F 통계량을 이용한 집단 계통관계 검정의 기초

변다현[1], 정충원[1,2*]

요약: 방대한 양의 유전체 염기서열 변이 자료를 이용하면 기존 계통학 방법론을 통해 검정하기 어려웠던 종내 집단 간 계통관계 연구가 가능하다. F 통계량은 유전체 염기서열 변이의 집단별 대립유전자 빈도 정보를 이용하여 주어진 소수의 집단들 사이의 관계를 설명하는 최적의 계통수를 찾고, 단순 계통수로 집단 간 관계를 설명하지 못하는 경우인 집단 혼합을 엄밀히 탐지할 수 있는 유용한 방법론으로 최근 집단유전학에서 널리 이용되고 있다. 본 논문에서는 F 통계량의 원리를 설명하고, F_2, F_3, F_4 통계량의 구체적인 계산법과, 이를 통해 집단 간 유전자 흐름과 혼합을 탐지하는 방법을 제시한다.

키워드: 소진화, 집단유전체학, 집단계통수, 유전자 흐름, 집단 혼합

[1]*School of Biological Sciences, Seoul National University, Seoul, Republic of Korea 08826*

[2] *Institute for Data Innovation in Science, Seoul National University, Seoul, Republic of Korea 08826*

*Corresponding author: cwjeong@snu.ac.kr

서론

생물 분류군 사이의 유연관계를 계통수(phylogenetic tree)의 형태로 추론하는 것은 진화생물학의 핵심 연구 분야이다. 찰스 다윈이 “변이를 수반한 상속(descent with modification)”이라는 개념으로 명료하게 정리한 진화의 기제를 고려할 때, 생물 분류군 사이의 유연관계는 반복적인 분기를 통한 종분화의 양상을 반영하여 계층적으로 나타날 수밖에 없기 때문에, 계통수는 생물 분류군 사이의 유연관계를 표현하는 좋은 모형으로 널리 받아들여지고 있다. 계통분류학에서는 전통적으로 계통수 추론의 대상으로 생식적 격리가 뚜렷한 종 단위 이상의 생물 분류군들에 집중하여 왔고, 이를 통해 종과 그 상위 분류군의 진화에 해당하는 대진화(macro-evolution)의 양상을 상세히 규명해 왔다. 특히, 최근 약 20년 사이에 유전체학 연구방법론이 급격히 발달함에 따라 대량의 유전체 염기서열 변이를 분석하여 종간 계통수를 정밀히 추정하는 계통유전체학(phylogenomics)이 대두되었다. 이러한 연구는 궁극적으로는 모든 생물을 아우르는 계통수인 “생명의 나무(tree of life)”(Maddison et al. 2007; Hug et al. 2016; The Darwin Tree of Life Project Consortium et al. 2022)를 재구성하고자 하는 야심찬 계획으로 진행되고 있다.

종 사이의 유연관계와 마찬가지로 한 종에 속하는 개체들 사이의 유전적 관계 역시 다양한 수준으로 계층적인 양상을 나타내는 것이 일반적이다. 한 종의 모든 개체들이 동일한 확률로 서로 교배하는 무작위 교배(random mating)는 현실적으로 일어나기 매우 어려운데, 무작위 교배가 일어나지 않는 경우 개체들 사이의 유전적 거리에 차이가 발생하기 때문이다. 한 예로, 거리에 따른

생식적 격리(isolation by distance)는 유한한 집단에서 일어나는 방향성 없는 중립 진화인 유전적 부동(genetic drift)과 결합하여 한 종 안에서 유전적으로 구분되는 소집단을 생성하는 것이 일반적이다(Irwin et al. 2008; Kuchta et al. 2009). 상이한 환경에 놓인 소집단들이 국지 환경에 대한 자연선택에 따라 유전적으로 달라지는 것 역시 한 종 안에서 계층적인 유전적 관계를 갖는 집단들을 생성하는 대표적 기제이다(Kawecki and Ebert 2004; Hendry et al. 2009). 한 종의 유전적 조성의 시간에 따른 변화, 즉 소진화(microevolution)를 연구하는 분야인 집단유전학은 종 내 집단들의 계층적 관계인 집단 구조(population structure)를 주요 연구 주제로 삼고 있다. 집단 구조가 매우 뚜렷한 경우에는 종 사이의 유연관계를 분석하는 것과 마찬가지로 계통수가 집단 간 관계를 설명하는 좋은 모형일 수 있기 때문에 미토콘드리아 *COI* 유전자 등의 염기서열에 대한 유전자 계통수 추론을 통한 연구가 꾸준히 수행되어 왔다(Richards et al. 2000; Torroni et al. 2006). 하지만 많은 경우 집단 간 분화 정도가 약하거나 집단 혼합(admixture)이 광범위하게 일어나기 때문에, 단일 유전자 계통수에 기반하여 집단 간 유연관계를 추론하는 것은 뚜렷한 한계를 지니고 있다(Reich et al. 2010; Toews and Brelsford 2012).

계통유전체학의 경우와 마찬가지로 집단유전학에서도 유전체학 기술의 발전을 적극 수용하여 대량의 유전체 염기서열 변이 자료를 이용하여 집단 간 유연관계를 추론하는 방법론이 급격히 발전해 왔다. 하지만 계통유전체학이 상동유전자(orthologous gene)와 같은 자연스러운 유전자 계통수 분석 단위를 갖는데 비해 집단유전체학은 지속적으로 일어나는 유전자 재조합(recombination) 때문에 유전체를 분절하여 분절 단위별로 유전자 계통수 분석을 하는 직관적인 접근이 불가능하다. 재조합하는 유전체의 전장유전체 변이 자료를 분석하여 집단의 유연관계를 추론하는 대표적인 방법으로는 주성분분석(principal component analysis; PCA)(Patterson et al. 2006)과 STRUCTURE(Pritchard et al. 2000), ADMIXTURE(Alexander et al. 2009) 등과 같은 군집분석이 있다. 해당 방법론들은 개체를 분석 단위로 하여 개체별 유전자 프로필을 적은 수의 변수로 요약하는 매우 유용한 방법들이지만 이 방법들로 확인한 유전적으로 동질적인 소집단 혹은 연속적인 유전자 경사(genetic cline)를 이를 설명하는 계층적 유연관계 모형으로 전환하는 것은 복잡한 작업이다. 상이한 집단 계통수 혹은 계통수에 유전자 흐름(gene flow)이 추가된 집단 그래프(population graph)가 주성분분석과 군집분석에서 동일한 양상으로 나타날 수 있기 때문이다(Li et al. 2008; Novembre et al. 2008; HUGO Pan-Asian SNP Consortium et al. 2009; Lazaridis et al. 2014; Raghavan et al. 2014; Lazaridis et al. 2016). 또한 집단별 개체수 및 집단의 유전적 고립 정도 등 집단 계통수의 형태에 영향을 주지 않는 요인들에 의해 주성분분석과 군집분석 결과가 크게 달라질 수도 있다. 따라서 2000년대 말에는 주성분분석과 군집분석을 통해 도출한 집단의 계통 가설을 정량적으로 검정할 수 있는 집단유전체학 방법론의 필요성이 절실히 대두되었다.

F 통계량으로 불리는 일군의 방법론은 이러한 필요성에 기반하여 개발되었고, 소수의 집단 사이의 유연관계를 주어진 계통수(들)로 충분히 설명할 수 있는지 아니면 집단 혼합이 반드시 필요한지 검정하는 통계 방법론이다. F 통계량은 분석 대상 집단들이 최근 공통조상(Most recent common ancestor; MRCA)에게서 물려받은 조상 다형성 변이(ancestral polymorphism)의 집단별 빈도를 이용하여 계산되며, 개체를 분석 대상으로 하는 주성분분석과 군집분석 등의 기존 방법론과 달리 집단을 분석 대상으로 하며 개체수나 집단의 고립 정도에 영향받지 않는 특성을 갖고 있어 개발 이후 광범위하게 사용되었다. F 통계량은 Reich et al. 2009, Green et al. 2010, Durand et al. 2011, Moorjani et al. 2011에 의해 그 개념들이 소개되었고, 이후 2012년에 응용 수학자 닉 패터슨(Nick

Patterson)이 ADMIXTOOLS라는 프로그램 출시와 함께 F 통계량의 개념을 총괄적으로 정리하여 출판하였다(Patterson et al. 2012).

F 통계량으로 집단의 혼합과 유전자 흐름을 검정할 수 있는 원리를 이해하기 위해서는 기무라 모토(Kimura Motoo)의 중립진화이론(neutral theory of molecular evolution)(Kimura 1968; Kimura 1983)에 대한 이해가 선행되어야 한다. 중립진화이론이란 생물체의 유전적 변이 대부분은 적응도(fitness)에 큰 영향을 미치지 않는 중립적 변이이며, 이는 자연선택에 의해 제거되거나 선택되지 않고, 대신 유전적 부동에 의해 무작위적으로 고정되거나 소실된다는 이론이다. 해당 이론은 유전자 변이의 주요 동력이 자연선택이 아닌 유전적 부동이라는 점을 강조하며, 생물의 진화가 선택압 없이도 확률적 사건에 의해 진행될 수 있다는 것을 제시한다. 이러한 점에서 중립진화이론은 집단유전학과 진화생물학의 많은 연구에서 유전자 변이의 분포와 빈도를 이해하고 계통수를 재구성하는 데 중요한 이론적 기반을 제공한다.

F 통계량은 위와 같은 중립진화이론을 바탕으로 자연선택의 영향 없이 유전적 부동에 의한 유전적 변이의 무작위적 변화만을 고려하기 때문에 계통수의 가지별로 일어나는 대립유전자 빈도의 변화가 서로 독립이라는 점을 이용한다. 따라서 제안된 계통수 상에서 서로 독립인 두 가지에서 일어난 대립유전자 빈도 변화가 실제 자료 상에서 상관성을 갖는다면 제안된 계통수가 자료에 부합하지 않음을 알 수 있는 것이다. 보다 구체적인 내용은 이하 본문에서 자세히 설명하도록 하겠다.

본 논문에서는 집단유전학에서 중요하게 사용되는 F 통계량의 계산법과 원리에 대해 설명하고 이어서 F 통계량이 어떻게 집단 간의 관계를 주어진 유전적 자료를 기반으로 이와 부합하게 혼합을 고려하여 모델링할 수 있게 해주는지에 대해 소개하겠다. F_2, F_3, F_4 통계량에 대해 각각 설명하고 F 통계량의 사용 방법에 대해서도 정리하였다.

본 논문의 전문에 걸쳐 사용하는 예시 집단의 변이들은 별다른 언급이 없는 이상 두 개의 대립유전자를 갖는 단일염기다형성(bi-allelic SNP; single nucleotide polymorphism)을 고려할 것이며, 이 두 대립유전자를 임의로 0과 1으로 지칭하도록 하겠다. 또한, 공통조상에게서 물려받은 조상 다형성에 적용된다는 가정을 감안하여 반복 돌연변이(recurrent mutation)와 역돌연변이(back mutation)는 무시한다.

F_2 통계량 (F_2 Statistics)

F_2 통계량은 두 집단 사이의 유전적 거리를 나타내는 값으로, 두 집단 사이의 대립유전자 빈도 차이를 제곱한 것의 기댓값으로 정의한다. 두 집단의 대립유전자 빈도 차이, 즉, 유전적 차이를 나타내는 F_2 통계량이 클수록 두 집단의 유전적 거리가 멀다고 할 수 있으며, F_2 통계량이 작을수록 두 집단의 유전적 거리가 가깝다고 이야기할 수 있다.

집단 A와 B에 대한 대립유전자 빈도를 각각 a, b라고 나타낼 때 F_2 통계량은 아래 수식 (1)과 같이 간단히 정의한다. 이하 서술하는 모든 기댓값들은 현대 집단의 공통조상 대립유전자 빈도가 주어졌을 때 각 집단의 기댓값(조건부 기댓값)을 의미하는 것으로 통일한다[1].

$$\mathrm{F_2(A,B)} \;=\; \mathrm{E}[(a-b)^2] \qquad (1)$$

[1] 예를 들어 $\mathrm{E}(a)$은 조상집단 R의 대립유전자 빈도가 r로 주어졌을 때 a의 기댓값, $\mathrm{E}(a|r)$으로 정의한다.

보다 구체적으로는 M개의 SNP에 대하여 F_2 통계량을 정의한다고 할 때, 두 집단 A와 B에서 j번째 SNP의 대립유전자 빈도가 각각 a_j, b_j라면 F_2 통계량을 아래 수식 (2)와 같이 정의한다.

$$F_2(A,B) = \frac{1}{M}\sum_{j=1}^{M}(a_j - b_j)^2 = \frac{1}{M}\sum_{j=1}^{M}(a_j^2 - 2a_jb_j + b_j^2) \quad (2)$$

두 집단 A, B에서 한 대립유전자의 빈도가 a, b로 주어진다면, 나머지 대립유전자의 빈도는 1-a, 1-b로 주어지기 때문에 어떤 대립유전자를 사용하여 정의하여도 F_2 통계량 값은 동일하다.

$$F_2(A,B) = E[((1-a)-(1-b))^2] = E[(b-a)^2] = E[(a-b)^2] \quad (3)$$

F_2 통계량은 공통조상 집단에서 분기한 후 두 집단 사이에 발생한 대립유전자 빈도 변화의 크기, 즉 대립유전자 빈도 변화의 분산을 나타내는 값으로, 계통수 해석의 관점에서 보자면 두 집단 사이의 유전적 부동 가지 길이(drift branch length)를 나타내는 개념이라고 정리할 수 있다.

F_2 통계량은 집단유전학에서 오랫동안 널리 사용되어 온 Wright's F 통계량 혹은 고정계수(fixation indice)(Wright 1922; Wright 1951; Cavalli-Sforza and Edwards 1967)와 밀접한 관련을 갖는 값이다. F_{ST}는 집단 간 대립유전자 빈도 차이에 의해 나타나는 이형접합도의 이론적 감소분의 비율을 나타내는 값으로, 두 집단이 주어졌을 때 아래와 같이 정의된다. 두 집단 A, B의 대립유전자 빈도를 각각 a, b라 할 때,

전체 집단의 대립유전자 빈도: $\bar{p} = \frac{a+b}{2}$

대립유전자 빈도 차이를 감안하지 않을 때 기대되는 이형접합도 값: $H_T = 2\bar{p}(1-\bar{p})$

대립유전자 빈도 차이를 감안할 때 기대되는 이형접합도 값: $H_S = \frac{1}{2}\{2a(1-a) + 2b(1-b)\}$

집단 간 대립유전자 빈도 차이에 의한 이형접합도의 이론적 감소분:

$$H_T - H_S = (a+b)\times(1-\frac{a+b}{2}) - a(1-a) - b(1-b) = \frac{(a-b)^2}{2} = \frac{F_2(A,B)}{2}$$

이형접합도 이론적 감소분의 비율로 정의한 F_{ST}: $F_{ST} = \frac{H_T - H_S}{H_T} = \frac{F_2(A,B)}{4\bar{p}(1-\bar{p})}$

따라서 F_{ST}는 F_2 통계량을 집단 간 대립유전자 빈도 분화가 일어나지 않았을 때 기대되는 이론적인 이형접합도 값($2\bar{p}(1-\bar{p})$)의 2배로 표준화한 통계량이다.

F_2 통계량이 F_{ST}와 구별되는 유용한 성질로는 F_2 통계량은 오른쪽 그림 1의 예시와 같이 다른 F_2 통계량의 합으로 나타낼 수 있다는 점이 있다.

$$F_2(A,C) = F_2(A,B) + F_2(B,C) \quad (4)$$

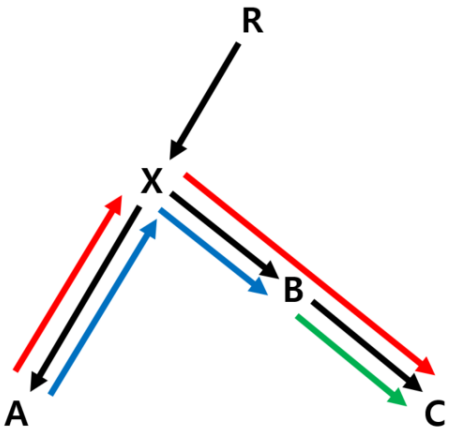

그림 1. F_2 통계량의 가산성(Additivity of F_2)

이를 F_2 통계량의 가산성(Additivity of F_2)이라고 한다(Patterson et al. 2012). 이와 같은 F 통계량의 가선성을 이해하기 위해서는 몇 가지 수학적 개념에 대한 이해가 필요하다. 본 논문에서는 공분산(covariance), 통계적 독립성(statistical independence), 비상관(uncorrelated), 직교성(orthogonality), 조건부 기댓값(conditional expectation)의 정의와 마팅게일(martingale) 속성에 대해서만 가볍게 설명하도록 하겠다.

공분산은 두 확률변수의 선형 관계를 나타내는 값으로, 한 변수의 값이 다른 변수 값에 대해 선형적으로 예측되는지 여부를 나타내며, 아래와 같은 식으로 나타낼 수 있다.

$$\mathrm{Cov(X,Y)} = E(XY) - E(X)E(Y)$$

통계적 독립성이란 두 확률 변수나 사건이 상호 간에 영향을 미치지 않는 관계를 의미한다. 확률 변수 X와 Y가 있다고 하였을 때, 확률 변수 X, Y가 통계적 독립이라면 아래 식이 필요충분조건으로 성립한다.

$$\mathrm{P(X = x \text{ and } Y = y)} = P(X = x)P(Y = y)$$

여기서 P(.)는 확률을 의미한다.
비상관이란 두 확률 변수 간의 선형 관계가 없음을 의미한다. 즉, 공분산 값이 0이라는 의미이므로 비상관 관계의 두 확률 변수 X, Y에 대해서 아래 식이 필요충분조건으로 성립한다.

$$\mathrm{Cov(X,Y)} = E(XY) - E(X)E(Y) = 0$$
$$\therefore E(XY) = E(X)E(Y)$$

두 변수가 **통계적 독립**이라는 것은 두 변수 사이에 어떠한 형태의 관계도 없음을 의미하는 반면, 두 변수가 비상관이라는 것은 두 변수 사이에 선형 관계가 없음을 의미한다. 따라서 통계적 독립인 두 변수는 항상 비상관이지만[2], 두 변수가 비상관이라고 하여 반드시 통계적 독립인 것은 아니다.
두 확률변수 X, Y가 **직교성**을 갖는다는 것은 이들의 곱의 기댓값이 0이라는 것을 의미한다. 이를 식으로 나타내면 아래와 같다.

$$\mathrm{E(XY)} = 0$$

조건부 기댓값이란 아래와 같이 표기하며, 다른 확률 변수 Y가 주어졌을 때 X의 기댓값을 나타낸다.

$$\mathrm{E[X|Y]}$$

마팅게일은 시간에 따라 변하는 확률 변수가 연속적으로 나열된 일련의 확률 과정으로, 현재

[2] 증명: $\mathrm{E(XY)} = \int_x \int_y xyp(xy)dxdy = \int_x \int_y xyp(x)p(y)dxdy = \int_x xp(x)dx \int_y yp(y)dy = E(X)E(Y)$

시점에서 다음 시점 값의 기댓값을 구할 때 조건부 기댓값을 사용하게 되는데, 마팅게일은 확률변수가 과거에 어떤 값을 가졌는지에 관계없이 미래 시점 값의 기댓값이 현재 값과 동일하다는 속성을 가진다.

$$\mathrm{E}[\mathrm{X_t}|X_0, X_1, X_2, \dots, X_s] = X_s \ \ (s \le t)$$

이제 다시 F_2 통계량의 가산성으로 돌아와서 해당 수식을 위의 수학적 개념들을 바탕으로 증명해 보도록 하겠다. 좌변을 F_2 통계량의 정의에 따라 전개하면 수식 (5)와 같다.

$$\mathrm{F_2}(A,C) = E[(a-c)^2] = E[(a-b+b-c)^2] = E[(a-b)^2] + E[(b-c)^2] + 2E[(a-b)(b-c)] \quad (5)$$

수식 (5)에서 3번째 등식을 4번째 등식과 같이 전개할 수 있는 이유는 기댓값의 선형성(linearity) 때문이다.

위 수식 (5)에서 4번째 등식의 마지막 항은 중립진화이론의 가정으로 인해 계통수 상의 독립적 가지에서 일어난 대립유전자 빈도 변화는 통계적 독립성을 가지기 때문에 $(a-b)$ 확률 변수와 $(b-c)$ 확률변수는 통계적 독립이자 비상관이다. 또한 중립진화이론의 방향성 없는 유전적 부동의 특성에 의해 대립유전자 빈도가 마팅게일 속성을 가지므로 수식 (5)의 4번째 등식의 마지막 항에 대해 아래 수식 (6)이 성립한다.

$$E[(a-b)(b-c)] = E[(a-b)]E[(b-c)] = [E(a)-E(b)][E(b)-E(c)] = 0 \quad (6)$$

따라서 아래 수식 (7)과 같이 F_2 통계량의 가산성이 성립할 수 있게 되는 것이다.

$$\mathrm{F_2}(A,C) = \ E[(a-b)^2] + E[(b-c)^2] = F_2(A,B) + \ F_2(B,C) \quad (7)$$

이를 직관적으로 이해하는 방법은 그림 1과 같이 유전적 부동의 경로를 계통수의 가지 길이로 가시적으로 생각하는 것이다. 그림 1에서와 같이 집단 A, C 사이의 유전적 부동(빨간색 경로)은 A, B 사이의 유전적 부동(파란색 경로)과 B, C 사이의 유전적 부동(초록색 경로)의 합으로 계산될 수 있음을 직관적으로 생각할 수 있다.

위에서 설명한 수학적 개념들을 적용하여 F_2 통계량의 가산성에 대해 증명한 일련의 과정은 이후 F_3, F_4 통계량을 이해하는 데에도 굉장히 중요한 비계로 작용하므로 잘 이해하고 넘어가는 것이 중요하다.

지금까지 소개한 F_2 통계량은 집단의 대립유전자 빈도를 알고 있다는 가정 하에 정의되었다. 실제 상황에서는 집단의 대립유전자 빈도를 각 집단에 속한 일정 수의 개체들의 유전자형을 관찰함으로써 추정하기 때문에, 이를 이용하여 실제 F_2 통계량과 일치하는 추정치, 즉 불편추정치(unbiased estimate)를 구성하는 것이 중요하다. 지금부터 집단의 대립유전자 빈도를 통해 정의한 F 통계량을 대문자 F로, 실제 자료에서 추정한 집단별 대립유전자 빈도를 이용해 구성한 통계량을 소문자 f로 나타낼 것이다. 두 집단 A와 B에서 j번째 SNP에 대하여 1 대립유전자를 각각 x_j, y_j개, 전체 대립유전자를 각각 n_{Aj}, n_{Bj}개 관찰하였다고 할 때, 집단별 대립유전자 빈도 a_j, b_j의

불편추정치는 각각 $a_j' = \frac{x_j}{n_{Aj}}$ 및 $b_j' = \frac{y_j}{n_{Bj}}$로 주어진다. 이를 그대로 대입한 값의 기댓값을 구해보면 수식 (8)과 같다.

$$\mathrm{E}\left(\left(a_j' - b_j'\right)^2\right) = \mathrm{E}\left(\left(\frac{x_j}{n_{Aj}} - \frac{y_j}{n_{Bj}}\right)^2\right) = E\left(\frac{x_j^2}{n_{Aj}^2} - \frac{2x_j y_j}{n_{Aj} n_{Bj}} + \frac{y_j^2}{n_{Bj}^2}\right) = \frac{E(x_j^2)}{n_{Aj}^2} - \frac{2E(x_j)E(y_j)}{n_{Aj} n_{Bj}} + \frac{E(y_j^2)}{n_{Bj}^2} \quad (8)$$

여기에서 x_j, y_j가 각각 $\mathrm{Binom}(n_{Aj}, a_j)$, $\mathrm{Binom}(n_{Bj}, b_j)$와 같은 이항분포를 따름을 이용하면,

$$\mathrm{E}(x_j) = n_{Aj} a_j$$
$$\mathrm{E}(y_j) = n_{Bj} b_j$$
$$\mathrm{E}(x_j^2) = Var(x_j) + (E(x_j))^2 = n_{Aj} a_j (1 - a_j) + (n_{Aj} a_j)^2$$
$$\mathrm{E}(y_j^2) = Var(y_j) + (E(y_j))^2 = n_{Bj} b_j (1 - b_j) + (n_{Bj} b_j)^2$$

이를 대입하여 정리하면,

$$\mathrm{E}\left(\left(a_j' - b_j'\right)^2\right) = a_j^2 + \frac{a_j(1 - a_j)}{n_{Aj}} - 2a_j b_j + b_j^2 + \frac{b_j(1 - b_j)}{n_{Bj}} = F_2(A, B) + \frac{a_j(1 - a_j)}{n_{Aj}} + \frac{b_j(1 - b_j)}{n_{Bj}} \quad (9)$$

따라서, 대립유전자 빈도의 제곱 값에 의해 발생하는 편향을 제거하기 위하여 F_2 통계량의 불편추정치 f_2는 아래 수식 (10)과 같이 정의된다.

$$f_2(A, B) = \frac{1}{M}\sum_{j=1}^{M}\left\{\left(a_j' - b_j'\right)^2 - \frac{x_j(n_{Aj} - x_j)}{{n_{Aj}}^2(n_{Aj} - 1)} - \frac{y_j(n_{Bj} - y_j)}{{n_{Bj}}^2(n_{Bj} - 1)}\right\} \quad (10)$$

F_3 통계량 (F_3 Statistics)

F_3 통계량은 다른 말로 세 집단 검정(three-population test)이라고도 불리는데, 이는 세 개의 집단의 관계를 검정하는 방법론이다. F_3 통계량은 특히 두 가지 목적으로 사용되는데, 첫 번째는 검정 대상인 타겟 집단 C와 두 참조 집단 A, B 사이의 관계를 충분히 설명하는 단순 계통수를 찾을 수 있는지 아니면 타겟 집단인 C가 집단 혼합을 겪었는지 검정하는 것이고, 두 번째는 타겟 집단 C를 참조 집단 A, B 공통의 외군으로 설정하여 A와 B가 공유하는 가지 길이, 즉 유전적 유사성을 측정하는 것이다. 전자를 혼합 F_3 통계량(admixture F_3 statistic)이라 하고, 후자를 외군 F_3 통계량(outgroup F_3 statistic)이라고 한다.

집단 A, B, C의 대립유전자 빈도를 각각 a, b, c라고 할 때 F_3 통계량은 아래 수식 (11)과 같이 간단히 정의된다.

$$\mathrm{F_3}(C; A, B) = \mathrm{F_3}(A, B; C) = E[(c - a)(c - b)] \quad (11)$$

앞서 설명한 F_2 통계량의 표기법과는 달리 세미콜론(;)이 추가된 것을 볼 수 있는데, 세미콜론을 기준으로 낱개의 집단이 위치한 쪽에서 두 개의 집단이 위치한 쪽으로 분배법칙 연산을 하듯이

계산한다고 이해하면 편하다[3].

위의 F_3 통계량 식을 전개하면 F_2 통계량의 선형결합으로 나타낼 수 있음을 확인할 수 있다.

$$\begin{aligned} F_3(C;A,B) &= E[(c-a)(c-b)] = E[c^2 - cb - ac + ab] \\ &= \frac{1}{2}E[(c-a)^2 + (c-b)^2 - (a-b)^2] \\ &= \frac{1}{2}[F_2(C,A) + F_2(C,B) - F_2(A,B)] \end{aligned} \quad (12)$$

보다 직관적으로 F_3 통계량을 이해하는 방법은 F_2 통계량의 그림 1과 유사한 방식으로 F 통계량 계산에 포함된 집단들의 계통수 상에서 가능한 모든 유전적 부동 경로를 그려본 뒤, 그중에서 겹치는 경로만을 추출하는 것이다. 이때 경로의 진행 방향에 대해서도 고려하여야 한다. 구체적인 예시는 아래 혼합 F_3 통계량과 외군 F_3 통계량 설명에서 자세히 다루도록 하겠다.

F_2 통계량에서와 동일하게 F_3 통계량은 타겟 집단 C의 대립유전자 빈도 제곱항을 포함하므로, 실제 자료를 이용하여 계산한 불편추정량 f_3는 아래 수식 (13)과 같이 정의된다.

$$f_3(C;A,B) = \frac{1}{M}\sum_{j=1}^{M}\{(c_j' - a_j')(c_j' - b_j') - \frac{z_j(n_{Cj} - z_j)}{{n_{Cj}}^2(n_{Cj} - 1)}\} \quad (13)$$

여기에서 z_j와 n_{Cj}는 각각 집단 C의 j번째 SNP에서 관찰한 1 대립유전자 및 전체 대립유전자의 개수를 나타내며, 집단 C의 대립유전자 빈도의 불편추정치인 $c_j' = \frac{z_j}{n_{Cj}}$로 정의한다.

혼합 F_3 통계량 (Admixture F_3 Statistics)

그림 2A는 세 집단 사이에 혼합이 없는 단순한 계통수를 나타낸 것이고, 그림 2B는 집단 A, B의 자매군인 G, H가 α : β (β=1-α) 의 비율로 혼합하여 혼합 조상집단 I를 형성하고, I가 유전적 부동을 겪어 현대 혼합집단 C가 형성되었음을 나타내는 집단 계통 그래프이다. 그림 2A와 같이 혼합이 없는 계통수에서 F 통계량을 계산할 때는 계통수 상에서 가능한 단일 경로로 유전적 부동 경로를 고려하면 되지만, 그림 2B와 같은 유전자 흐름이 있는 계통수의 경우 가능한 모든 경로를 고려함과 동시에 해당 경로로 유전적 부동을 겪었을 확률, 즉 혼합 비율을 고려하여 기여도를 계산에 포함하여야 한다.

[3] F_3 통계량 표기법에서 두개의 집단이 위치한 쪽의 집단 순서는 F 통계량 값에 영향을 미치지 않는다.

$$F_3(C;B,A) = E[(c-b)(c-a)] = E[(c-a)(c-b)] = F_3(C;A,B)$$

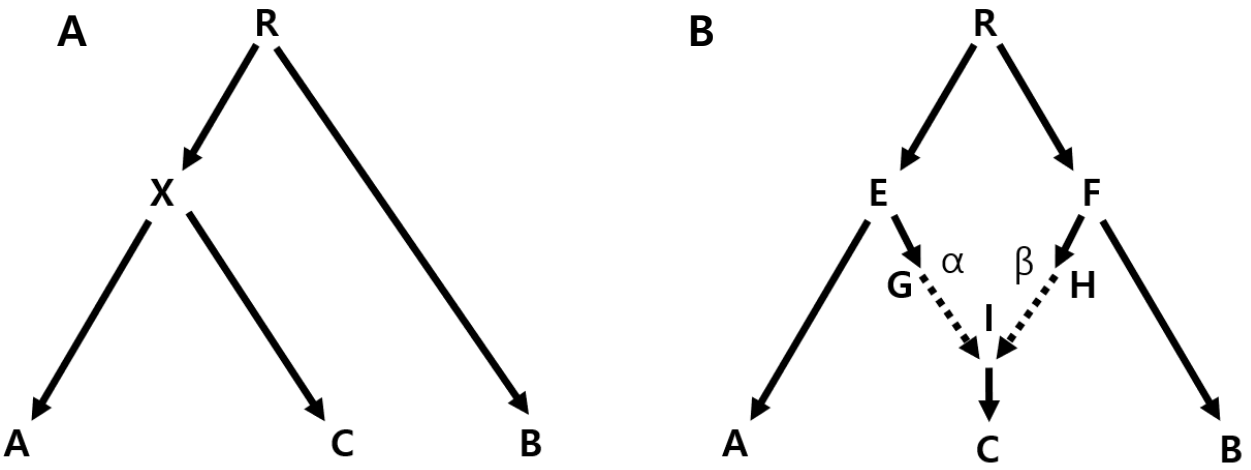

그림 2. 집단 계통수 표현 예시. (A) 혼합이 없는 단순 계통수. (B) 혼합이 포함된 집단 그래프

혼합이 없는 계통수에서 F_3 통계량을 계산하는 과정을 나타낸 그림 3A에서 볼 수 있듯이, 한 시작 집단에서 다른 끝 집단까지 연결하는 유전적 부동의 경로는 하나씩만 존재한다. 따라서 예시 그림 3A와 같이 $C \to A$, $C \to B$의 각각의 경로(파란색 경로와 초록색 경로)를 그린 뒤, 겹치는 경로만을 추출한 그림 3B의 경로(빨간색 경로), 즉, $(\mathbf{c}+\mathbf{d})$가 $F_3(C;A,B)$의 결과값이 되는 것이다.

$$F_3(C;A,B) = \mathbf{c}+\mathbf{d} \tag{14}$$

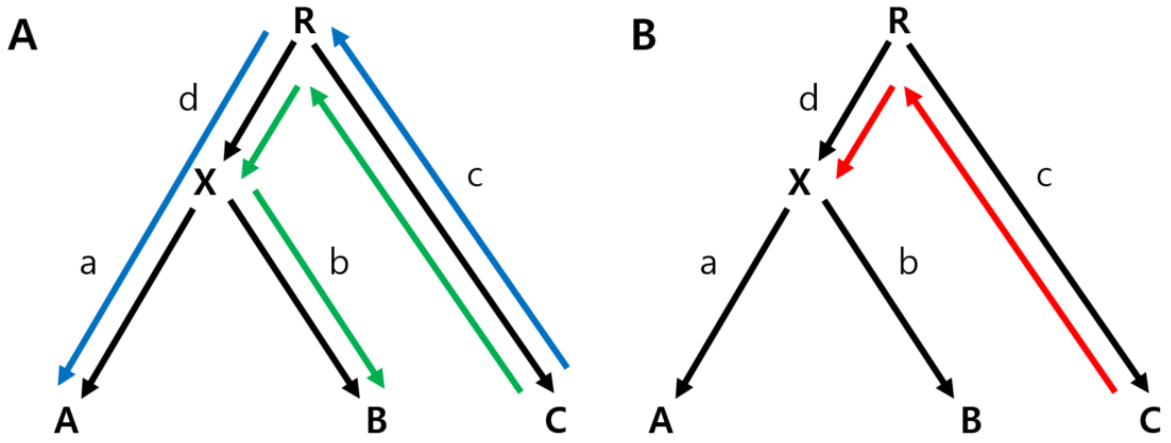

그림 3. 혼합이 없는 계통수에서 $\mathbf{F_3(C;A,B)}$를 계산하는 과정의 시각화. 파란색 경로와 초록색 경로가 겹치는 부분에서 신호가 생성된다. (A) $\mathbf{F_3(C;A,B)}$의 계산 과정. (B) $\mathbf{F_3(C;A,B)}$ 결과값의 시각화. a, b, c, d는 각각 인접한 집단 사이의 유전적 거리를 나타낸다. 예를 들어 $\mathbf{a} = \mathbf{F_2(A,X)}$가 된다.

반면에 혼합을 겪은 집단의 계통수를 나타내는 그림 2B의 경우에는 $C \to A$, $C \to B$ 경로가 여러 개씩 존재함을 알 수 있다. $C \to A$의 경우 그림 4의 파란색 화살표가 나타내는 경로와 같이 $C \to I \to G \to E \to A$와 $C \to I \to H \to F \to R \to E \to A$ 두 경로가 있으며, $C \to B$의 경우 그림 4의 초록색 화살표가 나타내는 경로와 같이 $C \to I \to H \to F \to B$ 와 $C \to I \to G \to E \to R \to F \to B$ 두 경로가 있다. 해당 경로들을 조합하는 방법은 그림 4와 같이 4가지가 나오는데, 해당 조합들에 대한 계산 값들을 각 조합 상단에 표시하였다.

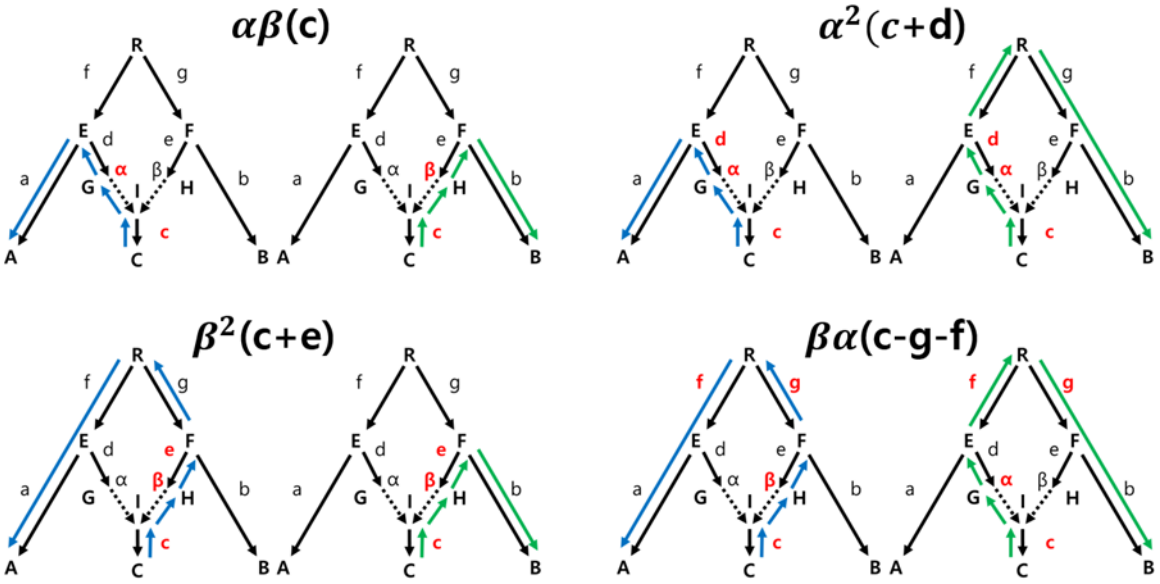

그림 4. 집단 관계 설명에 혼합을 필요로 하는 계통수에서 $\mathbf{F_3}(C;A,B)$ 계산의 시각화.

이 값들을 다 더한 값이 혼합을 포함한 계통수에서의 F_3 통계량이 된다.

$$\begin{aligned} \mathrm{F}_3(C;A,B) &= \alpha\beta(\mathbf{c}) + \beta^2(\mathbf{c}+\mathbf{e}) + \alpha^2(\mathbf{c}+\mathbf{d}) + \beta\alpha(\mathbf{c}-\mathbf{g}-\mathbf{f}) \\ &= \mathbf{c}(\alpha^2 + 2\alpha\beta + \beta^2) + \alpha^2\mathbf{d} + \beta^2\mathbf{e} - \alpha\beta(\mathbf{g}+\mathbf{f}) \\ &= \mathbf{c} + \alpha^2\mathbf{d} + \beta^2\mathbf{e} - \alpha\beta(\mathbf{g}+\mathbf{f}) \quad \because \alpha+\beta=1 \end{aligned} \tag{15}$$

위 수식 (15)를 보면 그림 3의 예시에서 보았던 수식 (14) $\mathbf{F_3}(C;A,B) = \mathbf{c}+\mathbf{d}$ 등식과 달리 4번째 등식의 마지막 항, $-\alpha\beta(\mathbf{g}+\mathbf{f})$이 음의 값을 갖는 것을 볼 수 있는데, 해당 항으로 인해 검정하고자 하는 집단들의 관계가 단순한 계통수로 나타내질 수 있는지, 아니면 혼합이 있는 계통수로 나타낼 수 있는지를 가늠할 수 있게 된다. 단순 계통수의 $\mathbf{F_3}(C;A,B)$은 반드시 양수 값이 나오나, 혼합이 있는 계통수의 $\mathbf{F_3}(C;A,B)$ 값은 음의 값이 나올 가능성이 있다는 것이다.

즉, 다음과 같은 명제가 성립한다.

참인 명제: "타겟 집단(C)가 혼합을 겪지 않았다면 F_3 통계량은 양수가 나온다."

이는 참인 명제이므로 다음과 같은 대우 명제도 성립하게 된다.

참인 명제의 대우: "F_3 통계량이 음수가 나왔다면 타겟 집단은 혼합을 겪었다."

특정 명제가 참이라고 하여 그 역과 이 또한 성립하는 것은 아니므로, F_3 통계량을 해석할 때 다음과 같은 거짓 명제를 참이라고 생각하지 않도록 유의해야 한다.

참인 명제의 역: "F_3 통계량이 양수가 나왔다면 타겟 집단은 혼합을 겪지 않았다."

참인 명제의 이: "타겟 집단이 혼합을 겪었다면 F_3 통계량은 음수가 나온다"

역과 이가 거짓 명제가 되는 이유는 혼합을 포함한 계통수의 $\mathbf{F_3}(C;A,B)$ 계산식인 수식 (15)에서 4번째 등식의 c항이 매우 큰 값을 갖는다면, 즉, C집단이 혼합을 겪은 후 해당 집단 특이적인 매우 강한 유전적 부동을 겪었다면, $-\alpha\beta(\mathbf{g}+\mathbf{f})$의 영향을 상쇄하여 4번째 등식이 양의 값을 가질 수 있기 때문이다. 실제 집단의 진화적 역사에서 유전적 부동을 겪은 경우 외에도 시퀀싱 과정에서의 플랫폼 편향(bias)이나, 분석 과정에서의 샘플링 오류(sampling error) 등과 같은 유전적 부동과 유사한 효과를 경험한 집단을 대상으로 한 경우에도 F_3 통계량 결과를 신뢰하기 어렵다는 점을 염두에

두어야 한다.

F_3 통계량 사용 시에 한 가지 더 고려해야 하는 사항은 F_3 통계량 값이 음수 값이 나왔다고 하더라도 타겟 집단 외의 나머지 두 집단이 타겟 집단의 혼합의 역사에 기여한 집단과 유연관계가 가깝다는 보장이 없다는 것이다. 아래 그림 5에는 이러한 상황의 예시로 그림 2B의 집단 그래프에 혼합에 관여하지 않은 외군 집단 D를 추가하여 보여준다. 그림 5의 계통수를 기반으로 D 집단을 포함한 다음 두 F_3 통계량을 계산하면 다음과 같은 결과가 나온다.

$$\mathrm{F}_3(C;D,A) = \mathbf{c} + \alpha^2\mathbf{d} + \beta^2(\mathbf{e}+\mathbf{g}) - \alpha\beta\mathbf{f} \quad (16)$$

$$\mathrm{F}_3(C;D,B) = \mathbf{c} + \alpha^2(\mathbf{d}+\mathbf{f}) + \beta^2\mathbf{e} - \alpha\beta\mathbf{g} \quad (17)$$

위의 $\mathrm{F}_3(C;D,A)$, $\mathrm{F}_3(C;D,B)$ 결과값은 $-\alpha\beta\mathbf{f}$와 $-\alpha\beta\mathbf{g}$ 값에 따라 음수가 나올 수도 있는 값이나 D 집단은 실제 C 집단의 혼합의 역사에는 직접적 기여를 하지 않은 집단이다. 따라서 F_3 통계량은 타겟 집단이 혼합으로 형성되었다는 여부만 검증할 수 있을 뿐, 타겟 집단 외의 두 집단이 실제 혼합의 역사에 기여하였음을 보장해주지는 않는다는 점을 염두에 두어야 한다.

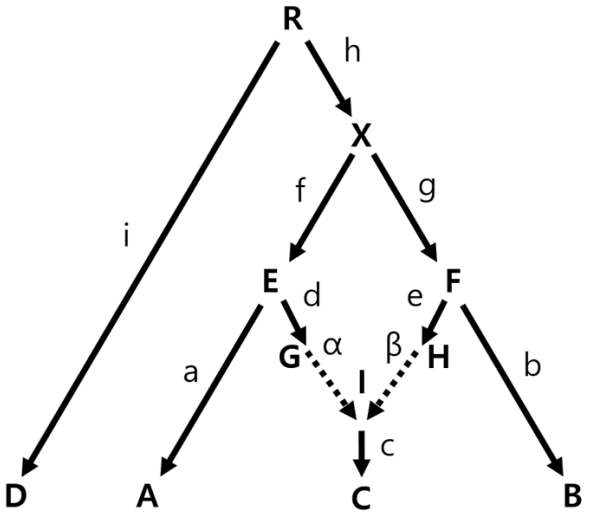

그림 5. 타겟 집단 C에 직접 기여하지 않은 계통을 나타내는 집단 D를 추가한 계통수

정리하자면 F_3 통계량이 양수 값이 나왔다고 하더라도 대상 집단들의 관계가 단순 계통수로 나타난다고 단정 지어서는 안되며, 음수 값이 나왔다고 하더라도 타겟 집단이 나머지 두 집단의 혼합으로 형성되었다고 생각하여서는 안 된다. 앞의 두 가지를 확인하기 위해서는 뒤에서 다룰 F_4 통계량 비교가 필요하다.

외군 F_3 통계량 (Outgroup F_3 Statistics)

외군 F_3 통계량의 계산 방법은 혼합 F_3 통계량과 동일하다. 차이점이라면 그림 3의 C집단에 나머지 집단 A, B와 분기한지 오래되어 두 집단과 혼합을 겪었을 가능성이 희박한 외군을 사용한다는 점이다. 외군 F_3 통계량은 그림 3의 빨간 경로로 표시한 바와 같이 C 집단에 외군을 넣어 C 집단과 나머지 두 집단 A와 B의 공통조상이 공유하는 유전적 부동을 수치화하여 두 집단 A, B의 유전적 거리를 추산한다. 외군 F_3 통계량의 값이 클수록 두 집단 A와 B가 서로 더 가까운 관계임을 의미한다.

F_4 통계량 (F_4 Statistics)

F_4 통계량은 F_3 통계량과 같이 집단의 혼합이 일어났는지에 대한 여부도 검정할 수 있을뿐더러, 더 나아가 유전자 흐름의 방향에 대한 정보도 제공하는 통계적 검정 방법이다. F_4 통계량은 네 집단 검정(four-population test)으로도 불리며, 분자가 동일하고 분모의 표준화 항만 달라 실질적으로 동일한 통계량인 Patterson's D 통계량 혹은 ABBA-BABA 검정으로도 불린다. 네 집단의 대립유전자 빈도를 통해 정의한 F_4 통계량은 아래 수식 (18)과 같다.

$$\mathrm{F}_4(A,B;C,D) = E[(a-b)(c-d)] \tag{18}$$

F_2 혹은 F_3 통계량과는 달리 F_4 통계량은 대립유전자 빈도의 제곱항을 포함하지 않기 때문에 F_4 통계량의 불편추정치 f_4는 단순히 집단별 대립유전자 빈도의 불편추정치로 집단별 대립유전자 빈도를 대체한 아래 수식 (19)와 같은 형태를 갖는다.

$$f_4(A,B;C,D) = \frac{1}{M}\sum_{j=1}^{M}\{(a_j'-b_j')(c_j'-d_j')\} \tag{19}$$

F_4 통계량의 표시법은 F_2, F_3 통계량의 표기법과 또 차이가 있는데, 세미콜론을 기준으로 양쪽에 포함되는 두 집단을 쌍으로 계산한다고 생각하면 편하다.

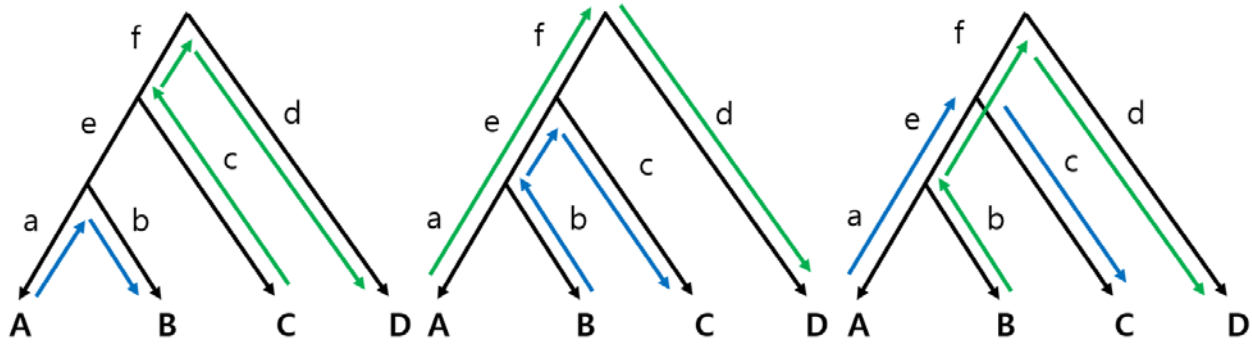

그림 6. 혼합이 없는 계통수에서 계산할 수 있는 F_4 통계량의 예시

그림 6에서 나타낸 것과 같이 혼합이 없는 계통수에서의 F_4 통계량을 계산하면 하나는 0, 나머지 두 개는 노드(node)와 노드를 연결하는 내부 가지(internal branch)의 길이의 값, e를 가진다[4].

[4] 그림 5의 계통수가 주어졌을 때 가능한 F_4 통계량의 개수를 A, B, C, D 4개의 집단을 순서를 고려하여 4개의 자리에 배치하는 순열로 생각하여 4!=24개라고 생각할 수 있다. 하지만 $\mathrm{F}_4(A,B;C,D) = \mathrm{F}_4(B,A;D,C) = \mathrm{F}_4(C,D;A,B) = \mathrm{F}_4(D,C;B,A)$라는 점을 고려하면 ${}_4C_2$=6개 조합만을 고려할 수 있게 되고, $\mathrm{F}_4(A,B;C,D) = -\mathrm{F}_4(A,B;D,C)$와 같이 세미콜론 뒤의 집단의 배치 순서는 F_4 통계량의 양수, 음수 여부에만 영향을 미친다는 점

F_4 통계량은 그림 6의 위의 $\mathrm{F}_4(A,B;C,D)=0$이 되는 점을 이용하여 주어진 집단 간의 관계에서 혼합의 여부와 유전자 흐름 방향에 대한 정보를 확인한다. 계통수에서 혼합이 없다면 A 집단과 B 집단 간의 대립유전자 빈도의 차이는 C 집단과 D 집단 간의 대립유전자 빈도 차이와 완전히 독립이므로 $\mathrm{F}_4(A,B;C,D)$ 값은 0이 되지만, A↔C, A↔D, B↔C, B↔D 등 표현식에서 세미콜론 좌우에 있는 두 집단 사이에 유전자 흐름이 있었다면 $\mathrm{F}_4(A,B;C,D)$ 값은 0이 아니게 되기 때문이다.
보다 구체적으로 보기 위해 F_4 통계량 기본식을 전개하면 아래와 같다.

$$\begin{aligned}\mathrm{F}_4(A,B;C,D) &= E[(a-b)(c-d)] = E[ac-ad-bc+bd] \\ &= \frac{1}{2}E[(a-d)^2+(b-c)^2-(a-c)^2-(b-d)^2] \\ &= \frac{1}{2}[F_2(A,D)+F_2(B,C)-F_2(A,C)-F_2(B,D)]\end{aligned} \quad (20)$$

위의 전개식을 기반으로 생각해보면, F_4 통계량 값이 유의미하게 음수가 나왔다는 말은 $F_2(A,D)$, $F_2(B,C)$의 값이 $F_2(A,C)$, $F_2(B,D)$ 값에 비해 유의미하게 작다는 이야기가 되고, 이는 B 집단과 C 집단 사이, 또는 A 집단과 D 집단 사이에서 유전적 흐름이 있어 두 집단 간의 유전적 거리가 다른 나머지 집단에 비해 가깝다는 것을 의미한다(그림 7A). 반대로 F_4 통계량 값이 유의미한 양수 값이 나온 경우 이는 $F_2(A,C)$, $F_2(B,D)$의 값이 $F_2(A,D)$, $F_2(B,C)$ 값에 비해 유의미하게 작다는 이야기가 되고, 즉, A 집단과 C 집단 사이 또는 B 집단과 D 집단 사이에서 유전자 흐름이 있었음을 의미한다(그림 7B).

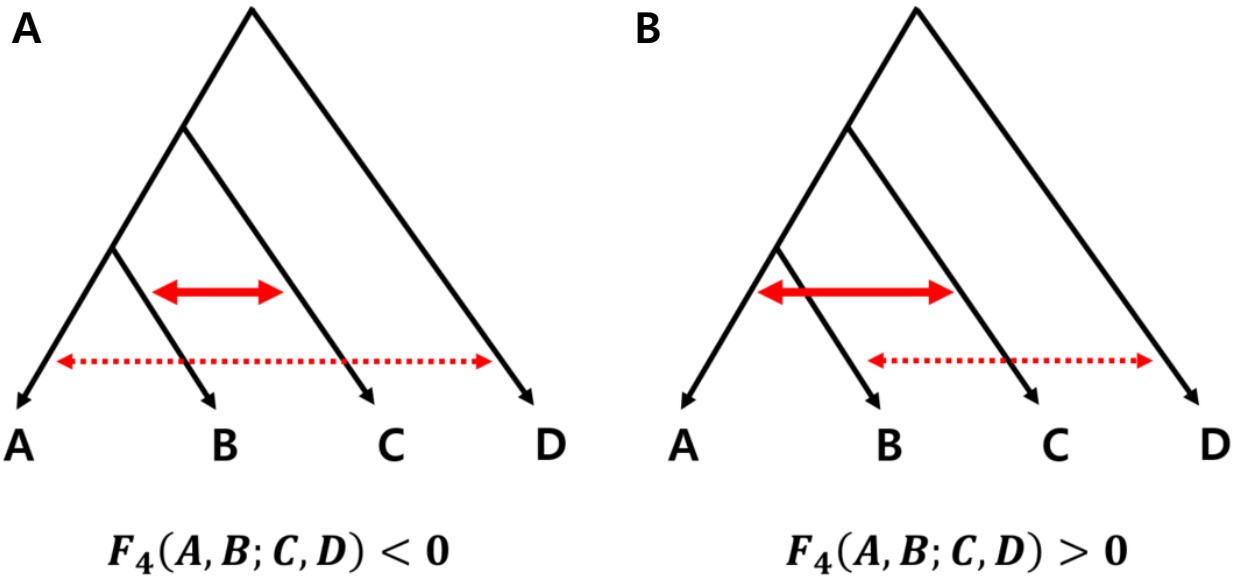

그림 7. 주어진 계통수에서 $\mathbf{F_4}(A,B;C,D)$ 값에 따른 집단 관계의 해석

해석의 가능성을 좁히기 위하여 보통 F_4 통계량을 계산할 때 나머지 세 집단의 공통의 외군에 해당하는 집단 하나를 포함시키는 것이 일반적이다. 이를테면 네 집단 중 D 집단이 A, B, C 세 집단의 명확한 외군이고, A, B, C는 D에 대하여 단계통군을 형성한다고 해 보자. 이 경우 A↔D 또는

을 고려하면, ${}_4C_2/2=3$개 조합만을 고려하면 된다. 따라서 위의 세 가지 F_4 통계량 값만 고려해도 무방하다.

B↔D 사이의 유전자 흐름이 존재하지 않기 때문에, F_4 통계량이 0에서 벗어날 경우 A↔C 사이의 유전자 흐름($f_4>0$) 혹은 B↔C 사이의 유전자 흐름($f_4<0$)이 있었음을 시사하게 된다.

외군인 D 집단을 제외했을 때, 세 집단 A, B, C 사이의 계통관계에 대한 사전 지식이 없는 경우, 한 F_4 통계량이 0에서 벗어난다고 하여 세 집단 사이의 관계를 단순 계통수로 설명할 수 없다고 단정해서는 안 된다. A, B, C 세 집단이 가질 수 있는 세 가지 가능한 계통수가 있기 때문이다. 따라서 이 세 가능성에 해당하는 3개의 F_4 통계량인 $F_4(A,B;C,D)$, $F_4(B,C;A,D)$, $F_4(A,C;B,D)$을 모두 확인하였을 때, 세 값이 모두 유의미하게 0에서 벗어난다면 세 집단 사이의 관계를 설명할 수 있는 단순한 계통수는 존재하지 않고, 집단 혼합이 필요함을 알 수 있다 (그림 8, 9).

집단 혼합의 신호를 탐지할 때 집단 혼합이 있었던 경우에도 세 가지 F_4 통계량이 반드시 0에서 유의미하게 벗어나지 않음을 주의해야 한다 (그림 9). 집단 C가 집단 A, B와 관련된 계통의 혼합이라고 할 때, $F_4(A,B;C,D) = \alpha\mathbf{g} - \beta\mathbf{h}$ 로 주어지는데, $\alpha\mathbf{g}$ 와 $\beta\mathbf{h}$ 의 크기가 비슷하여 0에서 유의미하게 벗어나지 않을 수 있기 때문이다. 이 경우 (((A,B),C),D)의 계통수가 자료를 충분히 설명할 수 있다는 잘못된 판단을 할 수 있는데, 나머지 두 통계량의 부호를 확인함으로써 이러한 잘못된 판단을 피할 수 있다. 즉, 집단 C가 A, B의 공통의 외군인 경우에는 C에 비해 A와 B가 서로 더 가깝기 때문에 나머지 두 통계량 $F_4(B,C;A,D)$과 $F_4(A,C;B,D)$가 모두 양수 값을 갖게 된다 (그림 8). 반면 집단 C가 A, B 계통의 혼합 집단인 경우 A와 B 사이의 유전적 거리보다 A와 C, B와 C 사이의 거리가 더 가깝기 때문에 나머지 두 통계량 $F_4(B,C;A,D)$과 $F_4(A,C;B,D)$가 모두 음수 값을 갖게 되기 때문에 두 경우를 명확히 구분할 수 있다 (그림 9).

따라서 위의 특성을 활용하여 $F_4(A,B;C,D)$, $F_4(B,C;A,D)$, $F_4(A,C;B,D)$ 결과값을 계산하고, 하나의 0값과 두 개의 내부 가지 값을 가지는지, 다른 값을 가지는지를 확인하여 대상 집단들의 혼합 여부와 유전자 흐름 방향의 정보를 얻을 수 있다.

	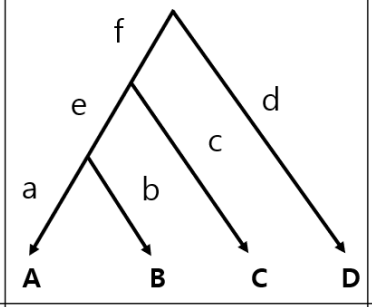	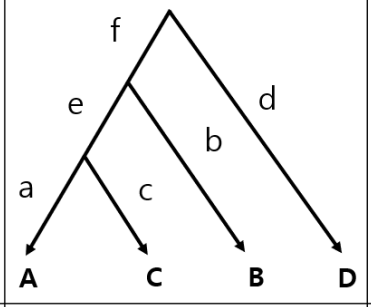	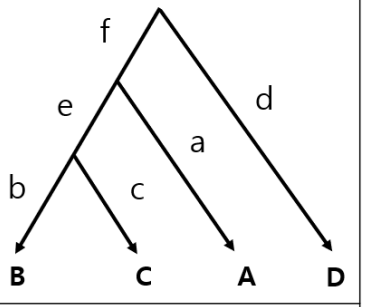
$F_4(A,B;C,D)$	0	e > 0	-e < 0
$F_4(B,C;A,D)$	e > 0	-e < 0	0
$F_4(A,C;B,D)$	e > 0	0	-e < 0

그림 8. 혼합이 없는 계통수에서 나올 수 있는 F_4 통계량의 값

	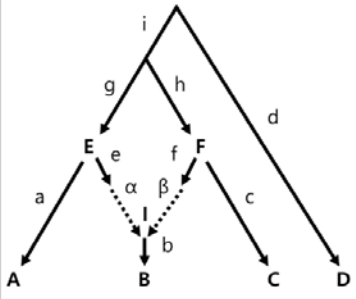	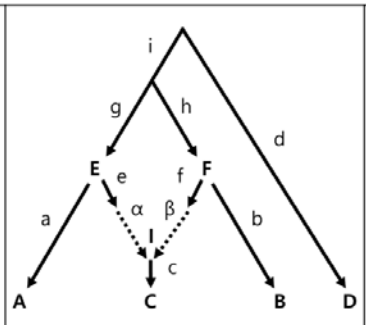	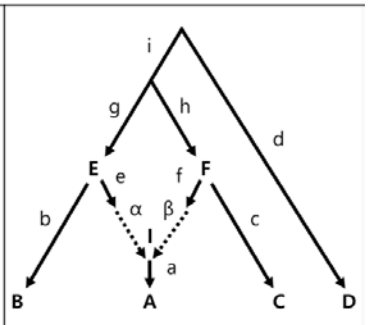
$F_4(A,B;C,D)$	-βh < 0	αg - βh	βh > 0
$F_4(B,C;A,D)$	αg > 0	-αg < 0	αg - βh
$F_4(A,C;B,D)$	αg-βh	-βh < 0	αg > 0

그림 9. 혼합이 있는 계통수에서 나올 수 있는 F_4 통계량의 값

결론

본 논문에서는 유전자 흐름이 지속적으로 있는 종 내의 집단 관계를 모델링하는 집단유전학에서 중요한 통계적 검증 도구로 사용되고 있는 F 통계량에 대해 설명하였다. F 통계량은 집단 간의 유전자 흐름과 혼합을 감지하는 강력한 도구로서, 유전체 데이터의 해석과 집단의 계통수 재구성에 중요한 역할을 한다. F_2 통계량은 두 집단 간의 유전적 거리를 측정하여 집단 간의 차이를 정량화하는 데 사용된다. 이를 통해 유전적 변이의 정도를 파악하고, 계통수 상에서 집단 간의 가지 길이를 추정할 수 있다. F_3 통계량은 세 집단 간의 관계를 분석하여 단순 계통수로 관계를 나타낼 수 있는지, 혼합이 추가된 계통수로 나타내지는지를 확인한다. F_3 통계량에는 특정 집단이 다른 두 집단의 혼합으로 형성되었는지를 검정하는 혼합 F_3 통계량과 세 집단 중 한 집단에 외군을 넣어 나머지 두 집단의 유전적 거리를 측정하는 외군 F_3 통계량이 있다. F_4 통계량은 네 집단 간의 상관관계를 분석하여, 혼합의 여부와 유전자 흐름의 방향성을 밝히는 데 사용된다. 이를 통해 집단 관계의 복잡한 혼합 패턴을 명확히 하고, 과거 집단의 이동 경로를 추적할 수 있다. 이러한 F 통계량은 집단유전학 연구에서 필수적인 도구로 자리잡았으며, 차세대 염기서열 분석 기술과 결합하여 더욱 정교한 집단 유전학 연구를 가능하게 한다. 본 논문이 F 통계량의 이론적 배경과 실질적 적용에 대한 이해를 높이는 데 기여하기를 기대한다.

사사

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. RS-2023-00212640) 및 2024년도 정부(교육부)의 재원으로 한국연구재단의 G-램프(LAMP) 사업 지원을 받아 수행된 연구임(No. RS-2023-00301976).

참고문헌

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19: 1655–1664.

Cavalli-Sforza LL, Edwards AW. 1967. Phylogenetic analysis: Models and estimation procedures. Evolution 21: 550–570.

HUGO Pan-Asian SNP Consortium, Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, Chaurasia A, Chen C-H, Chen J, et al. 2009. Mapping human genetic diversity in Asia. Science 326: 1541–1545.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28: 2239–2252.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A draft sequence of the Neandertal genome. Science 328: 710–722.

Hendry AP, Bolnick DI, Berner D, Peichel CL. 2009. Along the speciation continuum in sticklebacks. J. Fish Biol. 75: 2000–2036.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, et al. 2016. A new view of the tree of life. Nat. Microbiol. 1: 16048.

Kawecki TJ, Ebert D. 2004. Conceptual issues in local adaptation. Ecol. Lett. 7: 1225–1241.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217: 624–626.

Kimura M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press.

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, et al. 2016. Genomic insights into the origin of farming in the ancient Near East. Nature 536: 419–424.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513: 409–413.

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

Maddison DR, Schulz K-S, Maddison WP. 2007. The tree of life web project. Zootaxa 1668: 19–40.

Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D. 2011. The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet. 7: e1001373.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al. 2008. Genes mirror geography within Europe. Nature 456: 98–101.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient Admixture in Human History. Genetics. 192: 1065–1093.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus

genotype data. Genetics 155: 945–959.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford Jr TW, Orlando L, Metspalu E, et al. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505: 87–91.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468: 1053–1060.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature 461: 489–494.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, et al. 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. Am. J. Hum. Genet. 67: 1251–1276.

The Darwin Tree of Life Project Consortium, Blaxter M, Mieszkowska N, Di Palma F, Holland P, Durbin R, Richards T, Berriman M, Kersey P, Hollingsworth P, et al. 2022. Sequence locally, think globally: The Darwin Tree of Life Project. Proc. Natl. Acad. Sci. USA. 119: e2115642118.

Toews DPL, Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. Mol. Ecol. 21: 3907–3930.

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H. 2006. Harvesting the fruit of the human mtDNA tree. Trends Genet. 22: 339–345.

Wright S. 1922. Coefficients of inbreeding and relationship. Am. Nat. 56: 330–338.

Wright S. 1951. The genetical structure of populations. Ann. Eugen. 15: 323–354.

영문초록

Title: Reconstructing the Phylogenetic Relationship of Populations with F-Statistics: A Comprehensive Guide

Abstract: Traditional phylogenetic methods had a limited resolution for investigating the phylogenetic relationship of conspecific populations due to subtle nature of between population differentiation. However, recent advances on statistical methods, leveraging on large-scale genome variation data, have made noticeable progress in the study of intraspecific population relationship. Among these methods, F-statistics offer by far the simplest and most robust measures to test the topology of population graph and thus have been widely used in population genomics recently. Operating on a small number of populations, they test if a proposed population tree sufficiently explains the genetic relationship among these populations by examining a correlation between allele frequency changes along different branches. Additionally, F-statistics can rigorously detect population admixture when simple phylogenetic trees fail to explain the relationships between populations. This paper explains the principles of F-statistics, details the specific calculations of F_2, F_3, and F_4 statistics, and presents methods for detecting gene flow and admixture among populations using these statistics.

Authors: Dahyun Byeon[1], Choongwon Jeong[1,2*]

Affiliation: [1]*School of Biological Sciences, Seoul National University, Seoul, Republic of Korea 08826*

[2] *Institute for Data Innovation in Science, Seoul National University, Seoul, Republic of Korea 08826*

Corresponding author: cwjeong@snu.ac.kr