보상적 frameshift 돌연변이에 의한 RNA 바이러스의 진화

박동빈, 한윤수[1*]

요약: RNA 바이러스는 거의 모든 생명체에 감염하여 증식할 수 있으며, 그 놀라운 적응력은 매우 빠른 진화 속도에 기인한다. RNA 바이러스는, 그 복제 효소인 RNA 의존적 RNA 중합효소(RdRp)의 오류 교정 기능이 약하고, 유전체 크기가 작으며, 복제 속도가 빠르기 때문에 급속한 염기 서열 치환이 일어난다고 알려져 있다. 교정 기능이 약한 RdRp는 염기의 삽입 또는 결실에 의한 frameshift 돌연변이를 유발하여 기능이 상실된 변이체를 생성할 수 있다. RNA 바이러스는 다양한 변이체들이 집단으로 감염하고 전파하기 때문에, 기능이 상실된 변이체도 정상 바이러스의 도움을 받아 지속적으로 복제될 수 있다. 이때, 첫 번째 삽입/결실 위치 주변에서 두 번째 삽입/결실이 발생하면, open reading frame이 회복되는 보상적(compensatory) frameshift 돌연변이가 발생할 수 있으며, RNA 바이러스 유전체에서 흔하게 발견된다. 보상적 frameshift 돌연변이는 다수의 아미노산 서열이 동시에 치환되는 현상을 보이며, 서열 치환에 의존하는 기존의 계통 분석 방법을 사용하면 특정 계통에서의 급속한 아미노산 서열 치환 현상과 잘못된 계통군 추정 등을 유발할 수 있다. 보상적 frameshift 돌연변이는 RNA 바이러스의 진화 속도를 증가시키는 분자 진화의 주요 메커니즘 중 하나이며, RNA 바이러스의 계통학적 분석에 반드시 고려해야 할 현상이다.

키워드: RNA 바이러스, Frameshift 돌연변이, 보상적 돌연변이

[1]*06975 서울시 동작구 흑석로 84, 중앙대학교 생명과학과*

*Corresponding author: hahny@cau.ac.kr

RNA 바이러스의 빠른 진화 속도

RNA 바이러스는 단일가닥(ss) RNA 또는 이중가닥(ds) RNA로 구성된 유전체를 가지고 있으며, 인간과 동물, 균류 및 식물을 포함한 대부분의 진핵생물을 숙주로 삼아 증식할 수 있다(Gilbert et al., 2019; Roossink, 2012; Shi et al., 2016; Shi et al., 2018). 인간에 감염하여 질병을 일으키는 RNA 바이러스로는 소아마비를 일으키는 폴리오바이러스(poliovirus), 홍역(measles)의 원인이 되는 파라믹소바이러스(paramyxovirus), 독감을 유발하는 인플루엔자바이러스(influenza virus), 중증급성호흡기증후군(SARS)과 코로나바이러스감염증-19(Covid-19)의 원인이 되는 코로나바이러스(coronavirus) 등이 있다(Wu et al., 2020). RNA 바이러스에 의한 질병들을 통제하기 위하여 다양한 치료제와 백신을 개발하는 노력이 있지만, 지난 3년여 간 코로나바이러스감염증-19를 유발한 SARS-CoV-2 바이러스의 예에서 알 수 있듯이, RNA 바이러스는 짧은 기간에 다양한 새로운 변이체 바이러스들로 빠르게 진화하여 우리의 노력을 무력화시킨다(Cao et al., 2022; Starr et al.,

2022).

RNA 바이러스의 빠른 진화 속도를 설명하는 분자 메커니즘으로는, 매우 높은 염기 서열 치환 돌연변이율, 서로 다른 유전체 간의 재조합, 유전체를 구성하는 독립 분자들의 재구성 등이 있다(Bentley and Evans 2018; McDonald et al. 2016; Peck and Lauring, 2018). 매우 높은 염기 서열 치환 돌연변이율을 보이는 이유는, RNA 바이러스 유전체의 복제를 담당하는 RNA 의존적 RNA 중합효소(RNA-depenedent RNA-polymerase) 즉 RdRp는 복제 오류를 교정하는 기능(proofreading) 매우 약하기 때문이다(Barr and Fearns, 2010). RNA 바이러스는 다른 복제 단위에 비해 매우 높은 염기 치환 돌연변이율을 보이는데, 그 값은 염기 위치당 10^{-6}에서 10^{-4} 정도로 계산되었다(Drake, 1999; Sanjuán et al., 2010).

RNA 바이러스가 빠르게 진화할 수 있는 다른 요인으로는 RNA 유전체의 크기와 불안정성을 들 수 있다. RNA 분자는 화학적으로 불안정하고 RdRp의 교정 기능이 약하여 돌연변이가 쉽게 축적되므로, RNA 바이러스의 유전체는 일반적으로 약 10 kb 이하의 RNA 분자로 구성되어 있다(Sanjuán, 2012). 코로나바이러스는 RNA 바이러스로는 특이하게 약 30 kb 정도의 유전체 크기를 유지하는데, 그 이유는 상대적으로 우수한 교정 기능을 지니고 있기 때문이다(Robson et al., 2020). 유전체 크기와 염기 서열 치환 돌연변이율은, 그림 1과 같이, RNA 바이러스, DNA 바이러스, 박테리아(Bacteria) 및 진핵생물(Eukaryotes)에서 음의 상관관계를 나타내는 것으로 알려져 있다(Gago et al., 2009; Lynch, 2010). 이는 유전체의 크기가 작을수록 복제 속도가 빨라지고 재생산 시간이 짧아져 염기 서열 치환 돌연변이가 빠르게 선택되어 축적될 수 있기 때문으로 해석될 수 있다.

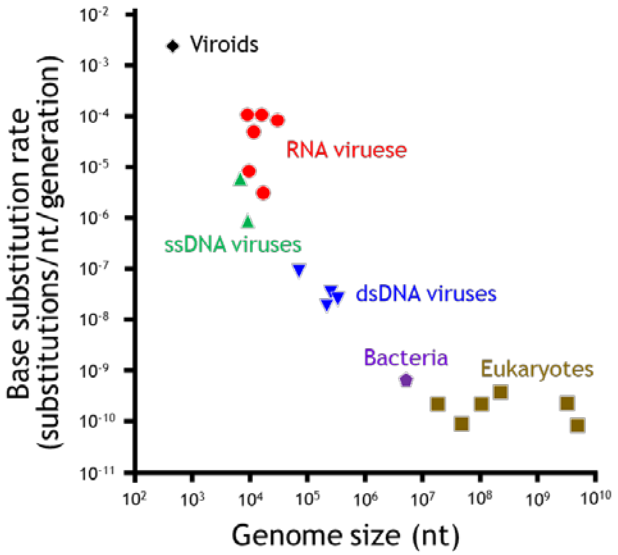

그림 1. 유전체의 크기와 염기 서열 치환 돌연변이율의 관계. (Gao et al., 2009의 그림을 수정)

보상적 frameshift 돌연변이

RNA 바이러스는 RdRp의 낮은 효율의 교정 기능과 RNA 유전체의 물리화학적인 특징 및 여러 분자적 기작에 의해 염기 서열 치환 돌연변이율뿐만 아니라 염기의 삽입(insertion) 및 결실(deletion) 돌연변이율 또한 높게 나타난다(Chrisman et al., 2021; Elena et al., 2008). 삽입/결실(indel) 돌연변이가 단백질 코딩 서열(coding sequence)에서 발생할 경우, open reading frame(ORF)이

변경되는 frameshift 돌연변이를 유발한다. 일반적으로 frameshift 돌연변이는 단백질 합성을 조기에 종결시키므로 정상 기능이 상실된 단백질 생산을 유발한다(그림 2). 그런데 만일 첫 번째 삽입/결실 돌연변이와 가까운 위치에서 두 번째 삽입/결실 돌연변이가 발생하는 경우, 두 돌연변이가 상쇄되어 ORF가 회복될 수 있다. 이때 두 삽입/결실 돌연변이 사이에 종결 코돈(codon)이 나타나지 않는다면 기능이 회복된 단백질을 생산할 수 있다. 이와 같이 둘 이상의 frameshift 돌연변이가 발생하여 결과적으로 ORF를 회복시키는 경우를 보상적(compensatory) frameshift 돌연변이라고 한다(Mbong et al., 2012; Park and Hahn, 2021; Yourno, 1970).

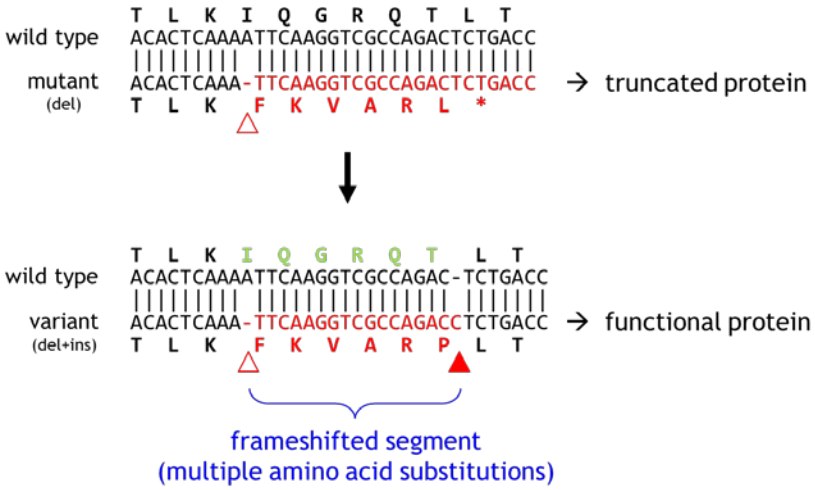

그림 2. 연속된 삽입/결실 돌연변이에 의한 보상적 frameshift 돌연변이의 발생. 빈 삼각형, 결실 돌연변이; 채운 삼각형, 삽입 돌연변이.

보상적 frameshift 돌연변이는 두 삽입/결실 돌연변이 위치 사이에서는 ORF가 변경되기 때문에, 여러 개의 아미노산이 동시에 치환되는 효과를 나타낸다. 보상적 frameshift 돌연변이는 단순히 결함 있는 삽입/결실 돌연변이를 상쇄시켜 기능성을 정상으로 회복시키는 과정일 뿐만 아니라, 다중 아미노산 서열 치환을 유발하여 매우 빠른 단백질 진화를 유도하는 수단일 수 있다(Hastings et al., 2004; Park and Hahn, 2021). 보상적 frameshift 돌연변이는 바이러스, 박테리아, 균류, 동물 등 다양한 생물에서 보고되었다(Biba et al., 2022; Colgrove et al., 2016; Jennings and Fane, 1997; Kihara et al., 1996; Sharma et al. 2016).

보상적 frameshift 돌연변이의 발생 메커니즘

보상적 frameshift 돌연변이가 발생하기 위해서는 최소 두 번의 삽입/결실 돌연변이가 발생하여야 한다(그림 3). RNA 바이러스의 경우 RdRp의 복제 오류의 교정 기능이 약하기 때문에 두 번의 삽입/결실 돌연변이가 발생하는 원스텝(one-step) 과정에 의해 발생할 가능성이 있으나, 확률적으로 매우 드물 것이다. 두 번의 삽입/결실 돌연변이가 순차적으로 발생하는 투스텝(two-step) 과정으로 보상적 frameshift 돌연변이가 발생하려면, 필연적으로 frameshift 돌연변이에 의해 기능을 상실한 RNA 바이러스 유전체가 충분한 시간 동안 생존하면서 복제할 수 있어야 한다.

일반적으로 바이러스는 동일한 유전체 서열을 가진 클론 집단이 아닌 다양한 변이체들이 혼합된 유사종(quasispecies) 상태로 숙주를 감염시킨다(Andino and Domingo, 2015; Domingo et al., 2012;

Sanjuán and Thoulouze, 2019). RNA 바이러스의 RdRp는 본질적으로 복제 오류가 발생하기 쉽고, 짧은 기간에 많은 수의 자손을 생산하기 때문에, 자손 바이러스의 유전체들은 다양한 변이체로 구성될 수 밖에 없으므로 자연스럽게 유사종의 상태를 유지하게 된다. 유사종 상태의 RNA 바이러스 변이체 간에는 긍정적 또는 부정적 상호작용이 일어나면서 상호보완(complementation) 또는 간섭(interference)을 일으킨다(Ciota et al., 2012; Perales et al., 2007; Vignuzzi and López, 2019). 돌연변이에 의해 기능을 상실한 RNA 바이러스 변이체는 동시에 감염한 정상 RNA 바이러스가 생산한 RdRp와 다른 단백질의 상호보완을 통해 유전체의 복제와 전파가 가능하다(Aaskov et al., 2006; Díaz-Muñoz et al., 2017). 따라서 한 번의 삽입/결실 돌연변이를 가진 바이러스 유전체는 도우미(helper) 바이러스의 상호보완에 의해 복제되는 동안 두 번째의 삽입/결실 돌연변이가 발생하여 기능을 회복한 보상적 frameshift 돌연변이체로 진화할 수 있다(그림 3).

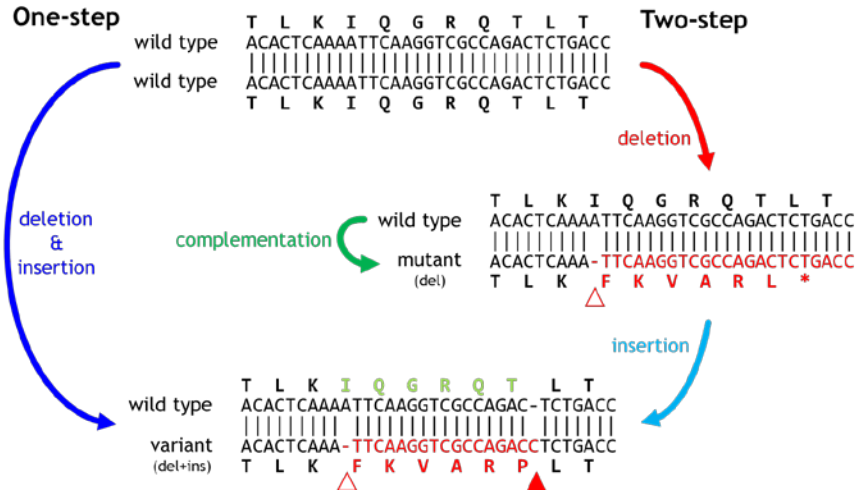

그림 3. 보상적 frameshift 돌연변이의 발생 메커니즘.

RNA 바이러스의 보상적 frameshift 돌연변이의 예

RNA 바이러스에서 보상적 frameshift 돌연변이에 의한 다중 아미노산 서열 치환 현상은 광범위하게 나타나는 것으로 보고되었다(Colgrove et al., 2016; Park and Hahn, 2021). 염기 서열 유사도가 97% 이상인 RNA 바이러스 단백질 유전자 서열을 수집하여 분석한 결과, 총 2,744개 유전자의 약 7.07%에 해당하는 194개에서 적어도 하나 이상의 보상적 frameshift 돌연변이가 발견되었다(Park and Hahn, 2021). 예를 들어, 돼지 루불라바이러스(Porcine rubulavirus)의 매트릭스(matrix) 단백질의 경우 RefSeq인 NC_009640.1 서열과 비교했을 때, 1-nt 결실과 1-nt 삽입에 의해 보상적 frameshift 돌연변이가 발생한 서열이 있다(그림 4A). 그 결과로 6개의 아미노산 서열이 동시에 치환되었다. 또 다른 예로는, 인플루엔자 A 바이러스(Influenza A virus)의 헤마글루티닌(hemagglutin) 단백질의 경우 RefSeq인 NC_007362.1 서열과 비교했을 때, 세 개의 서로 다른 위치에서의 1-nt 삽입에 의해 보상적 frameshift 돌연변이가 발생한 것으로 판정되었다(그림 4B). 그 결과로 12개의 아미노산 서열이 완전히 다른 13개의 아미노산 서열로 치환되었다. RNA 바이러스 단백질에서 발견된 보상적 frameshift 돌연변이에 의해 치환된 아미노산의 길이는 최소 5개에서 최대 87개까지였으며, 8개가 치환된 경우가 가장 빈도가 높았다(Park and Hahn, 2021)

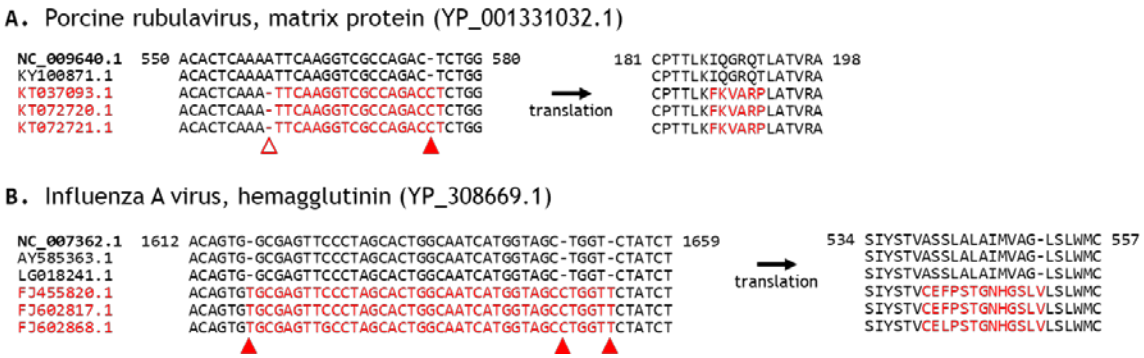

그림 4. RNA 바이러스의 보상적 frameshift 돌연변이의 예. (Park and Hahn, 2021의 그림을 수정)

보상적 frameshift 돌연변이에 의한 급속한 단백질 진화가 계통 분석에 미치는 영향

보상적 frameshift 돌연변이는 RNA 바이러스의 유전자 염기 서열은 거의 동일하더라도 다수의 아미노산 서열이 동시에 치환되는 급속한 단백질 진화를 유발하게 된다. 염기 서열의 치환 비율과 아미노산 서열의 치환 비율은 정비례하는 상관관계를 보인다. 일반적으로 아미노산 서열의 치환은 염기 서열의 치환에 의해 발생한다고 가정되며, RNA 바이러스의 계통 및 분자 진화 분석은 단백질의 아미노산 서열 치환 비율을 이용한다. 그러나 보상적 frameshift 돌연변이는 동시에 다중 아미노산 서열의 치환 돌연변이를 유발하므로 염기 서열 치환 비율을 크게 초과하는 아미노산 서열 치환 비율을 나타나게 된다. 만일 특정 계통에서 과거에 보상적 frameshift 돌연변이를 경험하였고, 그 역사를 알지 못한다면 모든 아미노산 서열 치환은 염기 서열 치환에 의한 것으로 판정될 것이므로, 그 계통의 아미노산 서열 치환 비율은 일반적인 염기 서열 치환에 의한 비율에 비하여 크게 과장된 값을 가지게 된다.

보상적 frameshift 돌연변이에 의한 과장된 아미노산 서열 치환 또는 급속한 단백질 진화는 인플루엔자 A 바이러스의 헤마글루티닌 단백질의 예에서 확인할 수 있다(그림 5). 그림 5A와 그림 5B는 헤마글루티닌 단백질 유전자의 RefSeq인 NC_007362.1 서열과 97%의 염기 서열 유사도를 보이는 바이러스 변이체 서열들의 염기 서열과 단백질 서열로 작성된 계통수이다. 보상적 frameshift 돌연변이를 가진 변이체들은 모두 하나의 계통을 이루고 있으며, 화살표로 표시된 가지에서 돌연변이가 발생한 것으로 판정되었다. 이 때 염기 서열 계통수와 아미노산 서열 계통수를 비교하면, 화살표로 표시된 가지의 길이가 아미노산 서열 계통수에서 상대적으로 매우 길게 표시되는 것을 볼 수 있다. 이는 보상적 frameshift 돌연변이가 유발한 다중 아미노산 변이가 순차적인 염기 서열 치환에 의한 아미노산 서열 치환으로 계산되었기 때문이다. 따라서 보상적 frameshift 돌연변이에 대한 사전 지식이 없는 경우에는, 해당 계통에서 특이적으로 서열 치환이 빠르게 일어났거나, 다른 계통과 분지한 시점이 실제보다 오래된 것으로 오인 판정될 수 있다.

단백질 계통수에서 보이는 급속한 아미노산 서열 치환 현상은 계통수의 변이체간 거리를 비교하여 확인할 수도 있다. 염기 서열 계통수와 단백질 서열 계통수에서, 기준이 되는 RefSeq NC_007362.1에서 각 변이체까지의 계통학적 거리, 즉 염기 서열 거리(nucleotide distance, Dn)와 단백질 서열 거리(protein distance, Dp)를 계산하여 상관관계를 그래프로 작성하면 그림 5C와 같이 나타난다. 보상적 frameshift 돌연변이가 없는 변이체들은 Dn과 Dp의 값이 일정한 비율로 비례하는 분포를 보인다(그림 5C의 검정 점들). 그러나 보상적 frameshift 돌연변이를 가진 변이체들은 매우

동떨어진 분포를 보이며, 유사한 Dn 값을 보이는 보상적 frameshift가 없는 변이체와 비교하였을 때 현저히 큰 Dp 값을 보인다(그림 5C의 빨강 점들). 즉, 과장된 아미노산 서열 치환 속도를 가진 것으로 판정된다. 만일, frameshift가 일어난 구간을 제외하고 계통수를 작성하면 그림 5D와 같이 모든 변이체가 염기 서열 치환과 아미노산 서열 치환이 비례하는 유사한 분포를 보인다. 즉, 과장된 아미노산 서열 치환 속도는 보상적 frameshift 돌연변이에 의한 것임을 알 수 있다.

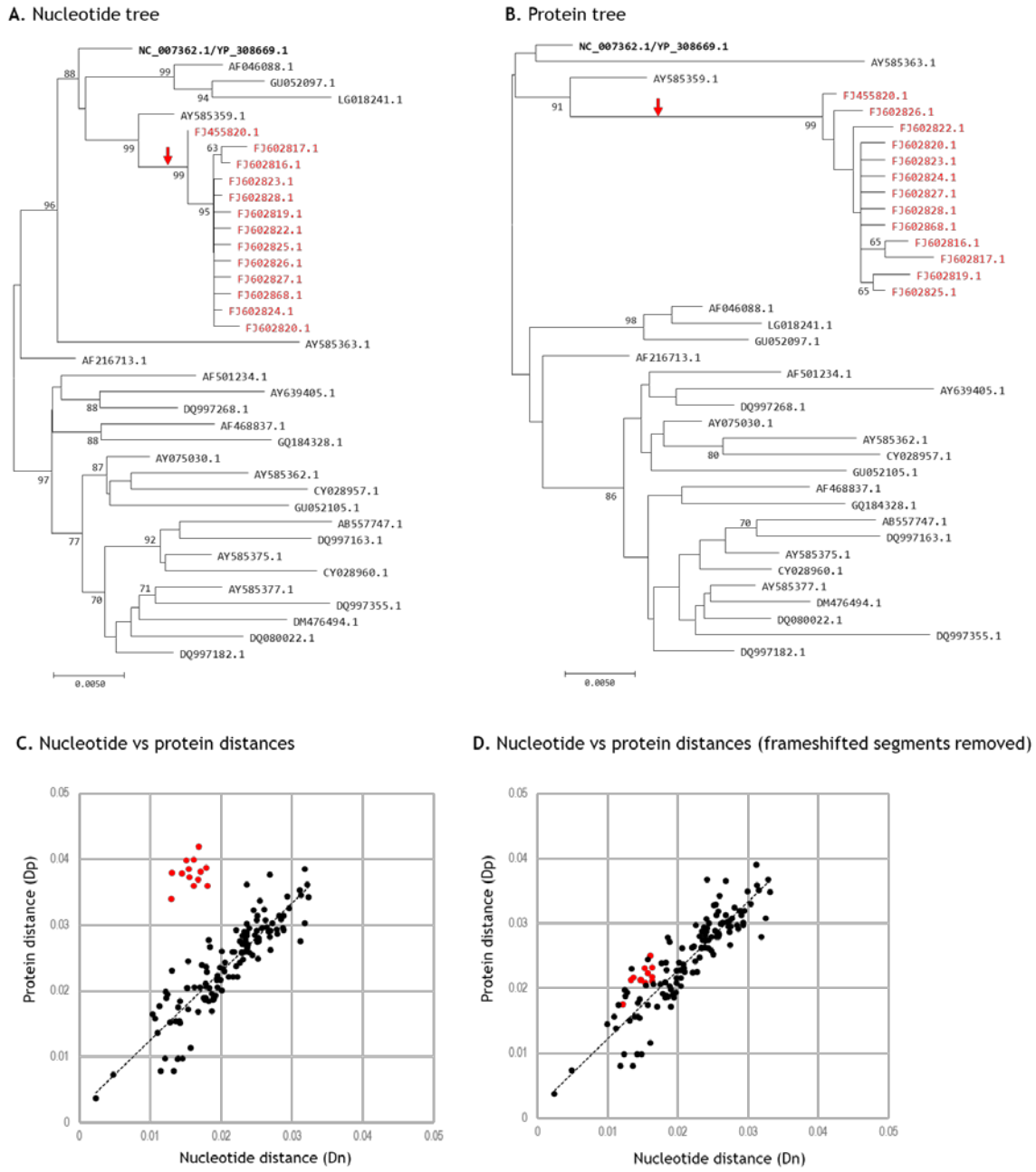

그림 5. 인플루엔자 A 바이러스의 헤마글루티닌 단백질의 보상적 frameshift 돌연변이. (A) 염기 서열 계통수. (B) 단백질 서열 계통수. (C) 염기 서열과 단백질 서열의 계통수에서 계산된 변이체들의 계통학적 거리의 분포도. (D) 보상적 frameshift 돌연변이 구간을 제외하고 작성한 계통수에서 계산된 변이체들의 계통학적 거리의 분포도. 빨강 화살표, 보상적 frameshift 돌연변이가 발생한 지점. (Park and Hahn, 2021의 그림을 수정)

보상적 frameshift 돌연변이에 의한 과장된 아미노산 서열 치환 속도는 단백질 서열 계통수를 이용한 RNA 바이러스의 계통 분석 과정에서 부정확한 계통군(clade) 판정을 유발할 수도 있다. 예를 들어 서울 오쏘한타바이러스(Seoul orthohantavirus)의 뉴클레오캡시드(nucleocapsid) 단백질에서 발견된 보상적 frameshift 돌연변이가 있다(그림 6A). 염기 서열 계통수와 단백질 서열 계통수를 비교하면 보상적 frameshift 돌연변이를 보이는 변이체들의 자매계통군이 달라짐을 알 수 있다. 염기 서열 계통수(그림 6B)에서는 보상적 frameshift 돌연변이체 계통군(노랑 상자)은 변이가 없는 RefSeq인 NC_005236.1 계통군(녹색 상자)의 자매군을 형성하며(붓스트랩 지지도 값은 97), 다른 보상적 frameshift 돌연변이가 없는 변이체들은 독립적인 계통군(파랑 상자)을 형성한다. 따라서 보상적 frameshift 돌연변이는 RefSeq인 NC_005236.1 계통과 분지된 계통에서 발생한 것으로 추정할 수 있다(빨강 화살표). 그러나 단백질 서열 계통수에서는 보상적 frameshift가 없는 모든 변이체들은 마치 하나의 계통군(녹색/파랑 상자)인 것처럼 나타나며, 보상적 frameshift 돌연변이체는 독립적인 계통군(노랑 상자)으로 판정된다(붓스트랩 지지도 값은 100).

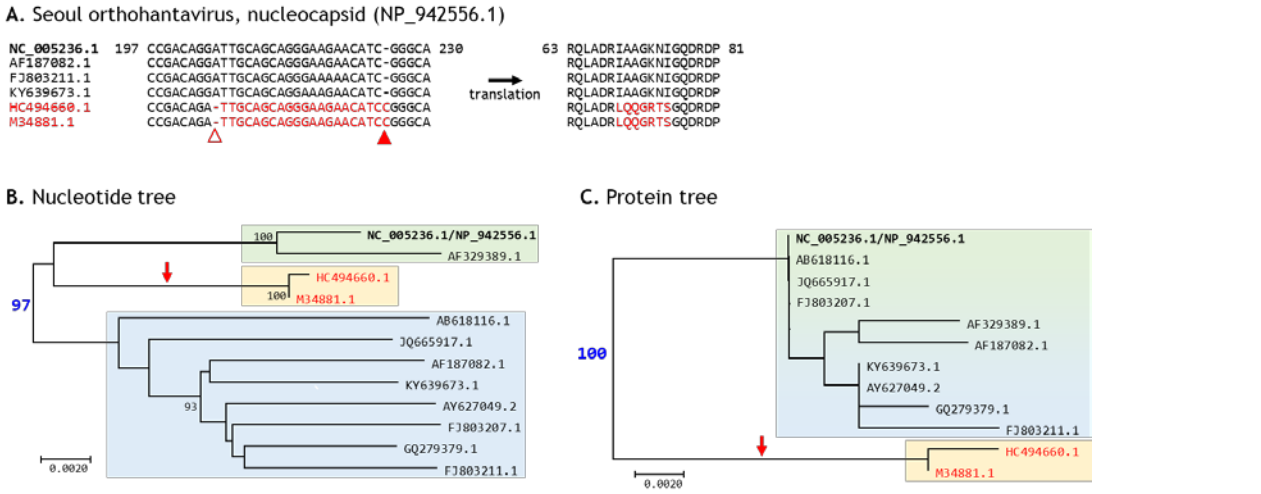

그림 6. 서울 오쏘한타바이러스의 뉴클레오캡시드 단백질의 보상적 frameshift 돌연변이. (A) 결실과 삽입에 의한 보상적 frameshift 돌연변이. (B) 염기 서열 계통수. (C) 단백질 서열 계통수. 빨강 화살표, 보상적 frameshift 돌연변이가 발생한 지점; 노랑 상자, 보상적 frameshift 돌연변이 계통군; 파랑 수, 언급된 붓스트랩 지지도 값. (Park and Hahn, 2021의 그림을 수정)

보상적 frameshif 돌연변이가 최근에 발생하여 염기 서열의 유사도가 충분히 유지되는 경우에는 염기 서열 계통수를 통해 정확한 계통군 분석이 가능할 수 있다. 그러나 보상적 frameshift 돌연변이 이후 오랜 시간이 경과하여 염기 서열 치환이 축적된 경우에는, 일반적인 염기 서열 치환에 의한 아미노산 서열 치환과 보상적 frameshift 돌연변이에 의한 다중 아미노산 서열 치환의 구분이 사실상 불가능하다. 따라서 보상적 frameshift 돌연변이의 존재 여부 자체를 판정할 수 없는 경우가 발생할 수 있으며, 이때 계통 간의 분지 양상에 오류가 있다 하더라도 전혀 인지할 수 없다.

염기 서열의 치환율을 정확히 측정하기 위한 방법으로 아미노산 서열 정렬 정보를 바탕으로 코돈 단위로 염기 서열을 정렬하는 방식이 이용된다(Libin et al., 2019). 이 방법은 축적된 염기 서열 치환으로 인하여 염기 서열 정렬이 불확실한 경우와, 아미노산의 삽입과 결실, 즉 코돈 단위의 삽입과 결실이 있는 경우에 정확한 염기 서열 정렬과 염기 서열 치환율 계산에 잘 활용될 수 있다. 그러나 최근에 발생한 보상적 frameshift 돌연변이에 적용한다면 frameshift가 일어난 구간에서의

아미노산 서열 정렬이 무의미하여, 코돈 단위의 정렬은 오히려 염기 서열의 부정확한 정렬을 유발하고 염기 서열 치환율을 실제보다 과장되게 계산되는 결과를 초래할 수 있다.

결론

보상적 frameshift 돌연변이는 동시에 많은 수의 아미노산 서열 치환을 유발하며, RNA 바이러스의 유전체에서 흔하게 발견된다. RNA 바이러스들은 다양한 변이체들이 유사종 상태의 집단으로 감염하고 전파하기 때문에, 삽입/결실 돌연변이에 의해 기능을 상실한 유전체들도 지속적으로 복제되면서 보상적 frameshift 돌연변이체가 될 기회가 있다. 다중 아미노산 서열 치환이 발생한 변이체의 계통 분석에는 주의를 기울여야 하며, 서열 치환에 근거한 기존의 계통 분석 방법을 이용하면 계통 특이적 급속한 아미노산 서열 치환 또는 잘못된 계통군 추정이라는 결론에 도달할 수 있다. 보상적 frameshift 돌연변이는 RNA 바이러스의 숙주 적응도를 빠르게 높일 수 있는 분자 진화의 주요 메커니즘으로 볼 수 있다.

참고문헌

Aaskov J, Buzacott K, Thu HM, et al. 2006. Long-term transmission of defective RNA viruses in humans and Aedes mosquitoes. Science 311:236-238.

Andino R, Domingo E. 2015. Viral quasispecies. Virology 479-480:46-51.

Barr JN, Fearns R. 2010. How RNA viruses maintain their genome integrity. J Gen Virol 91:1373-1387.

Biba D, Klink G, Bazykin GA. 2022. Pairs of mutually compensatory frameshifting mutations contribute to protein evolution. Mol Biol Evol 39:msac031.

Bentley K, Evans DJ. 2018. Mechanisms and consequences of positive-strand RNA virus recombination. J Gen Virol 99:1345-1356.

Cao Y, Wang J, Jian F, et al. 2022. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. Nature 602:657-663.

Chrisman BS, Paskov K, Stockham N, et al. 2021. Indels in SARS-CoV-2 occur at template-switching hotspots. BioData Min 14:20.

Ciota AT, Ehrbar DJ, Van Slyke GA, et al. 2012. Cooperative interactions in the West Nile virus mutant swarm. BMC Evol Biol 12:58.

Colgrove RC, Liu X, Griffiths A, et al. 2016. History and genomic sequence analysis of the herpes simplex virus 1 KOS and KOS1.1 sub-strains. Virology 487:215-221.

Díaz-Muñoz SL, Sanjuán R, West S. 2017. Sociovirology: conflict, cooperation, and communication among viruses. Cell Host Microbe 22:437-441.

Domingo E, Sheldon J, Perales C. 2012. Viral quasispecies evolution. Microbiol Mol Biol Rev 76:159-216.

Drake JW. 1999. The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. Ann N Y Acad Sci 870:100-107.

Elena SF, Agudelo-Romero P, Carrasco P, et al. 2008. Experimental evolution of plant RNA viruses. Heredity (Edinb) 100:478-483.

Hastings PJ, Slack A, Petrosino JF, et al. 2004. Adaptive amplification and point mutation are independent mechanisms: evidence for various stress-inducible mutation mechanisms. PLoS Biol 2:e399.

Gago S, Elena SF, Flores R, et al. 2009. Extremely high mutation rate of a hammerhead viroid. Science 323:1308.

Gilbert KB, Holcomb EE, Allscheid RL, et al. 2019. Hiding in plain sight: New virus genomes discovered via a systematic analysis of fungal public transcriptomes. PLoS One. 14:e0219207.

Jennings B, Fane BA. 1997. Genetic analysis of the phi X174 DNA binding protein. Virology 227:370-377.

Kihara M, Francis NR, DeRosier DJ, et al. 1996. Analysis of a FliM-FliN flagellar switch fusion mutant of Salmonella typhimurium. J Bacteriol 178:4582-4589.

Libin PJK, Deforche K, Abecasis AB, et al. 2019. VIRULIGN: fast codon-correct alignment and annotation of viral genomes. Bioinformatics. 35:1763-1765.

Lynch M. 2010. Evolution of the mutation rate. Trends Genet 26:345-352.

Mbong EF, Woodley L, Dunkerley E, et al. 2012. Duffy C. Deletion of the herpes simplex virus 1 UL49 gene results in mRNA and protein translation defects that are complemented by secondary mutations in UL41. J Virol 86:12351-12361.

McDonald SM, Nelson MI, Turner PE, et al. 2016. Reassortment in segmented RNA viruses: mechanisms and outcomes. Nat Rev Microbiol 14:448-460.

Park D, Hahn Y. 2021. Rapid protein sequence evolution via compensatory frameshift is widespread in RNA virus genomes. BMC Bioinformatics 22:251.

Peck KM, Lauring AS. 2018. Complexities of viral mutation rates. J Virol 92:e01031-17.

Perales C, Mateo R, Mateu MG, et al. 2007. Insights into RNA virus mutant spectrum and lethal mutagenesis events: replicative interference and complementation by multiple point mutants. J Mol Biol 369:985-1000.

Robson F, Khan KS, Le TK, et al. 2020. Coronavirus RNA proofreading: molecular basis and therapeutic targeting. Mol Cell 79:710-727.

Roossinck MJ. 2012. Plant virus metagenomics: biodiversity and ecology. Annu Rev Genet 46:359-369.

Sanjuán R. 2012. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. PLoS Pathog 8:e1002685.

Sanjuán R, Nebot MR, Chirico N, et al. 2010. Viral mutation rates. J Virol 84:9733-9748.

Sanjuán R, Thoulouze MI. 2019. Why viruses sometimes disperse in groups? Virus Evol 5(1):vez014.

Sharma V, Elghafari A, Hiller M. 2016. Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. Nucleic Acids Res 44:e103.

Shi M, Lin XD, Tian JH, et al. 2016. Redefining the invertebrate RNA virosphere. Nature 540:539-543.

Shi M, Lin XD, Chen X, et al. 2018. The evolutionary history of vertebrate RNA viruses. Nature 556:197-202.

Starr TN, Greaney AJ, Hannon WW, et al. 2022. Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. Science 377:420-424.

Vignuzzi M, López CB. 2019. Defective viral genomes are key drivers of the virus-host interaction. Nat Microbiol 4:1075-1087.

Wu F, Zhao S, Yu B, et al. 2020. A new coronavirus associated with human respiratory disease in China. Nature 579:265-269.

Yourno J. 1970. Nature of the compensating frameshift in the double frameshift mutant hisD3018 R5 of *Salmonella typhimurium*. J Mol Biol 48:437-442.

영문초록

Title: Evolution of RNA viruses via compensatory frameshift mutations

Abstract: RNA viruses can infect and replicate in almost all living organisms, thanks to their remarkable adaptability, which is driven by their rapid evolutionary rate. RNA viruses are known to undergo rapid nucleotide sequence substitutions due to the weak error correction function of the RNA-dependent RNA polymerase (RdRp), their small genome size, and fast replication speed. The error-prone RdRp can induce frameshift mutations caused by insertions or deletions of nucleotides, leading to the generation of non-functional variants. As RNA viruses infect hosts and spread as a population of diverse variants, non-functional variants can persistently replicate by help of functional viruses. In such cases, when a second insertion or deletion occurs near the first insertion or deletion site, a compensatory frameshift mutation can restore the open reading frame. Compensatory frameshift mutations are commonly found in RNA virus genomes and induce multiple amino acid sequence substitutions. The use of traditional phylogenetic analysis methods that rely on sequence substitutions can lead to erroneous estimation of rapid amino acid substitution rates and incorrect lineage inference. Compensatory frameshift mutations can be regarded as a major mechanism of molecular evolution that increases the evolutionary rate of RNA viruses.

Authors: Dongbin Park and Yoonsoo Hahn*

Affiliation: Department of Life Science, Chung-Ang University, 06974 Seoul, Republic of Korea

Corresponding author: hahny@cau.ac.kr