욱 향상될 가능성도 기대할 수 있을 것이다. 또한, 상염색체 SNP 중에서도 현재까지 밝혀지지 않은 아시아인에 특이적인 새로운 마커를 추가하거나, 상염색체 SNP가 아닌 다른 리니지 마커(Lineage marker)들과 통합해서 분류하는 모델의 개발도

의미가 있을 것으로 기대한다.

다시 한 번 강조하지만, 본 연구의 의미는 결과의 완성도라기보다는 아시아인을 대상으로 한 새로운 접근이라는 차원에서 의미가 있다고 생각한다. 초기단계의 예측 모델이라는 측면에서 향후 새로운 민족의 데이터가 추가될 때 보다 정교한 접근이 가능할 것으로 기대 된다. 아직 좀 더 보완이 필요한 부분이 적지 않고, 결과의 해석에 있어 세부 아시아인들의 유전적 특성에 대한 이해 또한 필요한데, 추가적인 연구를 통해 좀 더 세분화된 분류모형 구축으로의 확장 가능성을 확인한 것도 중요한 의미라고 본다. 또한 새로운 분석 알고리즘을 개발을 통해 의미 있는 수사도구로서의 실용적 활용성을 기대할 수 있을 것으로 본다.

참 고 문 헌

Ajay SS, Parker SC, Abaan HO, Fajardo KVF, Margulies EH (2011) Accurate and comprehensive sequencing of personal genomes. Genome research.

Allocco DJ, Song Q, Gibbons GH, Ramoni MF, Kohane IS (2007) Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms. BMC Genomics 8(1):68.

Barandiaran I (1998) The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8):832-844.

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263-265.

Børsting C, Morling N (2015) Next generation sequencing and its applications in forensic genetics. Forensic Science International: Genetics 18:78–89.

Breiman L (2001) Random Forests. Machine Learning 45(1):5–32.

Brown I (2012) An experimental comparison of classification techniques for imbalanced credit scoring data sets using SASÒ Enterprise Miner. In Proceedings of SAS Global Forum.

Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL (2015) The genetic ancestry of african americans, latinos, and european Americans across the United States. The American Journal of Human Genetics 96(1):37–53.

Butler JM (2011) Advanced topics in forensic DNA typing: methodology. Academic Press.

Chen J, Zheng H, Bei JX, Sun L, Jia WH, Li T, Zhang F, Seielstad M, Zeng YX, Zhang X, Liu J (2009) Genetic structure of the Han

Chinese population revealed by genome-wide SNP variation. The American Journal of Human Genetics 85(6):775-785.

Conomos MP, Miller MB, Thornton TA (2015) Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genetic epidemiology 39(4):276-293.

De la Puente M, Santos C, Fondevila M, Manzo L, Carracedo Á, Lareu MV, Phillips C, EUROFORGEN-NoE Consortium (2016) The Global AIMs Nano set: a 31-plex SNaPshot assay of ancestry-informative SNPs. Forensic Science International: Genetics 22:81-88.

Dreiseitl, S., Ohno-Machado, L (2002) Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics 35(5-6):352-359.

Earl DA, vonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and

implementing the Evanno method. Conservation genetics resources 4(2):359-361.

Enoch MA, Shen PH, Xu K, Hodgkinson C, Goldman D (2006) Using ancestry informative markers to define populations and detect population stratification. Journal of Psychopharmacology 20(4_suppl):19-26.

Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, Saxonov S, Avey L, Wojcicki A, Pe'er I, Mountain JL (2010) Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS Genetics 6(6):e1000993.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164 (4):1567-1587.

Freund Y, Schapire RE (1999) Large margin classification using the perceptron algorithm. Machine Learning 37(3):277-296.

Friedman JH (2001) Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistic 29(5):1189-1232.

Friedman JH (2002) Stochastic Gradient Boosting. Computational Statistics & Data Analysis 38(4):367-378.

Kamiński B, Jakubczyk M, Szufel P (2018) A framework for sensitivity analysis of decision trees. Central European Journal of Operations Research 26(1):135-159.

Karczewski K, Francioli L (2017) The genome Aggregation Database (gnomAD). MacArthur Lab.

Kidd KK, Kidd JR, Pakstis AJ, Speed WC (2012) Better SNPs for better forensics: ancestry, phenotype, and family identification. In Poster presented at the National Institute of Justice (NIJ) annual meeting, Arlington VA.

Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Seldin MF (2009) Ancestry informative marker sets for determining

continental origin and admixture proportions in common populations in America. Human mutation 30(1):69-78.

Lee HY, Lee SD, Shin KJ (2016) Forensic DNA methylation profiling from evidence material for investigative leads. BMB reports 49(7):359.

Li CX, Pakstis AJ, Jiang L, Wei YL, Sun QF, Wu H, Bulbul O, Wang P, Kang LL, Kidd JR, Kidd, KK (2016) A panel of 74 AISNPs: improved ancestry inference within Eastern Asia. Forensic Science International: Genetics 23:101-110.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100-1104.

Lin M, Chu CC, Lee HL, Chang SL, Ohashi, J, Tokunaga K., Akaza T, Juji T (2000) Heterogeneity of Taiwan's indigenous population:

possible relation to prehistoric Mongoloid dispersals. Tissue Antigens 55(1):1-9.

Majumder PP (1998) People of India: Biological diversity and affinities. Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews 6(3):100-110.

Mallick CB, Iliescu FM, Möls M, Hill S, Tamang R, Chaubey G, Goto R, Ho SYW, Romero IG, Crivellaro F, Hudjashov G, Rai N, Metspalu M, Mascie-Taylor CGN, Pitchappan R, Singh L, Mirazon-Lahr M, Thangaraj K, Villems R, Kivisild T (2013) The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. PLoS genetics 9(11):e1003912.

Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF (2009) An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC genetics 10(1):39.

Opitz D, Maclin R (1999) Popular ensemble methods: An empirical study. Journal of artificial intelligence research 11:169−198.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155(2):945-959.

Quinlan JR (1987) Simplifying decision trees. International Journal of Man−Machine Studies 27(3):221−234.

Quintana−Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, Santachiara−Benerecetti AS (1999) Genetic evidence of an early exit of Homo sapiens from Africa through eastern Africa. Nature genetics 23(4):437-441.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd K, Zhivotovsky LA, Feldman MW (2012) Genetic structure of human populations. Science 298(5602):2381-2385.

Rosenblatt F (1958) The perceptron: a Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review 65(6):386-408.

Royal CD, Novembre J, Fullerton SM, Goldstein DB, Long JC, Bamshad MJ, Clark AG (2010) Inferring genetic ancestry: opportunities, challenges, and implications. The American Journal of Human Genetics 86(5):661-673.

Santos C, Phillips C, Fondevila M, Daniel R, van Oorschot RA, Burchard EG, Schanfield MS, Souto L, Uacyisrael J, Via M, Carracedo Á, Lareu MV (2016) Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region. Forensic Science International: Genetics 20:71-80.

Sayasone S, Mak TK, Vanmany M, Rasphone O, Vounatsou P, Utzinger J, Akkhavong K, Odermatt P (2011) Helminth and intestinal protozoa infections, multiparasitism and risk factors in Champasack province, Lao People's Democratic Republic. PLoS neglected tropical diseases 5(4):e1037.

Stuart-Fox M (2008) Historical dictionary of Laos. Scarecrow Press.

Suo C, Xu H, Khor CC, Ong RT, Sim X, Chen J, Tay WT, Sim KS, Zeng YX, Zhang X, Liu J, Tai ES, Wong TY, Chia KS, Teo YY (2012) Natural positive selection and north-south genetic diversity in East Asia. European Journal of Human Genetics 20(1):102-110.

Themudo GE, Mogensen HS, Børsting C, Morling N (2016) Frequencies of HID-ion ampliseq ancestry panel markers among Greenlanders. Forensic Science International: Genetics 24:60-64.

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ (2000) Y chromosome sequence variation and the history of human populations. Nature genetics 26(3):358-361.

Wei YL, Wei L, Zhao L, Sun QF, Jiang L, Zhang T, Liu HB, Chen JG, Ye J, Hu L, Li CX (2016) A single-tube 27-plex SNP assay for estimating individual ancestry and admixture from three continents. International journal of legal medicine 130(1):27-37.

Xu S, Yin X, Li S, Jin W, Lou H, Yang L, Gong X, Wang H, Shen Y, Pan X, He Y, Yang Y, Wang Y, Fu W, An Y, Wang J, Tan J, Qian J, Chen X, Zhang X, Sun Y, Wu B, Jin L (2009) Genomic dissection of population substructure of Han Chinese and its implication in association studies. The American Journal of Human Genetics 85(6):762-774.

Xu S (2012) Human population admixture in Asia. Genomics & informatics 10(3):133-144.

Zhang X, Kampuansai J, Qi X, Yan S, Yang Z, Serey B, Sovannary T, Bunnath L, Aun HS, Samnom H, Kutanan W, Luo X, Liao S, Kangwanpong D, Jin L, Shi H, Su B (2014) An updated phylogeny of the human Y-chromosome lineage O2a-M95 with novel SNPs. PLoS One 9(6):e101020.

Zhong H, Shi H, Qi XB, Duan ZY, Tan PP, Jin L, Su B, Ma RZ (2011) Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route. Molecular biology and evolution 28(1):717-727.

Abstract

Introduction: The estimation of ethnic origin from biological derivatives obtained at crime scenes could be usefully used to solve the crime efficiently by reducing the scope of the suspect related to the case. STR (short tandem repeats) markers used for individual identification in the forensic field are not related to human phenotypes, the information available is limited, and each country has a strict database management, making it difficult to use data from other countries. It is expected that there will be an unsatisfactory accuracy in obtaining useful information from paternal or maternal ethnic prediction using Y chromosome and mitochondrial DNA in the situation of increasing international marriage. In this study, the utility of prediction model using autosomal single nucleotide polymorphism marker for Asians were constructed and the factors that can predict the Asian population in detail were examined.

Methods: DNA samples from 750 Asians and the autosomal single nucleotide polymorphism (SNP) analysis was performed. For SNP analysis, commercialized panels using massively parallel sequencing devices were used. This panel was developed for ethnicity estimation that includes 165 SNP markers known to have

significant differences between ethnic groups. Based on the obtained experimental results, a statistical method was applied to develop an ethnic prediction model.

Results: Nine statistically significant SNP markers were selected to construct an ethnicity prediction model for predicting the Asian population to one of Northeast Asia, Southeast Asia, and Southwest Asia. The misclassification rate of Northeast / Southeast Asia prediction model was 0.277, misclassification rate of Northeast / Southwest Asia model was 0.015, and misclassification rate of Southeast / Southwest Asia model was 0.027 was confirmed.

Conclusions: The model for ethnicity prediction using autosomal SNP was established in this Asian group divided into Northeast / Southeast / Southwest Asia. This model was constructed to be predicted step by step by dividing it into two ethnic groups. Based on the results of this study, it is expected that the forensic application would be expanded by reflecting the data of the various Asian ethnic groups not included in this study

---

Keywords: forensic science, massively parallel sequencing, ethnicity prediction model, SNP, Asian

Student number: 2013 - 23510