왜 유전자 카세트가 고아 역전사효소와 함께 다닐까? 5대륙의 Pollutibacter

표철민

AI3 Discovery, 한국

교신저자: discovery@ai3.kr

초록

세균 역전사효소(RT) 가운데 여럿은 파지로부터 세포를 보호하지만, 알려진 어느 계열에도 들어맞지 않는 RT도 있다. 우리는 게놈 189,630개를 조사해, 우리의 어떤 비교(알려진 RT 유형, 계통수, 예측 구조)로도 알려진 계열에 놓이지 않는 RT(발표된 라이브러리에서 가장 가까운 RT와의 동일성은 최대 34.6%)를 찾았고, 이를 ESOL RT라 부른다. 이 RT는 게놈 16개에서 유전자 카세트 곁에 자리한다. 이 게놈은 카세트를 지녔다는 이유로 고른 것이며, 5개 대륙 9개국의 호수, 강, 연못과 그 밖의 수역에서 왔고, 모두 Bacteroidota의 한 속인 Pollutibacter에 속한다. 카세트에는 보존된 상류 요소와 고정된 위치의 유전자 네 개가 있다. 이 유전자 가운데에는 MIDAS 모티프를 지닌 단백질이 있고, 메탈로프로테아제를 닮은 것으로 예측되지만 H[EDQ]xxH 모티프가 없는 단백질도 있다. 추가 Pollutibacter 좌위 27개에서는 RT 곁의 관측 가능한 위치 99개 중 97개에 카세트 구성 요소가 들어 있고, 두 예외는 모두 한 좌위에 있다. 관측 가능한 이웃 유전자 네 개의 사본은 하나도 파괴되지 않았고, 코돈 모형은 정화 선택을 시사하며, 다른 담수 시료에서 얻은 긴 읽기는 모듈을 최대 여섯 개까지 걸친다. 게놈 794개를 대상으로 한 사후(post hoc) 탐색에서 요소는 히트 19개 중 18개에서 RT의 상류에 놓이고, 19번째 히트는 콘티그 끝에서 잘려 있다. 그러나 −1 유전자와 +3 유전자의 상동 유전자는 +3형 유전자를 가진 속 149개 중 54개에서 서로 곁에 놓인 쌍으로 나타나며, 그 근처에서는 RT가 검출되지 않는다. 그렇다면 카세트는 왜 이 RT와 함께 다니는가? RT와 카세트가 함께 다니는 것은 관측되지만, 두 질문이 남는다. RT가 기존 모듈에 합류했는가, 그리고 DRT10이 RNA 짝, 단백질 짝과 함께 작동한다고 보고된 것처럼 이 RT도 그 모듈과 함께 작동하는가?

1 서론

역전사효소(RT)는 세균에 흔하며, 그룹 II 인트론, 레트론, 다양성 생성 레트로요소(DGR), UG/Abi 계열(기능을 모르는 RT 그룹과 감염 중단 시스템의 RT 그룹)과 그 밖의 여러 그룹으로 나타난다 [1–4]. 세균 게놈을 조사한 연구는 이 RT를 17개 주요 그룹으로 나누었고 [1], UG/Abi 계열만 해도 이후 42개 그룹으로 나뉘었으며 [4], CRISPR–Cas 시스템과 연관된 RT는 15개 단계통군으로 나뉘었다 [5]. 이 수는 사용한 분류에 따라 달라지며 서로 더할 수 없다. 이들 RT 가운데 다수는 숙주의 기능을 맡는다. 레트론은 파지로부터 세포를 보호하고 [6], 그 부속 유전자는 계산으로 예측되고 분류되었으며 [7], 여러 UG/Abi 및 방어 관련 RT(DRT) 그룹도 세포를 보호하고 [4,8], DGR은 표적 단백질에 다양성을 만들며 [9], RT–Cas1 융합체는 RNA에서 CRISPR 스페이서를 얻는다 [10].

세 저자 팀의 프리프린트 네 편(Sánchez-Nieto et al. 2026; Tang et al. 2025 및 Ramirez et al. 2026, 두 편은 책임저자가 같다; Chen et al. 2026)은 DRT10 (UG17)을 구조를 가진 비암호화 RNA(ncRNA), SLATT(SMODS and SLOG-associating two-transmembrane) 이펙터 단백질, RT의 세 부분으로 이루어진 시스템으로 기술한다. 이 가운데 Sánchez-Nieto et al.과 Tang et al.은 이들이 시험한 E. coli 시스템에서 세 구성 요소가 모두 방어에 필요하다고 보고한다 [11–14]. Wang et al. (Cell 2026; 저자 원고)은 DRT3과 RecBCD 소실(ΔrecB) 사이, DRT9와 ExoI 소실

(ΔsbcB) 사이에 선택적 합성 치사가 있다고 보고하고, 파지의 RecBCD 억제자 Gam이 DRT3 매개 감염 중단을 촉발한다고도 보고한다 [8]. van den Berg et al. (bioRxiv 프리프린트, 2025)은 Pseudomonas aeruginosa 유전자 클러스터를 다루며, 파지로부터 보호하는 세균 시스템이 모듈성이 매우 높고 감지, 신호 전달, 이펙터 구성 요소가 유전자 클러스터 사이에서 자주 교환된다고 기술한다 [15].

Bacteroidota의 경우, Toro (mSystems 2026)는 SPIRE(Searchable planetary-scale microbiome REsource) 대표 컬렉션의 XI형 유사 레트론 후보 232개 중 170개가 Bacteroidota에 있으며, 그 232개 중 167개(72%)는 신뢰도가 낮은 판정이라고 보고한다. 또한 Toro는 후보 계통 두 개의 좌위 일부에서 RT 상류의 msr/msd 유사 ncRNA 모티프(msr과 msd는 레트론 ncRNA의 두 요소이다) 후보를 계산으로 예측했으나, 실험적 검증은 보고되지 않았다 [16]. 우리는 여기서 기술하는 RT가 이 레트론 유사 후보나 UG17 아형에 속한다고 주장하지 않으며, 이 RT를 비교, 구성, 동반 출현만으로 기술한다.

어떤 RT가 알려진 유형 밖에 있음을 보이기는 어렵다. RT 서열은 깊이 분기한다. 유사도 검색으로는 분기한 RT가 알려진 계열의 먼 구성원인지 별개의 계통인지 가릴 수 없고, RT 핵심부로 만든 계통수는 포화한다. RT가 미배정이라는 주장은 어떤 라이브러리를 점검했는지에도 달려 있다. 이 프로젝트를 진행하는 동안 미배정으로 보였던 후보 그룹이 우리 라이브러리에 없던 발표된 계열과 일치한 경우가 있었다. 그러므로 우리는 미배정을 명시한 비교에 관한 진술로 다루며, 어떤 계통이 새롭다는 주장으로는 다루지 않는다. 그리고 그 답에 달려 있지 않은 두 번째 질문을 던진다. 이런 RT의 곁에는 무엇이 있으며, 이를 공개 서열 자료로 어디까지 보일 수 있는가?

우리는 Genome Taxonomy Database(GTDB) R232의 세균 종 대표 게놈 189,630개에서 RT 좌위를 총조사하는 데서 출발했다. 후보 그룹 98개 중 32개는 당시 입수할 수 있던 발표된 라이브러리에 견주어 미배정으로 남았고, 98개 중 하나가 되풀이해 나타나는 같은 종류의 이웃으로 눈에 띄었다. 우리는 이 그룹의 RT를 ESOL RT라 부른다. 이 이름은 작업용 약칭 ESOL(“enigmatic standalone ORF lineage”)에서 왔으며, 이 RT가 하나의 계통이라는 주장을 담지 않는다. 별개의 계통인지 알려진 계통의 크게 갈라진 구성원인지는 열려 있으며, 우리는 이를 미배정이라고 기술한다. 이 RT는 곁에 무리를 두고 있다. 우리는 그 이웃을 ESOL 카세트라 부른다. ESOL 카세트는 상류의 비암호화 요소와, RT 오픈 리딩 프레임(ORF)을 기준으로 −1, +1, +2, +3 위치에 놓인 단백질 슬롯으로 이루어지며, 오직 위치로만 정의되고 이동성이나 기능에 대한 주장을 담지 않는다. 이 논문에서 ESOL RT는 카세트 곁에 서 있는 RT 하나를 뜻하며, 그림에서는 이를 카세트 RT라 부른다.

우리의 어떤 비교로도 이름 있는 계통에 배정되지 않는 RT 하나와 그 곁의 카세트를 기술한다. 또한 카세트를 지닌 메타게놈 조립 게놈(MAG)이 나온 물 시료 16개(5개 대륙 9개국의 호수, 강, 연못과 그 밖의 수역; 대륙은 국가 항목에서 도출했다)와, 우리가 탐색한 범위 안에서 카세트가 발견되는 Bacteroidota의 속 Pollutibacter를 기술한다. 이 연구의 중심에는 수수께끼가 있다. 카세트의 구성 요소는 다른 속에서도 RT 없이 나타나는데, 카세트가 발견되는 단일 속에서는 RT와 함께 다닌다. 그래서 우리는 그 까닭을 묻는다. 우리는 네 가지 대안 설명을 시험했다. 알려진 RT 유형: 어떤 관문도 이 RT를 그 유형에 배정하지 않는다(구성원 14개 중 12개의 콘티그에서 반복 쌍 9개 중 DGR형 0개, 구성원 17개 중 그룹 II 마투라제 도메인을 가진 것 0개, ESOL 질의 30개 가운데 발표된 라이브러리에서 가장 가까운 RT와의 동일성 최대 34.6%). 퇴화하는 유물: 관측 가능한 모듈 사본은 하나도 파괴되지 않았다. 조립 인공물: 우리 참조에 게놈을 내놓지 않은 시료와 연구에서 나온 읽기 가운데 36개가 인접한 카세트 모듈 사이의 접합부를 가로지른다(11개는 첫 히트를 본 뒤에 고정한 동결 층위에 있고, 25개는 나중에 추가한 사후(post hoc) 층위에 있다). 검출 공백: 동일성은 맞추었으나 정렬 덮임은 맞추지 않은 참조 풀 모사에서, 바꾸지 않은 도구 문턱값으로 알려진 단위는 49개 중 49개가 검출되었으나 카세트

RT 단위는 16개 중 0개가 검출되었는데, 이 대비는 어느 방향의 증거도 아니므로 이 설명은 열려 있다. 이 시험을 거친 뒤에 남는 것은 수수께끼이며, 우리는 이를 별개의 두 질문으로 제시한다. 하나는 RT가 이미 존재하던 모듈에 합류했는지이고, 다른 하나는 RT가 그 모듈과 함께 작동하는지이다. 각 질문마다 그것을 가려 줄 관찰을 명시한다. 우리는 증거가 보여 주는 것(카세트의 구성 요소, 그 배열과 온전한 상태, RT가 곁에 두는 무리)과 열어 두는 것(그 무리가 어디서 왔는지와 무엇을 하는지)을 밝히며, 습식 실험 결과는 이 논문에 보고하지 않는다. 다만 습식 실험은 진행 중이다.

2 결과

우리의 어떤 비교로도 알려진 계열에 배정되지 않는 역전사효소가 16개 게놈에서 같은 유전자 무리와 함께 다닌다(그림 1). 이 게놈은 5대륙 9개국의 물 시료 16개(호수, 강, 연못과 그 밖의 수역)에서 나왔는데도 16개 모두 한 속, Pollutibacter에 속한다. 따라서 카세트는 다섯 대륙에서 발견되지만 좁은 계통 안에 있다. 우리가 탐색한 범위 안에서 카세트는 한 속, Pollutibacter에 있으며, 이 속과 가까운 근연 생물은 독립적 반복 출현을 따지는 어떤 검정에서도 하나의 계통으로 센다(그림 2). 열린 질문은 이 RT와 이 무리가 왜 함께 머무는가이다. 결과는 질문의 논리를 따른다. 이 RT는 무엇인가, 그 곁에는 무엇이 서 있는가, 그 무리는 실재하며 온전한가, 세계 어디에서 어느 속에 나타나는가, 그리고 왜 거기에 있을 수 있는가. 우리가 쓰는 용어는 표 1에서 정의한다.

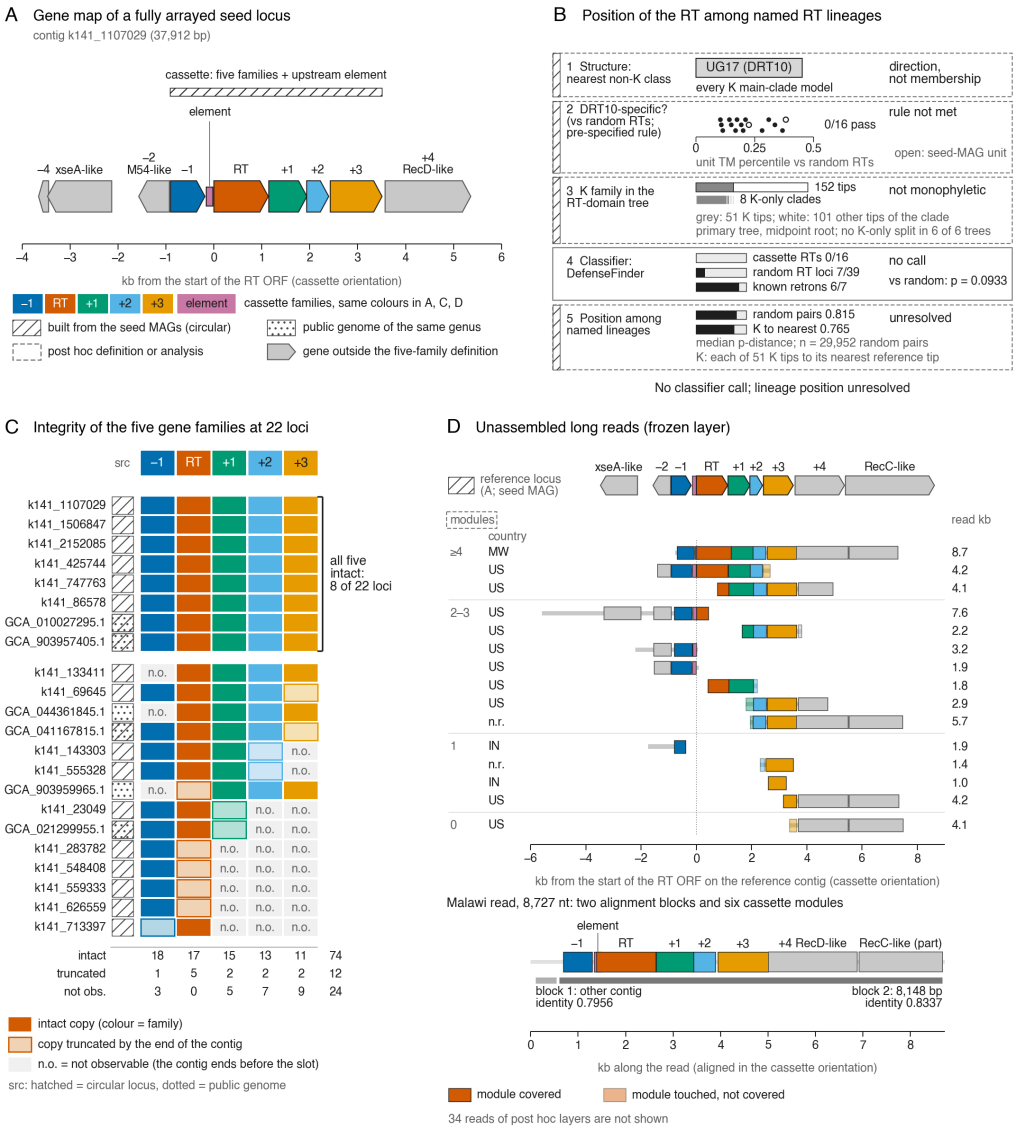


그림 1. 카세트 RT와 그 옆의 유전자 계열: 유전자 지도, 이름 있는 RT 계통 사이에서 RT의 위치, 조립된 좌위에서 유전자 계열의 온전성, 조립되지 않은 긴 읽기에서 카세트 특이적 서열. A, 콘티그가 배열 전체에 걸쳐 있는 시드 MAG 좌위 하나(콘티그 k141_1107029, 37,912 bp)의 유전자 지도로, 주석 표에서 카세트 방향을 기준으로 그렸고 위치는 RT ORF 시작점에서 kb 단위로 나타냈다. 빗금은 계열 모형을 만든 바탕인 시드 MAG 16개에 속한 요소를 나타내므로 이 좌위의 유전자는 구성상 순환적이며, 이 16개 MAG는 카세트를 가졌다는 이유로 선별되었고 한 속에 속한다. B, RT를 이름 있는 RT 계통과 다섯 경로로 비교한 것과 각 경로의 결과이다. 파선 윤곽은 사후(post hoc) 정의나 분석을, 빗금은 시드 MAG로 만든 비교 집합을 나타낸다. C, 한 속의 좌위 22곳(시드 MAG 좌위 16곳과 공개 게놈 6개)에서 유전자 계열 다섯 개(−1, RT, +1, +2, +3)의 상태이며, 이 중 20곳은 보충 데이터 S1에서 순환적인 좌위로 표시되어 있다(시드 MAG 좌위 16곳과 공개 게놈 4개; src 열에서 빗금으로 나타냈다). 슬롯 110개 중 74개는 온전한 사본을 가지고, 12개는 콘티그 끝에서 잘린 사본을 가지며, 24개는 슬롯 앞에서 콘티그가 끝나 관측할 수 없다. D, 조립되지 않은 긴 읽기이다. 동결 층위의 읽기 15개를 하나씩, 그 주 정렬 블록이 정렬되는 기준 콘티그의 카세트 좌표 위에 RT ORF 시작점을 기준으로 그렸고, 그 블록이 덮는 카세트 모듈의 수에 따라 묶었다. 7.6 kb 읽기 하나는, 그리지 않은 다른 콘티그의 두 번째 정렬 블록이 네 번째 모듈을 더하므로 2–3개 군에 놓였다. 그래서 읽기 15개 중 4개가 모듈 넷 이상을 덮는다(그림 7C; 보충 데이터 S2; 국가 코드: MW 말라위, US 미국, IN 인도, n.r. 기록되지 않음).

대체 텍스트: 카세트 RT를 다룬 4개 패널 그림. A, 콘티그 k141_1107029(37,912 bp)의 시드 좌위 하나에 대한 유전자 지도로, 축은 RT ORF 시작점에서 잰 kb이고 유전자는 xseA 유사 유전자에서 RecD 유사 유전자까지 이어지며, 다섯 계열에 색을 입혔고 빗금 친 카세트 막대는 시드 MAG로 만든 대상(순환적)을 나타낸다. B, RT를 이름 있는 RT 계통과 견준 다섯 개의 상자 비교로, 결과는 ‘direction, not membership’(방향이지 소속이 아님)에서 ‘unresolved’(미해결)까지 있으며, 파선 윤곽은 사후(post hoc) 분석을 나타낸다. C, 좌위 22곳과 계열 다섯 개의 격자로, 온전함 74, 잘림 12, 관측 불가 24이다. D, 덮은 모듈 수(≥4, 2–3, 1, 0)별로 묶은 긴 읽기 15개.

2.1 우리의 비교로 본 고아: 어느 비교로도 알려진 계열에 배정되지 않는 RT

우리는 알려진 RT 계열이 남기고 가는 것이 무엇인지 물어서 이 RT를 찾았다. GTDB R232 [17]의 세균 종 대표 게놈 189,630개에 있는 RT 좌위 122,323개 가운데 109,524개는 알려진 RT 계열과 일치했고, 8,178개는 측정할 수 없었으며, 4,621개는 라이브러리 항목에는 이르렀으나 어떤 계열에도 이르지 못했다(갈라짐). 이 가운데 3,750개가 후보 그룹 98개를 이루었고, 그중 32개는 총조사 분할이 미배정으로 남겼으며, 98개 가운데 하나인 ESOL은 되풀이해 나타나는 같은 종류의 이웃 유전자로 눈에 띄었다(보충 표 S1, 표 H; 보충 표 S1, 표 I).

레트론, UG/Abi, Toro 2014 RT [1,4,7]의 라이브러리와 견주면 ESOL 질의 30개(총조사 그룹 구성원 14개, 추가 게놈 구성원 3개, 메타게놈 RT 13개)의 최근접 동일성은 최대 34.6%(총조사 그룹 구성원만으로는 31.7%)이고, 최근 메타게놈 조사에서 Bacteroidota에 배정된 레트론 RT 후보 306개(그중 39%는 신뢰도가 높다) [16]에 견주어도 어떤 비교도 ESOL 구성원을 레트론 그룹에 두지 않는다(보충 표 S2, 표 M).

계통수와 예측 구조도 더 나은 답을 주지 못한다. 서열 계통수에서 K 계열 RT는 단계통이 아니다. K 계열은 ESOL RT를 포함하는, 관련 RT의 더 넓은 사후(post hoc) 집합이다(표 1). 예측 구조에서 K 계열의 주 단계통군은 UG17 (DRT10) 쪽을 가리키되, 그것은 소속이 아니라 방향이다. 그리고 서열 계통수에서 이 RT의 위치는 해소되지 않았다(그림 3). DRT10에 가깝다는 점은 특이적이지 않다. 무작위 세균 RT 귀무에 견주면 사전에 정한 규칙(템플릿 모델링 점수, TM-score에 대한 규칙)을 통과하는 구조 단위가 없지만, 양성 대조 지도는 레트론이 대부분을 차지하고 이 규칙은 민감도가 낮다.

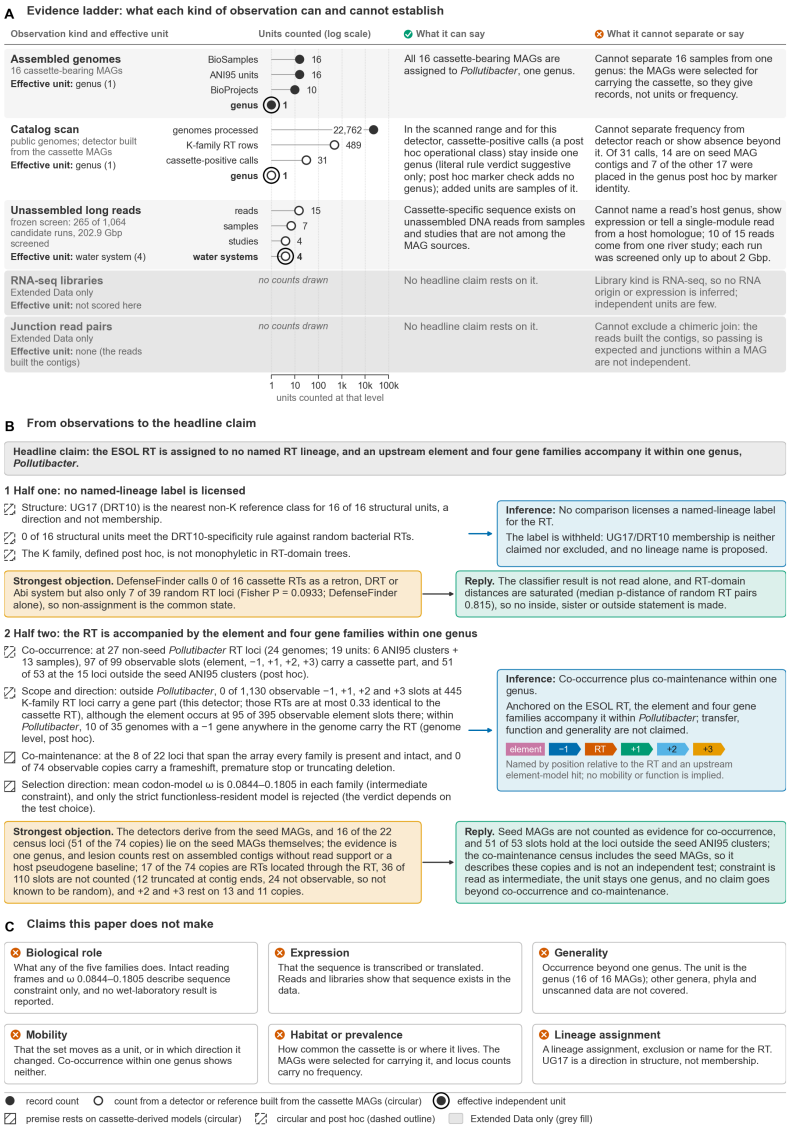
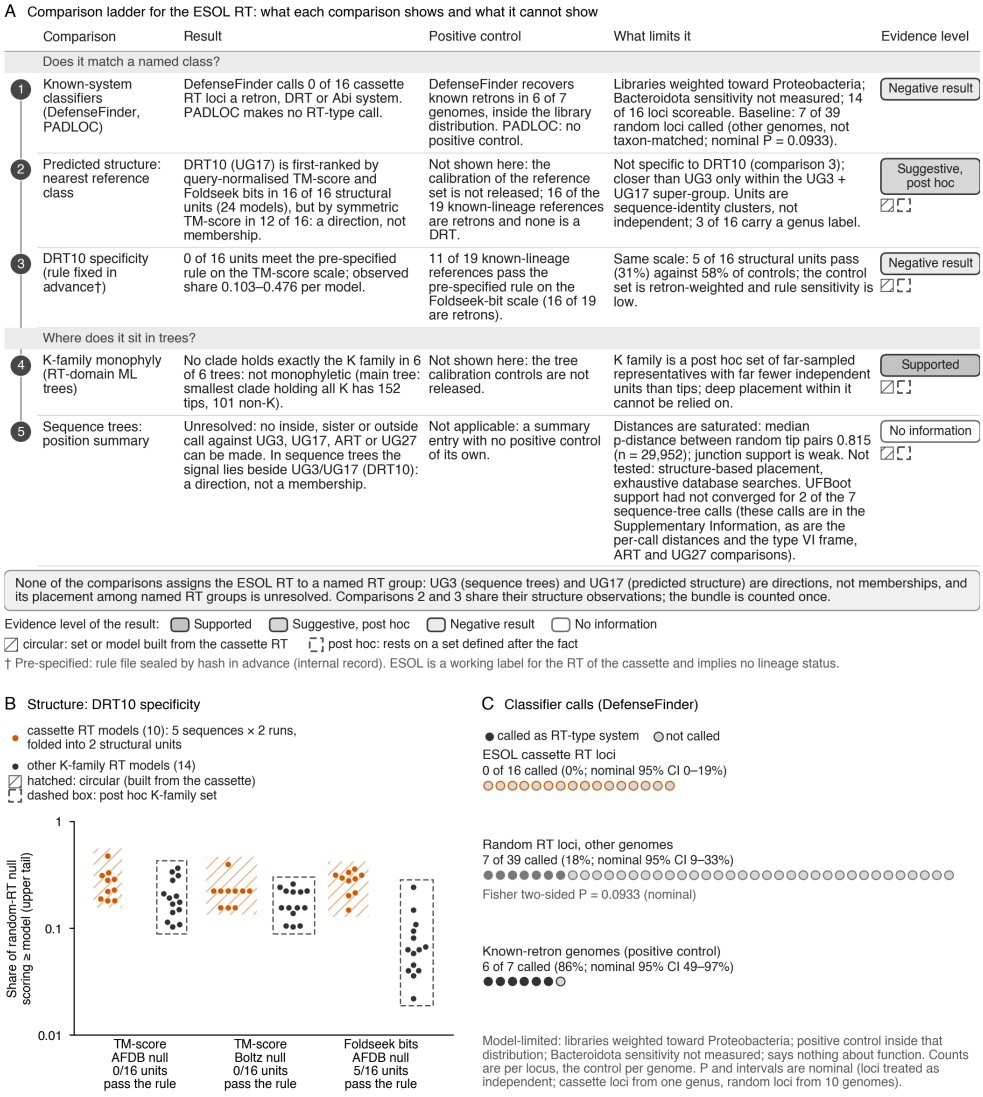


그림 2. ESOL RT와 그 카세트의 증거 사다리: 각 관측이 말해 줄 수 있는 것과 가려내지 못하는 것, 그리고 핵심 주장이 거기서 어떻게 따라 나오는지. 패널 A에서 각 사다리는 한 종류 관측의 계수 단위를 로그 눈금으로 나타내며, 수는 층별로 따로 세고 층 사이에 합산하지 않으며, 카세트 양성 판정은 사후(post hoc) 조작적 분류이다(K 계열 RT와 같은 가닥에 놓인 −1 유전자). 채운 기호는 기록 수이고, 속이 빈 기호는 카세트를 가진 MAG로 만든 검출기나 기준으로 센 수이며, 고리는 유효 독립 단위를 나타낸다. 수가 없고 회색으로 채운 두 단에는 '확장 데이터에만'이라고 표시했다. RNA-seq 계수는 확장 데이터와 2.6절에 제시하며 어떤 핵심 주장도 이 계수에 근거하지 않고, 접합부 읽기 쌍은 확장 데이터에만 제시한다. 패널 B는 핵심 주장의 각 부분마다 사슬 하나를 따라 주장을 추적하고, 각 사슬 아래에 가장 강한 반론과 우리의 답을 둔다. 빗금 친 사각형은 카세트를 가진 MAG로 만든 모형에 의존하는 전제를, 파선 윤곽은 사후(post hoc) 전제를 나타낸다. 패널 C는 이 논문이 하지 않는 주장 여섯 가지, 곧 생물학적 역할, 발현, 한 속 너머의 일반성, 단위로서의 이동성, 서식지 또는 유병률, 계통 배정이나 이름을 나열한다.

대체 텍스트: 증거를 보인 3개 패널 그림. A, 관측 종류의 표로, 계수 단위를 1에서 100k까지 로그 눈금으로 나타내고 각 관측이 말해 줄 수 있는 것과 가려내지 못하는 것을 적었다. 조립된 게놈과 카탈로그 탐색의 유효 단위는 속(1), 긴 읽기의 유효 단위는 수계(4)이며, 확장 데이터에만 있다고 표시한 두 행은 회색으로 채웠고 수가 없다. RNA-seq 계수는 확장 데이터 그림과 2.6절에 제시하며 어떤 핵심 주장도 여기에 근거하지 않는다. 채운 점은 기록 수이고, 속이 빈 원은 카세트

MAG로 만든 검출기나 기준으로 센 수이며, 고리는 유효 독립 단위를 나타낸다. B, 핵심 주장에 이르는 전제들로, 각 부분마다 가장 강한 반론과 그에 대한 답이 붙어 있다. 빗금 친 사각형은 순환적 전제를, 파선 윤곽은 사후(post hoc) 전제를 나타낸다. C, 이 논문이 하지 않는 주장 여섯 가지를 담은 상자들.



A Comparison ladder for the ESOL RT: what each comparison shows and what it cannot show

	Comparison	Result	Positive control	What limits it	Evidence level
	Does it match a named class?				
1	Known-system classifiers (DefenseFinder, PADLOC)	DefenseFinder calls 0 of 16 cassette RT loci a retron, DRT or Abi system. PADLOC makes no RT-type call.	DefenseFinder recovers known retrons in 6 of 7 genomes, inside the library distribution. PADLOC: no positive control.	Libraries weighted toward Proteobacteria; Bacteroidota sensitivity not measured; 14 of 16 loci scoreable. Baseline: 7 of 39 random loci called (other genomes, not taxon-matched; nominal P = 0.0933).	Negative result
2	Predicted structure: nearest reference class	DRT10 (UG17) is first-ranked by query-normalised TM-score and Foldseek bits in 16 of 16 structural units (24 models), but by symmetric TM-score in 12 of 16: a direction, not membership.	Not shown here: the calibration of the reference set is not released; 16 of the 19 known-lineage references are retrons and none is a DRT.	Not specific to DRT10 (comparison 3); closer than UG3 only within the UG3 + UG17 super-group. Units are sequence-identity clusters, not independent; 3 of 16 carry a genus label.	Suggestive, post hoc ◪ ⬚
3	DRT10 specificity (rule fixed in advance†)	0 of 16 units meet the pre-specified rule on the TM-score scale; observed share 0.103–0.476 per model.	11 of 19 known-lineage references pass the pre-specified rule on the Foldseek-bit scale (16 of 19 are retrons).	Same scale: 5 of 16 structural units pass (31%) against 58% of controls; the control set is retron-weighted and rule sensitivity is low.	Negative result ◪ ⬚
	Where does it sit in trees?				
4	K-family monophyly (RT-domain ML trees)	No clade holds exactly the K family in 6 of 6 trees: not monophyletic (main tree: smallest clade holding all K has 152 tips, 101 non-K).	Not shown here: the tree calibration controls are not released.	K family is a post hoc set of far-sampled representatives with far fewer independent units than tips; deep placement within it cannot be relied on.	Supported ◪ ⬚
5	Sequence trees: position summary	Unresolved: no inside, sister or outside call against UG3, UG17, ART or UG27 can be made. In sequence trees the signal lies beside UG3/UG17 (DRT10): a direction, not a membership.	Not applicable: a summary entry with no positive control of its own.	Distances are saturated: median p-distance between random tip pairs 0.815 (n = 29,952); junction support is weak. Not tested: structure-based placement, exhaustive database searches. UFBoot support had not converged for 2 of the 7 sequence-tree calls (these calls are in the Supplementary Information, as are the per-call distances and the type VI frame, ART and UG27 comparisons).	No information ◪ ⬚

None of the comparisons assigns the ESOL RT to a named RT group: UG3 (sequence trees) and UG17 (predicted structure) are directions, not memberships, and its placement among named RT groups is unresolved. Comparisons 2 and 3 share their structure observations; the bundle is counted once.

Evidence level of the result: Supported · Suggestive, post hoc · Negative result · No information
◪ circular: set or model built from the cassette RT ⬚ post hoc: rests on a set defined after the fact
† Pre-specified: rule file sealed by hash in advance (internal record). ESOL is a working label for the RT of the cassette and implies no lineage status.

B Structure: DRT10 specificity

cassette RT models (10): 5 sequences × 2 runs, folded into 2 structural units
other K-family RT models (14)
hatched: circular (built from the cassette)
dashed box: post hoc K-family set

C Classifier calls (DefenseFinder)

called as RT-type system · not called

ESOL cassette RT loci
0 of 16 called (0%; nominal 95% CI 0–19%)

Random RT loci, other genomes
7 of 39 called (18%; nominal 95% CI 9–33%)
Fisher two-sided P = 0.0933 (nominal)

Known-retron genomes (positive control)
6 of 7 called (86%; nominal 95% CI 49–97%)

Model-limited: libraries weighted toward Proteobacteria; positive control inside that distribution; Bacteroidota sensitivity not measured; says nothing about function. Counts are per locus, the control per genome. P and intervals are nominal (loci treated as independent; cassette loci from one genus, random loci from 10 genomes).

그림 3. 이름 있는 RT 그룹 사이에서 ESOL RT가 놓이는 위치에 관해 각 비교가 보여 주는 것과 보여 주지 못하는 것. 패널 A는 번호를 붙인 항목 다섯 개를 나열한다. 네 개는 ESOL RT(카세트 RT 좌위 16곳의 RT)를 이름 있는 RT 그룹과 견준 비교이고 다섯째는 이 RT의 서열 계통수 위치를 요약한 것이며, 항목마다 결과, 이 그림에 있는 경우 그 양성 대조, 그리고 그 결과를 제한하는 요인을 제시한다. 각 항목에는 그 항목이 근거하는 결과의 증거 수준이 표시되며, 결과가 순환적인 곳에는 빗금이, 사후에 정의한 집합에 근거하는 곳에는 파선 윤곽이 있다. 카세트 RT가 사후(post hoc) 정의로 속하게 되는 K 계열은 RT 도메인 최대우도 계통수에서 단계통이 아니다. K 계열의 구조 단위 16개 중 어느 것도 TM 점수 척도 두 가지에서는 DRT10 특이성 규칙을 충족하지 못하며(패널 B), 이차 Foldseek 비트 척도에서는 16개 중 5개가 통과하지만 그중 카세트 계통의 단위는 없다. DefenseFinder는 카세트 RT 좌위 16곳 가운데 어느 것도 레트론, DRT, Abi 시스템으로 판정하지 않으며(패널 C), PADLOC도 RT 유형을 판정하지 않는다. 서열 계통수에서 신호는 UG3 및 UG17 (DRT10) 곁에 놓이고, 예측 구조에서는 UG17이 가장 가까운 비K 부류이며, 둘 다 방향이지 소속이 아니다.

대체 텍스트: 이름 있는 RT 그룹 사이에서 ESOL RT가 놓이는 위치를 다룬 3개 패널 그림. A, 번호를 붙인 항목 다섯 개의 사다리로, 'Does it match a named class?'(이름 있는 부류와 일치하는가?)와 'Where does it sit in trees?'(계통수에서 어디에 놓이는가?)의 두 묶음으로 나뉘며, 항목마다 결과, 양성 대조, 그 결과를 제한하는 요인, 증거 수준 표지('Negative result'(음성 결과), 'Suggestive'(시사적), 'post hoc'(사후), 'Supported'(뒷받침됨) 또는 'No information'(정보 없음))를 제시한다. 빗금 친 사각형은 순환적 항목을, 파선 사각형은 사후(post hoc) 항목을 나타낸다. B, 세 가지 규모에 대해 0.01에서 1까지 로그 축에 그린 점 도표로, 규칙을 통과하는 단위는 0/16, 0/16, 5/16이며 카세트 RT 모형은 주황색이다. C, 분류기 판정을 나타내는 원의 행으로, 16개 중 0개, 39개 중 7개, 7개 중 6개이다.

DRT10 해석에는 두 면이 있으며, 우리는 두 면을 모두 적는다. GTDB R232 대표 프로테옴에서 K 계열 RT 좌위를 가장 가까운 참조 클래스(RT 서열 동일성만으로 배정)별로 나누어 사후(post hoc)에 보면, RT 곁의 같은 가닥 SLATT형 유전자는 가장 가까운 클래스가 UG17 (DRT10)인 좌위에서 다른 K 계열 단위보다 훨씬 자주 나타났고, 맞춘 귀무 창이나 K가 아닌 RT 곁에서는 보이지 않았다. 다만 K 계열 목록은 그 자체가 카세트에서 유래했고, DRT10 RT를 SLATT 동반으로 정의하는 문헌의 RT 집합은 독립된 뒷받침이 되지 못한다. 이 해석에 맞서는 사실도 있다. 카세트가 든 K 주 단계통군에서는 평가할 수 있었던 소수의 좌위와 속 단위 곁에서 SLATT형 유전자가 검출되지 않았고(검정력이 낮은 음성), 카세트를 가진 MAG의 프로테옴에서도 SLATT 계열 단백질이 나오지 않았으며, ESOL RT에는 프리프린트가 DRT10 RT에서 보고한 종류의 C 말단 세린·티로신 프라이밍 잔기가 없다[12–14]. 다만 불완전한 조립에서 보이지 않는다고 해서 서열이 크게 갈라진 짝이 없다고 단정할 수는 없다. 이들은 한 속 안의 기술적 비교이므로, 카세트 RT를 DRT10에 넣지 못하며 이 RT가 그런 모듈과 함께 작동한다는 것도 보이지 않는다. 그래서 이 RT는 고아이다. K 주 단계통군의 예측 구조는 UG3보다 UG17에 가까이 놓이지만, 관문, 계통수, 구조 어느 것도 이 RT를 이름 있는 계통에 배정하지 않으며, 우리는 이들로부터 새로운 계통을 추론하지 않는다(질문 1). 이 RT가 곁에 두는 무리가 더 나은 단서이다. 총조사 그룹, 사후(post hoc) 자매 후보 비교, 상류 요소에 관한 더 자세한 내용은 보충 결과 S1.1에 있다.

2.2 RT는 무리와 함께한다: 상류 요소와 고정된 위치의 유전자 네 개

많은 RT 시스템은 짝이 되는 요소와 함께 암호화되어 있으며, DRT10과 그 ncRNA, SLATT 단백질이 한 예이다. 그래서 우리는 이 고아의 곁에 무엇이 서 있는지 물었고, 무리를 발견했다. 우리는 ESOL 카세트를 위치만으로 정의한다. RT를 기준으로 −1, +1, +2, +3 위치에 놓이는 상류 비암호화 요소와 단백질 슬롯이 그것이다. 배열이 완전한 시드 좌위 하나는 그림 1A에 그렸고, 모듈별 구성은 그림 4 및 표 2에 나타냈다(구성 요소 카드는 보충 표 S3에 있다). +4 위치에는 RecD 유사 유전자가 있으나 우리는 이를 모듈로 세지 않는다. RT 둘레에는 보존된 상류 요소가 서 있다. −1에는 단백질이 있고, 우리가 평가할 수 있었던 −1 ORF는 모두 MIDAS(metal-ion-dependent adhesion site) 모티프를 지닌다. +1, +2에도 단백질이 있고, +3에는 메탈로프로테아제처럼 접히는 것으로 예측되는 단백질이 있다. +4에는 RecD 유사 헬리카제가 있고, 우리가 평가할 수 있었던 섬 사본은 모두 이 헬리카제의 Walker A/B 모티프가 온전하다(완전한 +4 ORF를 가진 Pollutibacter 게놈 8개에서 이 헬리카제는 조립체에서 발견된 유일한 RecD 유사 단백질이며, 조립되지 않은 콘티그에 있을 수 있는 두 번째 사본은 배제되지 않는다). 시드 콘티그 16개의 완전한 ORF 가운데 +1, +2, +3 단백질은 어느 것도 Pfam, NCBIfam, CDD 라이브러리의 항목과 일치하지 않는다. 이는 주석 라이브러리 수준의 결과이며, 이 단백질들의 폴드에 관한 진술이 아니다(그림 4와 표 2의 세 번째 층인 3Di는 3Di 구조 알파벳으로 쓴 예측 구조의 Foldseek 탐색이고, 코돈 모형 열은 부위별 검정 FEL(고정 효과 우도법)과 재조합 선별 GARD(재조합 검출 유전 알고리즘)를 쓴다).

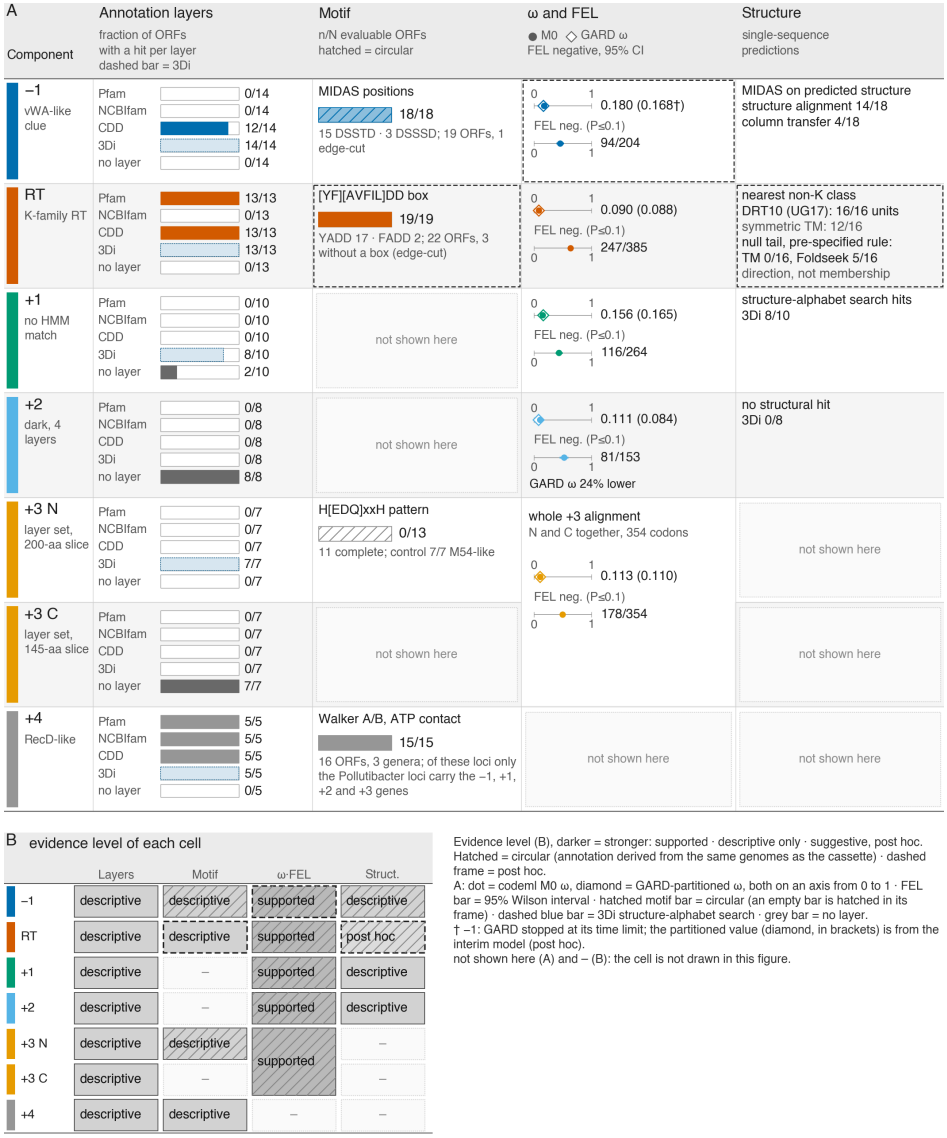


그림 4. ESOL 카세트의 구성 요소 카드 모자이크로, −1, RT, +1, +2 유전자, +3 유전자의 N 부분과 C 부분, RecD 유사 +4 유전자에 행을 하나씩 둔다. 패널 A는 행마다 탐색한 주석 층, 모티프 개수, 코돈 모형 ω와 FEL, 구조 수준의 해석을 제시한다. 패널 B는 패널 A의 모든 칸의 증거 수준을 순환 표지 및 사후(post hoc) 표지와 함께 제시한다. 증거 수준은 강한 수준일수록 짙어지는 회색 음영으로 나타내며(세 수준: 뒷받침됨, 서술에 그침, 시사적이며 사후), 빗금은 순환적인 값을, 파선 테두리는 사후(post hoc) 결과에 근거한 값을 나타낸다. 패널 A에서 파선 테두리는 RT 구조 칸(K 계열은 사후(post hoc) 집합이다), RT 촉매 상자(사본 두 개를 본 뒤에 패턴을 넓혔다), −1 ω 칸을 나타내며, −1 ω 칸의 분할된 값(†)은 잠정 재조합 모형에서 나온 것이다. 패널 A에서 파선 막대는 3Di 구조 알파벳 탐색이고, 값이 순환적일 때는 빈 모티프 막대의 테두리에 빗금을 넣었으며, 'not shown here'(여기에 표시하지 않음)로 표기된 칸은 이 그림에 그리지 않았다.

대체 텍스트: ESOL 카세트의 구성 요소 카드 모자이크를 담은 2개 패널 그림. A, 구성 요소마다 행 하나(−1, RT, +1, +2, +3 N, +3 C, +4)를 둔 표로, 열은 네 가지이다. 주석 층(히트가 있는 오픈 리딩 프레임의 비율을 막대로 나타내며, 예: 12/14), 모티프 개수(예: 19/19), 코돈 모형 ω와 FEL(M0은 점, GARD ω는 마름모로 그렸으며, 예: −1은 0.180 (0.168†)), 단일 서열 구조 예측이다. B, 칸마다 증거 수준(서술, 뒷받침됨, 사후(post hoc))을 나타낸 격자로, 수준이 강할수록 더 짙은 회색이며 순환적인 값에는 빗금이 있다. 파선 테두리는 사후(post hoc) 결과, 곧 RT의 모티프 칸과 구조 칸, 그리고 −1 ω 칸을 나타낸다.

다른 속과 메타게놈 조립체에서 온 상류 서열 15개의 정렬에서, 연속한 60개 열로 이루어진 가장 보존된 구간은 RT 시작점의 상류 25에서 84 위치에 놓이며, 가장 흔한 염기의 평균 빈도는 이 구간에서 0.77이고 열 141개 전체에서는 0.72이다(확장 데이터 그림 1). 이 요소의 정체는 미해결이며(질문 6) 그 형태 기술은 진단적이지 않다. 보충 결과 S1.2에 기술한 숙주 계통 게놈 스캔에서 요소와 RT가 짝을 이룬 경우는 게놈 20개, 종 15개, 속 4개에서 발견되며(보충 표 S4, 표 A20), 이 수는 그 스캔에서 나온 것이지 뒤에서 서술하는 게놈 794개 탐색에서 나온 것이 아니다.

표 1. RT, 카세트, 관련 RT 집합에 쓰는 용어.

용어	가리키는 대상	정의 기준
ESOL RT	ESOL 카세트의 기준점이 되는 RT(일부 패널과 범례에서는 카세트 RT라고 부른다)	카세트를 가진 MAG에서 카세트 옆에 암호화된 RT
ESOL 카세트	그 RT 주변의 배열	위치만으로 정의한다. RT를 기준으로 −1, +1, +2, +3 위치의 상류 비암호화 요소와 단백질 슬롯이며, +4 위치의 RecD 유사 유전자는 언급하되 카세트 구성 요소로 세지 않는다
모듈(구성 요소)	ESOL 카세트의 위치로 정의되는 한 부분: 상류 요소, 또는 −1, +1, +2, +3 위치의 단백질	위치만으로 정의한다. '모듈'과 '구성 요소'는 같은 것을 가리키는 말이며 함께 이동하거나 함께 기원했거나 함께 기능한다는 뜻을 담지 않는다
카세트를 가진 MAG(시드 MAG)	카세트 모형을 만든 바탕인 MAG 16개 중 하나	이 연구의 모듈 표에 있는 카세트 콘티그를 조립체가 지녔다는 이유로 선별했다. 검출이 아니라 선별이므로 순환적이며, 16개 중 14개는 카세트 양성 판정이기도 하다
카세트 양성 판정	카탈로그 스캔에서 −1 유전자가 같은 가닥에 놓인 K 계열 RT 좌위(뉴클레오타이드 창 카탈로그의 K 계열 RT 행 489개 중 31개; 그림 8, 방법). 보충 표에서는 카세트형이라고 부른다	K 계열 RT와 같은 가닥에 놓인 −1 유전자. 상류 요소나 +1, +2, +3 유전자를 요구하지 않는 사후(post hoc) 조작적 분류이므로, 이 판정이 완전한 카세트를 뜻하지는 않는다
카세트 RT 단계통군	ESOL RT와 가깝게 관련된 RT	ESOL RT와의 동일성. 우리가 탐색한 범위에서는 한 속인 Pollutibacter에 한정된다
K 계열(K 계열 RT)	관련 RT의 더 넓은 사후(post hoc) 집합으로, 대부분 먼 친척이며 ESOL RT를 포함한다. 단계통군이 아니다	프로파일 점수와 쌍 동일성 규칙. K는 우리 분석 표의 임의 라벨이며 계통 이름이 아니다
ESOL RT 군	우리의 총조사에서 ESOL RT 프로파일에 히트한 RT	점수와 길이 문턱값을 넘는 프로파일 히트

이 집합들은 범위가 서로 다르며 하나를 가리키는 다른 이름이 아니다. ESOL RT는 RT 하나이고, 카세트 RT 단계통군은 그 가까운 친척이며, K 계열은 이 둘을 모두 포함하는 훨씬 더 넓은 사후(post hoc) 집합이다. ESOL RT 군과 K 계열은 따로 정의되었고 식별자 매핑으로만 겹친다.

평가 가능한 −1 ORF는 모두 MIDAS(metal-ion-dependent adhesion site) 모티프를 지니며(18개 중 18개), 14개에서는 ESMFold로 예측한 모형 [18]에 구조를 대응시켜 다섯 위치를 읽었고 4개에서는 그 대응에서 정렬된 열을 옮겨 읽었다(확장 데이터 그림 2). +3 단백질은 메탈로프로테아제처럼 접히는 것으로 예측되지만, 이 단백질의 ORF 13개 전부(그중 11개는 완전한 ORF)에는 표준 HExxH 모티프(H[EDQ]xxH로 검색)가 없다. 반면 같은 좌위에서 나온 Peptidase_M54 유사 ORF 7개 전부와 사람 MMP1에는 같은 모티프가 있다. 이 모티프의 부재는 +3 계열 전체의 성질이다(단독으로 존재하는 +3 유사 서열도 거의 모두 이 모티프가 없고, 127개 중 2개만 속마다 하나씩 지닌다; 보충 표 S5, 표 T10, 행 H-8 및 H-9). 이 단백질이 활성이 있는 프로테아제인지는 열려 있는 질문이다(질문 4). 프로파일 대 프로파일 검색은 +3 N 말단 도메인이 M54 펩티다제의 촉매 부위 상실형이라는 해석을 지지하지 않았고(이 검색은 −1 유전자와 vWA형 단백질 사이와 비슷한 정도로 가까운 관계만 본다), 예측 구조를 사후(post hoc)에 비교하면

archaemetzincin형 접힘을 닮는다. 따라서 +3 단백질은 배정된 계열이 아니라 예측 구조의 닮음으로 기술한다. 사전에 정한 구조 수준 비교에서는 카세트의 −1과 +3 단백질이 RT 없는 −1형·+3형 단백질의 자유-자유 쌍 점수 가운데 하위 사분위 아래의 점수를 받아 사전 규칙 둘이 모두 부분적으로만 충족되었으나, 점수는 서로 무관한 접힘의 점수를 웃돌았다. 상류 요소의 구조와 공분산 모형 스캔, 촉매 모티프와 헬리카제 모티프의 온전함, −1 단백질의 구조 주석에 관한 더 자세한 내용은 보충 결과 S1.2에 있다.

표 2. RT 오픈 리딩 프레임(ORF)을 기준으로 한 위치별 ESOL 카세트의 구성으로, RecD 유사 +4 유전자는 보조 유전자(모듈 아님)로 나타내고 카세트 밖의 이웃 유전자 두 개는 위치를 가늠하도록 함께 나타낸다.

위치	유전자 계열(주석 층 Pfam / CDD / 3Di)	단백질 길이 범위(aa, ORF 수 n)	관측 가능한 사본(온전 / 전체)	모티프 결과(n/N)	코돈 모형 오메가 범위
상류 요소	요소 모형 히트(뉴클레오타이드 프로파일); 평가한 ORF 없음	해당 없음(ORF 없음); 요소 모형 히트 138–141 nt, 시드 창 16개 각각(동반 출현 총조사의 시드 좌위에서 판정하면 보충 표 S3에 48–141 nt로 나오며, 히트 하나는 창에 잘렸다)	해당 없음(ORF 없음)	해당 없음	해당 없음(ORF 없음)
−1	vWA 유사(단서); Pfam 0/14, CDD 12/14, 3Di 14/14	198–239 (15개 중 14개)	18/18 (좌위 22개)	MIDAS D-S-S-[T/S]-D 18/18 (DSSTD 15, DSSSD 3)	0.168–0.180 (서열 15개)
RT(기준점)	RT (RVT_1); Pfam 13/13, CDD 13/13, 3Di 13/13	377–425 (16개 중 12개)	17/17 (좌위 22개)	방향족-x-DD 박스 19/19 (YADD 17, FADD 2; FADD 사본을 본 뒤에 패턴을 넓혔다)	0.088–0.090 (서열 16개)
+1	배정된 계열 없음; Pfam 0/10, CDD 0/10, 3Di 8/10	261–294 (11개 중 10개)	15/15 (좌위 22개)	검정 불가(참조 모티프 없음)	0.156–0.165 (서열 14개)
+2	배정된 계열 없음; Pfam 0/8, CDD 0/8, 3Di 0/8	152–163 (10개 중 8개)	13/13 (좌위 22개)	검정 불가(참조 모티프 없음)	0.084–0.111 (서열 12개)
+3, N 말단 부분	metzincin 유사(단서); Pfam 0/7, CDD 0/7, 3Di 7/7	355–372 (8개 중 7개; +3 ORF 전체)	11/11 (좌위 22개; +3 ORF 전체)	N 말단 검사에서 H[EDQ]xxH 없음 12/12 (전체 ORF에서는 없음 13/13)	0.110–0.113 (서열 11개; +3 전체 정렬)
+3, C 말단 부분	배정된 계열 없음; Pfam 0/7, CDD 0/7, 3Di 0/7	355–372 (8개 중 7개; +3 ORF 전체)	11/11 (좌위 22개; +3 ORF 전체)	해당 없음	0.110–0.113 (서열 11개; +3 전체 정렬)
+4, RecD 유사(보조, 모듈 아님)	RecD 유사; Pfam 5/5, CDD 5/5, 3Di 5/5	592–622 (6개 중 5개)	8/8 (좌위 22개)	Walker A/B 및 ATP 접촉부류 온전 8/8	0.084–0.086 (서열 8개)
카세트 밖의 인접 유전자 (슬롯 −2 M54 유사; 슬롯 −3 xseA 유사)	−2 M54 유사(Pfam 5/5, CDD 5/5, 3Di 5/5); −3 xseA 유사 (단서일 뿐; 층 탐색 없음)	−2: 208–213 (ORF 5개); −3: 기록하지 않음	기록하지 않음	−2: H[EDQ]xxH 있음 7/7(양성 대조); −3: 기록하지 않음	기록하지 않음

위치는 조작적 정의이며 어떤 활성이나 생물학적 역할도 부여하지 않는다. vWA 유사나 메탈로프로테아제 유사 같은 표현은 단서일 뿐 배정이 아니며, xseA 유사라는 표현은 라이브러리 모형에서 읽은 주석 라벨이고 시험하지 않았다. 관측 가능한 사본과 모티프 수는 한 속의 조립된 좌위를 가리킨다. 열마다 가리키는 집합이 달라 분모가 다르며(카세트 위치의

길이 범위, 층 수, 요소 길이는 시드 콘티그 16개의 주석에서, 관측 가능한 사본은 좌위 22개에서 나온다), ω 범위는 분할하지 않은 codeml M0 추정값과 재조합으로 분할한 추정값을 모두 아우르며, −1 범위의 아래쪽 끝(0.168)은 잠정 재조합 모형의 분할된 값이고 사후(post hoc)이다. 세 층에서 모든 ORF의 호출은 보충 표 S3의 파일 SI-04_layer_calls_per_orf에 있다.

2.3 한 속 안에서는 곁의 무리가 RT와 함께 발견되고, 우리가 볼 수 있는 다른 어디에서도 그렇지 않다

곁의 무리가 RT와 한 단위를 이룬다면, RT가 나타나는 곳마다 그 곁에 서 있어야 한다. 그래서 RT를 기준점으로 삼아 슬롯마다 무엇이 서 있는지를 먼저 Pollutibacter 안에서, 다음에 그 밖에서, 그리고 구성 요소 쪽에서 세었다(그림 5). Pollutibacter에서 RT는 홀로 서 있지 않다. 시드 조립체 16개 밖의 추가 RT 좌위 27곳(게놈 24개)에서 관측 가능한 이웃 슬롯(상류 요소, −1, +1, +2, +3; 콘티그 끝 너머의 슬롯은 점수를 매기지 않는다) 99개 중 97개가 카세트 구성 요소를 가지며, 부재 2건은 한 좌위에 몰려 있다. 시드 MAG도 포함하는 ANI95 군집(평균 뉴클레오타이드 동일성 ANI로 묶은 게놈 군집; 4.2절)의 게놈을 모두 제거한 뒤에도, 남은 좌위 15곳(게놈 14개, 단위 14개)에서 요소는 9개 중 9개, −1 슬롯은 7개 중 7개, +1 슬롯은 15개 중 14개, +2 슬롯은 12개 중 11개, +3 슬롯은 10개 중 10개에서 여전히 발견된다(사후(post hoc)). 부재 두 건이 있는 좌위의 RT는 카세트 RT와 동일성이 0.388에 불과하고, 이 좌위의 상류 요소, −1, +3 위치는 관측할 수 없으며 숨은 결실도 점검할 수 없었다.

Pollutibacter 밖에서는, 이 검출기와 이 스캔 범위에 한해, K 계열 RT 곁에서 카세트 구성 요소가 발견되지 않는다. K 계열 RT 좌위 445곳에서 관측 가능한 −1 슬롯 334개 중 0개, +1 슬롯 305개 중 0개, +2 슬롯 258개 중 0개, +3 슬롯 233개 중 0개가 카세트 구성 요소를 가진다(전체 슬롯 1130개 중 0개). 밖에서 센 0은 부재의 증거가 아니다. 밖의 좌위 445곳은 거의 모두 RT와 카세트 RT의 동일성이 낮다. 그러므로 그 좌위의 0은 검출기가 그렇게 낮은 동일성에 도달하지 못한 경우와 구별할 수 없다. 카세트의 프로파일 은닉 마르코프 모형(HMM)을 쓰지 않는 두 프로파일 검출기는 밖의 좌위 445곳 어디에서도, K가 아닌 RVT_1 RT 곁의 대조 창 1,127개 어디에서도 +3형 단백질을 찾지 못한다. 하나 빼기(leave-one-out) 계열 프로파일은 두 동일성 구간에서 시험한 실제 +3형 쌍 계열 구성원을 모두 검출했다(18개 중 18개, 39개 중 39개; 보충 표 S4, 표 R01; 보충 표 S4, 표 R02). 그러므로 +3 부분에 대해 이 0은 이 검출기들, 이 좌위, 이 창에 한해 정보가 있는 부재이다. 상류 요소 모형의 히트는 Pollutibacter 밖의 K 계열 RT 곁에도 나타나기는 하지만(요소 창을 관측할 수 있는 밖의 좌위 395곳 중 95곳), 그 95곳에서 관측 가능한 −1, +1, +2, +3 슬롯 중 카세트 구성 요소를 가진 것은 하나도 없다(관측 가능한 슬롯 54개, 59개, 45개, 39개 중 각각 0개). 그래서 요소와 RT를 합친 층과 카세트 구성 요소 층은 따로 센다.

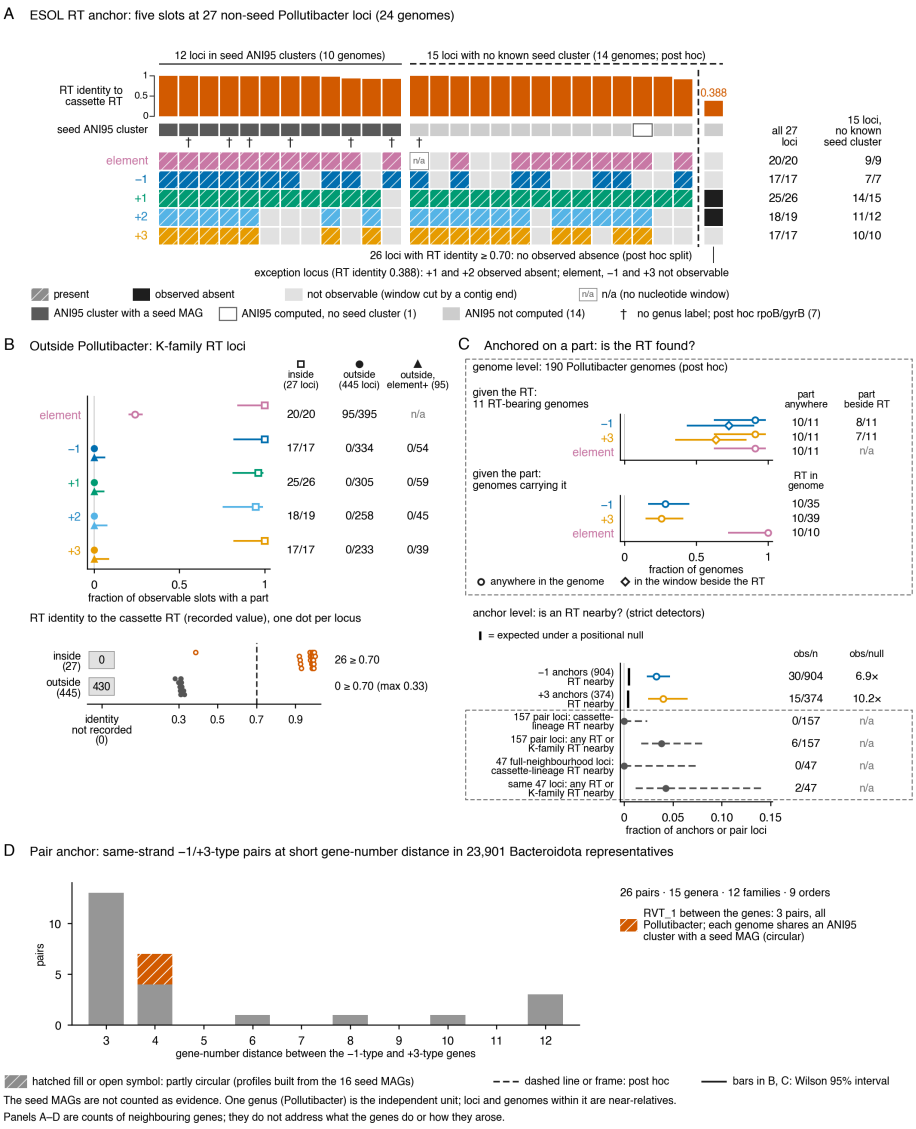


그림 5. ESOL RT의 동반 요소를 Pollutibacter 안(A), Pollutibacter 밖(B), 구성 요소 쪽(C), −1/+3형 쌍의 스캔(D)에서 세었으며, 층은 나란히 보이고 합산하지 않았다. 빗금 채움과 속이 빈 기호는 검출기 프로파일을 시드 MAG 16개로 만들었기 때문에 부분적으로 순환적인 값을 나타내고, 파선과 테두리는 사후(post hoc) 분석을 나타내며, 패널 B와 C의 가로 막대는 Wilson 95퍼센트 구간이다. (A) 비시드 Pollutibacter 좌위 27곳(게놈 24개)에서 관측 가능한 슬롯에서 상류 요소는 20개 중 20개, −1 유전자는 17개 중 17개, +1 유전자는 26개 중 25개, +2 유전자는 19개 중 18개, +3 유전자는 17개 중 17개에서 발견된다. (B) 이 검출기와 이 스캔 범위에서 Pollutibacter 밖의 K 계열 RT 옆에서는 카세트 구성 요소가 발견되지 않는다. 좌위 445곳에서 관측 가능한 −1 슬롯 334개 중 0개, +1 슬롯 305개 중 0개, +2 슬롯 258개 중 0개, +3 슬롯 233개 중 0개가 카세트 구성 요소를 가진다. 패널 B 아래의 띠는 각 좌위의 카세트 RT에 대한 기록된 동일성을 나타낸다. Pollutibacter 안의 좌위 27곳 중 26곳이 0.70 이상이고 밖의 좌위 445곳은 하나도 그렇지 않으며(최대 0.33), 밖의 좌위 430곳은 기록된 동일성이 0인데 이는 보고되지 않았음을 뜻할 수 있다. (C) Pollutibacter 게놈 단위(게놈 190개: GTDB R232의 Pollutibacter 게놈 186개와 라벨이 붙지 않은 근연 게놈 4개; 사후(post hoc))에서 RT를 가진 게놈 11개 중 엄격한 −1 유전자, 엄격한 +3 유전자, 게놈 내 어딘가의 요소를 가진 것은 각각 10개, 10개, 10개이다. 앵커 수준에서는 엄격한 −1 또는 +3 유전자 가까이에 어떤 종류의 RT든 있는 경우가 앵커 904개 중 30개와 374개 중 15개이며, 이는 위치 귀무의 6.9배와 10.2배이다. 다시 검출한 −1/+3 쌍 좌위 157곳 중 카세트 계통의 엄격한 RT를 가진 곳은 하나도 없다(주변이 완전한 47곳 중 0곳; 어떤 RT든 있는 곳은 157곳 중 6곳과 47곳 중 2곳). 쌍 좌위의 행은 사후(post hoc)이다(파선). (D) GTDB R232의 Bacteroidota 종 대표 게놈 23,901개에서, 사전 지정된 엄격한 검출기는 같은 가닥의 −1/+3형 쌍 26개를 찾으며(15개 속, 12개 과, 9개 목), 이 중 3개는 두 유전자 사이에 RVT_1 도메인이 있다.

대체 텍스트: ESOL RT의 동반 요소를 센 4개 패널 그림. A, Pollutibacter 좌위 27곳을 요소와 −1, +1, +2, +3 슬롯별로 나눈 격자로, 시드 군집에 속한 좌위 12곳의 묶음과 사후(post hoc) 묶음 15곳으로 나뉜다. 빗금 친 칸은 존재함, 검은 칸은 관측했으나 없음, 옅은 회색 칸은 관측 불가를 나타내며, 합계에는 요소 20/20과 +1 25/26이 들어 있다. B, Pollutibacter 밖에서는 좌위 445곳에서 관측 가능한 −1 슬롯 334개 중 구성 요소를 가진 것이 0개이고, 안에서는 17/17이며, 기록된 동일성의 띠는 밖의 좌위 430곳을 'identity not recorded (0)'(동일성 기록 없음 (0))로 표시한다. C, 사후(post hoc) 게놈 수준 블록과 앵커 수준 행들로, 파선 테두리의 쌍 좌위 행이 네 개 있다(0/157, 6/157, 0/47, 2/47). D, 쌍 26개를 유전자 수로 잰 거리(3에서 12까지)별로 나타낸 막대그래프.

Pollutibacter 안에서는 RT를 가진 게놈이 구성 요소를 가지는 빈도가, 구성 요소를 가진 게놈이 RT를 가지는 빈도보다 높다. 조사한 게놈 190개 중 RT를 가진 게놈은 11개이고, 그중 10개가 엄격한 −1 유전자를 가진다. 반면 엄격한 −1 유전자를 가진 게놈 35개 중 RT를 가진 것은 10개뿐이다. +3도 같은 비대칭을 보인다(+3 유전자를 가진 게놈 39개 중 10개가 RT를 가진다). 이 게놈 수준의 개수는 RT를 기준점으로 삼은 총조사와 같은 게놈에서 나온 것이므로, 같은 관찰을 다시 센 것이지 독립적인 증거가 아니다. 총조사의 범위, RT 동일성에 따른 층화, 검출기 도달 범위와 대조 창, 게놈 수준 공존 집계에 관한 더 자세한 내용은 보충 결과 S1.3에 있다.

2.4 우리가 볼 수 있는 사본은 하나도 파괴되지 않았고, 코돈 모형은 정화 제약과 일치한다

퇴화하는 카세트라면 망가진 사본을 지니고 있을 것이고, 유지되는 카세트라면 그렇지 않을 것이다. 그래서 우리는 다섯 계열의 관측 가능한 사본을 모두 프레임이동, 조기 종결 코돈, 절단 결실 여부로 판정했고, 선택이 그 코돈을 얼마나 강하게 제약하는지 물었다(그림 6; 변형과 대조는 확장 데이터 그림 3, 확장 데이터 그림 4, 확장 데이터 그림 5에 있다). 관측 가능한 모든 모듈 사례에서 조립된 콘티그는 프레임이동도, 조기 종결 코돈도, 절단 결실도 보이지 않는다. 관찰된 결함 수는 중립 부동 기대값보다 훨씬 낮고 검출 양성 대조는 통과했으므로, 퇴화 예측의 엄격한 형태는 기각된다(그림 6; 보충 데이터 S1). 이 기각은 관측한 계통수 위에서 조기 종결 코돈을 정확히 계산한 결과에 기대며, 종결 코돈 빈도가 시리즈에 따라 0.29~1.10퍼센트를 넘을 때에만 성립한다(보충 표 S6, 파일 SI-10_A6_break_even_stop_frequency). 종결 코돈 빈도를 관측한 코딩 조성에서 정하면(종결 코돈이 관측되지 않음; 0.022~0.074퍼센트) 모든 잎에 조기 종결 코돈이 없을 확률이 20개 칸 모두에서 0.35~0.75이고 아무것도 기각되지 않는다. 따라서 이 계산은 종결 코돈 획득이 선택으로, 또는 조성이 담지 못하는 돌연변이 편향으로 억제되는 모형에 대해서는 정보를 주지 못한다. 사후(post hoc) 민감도 점검에서 MAG 좌위 16개 뒤에 있는 연구 10개 가운데 어느 하나를 빼더라도 관측 가능한 이웃 모듈 사본 57개 중 최소 48개가 남고 그중 파괴된 것은 하나도 없으며(양측 정확 상한은 48개에서 0.0740), 연구당 좌위 하나만 남기고 공개 게놈 좌위 6곳을 더하면 사본 37~46개가 남고 파괴된 것은 하나도 없다(상한은 37개에서 0.0949). 이 점검은 분모만 바꿀 뿐 파괴된 사본의 수는 바꾸지 않는다.

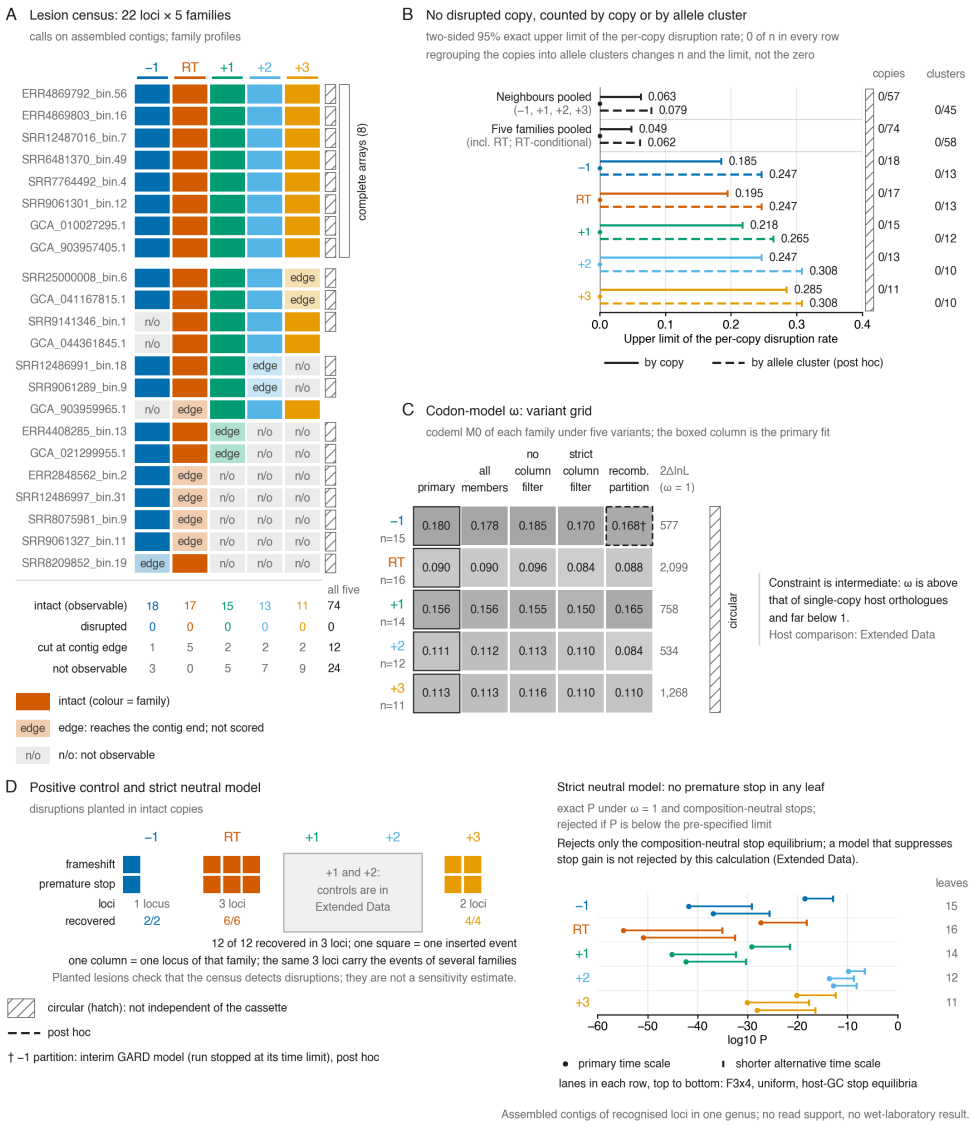


그림 6. 한 속의 조립된 좌위 22곳에서 5개 카세트 계열의 관측 가능한 사본 중 어느 것에도 프레임이동, 조기 종결 코돈, 절단 결실이 보이지 않으며, 5개 계열의 코돈 모형 ω(codeml M0) 값은 C에 제시한 5개 변형 각각에서 0.084에서 0.185 사이에 있다. A, 좌위 22곳(MAG 좌위 16곳과 공개 게놈 좌위 6곳; 행, 완전한 배열이 먼저)을 5개 카세트 계열(열)별로 온전함, 콘티그 끝에서 잘림(edge), 관측 불가(n/o) 중 하나로 점수화했다. B, n개 사본 중 0개가 파괴된 경우 사본당 파괴율의 정확한 양측 95% 상한이며, 사본 기준은 실선, 대립유전자 군집 기준은 파선으로 그렸다(사후(post hoc)). C, ANI95 대표 집합에서 점유율이 50%를 넘는 정렬 열만 남긴 경우(trim50; 주 분석, 상자로 표시, 왼쪽에 서열 수 n), 모든 구성원, 열 필터 없음, 틈 없는 열만 쓴 경우(strict), GARD 재조합 분할(분절별 ω의 코돈 길이 가중 평균; −1의 분할은 실행이 시간 제한에서 멈췄기 때문에 잠정 GARD 모형에서 가져온 것이며 사후(post hoc)이다, †) 각각에서 계열별 코돈 모형 ω(codeml M0). A, B, C의 빗금 무늬는 순환적 항목을 나타내며, 계열 프로파일을 카세트 MAG에서 만들었기 때문이다. A에서 이 표시는 보충 데이터 S1에서 순환적이라고 표지된 좌위 22곳 중 20곳(시드 MAG 좌위 16곳과 공개 게놈 4개)에 있고, 괄호는 완전한 배열 8개를 나타낸다. D 왼쪽, 6프레임 검출기는 심어 넣은 병변 12개 중 12개를 회수했다. 이는 검출기 배선을 점검할 뿐 다른 위치에서의 민감도는 추정하지 않으며, +1과 +2의 대조는 확장 데이터 그림 5에 있다. D 오른쪽, 각 시리즈의 관측 트리에 대한 정확한 가지치기 계산(모의실험이 아님)은 엄격한 중립 모형(종결 코돈 획득을 포함해 모든 비동의 변화에서 ω = 1, 종결 코돈은 조성 중립 평형; 칸 30개의 값은 이 그림의 원천 데이터에 있다)에서 모든 잎의 모든 위치에 조기 종결 코돈이 없을 확률 P를 산출한다.

대체 텍스트: 5개 카세트 계열의 병변을 다룬 4개 패널 그림. A, 좌위 22곳과 −1, RT, +1, +2, +3의 격자로 온전한 칸에 색을 입혔고, edge는 콘티그 끝에서 잘림, n/o는 관측 불가를 나타낸다. 괄호는 완전한 배열(8개)을 나타내며, 합계는 온전함 74, 파괴됨 0, 콘티그 끝에서 잘림 12, 관측 불가 24이다. B, 사본당 파괴율 상한의 막대로 0.049에서 0.308까지이며, 사본 기준은 실선, 대립유전자 군집 기준은 파선(사후(post hoc))이다. C, 변형 다섯 개의 코돈 모형 ω로 0.084에서 0.185까지이며, −1 recomb. 분할의 0.168†에 파선 테두리가 있다. D, 심어 넣은 병변 12개 중 12개를 회수했고, log10 P 구간은 −60에서 0까지이다.

코돈 모형에서 모든 모듈의 ω는 중립값에서 정화 선택 방향으로 멀리 떨어져 있어 정화 선택과 일치하며, 많은 코돈에 음성 선택 부위가 있으며, 동의 치환 속도 변이가 없는 FEL에서는 양성 선택 부위가 많아야 하나이다(−1에서, 분할하지 않은 분석에서만; 동의 치환 속도 변이가 있으면 계열당 많아야 둘이다). 촉매 H[EDQ]xxH 모티프가 없는 +3 계열도 같은 방향을 보이지만 그 기능은 정해지지 않은 채 남아 있다(보충 표 S6; 확장 데이터 그림 3). +3 정렬에 대한 FEL 출력을 사후(post hoc)에 점검하면, 동의 치환 속도 α는 두 HyPhy 빌드 모두에서 354개 부위 중 349개에서 정확히 1.000이므로 여기서는 부위별 동의 치환 속도가 풀리지 않으며, 음성 판정은 α를 1.000으로 고정한 상태에서 β가 1.000보다 작다는 뜻이다. α 패턴의 원인은 규명되지 않았다. 제약은 중간 정도이다.

2.5 조립되지 않은 단일 긴 읽기에 카세트의 모듈이 여럿 실려 있다

조립체에서만 보이는 서열은 조립 과정의 인공물일 수 있다. 그래서 우리는 카세트가 공개 시퀀싱 런에서 나온 조립되지 않은 단일 긴 읽기에 존재하는지, 그런 읽기가 키메라일 수 있는지를 물었다(그림 7; 확장 데이터 그림 6, 패널 D). 공개 담수 긴 읽기 자료의 동결된 층에서, 4개 연구의 시료 7개에서 나온 조립되지 않은 읽기 15개가 카세트와 일치한다. 이 시료와 연구는 어느 것도 참조에 게놈을 제공하지 않았고, 서열을 뒤집은 디코이에서는 히트가 나오지 않았으며, 읽기는 BLASTN, minimap2, 엄격한 단백질 blastx로 다시 확인했고, 이 읽기의 숙주 속은 정해지지 않았다. 모든 층위를 합치면 공개 단일 긴 읽기가 카세트 모듈 여섯 개를 모두 덮고 recC/recD 쪽에서는 카세트 너머까지 이른다. 이 사후(post hoc) 층위는 히트를 본 뒤에 정했으며 무작위 표본 통계에 들어맞지 않는다.

두 분자가 무작위로 결합할 때 나오는 다중 모듈 읽기(모듈 둘 이상을 덮는 읽기)는, 카세트 읽기가 모두 키메라라 해도 최대 1.61×10^{-4}개로 예측된다. 그런데 공개 긴 읽기 자료에는 다중 모듈 읽기가 34개 있고, 그중 10개는 동결 층위에 있다(그림 7; 보충 데이터 S2, 파일 SI-12_RANDOM_PAIRING_TOTALS). 이 비교는 무작위 짝짓기에서만 성립하며, PCR로 선택된 라이브러리의 비무작위 짝짓기는 시험하지 않았다. 이 긴 읽기는 4개 수계의 런 9개에서 나왔고, PCR로 선택된 연구 하나가 다중 모듈 읽기 34개 중 31개와 카세트 읽기 49개 중 43개를 제공한다(보충 데이터 S2, 파일 SI-12_RANDOM_PAIRING_BY_RUN). 사전 지정된 엄격 변형은 카세트를 가진 콘티그 16개보다 많은 콘티그를 담은 참조에서 같은 MAG의 콘티그 둘에 정렬되는 모든 읽기를 참조와 어긋난 것으로 세는데, 어긋난 비율은 다중 모듈 읽기 34개 중 2개(0.059), 어느 종류든 모듈 둘 이상을 덮는 읽기 46개 중 5개(카세트 특이적 모듈, 숙주 인접 모듈, 또는 −2와 −4 곁 유전자; 0.109), 읽기 49개 전체 중 5개(0.102)이다. 이것은 키메라에 관한 발견이 아니라 민감도 분석이며, 이 5개 읽기는 모두 한 연구의 시료 4개에서 나왔다(보충 데이터 S2, OBS_READS). 다중 모듈 읽기, 참조 밖의 시료와 연구에서 나온 접합부 통과 읽기, 무작위 짝짓기와 동결 층위의 비교에 관한 더 자세한 내용은 보충 결과 S1.5에 있다.

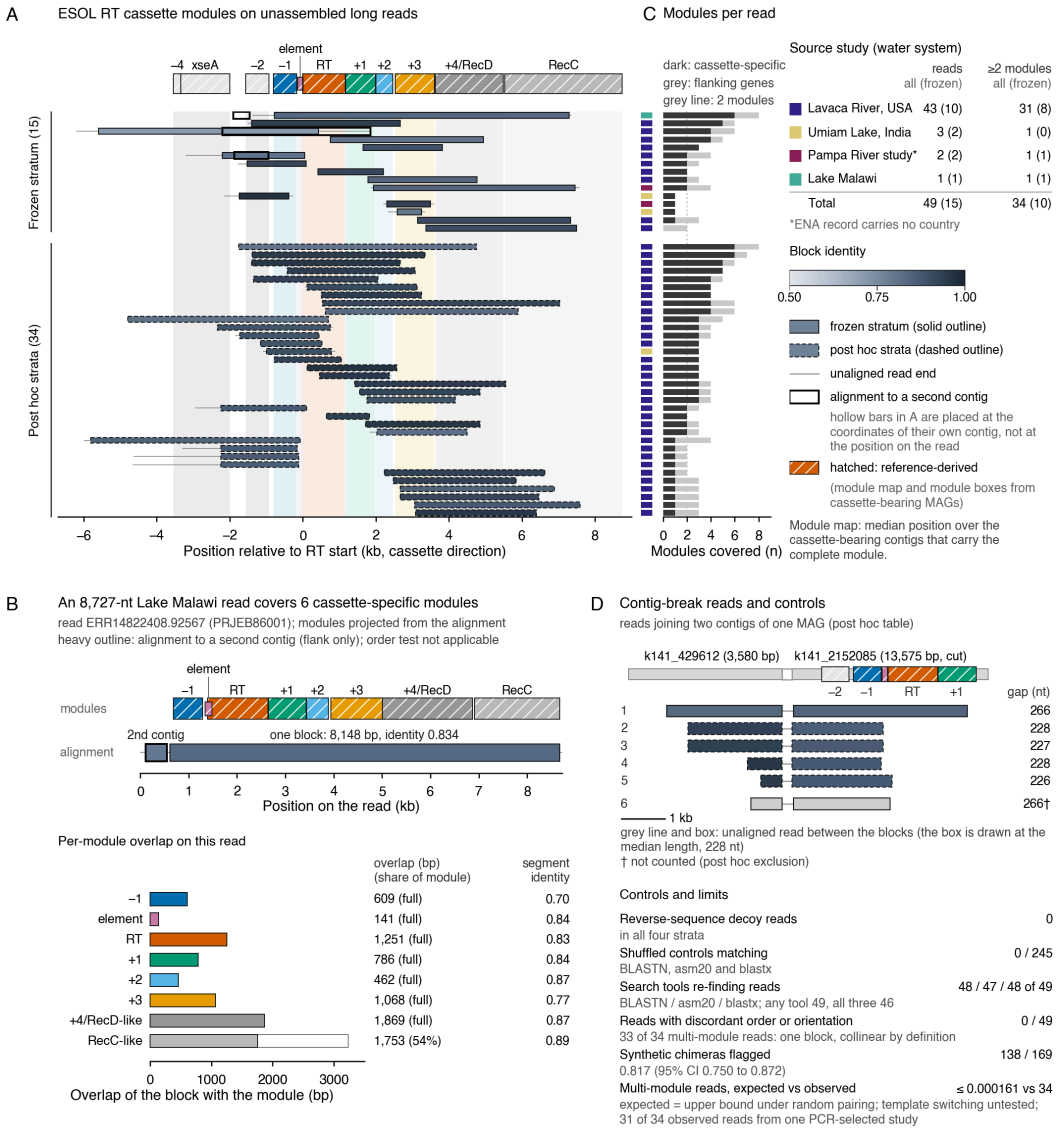


그림 7. 조립되지 않은 단일 긴 읽기 위의 ESOL 카세트. (A) 각 행은 공개 Oxford Nanopore 읽기 49개(동결 선별에서 15개, 사후(post hoc) 층위에서 34개) 중 하나이다. 이 읽기의 유지된 정렬 블록은 카세트를 가진 메타게놈 조립 게놈(MAG)에서 나온 참조 콘티그 8개 위에 놓인다. 읽기는 카세트 방향으로 그렸고, 위치는 RT 시작점 기준의 kb로 나타냈으며, 각 블록은 동일성에 따라 색을 입혔다. 행 위의 모듈 지도와 B, D의 모듈 상자에 빗금을 친 것은, 이들이 카세트를 가진 MAG에서 나온 참조 콘티그의 주석이기 때문이다. (B) 8,727 nt 길이의 말라위 호수 읽기에는 동일성 0.834의 주 정렬 블록이 하나 있다. 이 블록은 8,148 bp(정렬 열로 센 값이며, 읽기 8,062 nt 대 참조 8,067 bp)로, 카세트 특이 모듈 여섯 개(색), +4/RecD 유사 유전자, RecC 유사 유전자의 일부(회색)를 덮는다. (C) 막대는 각 읽기가 덮는 카세트 특이 모듈(진한 색)과 측면 유전자(회색)를 A의 행 순서대로 보여 준다. 색 칩과 표는 출처 연구, 읽기 수, 다중 모듈 읽기 수를 모든 층위와 동결 층위(괄호 안)에 대해 보여 준다. (D) 한 MAG의 콘티그 두 개를 잇는 읽기는 두 블록 사이의 정렬되지 않은 구간과 함께 그렸다(사후(post hoc) 표). 대조와 한계는 역서열 디코이 읽기, 섞은 읽기 대조, 검색 도구 세 가지로 읽기를 다시 찾은 결과, 순서나 방향이 어긋난 읽기, 표지된 합성 키메라, 다중 모듈 읽기의 기대 수 대 관측 수를 보고한다.

대체 텍스트: 조립되지 않은 긴 읽기 위의 카세트를 보인 4개 패널 그림. A, 읽기 49개를 읽기당 가로 막대 하나로 그린 것으로, RT 시작점 기준 −6에서 8 kb 축 위에 빗금 친 모듈 지도와 함께 놓았다. 동결 층위의 읽기 15개는 실선 윤곽, 사후(post hoc) 읽기 34개는 파선 윤곽이며, 블록 동일성(0.50에서 1.00)에 따라 음영을 넣었다. B, 8,727 nt 길이의 말라위

호수 읽기 하나로, 동일성 0.834의 주 블록 하나(8,148 bp; 두 번째 콘티그 위에 짧은 블록이 하나 더 있다)가 카세트 특이 모듈 여섯 개를 덮으며, 모듈별 겹침을 막대로 나타냈다. C, 읽기당 덮은 모듈 수의 막대와 출처 연구의 표로, 합계는 읽기 49(15)개이다. D, 틈이 266, 228, 227, 228, 226, 266† nt인 콘티그 끊김 읽기 6개와 대조의 표.

2.6 공개 RNA와 배양 DNA의 흔적: 있기는 하지만 RNA와 DNA는 구분되지 않는다

Lake Erie, 베이징의 배양 계통, 그리고 결과를 본 뒤에 만든 Sandusky Bay와 Wyoming 연장의 공개 RNA 시퀀싱 런에는, 미리 고정한 계수 규칙 아래 카세트 창과 일치하는 읽기가 들어 있다. 이것이 보여 주는 것은 그 라이브러리에 일치하는 서열이 있다는 점이지 카세트가 전사된다는 점이 아니다(확장 데이터 그림 6, 패널 A와 B). 첫 표적 신호를 본 뒤 공개 담수 메타전사체 런 하나(SRR6466487)를 전체 깊이로 다시 사상한 사후(post hoc) 분석에서는, 카세트와 높은 동일성을 보이는 서로 다른 세 단편(읽기 여섯 개)을 얻었고 디코이 읽기는 하나도 없었다. 이것이 보여 주는 것은 그 런에 카세트를 닮은 서열이 있다는 점뿐이다. 같은 런을 보충 표 S7의 계수 규칙에 따라 전체 깊이로 다시 훑으면 일치하는 단편이 여섯 개로 세어지는데, 두 수는 비교할 수 없다(보충 표 S8). 이들은 한 시료의 서로 다른 분자일 뿐 독립 시료, 지점, 계통이 아니다. 한 호수 시리즈(Beaver Lake)의 가닥 특이적 총 RNA 라이브러리를 사후(post hoc)에 보면, 카세트 특이 읽기가 숙주 mRNA와 같은 가닥에 놓이는데, 이는 이중가닥 DNA 오염만으로는 생기지 않는 불균형이다. 그러나 이것으로 RNA와 단일가닥 DNA를 가를 수는 없고, 이 가닥 편향은 라이브러리 키트에 좌우되며 다른 런에서도 보이고, 증거는 호수 시리즈 하나에서 나온다. 한 연구의 단일 배양 시리즈인 베이징 시리즈의 가닥 비특이적 라이브러리에서는 숙주 읽기에 대한 카세트 읽기의 밀도가 DNA만으로 기대되는 수준을 넘지 않는다. 그러므로 공개 시퀀싱 읽기로는 카세트와 일치하는 읽기가 RNA에서 왔는지 DNA에서 왔는지 지금은 가릴 수 없다.

한 연구의 실험실 배양 시리즈(전체 게놈 샷건(WGS) 런 4개)의 DNA에서, 별도로 조립한 결과물이 4,566 nt 콘티그 위에 카세트 RT 단계통군 좌위 하나를 지닌다. 이 콘티그의 모듈 단백질 다섯 개는 RT를 포함해 모두 런 SRR9061301의 시드 MAG 단백질과 아미노산 서열이 100.0퍼센트 동일하다(모듈 콘티그에서 뉴클레오타이드 동일성 99.93퍼센트; 사후(post hoc) 비교; 보충 표 S7, 표 T12). 이것은 한 연구, 한 속에서 나온 존재 관찰이며, 검출기는 시드 MAG에서 만들었고, 숙주, 기원, 기능은 배정하지 않는다.

공개 긴 읽기 RNA 런을 거른 결과는 어느 방향으로도 정보를 주지 못한다. 후보 런은 앰플리콘 라이브러리와 작은 자료 집합이 대부분이었고, 숙주 계통의 읽기를 담은 것은 거의 없었으므로, 그 안에 카세트 읽기가 없다는 사실은 보이지 않은 것과 없는 것을 가르지 못한다. 카세트를 가진 MAG를 낸 연구 10곳에서는 RNA 시퀀싱 런이 발견되지 않았고(검색 시점에 유럽 뉴클레오타이드 아카이브(ENA)에서 조회), 표본에 포함된 어느 카세트 좌위의 원천 시료와도 BioSample을 공유하는 RNA 런이 없었으므로, 발현은 여기서 공개 자료로 시험되지 않았다. 이 조회가 다른 곳에 RNA 자료가 없다는 것을 보여 주는 것은 아니다. 그러므로 우리가 거를 수 있었던 공개 자료는 카세트가 전사되는지 말해 주지 못한다. 다음 질문은 카세트가 세계의 어디에서 발견되었는가이다. 공개 RNA 자료를 찾은 이유, 연못 RNA 읽기 검색, 원천 연구와 시료의 RNA 런에 대한 보관소 조회에 관한 더 자세한 내용은 보충 결과 S1.6에 있다.

2.7 세계의 어느 곳에서: 다섯 대륙의 호수, 강, 연못, 그리고 한 속

곁의 무리는 어디에 나타나는가? 우리는 이 무리를 지닌 게놈과 읽기의 공개 기록으로 눈을 돌렸다(그림 8). 카세트를 가진 게놈은 5대륙 9개국의 호수, 강, 연못에서 나타난다. 서로 다른 시료 16개에서 나온 게놈 16개이며, 이 시료는 제출 프로젝트 10개에 속한다(캐나다, 스위스, 한국, 칠레, 르완다, 이스라엘, 미국, 인도, 중국; 대륙은 기록된 국가에서 도출했다) (그림 8). 정식 GTDB-Tk 분류는 카세트를 가진 MAG 16개를 모두 Pollutibacter에 배정하며(이 MAG는

조립체가 카세트 콘티그를 지녔다는 이유로 선별했고, 그중 14개는 카탈로그 스캔에서도 카세트 양성으로 판정되었다; 4.2절; 보충 표 S9, MAG_SOURCES), 이 16개는 ANI 수준에서 구별되는 단위이고 별개의 시료 16개에서 나왔으며, 고품질 조립체는 그중 절반뿐이다. 이 기록이 보여 주는 것은 카세트를 가진 게놈이 어디에서 발견되었는가이지 카세트가 얼마나 흔한가가 아니다.

16개 기록 가운데 어느 것도 자유 텍스트에 토양, 퇴적물, 숙주, 폐수, 해수를 적지 않았으며, 이는 이 기록들을 기술할 뿐 다른 서식지에 대해서는 아무것도 말해 주지 않는다. 더 넓은 Pollutibacter 집합에서는 기록이 모두 환경 시료 기록은 아니다. 조사한 게놈 뒤의 고유 BioSample 190개 중 114개가 'sample type = metagenomic assembly'를 달고 있고, 기록 16개에는 하나도 없다. 그래서 더 넓은 집합에서는 국가, 좌표, 환경 서술이 MAG 제출 기록에서 온 것이며 원래의 시료 기록은 추적하지 않았다. 이 속의 바탕 서식지는 이미 담수이다(GTDB R232의 Pollutibacter 게놈 186개 중 152개가 담수 표지를 달고 있는데, 이 표지는 용어 사전으로 읽어 붙인 것이고 그 판독은 수동 예비 점검을 통과하지 못했으므로 여기서 서식지에 관한 진술은 모두 사후(post hoc)이다). 그러므로 카세트가 이 속의 다른 서식지보다 담수를 선호하는지는 이 자료로 검정할 수 없다(확장 데이터 그림 7).

우리가 탐색한 범위 안에서 이 검출기로 볼 때, 우리가 소속을 정할 수 있는 카세트를 가진 게놈은 모두 Pollutibacter이며, 이것은 카세트가 다른 곳에 없다는 증거가 아니다. RT와 카세트의 짝은 한 속에 묶여 있는가, 아니면 그 밖에서 보이지 않았을 뿐인가? 일치하는 공개 긴 읽기, 출처 시료 기록, 카탈로그 스캔 좌위에 관한 더 자세한 내용은 보충 결과 S1.7에 있다.

2.8 한 속에서 만나는 두 조각: −1/+3 쌍은 RT 없이도 되풀이되고, 요소는 RT와 함께 발견된다

반대 방향의 물음도 그만큼 중요하다. 이 부분들은 존재하는 데 RT가 필요한가? 우리는 −1 유전자와 +3 유전자가 주변 창에 RT 히트 없이 한 쌍으로 나타나는지, 그리고 그 쌍의 상류에 공유된 조절의 표지가 될 만한 것이 있는지 물었다(확장 데이터 그림 8). 이 검출기의 범위 안에서는 필요하지 않다. −1/+3 유전자 쌍은 확증 집단의 앵커 속 149개(+3형 유전자를 가진 속) 중 54개에서 창 안에 RT 없이 나타난다(사전 지정된 규칙 충족). "강하다"는 말은 되풀이가 존재한다는 사실에만 붙는다. 그 존재는 주석이 달린 공개 Bacteroidota 게놈 중 치우친 비무작위 부분 집합에서 본 것이며, 기능과 기원은 시험하지 않았다(확장 데이터 그림 8; 보충 표 S10).

더 엄격한 탐색이 같은 질문을 GTDB R232의 Bacteroidota 종 대표 23,901개 전체에 대해, 미리 고정한 검출기로 던졌다. 그 결과 같은 가닥의 −1/+3형 유전자 쌍 26개가 나왔고(15개 속, 12개 과, 9개 목; 쌍 중 23개는 카세트 계통 속 바깥의 14개 속에 있다), 두 유전자 사이에 RVT_1 도메인이 있는 쌍 3개는 모두 Pollutibacter에서 카세트 배열로 나타난다. 이 개수는 이 검출기들에 속한 것이며, 검출기는 카세트에서 유래했고(더 넓은 프로파일로는 26개 속에서 쌍 55개가 나온다; 보충 표 S4, 표 B02 (config_id), 검출기 변형 4), 이 탐색은 단일 속 결과를 독립적으로 확인한 것이 아니다. 인접 쌍의 되풀이를 이 더 넓은 간격으로 사후(post hoc)에 확장하면, 카세트 계통 속 바깥의 속 수는 회전 귀무에서 기대되는 1.26을 11.1배 넘는다. 그러나 귀무는 짧은 거리에서만 보정되었고, 쌍 가운데 다수는 또 다른 인접 쌍과 구성원을 공유하며, 속은 독립 단위가 아니다. 또 다른 인접 쌍과 구성원을 공유하지 않는 쌍으로 한정해도 카세트 계통 바깥에서 속 9개(쌍 10개)가 남는다.

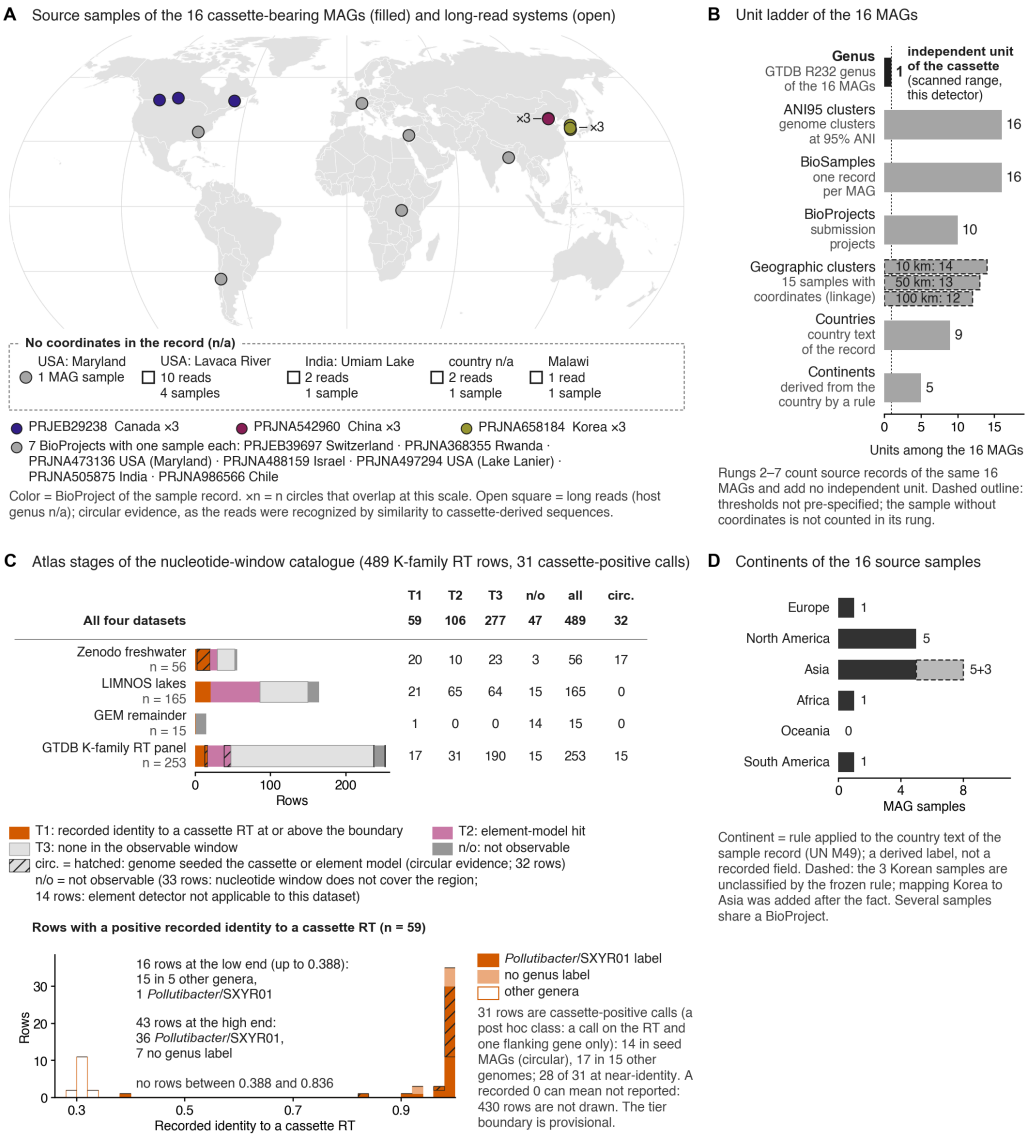


그림 8. 카세트를 가진 MAG 16개의 출처 기록(지도, 단위 사다리, 대륙)과 뉴클레오타이드 창 카탈로그의 K 계열 RT 행 489개의 층위별 분류. A, 카세트를 가진 MAG 16개의 출처 시료를 공개 BioSample 기록에 적힌 좌표에 그렸다(시료 16개 중 15개). 시료마다 원 하나를 그렸다. 시료를 3개씩 제공하는 BioProject 3개는 BioProject별로 색을 달리하고, 시료를 1개 제공하는 BioProject 7개는 회색으로 칠했다(BioProject 10개). 메릴랜드 시료는 좌표 쌍이 없어 지도 아래의 파선 띠에 표시했으며, 이 띠에는 우리가 가진 기록에 좌표가 없는 긴 읽기 수계 4곳도 함께 두었다. B, MAG 16개를 일곱 가지 입도로 센다: 속 하나, ANI95 게놈 군집, BioSample, BioProject, 좌표가 있는 시료 15개의 지리적 군집, 국가, 대륙. C, 위쪽에는 뉴클레오타이드 창 카탈로그의 K 계열 RT 행 489개(행은 원천 표 하나에 있는 K 계열 RT 유전자 하나이며, 행 중 31개는 카세트 양성 판정으로, 사후(post hoc) 조작적 분류이고 전체 배열이 아니라 RT 가닥의 −1 유전자로 정의되며 31개 모두 T1에 있다)를 데이터셋과 층위별로 나타냈다. T1, 카세트 RT에 대한 기록된 동일성이 층위 경계 이상인 경우(행 59개, 그중 15개는 다른 속에 있고 동일성이 0.33 이하이므로 층위는 단계통군이 아니다); T2, T1 없이 RT 상류에 요소 모형 히트가 있는 경우(행 106개); T3, 관측 가능한 창에서 이 둘 다 없는 경우(행 277개); 그리고 관측할 수 없는 행 47개(n/o); circ.는 시드 MAG의 행이나 검출기가 순환적인 게놈의 행(32개)을 나타낸다. 아래쪽에는 기록된 동일성이 있는 행 59개의 카세트 RT와의 기록된 동일성을 속 표지별로 색을 달리해 나타냈다. 기록된 동일성이 0인 행은 보고되지 않았음을 뜻할 수 있어 그리지 않았으며, 층위 경계는 잠정적이다. D, 카세트를 가진 MAG 시료 16개를 BioSample 기록의 국가 문구에서 도출한 대륙별로 센다. 이렇게 세면 이름이 붙은 대륙은 다섯이며, 한국 시료 3개는 동결한 규칙으로는 분류되지 않고 한국을 아시아에 넣은 것은 나중에 덧붙였다(파선 구간, 아시아 5+3).

대체 텍스트: 카세트를 가진 MAG 16개의 출처를 보인 4개 패널 그림. A, 출처 시료를 BioProject별로 색을 달리해 그린 세계 지도(캐나다 ×3, 중국 ×3, 한국 ×3, 시료가 하나뿐인 BioProject 7개는 회색)로, 긴 읽기는 속이 빈 사각형, 좌표가 없는 시료는 파선 상자로 나타냈다. B, 단위별 막대로 속 1, ANI95 군집 16, BioSample 16, BioProject 10, 국가 9, 대륙 5이며, 10, 50, 100 km 기준 지리적 군집 14, 13, 12는 파선이다. C, K 계열 RT 행 489개를 층위별로 센 막대(T1 59, T2 106, T3 277, n/o 47)와 기록된 동일성 59개의 히스토그램. D, 대륙별 수: 유럽 1, 북아메리카 5, 아시아 5+3(파선 부분: 한국 시료 3개), 아프리카 1, 오세아니아 0, 남아메리카 1.

상류 요소는 반대로 행동한다. 요소를 기준점으로 삼아 게놈 794개(Pollutibacter 게놈 190개 전부, 요소를 가진 속의 대표, 그리고 게놈 150개와 139개로 이루어진 음성 집단 둘; 이 선택은 보충 결과 S1.2에서 스캔한 게놈과 일부만 겹친다; 보충 표 S4, 표 A13)에서 사후(post hoc)에 찾았을 때, 요소 히트 19개가 판정되었다. 이 중 18개는 센스 방향으로 RT(카세트 RT에 한하지 않고 어떤 RT 히트든)의 상류에 놓이고 1개는 콘티그 끝에 RT 쪽이 잘려 있다. 따라서 결과를 본 뒤에 추가한 규칙을 적용하면 RT 없이 관측 가능한 요소 히트는 없다(그 규칙 전에는 1개였다). 음성 집단 게놈 289개에서는 히트가 없었다. 요소 모형은 카세트에서 유래했고, 집단은 요소를 가진 속을 중심으로 하므로, 이 결과는 이 게놈들과 이 검출기에 대한 서술이다.

RT 없는 쌍은 카세트와 배치가 다르다. 게놈에서 다시 센 결과, 창 안에 어떤 종류의 RT도 없는 강한 −1/+3형 쌍 172개 가운데 151개는 두 유전자가 인접해 있고, −1형에서 +3형 유전자까지 간격의 중앙값은 24 nt이다. 반면 주 카세트 좌위 30곳(뉴클레오타이드 창 카탈로그의 카세트 양성 판정 31개에서, 단백질 집합으로만 알려져 곁 서열이 없는 좌위 하나를 뺀 것)에서 −1과 RT 사이 간격의 중앙값은 189 nt이다. 정확히 4 nt가 겹치는 경우는 쌍을 이루는 유전자 사이에서는 드물고(151개 중 9개) +1과 +2 사이에서는 흔하며(19개 중 17개), 명목 P 값은 배치를 기술할 뿐 기원을 검정하지는 않는다(보충 표 S10, 시트 F2; 보충 표 S11, 파일 SI-07_spacing_all_loci).

1차 결과를 본 뒤 추가한 사후(post hoc) 비교에서, RT 없는 쌍 곁의 서열은 근연 Pollutibacter 게놈의 카세트 곁 서열과 한 번도 일치하지 않았다(곁 서열 비교 627건 중 0건, 게놈 전체 뉴클레오타이드 동일성이 91.96~99.71퍼센트인 게놈 사이의 27건 중 0건, 한 종 수준 단위 안의 10건 중 0건). 같은 방법은 이 동일성에서 쌍끼리의 비교 17건 중 15건, 카세트끼리의 비교 22건 중 22건에서 곁 서열을 일치시켰다(보충 표 S10, 표 T). 이것은 이 게놈들에서 쌍과 카세트가 서로 다른 자리에 있음을 시사하며, 쌍과 카세트를 함께 가진 게놈이 3개 있으며, 각 게놈에서 둘은 서로 다른 콘티그에 있다. 다만 이 27건은 쌍을 가진 게놈 6개와 카세트를 가진 게놈 11개로 이루어진, 순서 없는 게놈 쌍 24개(쌍을 가진 게놈과 카세트를 가진 게놈의 조합 25개)이고 이 비교 가운데 26건은 한쪽 곁 서열만 비교했으며, 비교를 독립으로 취급한 Fisher P 값은 증거를 부풀린다. 사전 지정된 이웃 유전자 계열 비교 하나(조합 255개)에서는 귀무 평균 1.35에 비해 29개 조합이 공유 유전자 계열을 가졌다. 그러나 29개 중 26개는 작은 단백질 계열 하나에서 나왔고, 창을 좁히거나 동일성 기준을 높이면 개수가 0이며, 그 계열을 빼면(사후(post hoc)) 귀무 1.39에 대해 조합 3개가 남는다. 그러므로 사전 지정된 결과는 쌍과 카세트가 곁 자리를 공유한다는 증거도, 공유하지 않는다는 증거도 아니다.

숙주 트리도 방향을 주지 못한다. GTDB R232 Bacteroidota 종 트리의 팁 23,901개 가운데 RT 없는 인접 쌍은 팁 362개가, 카세트 배열은 팁 3개가 지니며, 이 3개는 모두 Pollutibacter에 있다. 쌍의 상태는 라벨 순열보다 트리 위에서 더 뭉쳐 있다(변화 265회 대 순열 평균 360.3회; 보충 표 S4, 표 O13). 그러나 조상 상태 복원은 모형에 좌우된다. 순서 모형은 Pollutibacter 조상에게 쌍 상태를 확률 0.68로 주고 모든 속도 상이 모형은 0.98로 주는 반면, 아카이케 정보 기준(AIC)이 가장 낮은 은닉 속도 모형은 쌍 상태가 없다는 쪽에 0.93을 준다. 카세트 배열을 가진 독립적인 속 수준의 팁은 1개뿐이다. Pollutibacter의 RT 없는 쌍 유전자 22개, 공개 Pollutibacter 게놈의 카세트 유전자 6개, 다른 속의 쌍

구성원으로 만든 유전자 트리에서 카세트 유전자는 −1형 트리와 +3형 트리 모두에서 초고속 부트스트랩 지지 100의 별도 가지를 이루며, 이 가지에는 두 부류 어느 쪽에도 속하지 않는 Pollutibacter 유전자 5개도 들어 있다(카세트 유전자에서 Pollutibacter의 자유 쌍 유전자까지의 거리 중앙값은 −1형 유전자에서 자리당 치환 6.17개로 그 자유 쌍 유전자들 사이의 1.27개에 대비되고, +3형 유전자에서는 3.41개 대 1.43개이다; 보충 표 S4, 표 O07). 포화 범위의 거리는 분기 시기에 관한 정보를 주지 못하므로, 이 비는 카세트 유전자가 길고 별도인 가지에 놓여 있음을 보일 뿐 그 가지들이 얼마나 오래전에 갈라졌는지는 보이지 못하며, 어느 쪽이 먼저인지도 말해 주지 못한다. 한 가지 해석은, 카세트가 다른 곳에서는 따로 발견되는 두 조각, 곧 RT 곁의 요소와 −1/+3 쌍으로 이루어져 있고, 우리가 두 조각을 함께 발견하는 속은 Pollutibacter뿐이라는 것이다. 우리가 돌린 어떤 분석도 어느 쪽이 먼저인지는 말해 주지 않는다(질문 3). 다른 속의 같은 가닥 쌍, 상류 회문, 접합부 반복 탐색, 유전자 트리와 구조 기반의 방향 분석에 관한 더 자세한 내용은 보충 결과 S1.8에 있다.

2.9 이 고아는 특이한가, 그리고 이 탐색을 어디까지 믿을 수 있는가?

ESOL RT는 최근접 동일성으로 보면 자신을 표적으로 삼지 않은 비표적 탐색의 후보보다 더 고립되어 있지만, 정렬의 덮임은 서로 다르다(보충 결과 S1.9). 발표된 RT와의 최근접 동일성은 ESOL RT의 질의 30개(그중 13개는 메타게놈 RT이고 29개는 발표된 라이브러리 히트가 있다)에서 중앙값 29%(최댓값 35%)인 데 비해, DefensePredictor [19]로 점수를 매긴 우리 자신의 탐색에서 나온 미배정 후보 군집 15개 가운데 레트론 계통 곁에 놓인 9개는 군집 중앙값이 32~50%이며, 나머지 6개는 알려진 유형으로 귀착되었다.

우리 자신의 파이프라인은 발표된 근연체가 있는 계통을 거의 보지 못한다. 파이프라인의 라이브러리에서 제거한 발표된 RT 계통 37개 중 다시 찾아낸 것은 4개뿐이었다(11%; 신뢰구간 4~25%; 보충 표 S12, 파일 SI-28_removed_lineage_recall). 그러므로 이 파이프라인은 발표된 라이브러리에 근연체가 없는 계통은 찾지만, 발표된 계통의 가까운 근연체는 알려진 것으로 분류하는 경향이 있다. 따라서 다른 후보가 없다는 사실은 후보가 존재하지 않는다는 증거가 아니다. 비표적 탐색, 이웃 도메인 탐색, 메타게놈 조립 게놈 비교, 서열 보관소 군집화에 관한 더 자세한 내용은 보충 결과 S1.9에 있다.

2.10 미리 정한 두 예측: 동일성을 맞춘 대조, 그리고 아무것도 알려 주지 못한 트리 시험

아미노산 동일성을 맞추어 기존 검출 도구를 모사한 시험에서, 참조 풀로 다시 만든 프로파일은 알려진 UG17, UG3, UG8 단위 49개 중 49개(전장 카세트 RT 12개가 풀 안 가장 가까운 참조에 대해 보이는 동일성 범위 24.8~33.3%에 맞춘 것이며, 맞춘 단위는 26.9~33.2%에 놓인다; 49개 중 2개가 UG17 단위이다)를 검출했으나 카세트 RT 단위는 16개 중 0개를 검출했다(보충 표 S2, 표 Q1). 이는 한 속 안에서 동일성만 맞춘 참조 풀 모사이며, 도구 자체를 돌린 홀드아웃 실행은 아니다. 이 비교는 정렬이 아니라 동일성으로 맞춘 것이다. 동일성을 맞춘 알려진 구성원은 가장 가까운 참조와 질의 덮임 중앙값 약 98%(보충 표 S2, 표 Q1)로 정렬되는 반면 전장 카세트 RT 12개는 45.2~70.7%로 정렬되므로, 카세트 RT가 보이는 정렬 거리에서 도구가 어떤 호출을 할지는 알려진 구성원으로 전혀 시험할 수 없고, 49개 중 49개 대 16개 중 0개의 대비는 어느 방향의 증거도 아니다. 사후(post hoc) 탐색 분석에서, 카세트 RT가 정렬되는 프로파일 열만 남기도록 자른 알려진 UG17 구성원(UG17은 전장 카세트 RT 12개 중 11개의 가장 가까운 참조 계통이다)은 배포된 문턱값에서 336개 중 0개가 호출되었고 전장 구성원은 336개 중 335개가 호출되었다. 따라서 카세트 RT가 보이는 덮임에서는 호출이 없다는 사실을 그 RT 계통의 성질로 읽을 수 없다(보충 표 S2, 표 Q4).

이웃한 카세트 계열 4개의 유전자 트리는 카세트 RT 트리와의 사전 지정된 일치 조건을 4개 계열 중 4개 모두에서 충족하지 못했다. 그러므로 사전 지정된 규칙에 따르면 어느 계열도 일치 조건을 충족하지 않는다. 이에 따라 사전 지정된 문턱값에서 이 트리 일관성 시험은 이웃한 계열이 RT와 같은 유래를 공유하는지 알려 주는 정보를 내지 못한다. 4개 중 4개라는 이 음성 결과는 공통 유래를 반박하지도 뒷받침하지도 않는다. 이 시험 자체의 양성 대조가 실패했고, 또 같은 유전자의 열 절반 두 개를 비교한 사후(post hoc) 대조에서, 공통 유래가 구성상 성립하는데도 같은 규칙이 반복 15개 중 7개에서 지지된 충돌을 내놓았기 때문이다(보충 표 S12, 시험 T15). 미리 정한 예측, 참조 풀 프로파일 비교, 이웃한 카세트 계열의 유전자 트리에 관한 자세한 내용은 보충 결과 S1.10에 있다.

2.11 자료가 열어 둔 가설과, 이를 가려낼 관찰

질문 1 — 이 RT는 이름으로 말하면 무엇인가? 알려진 유형 관문은 모두 ESOL RT를 배정하지 못했다. 총조사 그룹 구성원 14개 중 12개의 콘티그에서 반복 쌍 9개 중 DGR형은 0개, 구성원 17개 중 마투라제 도메인을 지닌 것은 0개, 주석이 달린 이웃이 있는 총조사 그룹 좌위 8개 중 근처에 Cas 도메인이 있는 것은 0개였고, 메타게놈 RT 13개를 포함한 ESOL 질의 30개의 라이브러리와의 최근접 동일성은 최대 34.6%(히트가 있는 질의 29개에서 중앙값 29.3%; 총조사 그룹 구성원 14개만으로는 31.7%와 28.5%; 보충 표 S2, 표 M 및 보충 표 S2, 표 N)였다. 서열 트리에서는 거리가 포화되고 연결이 약해서 이 RT를 UG3/UG17 인접 그룹에도, 어레이 연관 RT(ART) [20] 또는 UG27 [4]의 근연체로도 배치할 수 없다. 그래서 결과는 미배정이다(확장 데이터 그림 9).

질문 2 — 카세트는 숙주 뉴클레아제에 의존하는가? Wang 등(Cell 2026, 저자 원고) [8]은 RecBCD 상실과 DRT3 사이, ExoI 상실과 DRT9 사이에 선택적 합성 치사가 있다고 보고한다. 이 보고는 유추를 거쳐 +4 부분이 RecD 유사인 카세트에도 같은 질문을 던질 동기가 되지만, 뒷받침은 아니다. 실험적 시험은 없고, RecBCD 의존성과 SbcB 의존성의 일반 규칙을 게놈 연관으로 검정한 우리의 시험도 음성이거나 정보를 주지 못했다(한계 참조).

질문 3 — 이 검출기와 이 탐색 범위에서는 Pollutibacter 밖의 K 계열 RT 좌위 445곳 곁에 있는 관측 가능한 슬롯 1,130개 중 카세트 구성 요소를 가진 것은 하나도 없었는데, 그 구성 요소는 다른 속에서 RT 없이 되풀이해 나타난다. 왜 그런가? 가설 1에서는 한 속이라는 결과가 표집과 검출기 도달 범위의 제약을 받는다. 속 밖에서 나온 0은 검출기 실패와 구별할 수 없고, 실제 분포는 아틀라스보다 넓을 수 있다. 가설 2에서는 이 조합이 한 번 생겨 Pollutibacter에 머물렀고, 구성 요소는 다른 곳에서 RT 없이 함께 존재한다. 지금까지 이 같은 내력에 방향을 주는 것은 없다. Bacteroidota 종 트리에서 Pollutibacter 조상의 상태를 복원한 결과는 모형에 좌우되며, 외군으로 −1 유전자 트리의 뿌리를 정하는 것은 사전 지정된 포화 규칙에 따라 시험할 수 없었고(확장 데이터 그림 10), RT 없는 쌍 곁의 서열은 근연 Pollutibacter 게놈의 카세트 곁 서열과 한 번도 일치하지 않았는데(사후(post hoc) 비교 27건 중 0건이며 같은 방법은 카세트끼리의 비교 22건 중 22건에서 곁 서열을 일치시켰다), 이는 어느 쪽이 먼저인지는 말하지 않은 채 서로 다른 자리를 시사한다. Pollutibacter 밖에서, Ferruginibacter RT 섬의 연속한 세 ORF(RT 상류의 ORF 하나, RT, RecD)가 서로 다른 연구와 시료에서 나온 Ferruginibacter 게놈과 CAINEC01 게놈 사이에서 동일한데, 이를 읽기 수준에서 사후(post hoc)에 다시 조사하면, 두 연구의 공개 빈 조립에 이 동일한 블록 전체를 가진 콘티그는 정확히 네 개이다. 셋은 서로 거의 동일한 메소코즘 빈에서 나와 한 클론 계통을 이루고, 하나는 다른 어느 빈에도 나타나지 않으며 출처를 귀속할 수 없는 짧은 CAINEC01 콘티그이다. 그러므로 둘째 숙주는 확립되지도 배제되지도 않으며, 방향이나 공여자는 추론하지 않는다.

질문 4 — +3 단백질은 촉매 활성이 있는가? +3 단백질은 회수한 모든 ORF에서 표준 아연 모티프 HExxH(H[EDQ]xxH로 검색)가 없는데도 코돈 모형은 그 유전자를 정화 선택의 방향에 놓는다. 그래서 그 제약이 기능을 반영하는지는 열려 있다. 이 ORF들은 한 속 안에서 서로 가까운 근연 관계에 있고, 이 모티프의 부재는 카세트에 국한되지

않으며, 조성을 맞춰 뒤섞은 서열은 이만한 크기의 집합에서 약 절반의 경우(N 말단 부위를 뒤섞느냐 ORF 전체를 뒤섞느냐에 따라 0.44에서 0.51; 보충 표 S5, 표 T10, 행 H-8 및 H-9) 이 모티프가 없을 것이다. 따라서 이는 서열을 기술한 것일 뿐 활성을 시험한 것이 아니며, 활성이 있는 경우와 없는 경우 모두 열려 있다.

질문 5 — +3 단백질이 이 회문에 결합하는가? 여기서는 시험하지 않았다. 이 회문(중심부 CAAGCTTG, 자유 −1/+3 단위에서 −1 유전자 시작점의 상류)은 홀드아웃 속에서 사전 설정 규칙을 적힌 그대로 충족했다. 그러나 이 속들은 독립된 계통이 아니고(훈련에서 보지 못한 목에 속한 홀드아웃 속의 대부분은 Chitinophagales목에 속한다), 엄격한 부위는 103개 속 중 78개에 있을 뿐 전부에 있지는 않으며, 그 비암호화 위치는 탐색적 판정이고, 이동성 좌위의 공동 유전은 배제되지 않았으므로 결합과 공동 유전 모두 열려 있다.

질문 6 — 이 요소는 무엇인가? 레트론 msr/msd 쌍인가, 다른 종류의 비암호화 RNA인가, 아니면 주석이 달리지 않은 다른 무엇인가? Toro(2026)는 후보 XI형 유사 계통 둘의 좌위 일부에서 RT 상류에 있는, 계산으로 예측된 msr/msd 유사 ncRNA 모티프를 보고하며, 실험적 검증은 없다. 그 요소와 우리 요소가 동일하다고 주장하지 않으며, 우리 요소와 목록에 오른 레트론 msr/msd 쌍 사이의 거리는 여기서 해결하지 않았다. RecBCD 짝짓기와 짝 비교, 검출기 도달 범위와 뿌리 설정 분석, 한 섬의 읽기 수준 재조사, 단백질 구조 예측에 관한 더 자세한 내용은 보충 결과 S1.11에 있다.

3 논의

3.1 우리가 찾은 것과 그것이 뜻할 수 있는 것

게놈 189,630개의 RT 좌위 122,323개 가운데 크게 갈라진 RT의 후보 그룹 하나가 되풀이해 나타나는 같은 종류의 이웃 때문에 눈에 띄었고, 그 이웃으로 구성한 카세트는 5대륙 9개국의 서로 다른 시료 16개에서 나온 게놈 16개(카세트를 지녔다는 이유로 고른 것)에서 RT 곁에서 발견된다(대륙은 기록된 국가에서 도출했다). 중심 주장은 구성과 동반 출현이다. 카세트 단계통군을 정의하는 조립체는 5대륙의 물 시료 16개(호수에서 6개, 한 강에서 3개, 연못에서 2개, 그리고 담수, 지표수 또는 수생 환경의 물이라고만 기재된 5개)에서 왔고, 이 16개 기록의 자유 텍스트에는 토양, 퇴적물, 숙주, 폐수, 해수가 하나도 적혀 있지 않다. 우리가 탐색한 범위 안에서, 우리의 어떤 비교로도 이름 있는 계통에 배정되지 않는 이 단계통군의 RT는 한 속에 있으며, 이를 정의하는 게놈에서도, 그 정의에 쓰이지 않은 추가 좌위에서도 카세트 곁에 있다. 이 속의 추가 좌위 27곳에서는 RT 곁의 슬롯이 관측 가능한 99건 중 97건에서 카세트 구성 요소를 가지며(비어 있는 2건은 한 좌위에 있고, 이 좌위의 RT는 카세트 RT와 동일성이 0.388에 불과하다), 우리가 볼 수 있는 사본은 온전하고 정화 선택의 방향을 보인다. 카세트 특이적 모듈을 둘 이상 걸치는 공개 긴 읽기는 두 분자의 무작위 짝짓기가 허용하는 수보다 훨씬 많다. 구성 요소는 다른 속에서도 RT 없이 되풀이해 나타난다(그림 5에서 그림 7까지; 확장 데이터 그림 8).

Sánchez-Nieto 등은 프리프린트에서 DRT10 시스템을 지닌 생존 가능한 형질전환체가 대장균 ΔsbcB 균주에서 하나도 회수되지 않았다고, 곧 SbcB(ExoI)가 없으면 DRT10 활성이 독성이라고 보고한다[11]. DRT3과 DRT9 보고와 마찬가지로 이는 RecD 유사 부분을 가진 카세트에 관한 가설의 원천일 뿐 짝짓기를 설명하는 것은 아니다.

수수께끼는 RT가 왜 이 무리와 어울려 다니는가 하는 점이다. Pollutibacter에서는 이 무리가 관측 가능한 슬롯 99개 중 97개에 있다. 그런데 이 속 밖의 K 계열 RT 445개 곁에서는 검출기가 이 무리를 하나도 찾지 못하며, 그 거리에서는 검출기가 이를 볼 수 없었다(445개 거의 전부에서 검출률이 0에 가깝다). 더 낮은 곳까지 닿는 프로파일 검출기도 그 RT들 곁에서 +3형 단백질을 찾지 못한다(445곳 어디에서도 없다). 이는 그 부분에 대해 정보가 있는 부재이므로, 이 0은 도달 범위의 실패만은 아니다. 한편 −1 유전자와 +3 유전자의 근연 유전자는 확증 집단에서 +3형 유전자를 가진 속 149개 중

54개에서 창 안에 RT 없이 서로 곁에 놓인 쌍으로 나타나며, 이 RT 없는 되풀이의 존재는 보정되지 않은 위치 귀무에 견주어 사전 지정된 규칙 아래 뒷받침되지만(보충 결과 S1.8), 그 기능과 기원은 시험하지 않았다. 반대로 요소는 사후(post hoc) 탐색에서 관측 가능한 요소 히트 전부가 RT의 상류에서 발견되었다. 그러므로 RT 없는 쌍과 RT 곁의 요소는 따로 존재하는 두 조각처럼 보이며, 우리는 이 둘을 Pollutibacter에서만 함께 발견한다(질문 3).

여기서 두 질문이 나오며, 둘은 별개이다. 기원: 이미 RT 없이 존재하던 모듈에 RT가 합류했는가, 아니면 이 단위가 함께 생겨났는가? 짝을 이룬 구성 요소는 주석이 달린 공개 Bacteroidota 게놈 중 비무작위 부분 집합에서 RT 없이 되풀이해 나타나지만, 극성이 규명된 트리 없이는 획득인지 상실인지 그 방향을 가릴 수 없다. 기능: RT는, 프리프린트가 구조를 가진 ncRNA, SLATT 단백질, RT로 이루어진 세 부분 시스템인 DRT10에 대해 기술하듯, 곁의 무리와 함께 작동하는가? 여기서는 어떤 것도 이를 시험하지 않는다. 두 답은 서로 독립적이다. 더 오래된 모듈에 합류한 RT가 그 모듈에 필수가 되었을 수 있고, 이 단위와 함께 도착한 RT가 여전히 우연한 동승자일 수 있기 때문이다.

우리가 주장하지 않는 것은 기능, 이동 기전, 수직 또는 수평 기원, 그리고 탐색한 속과 범위를 넘는 일반성이며, 이것들은 아래의 시험이 다룰 질문이다. 가장 강력한 반론은 카세트 MAG로 만든 검출기가 자신을 만든 바로 그 대상을 다시 찾아낸다는 것이다. 모형에 쓰이지 않은 게놈에서의 히트는 새로운 관찰이고 점검들은 분류군 의존성을 살피지만, 감도는 같은 모형으로 판단할 수 없으므로 일부 순환성이 남는다.

우리가 열어 둔 각 해석에는 그 해석을 끝낼 관찰이 따른다(질문 1에서 질문 6까지; 확장 데이터 그림 9). K 계열 RT 곁에 카세트 구성 요소를 가진 두 번째 속, RT를 알려진 계통에 강한 지지로 귀속시키는 구조 기반 위치 배정, 외군으로 뿌리를 정해 극성을 시험할 수 있는 트리, RT 전사물의 5′ 말단 지도, +3 N 말단 도메인의 활성 분석이 그것이다. 이 시험을 수행할 습식 실험 프로그램은 진행 중이다. 컨스트럭트, 프라이머, 대조는 in silico로 설계하고 점검했으며(보충 노트 S1), 여기에는 어떤 습식 실험 결과도 보고하지 않는다. 관련 문헌, 근연 게놈에서의 곁 서열 비교, 미배정 후보 군집 탐색에 관한 더 자세한 내용은 보충 논의 S2.1에 있다.

3.2 한계

우리가 훑은 범위에서 카세트는 한 속, Pollutibacter에서 나타나며, 카세트를 가진 게놈은 10개 프로젝트, 9개국의 서로 다른 시료 16개에서 왔다. 이 게놈들은 서로 가까운 근연 관계이므로 독립적 반복 출현을 따지는 어떤 검정에서도 하나의 계통으로 센다. 따라서 이 자료는 카세트의 존재와 배열에 관한 것이지, 아직 숙주 범위나 카세트가 몇 번 생겨났는지에 관한 것이 아니다. RT가 미배정이라는 말은 여기에 열거한 비교와 라이브러리에 관한 진술이고, 카세트는 위치만으로 정의되며, 시드 MAG 16개와 그로 만든 모든 검출기는 순환적이다. 또한 게놈 배열은 어느 쪽이 먼저인지도 말해 주지 못한다. 주 카세트 좌위 30곳(뉴클레오타이드 창 카탈로그의 카세트 양성 판정 31개에서 단백질 집합으로만 알려진 하나를 뺀 것; 2.8절) 가운데 좌위 곁 서열이 양쪽 모두 정보를 갖는 곳은 6곳뿐이고, 카세트가 없는 게놈 중 카세트 게놈과 가까운 근연 관계인 것은 하나뿐인데 그 게놈의 곁 서열 히트 6건은 모두 콘티그 가장자리에서 끝나며, 사전 지정된 곁 서열 매핑 분석에서 카세트가 비어 있는 자리나 쌍으로 채워진 자리는 하나도 관측되지 않았으므로, RT가 모듈에 합류했는지 모듈을 떠났는지는 말할 수 없다.

사전 지정된 구성 요소 검출기는 이 속 밖 대부분의 RT가 가진 동일성보다 훨씬 높은 동일성에서 감도를 잃는다. 그러므로 더 낮은 동일성에서 카세트 구성 요소가 없다는 사실은 정보를 주지 못하며, 한 속이라는 관측은 생물학뿐 아니라 검출기 도달 범위에도 묶여 있다(질문 3). 또 다른 두 프로파일 검출기는 −1 구성 요소보다 +3 구성 요소에서 더 낮은 곳까지 닿으며, 속 밖 좌위 445곳 어디에서도 +3형 단백질을 찾지 못한다. 그러므로 속 밖의 0은 +3에서는 이 검출기와 좌위

안에서 정보가 있고 −1에서는 검출기 도달 범위에 한정된다. 다만 모든 검출기 계열은 여전히 한 속의 시드 MAG에서 유래하므로, 시드와 닮지 않은 변이형은 보이지 않는다. 검출기 도달 범위의 추정은 하나의 서열 진화 모형과 카세트에서 유래한 검출기에 의존하며, 모형과 무관한 점검으로 재현된 적이 없다.

습식 실험 결과는 보고하지 않으므로, 여기서는 카세트나 그 구성 요소가 무엇을 하는지 보여 주는 것이 없다. 온전한 이웃 유전자와 코돈 모형의 정화 방향은 제약과 일치하지만 기능의 증거는 아니다. RecBCD 의존성의 일반 규칙을 검정한 우리의 게놈 연관 시험은 그런 규칙을 뒷받침하지 못했거나 정보를 주지 못했다. 곧 K 계열을 가진 게놈은 RecBCD가 전혀 없는 속에서 나타나고, 요소 기반 K 정의로 수행한 사전에 정한 RecC 동반 시험은 정보를 주지 못했다. 점검에 앞서 고정한 규칙에 따르면 후속 연구 후보가 되는 배양 분리주는 없다. 스캔한 범위에서 상류 요소가 K 양성 RT 좌위에서 호출된 서로 다른 게놈은 모두 메타게놈 조립 게놈(MAG)이며, 이는 시드에 의존한 요소 호출이 이루어진 게놈에 대한 서술일 뿐 그 범위에 한정되고, 배양 균주가 이 요소나 카세트를 지니는지는 보여 주지 않는다. 구성 요소 검출기의 순환성, 유병률 추정, 부재 호출의 위음성률, 긴 읽기 검출 한계에 관한 더 자세한 내용은 보충 논의 S2.2에 있다.

4 방법

4.1 총조사 자료, RT 검출과 군집화

총조사는 Pfam RVT_1 모형으로 세균 종 대표 게놈 189,630개에서 RT 좌위를 찾은 뒤 군집화와 알려진 유형 관문을 거치는 탐색이고, 이후의 탐색은 게놈 199,923개의 단백질 모음을 대상으로 한 비표적 탐색이다. 둘 다 결과에서 설명한다. 두 단계 모두에서 카세트 관찰의 독립 단위는 속이다. 총조사 그룹과 K 계열 좌위는 식별자 대응으로만 연결되며, 이 대응을 맺을 때 서열은 정렬하지 않았다.

총조사는 NCBI 어셈블리가 있는 세균 대표 게놈 189,630개를 대상으로 했다(고세균은 포함하지 않았고, 이 밖의 세균 대표 게놈 171개는 내려받을 수 있는 어셈블리가 없다). 이후의 탐색은 모든 종 대표 게놈 199,923개(세균 189,801개, 고세균 10,122개)의 예측 단백질을 담은 GTDB 타르볼을 읽었으며, 여기에는 총조사의 모든 게놈이 들어 있다. Genomes from Earth's Microbiomes(GEM) 카탈로그는 Prodigal 좌표가 붙은 공개 단백질 타르볼로 사용했다.

RT는 Pfam RVT_1 모형을 gathering 문턱값에서 적용해 판정했고(HMMER [21], Pfam [22]), 단백질 길이 필터를 적용했다.

군집은 동일성 50%, 커버리지 80%로 만들었다(GTDB 탐색은 DIAMOND 기반 군집화, GEM은 MMseqs2 [23]). 공개 라이브러리 참조는 레트론과 UG/Abi 시스템의 RT 집합 [4,7]에 Toro와 Nisa-Martínez의 세균 RT 분류 [1]를 합친 것이다.

후보 군집은 모두 아래의 관문으로 점검했고, 그 결과를 후보 표의 열로 기록했다. G1(DGR) 관문은 bAvd Pfam 모형 [22]과, 아데닌 편향 불일치를 가진 주형/가변 반복 쌍 [9]을 찾았다. G2(그룹 II 인트론) 관문은 마투라제와 HNH 도메인, YADD 모티프, 유전자를 가르는 배치를 찾았다. G3(발표된 라이브러리) 관문은 라이브러리에 대한 최근접 동일성을 구했다. G4(Toro 레트론) 관문은 같은 비교를 Bacteroidota 레트론 RT 후보 306개에 대해 되풀이한 것으로, Bacteroidota 문에 한정된다. G5(CRISPR–Cas1) 관문은 RT에 융합되어 있거나 RT의 이웃 유전자 가운데 있는 Cas1 도메인을 찾았다. G6(방어 시스템 이름) 관문은 가장 가까운 라이브러리 라벨을 DefenseFinder와 PADLOC의 이름과 대조하며, 이름 수준에서만 작동한다.

4.2 자료 출처, 분모, 시료의 유래

우리는 공개 게놈 카탈로그에서 K 계열 역전사효소(RT) 좌위를 훑었으며, 포괄 범위는 처리한 게놈 수로 보고하고 완결성으로는 결코 보고하지 않는다. −1 유전자가 K 계열 RT와 같은 가닥에 놓인 좌위를 카세트 양성이라 부르며, 이는 사후(post hoc) 조작적 분류이지 단계통군이 아니고 상류 요소나 +1, +2, +3 유전자를 요구하지도 않는다.

ANI는 skani 0.3.2로 계산했고, 두 게놈의 ANI가 95% 이상이면서 두 정렬 분율 중 큰 값이 50% 이상일 때 단일 연결 방식으로 두 게놈을 하나의 ANI95 단위로 묶었다.

우리는 게놈 집합과 그 원천 시료를 최대 일곱 단(rung), 곧 속, ANI95 단위, BioSample, BioProject, 좌표 군집, 국가, 대륙에서 세며, 서로 다른 단의 수를 결코 더하지 않는다. 카세트의 독립 단위는 속이고, 다른 단의 수는 같은 게놈의 원천 기록을 기술한다. ANI95 단위의 수는 skani 모드와 정렬 분율 규칙에 따라 달라지고, 시드 MAG 16개는 이 좌위를 지녔다는 이유로 고른 표본이므로 보유율을 추정하는 데 써서는 안 된다. 스캔은 시드 MAG 16개 중 14개에서 RT와 같은 가닥의 −1 유전자를, 16개 모두에서 K 계열 RT를 판정했다. 나머지 두 MAG에서는 판정할 수 없었는데, 하나는 RT가 콘티그의 첫 ORF이고 다른 하나는 상류 ORF가 콘티그 끝에서 잘렸기 때문이다. 검출기를 이 MAG로 만들었으므로 14개는 배선 점검이지 민감도의 척도가 아니다.

4.3 좌위 정의, 카세트 유형 판정, 아틀라스 상태

좌위의 각 구성 요소는 칸 하나이며, 칸은 존재함, 없지만 관측 가능함, 관측 불가능함, 해당 없음(NA)의 네 상태 가운데 하나를 가진다. 우리는 "검출되지 않음"을 없지만 관측 가능한 칸에 대해서만 보고하며, 이는 관측 가능한 창 안에서 이 검출기가 찾지 못했다는 뜻이다.

4.4 검출 모형

우리는 카세트를 가진 MAG 16개에서 카세트 자체의 ORF로 만든 프로파일 HMM으로 K 계열 RT 주변의 모듈을 검출한다. K 계열 RT 후보와 그 주변의 판정은 고정된 규칙을 따르며, 그 규칙은 모형, 씨앗 목록, 문턱값, 도구 버전과 함께 보충 표 S13과 보충 데이터 S3에 제시하고, 모든 판정은 기탁한 표에 결과로 보고한다.

4.5 모티프, 구조 예측, 구조 기반 배치

콘티그 끝에서 잘린 ORF는 모티프 판정에서 제외하고 평가 불가능한 것으로 세었으며, 부재로는 결코 세지 않았다.

−1 유전자에는 사람 ITAL(UniProt P20701)의 I 도메인에서 금속 이온 의존 부착 부위(MIDAS) 위치(D162, S164, S166, T231, D264)를 가져왔으며, 이는 PDB 1T0P 사슬 A의 D137, S139, S141, T206, D239에 해당한다. 다섯 잔기가 차례로 D 또는 E, S 또는 T, S 또는 T, S 또는 T, D 또는 E 부류에 속하면 −1 ORF를 MIDAS 유지(MIDAS-retained)로, 세 개 또는 네 개가 속하면 MIDAS 부분 유지(MIDAS-partial)로, 두 개 이하가 속하면 MIDAS 소실(MIDAS-lost)로 판정했다.

K 단위는 ESOL RT 말단을 포함하는 가장 작은 단계통군이다. K 단계통군은 K 단위에서 위로 올라가되, 더해지는 자매 단계통군이 각각 K 계열 말단 50% 이상으로 이루어지는 동안 도달하는 가장 큰 단계통군이다.

4.6 방어 시스템 분류기 비교와 간극 검정

두 도구 중 어느 한쪽이 ESOL RT를 Retron, DRT 또는 Abi(AbiA, AbiK, AbiP2) 시스템으로 분류하면 그 RT를 판정된 것으로 셌다.

판정이 없는 것이 RT 계통을 반영하는지, 아니면 ESOL RT가 알려진 RT에서 떨어진 거리에서 분류기 프로파일의 민감도가 떨어지는 것을 반영하는지 묻기 위해, 우리는 간극 검정을 수행했다. 규칙은 이 검정의 판정, 회수, 홀드아웃 결과 중 어느 것도 보기 전에 봉인했다.

4.7 온전성 총조사와 선택 분석

온전성 총조사는 카세트 모듈의 암호화 서열이 조립된 콘티그에서 프레임이동, 조기 종결 코돈 또는 절단 결실을 지니는지 묻는다. 대상은 좌위 22개, 곧 카세트를 가진 MAG 빈의 좌위 16개와, RT가 ESOL RT와 50% 이상 동일한 공개 Pollutibacter 게놈 6개의 좌위 6개이며(공개 게놈을 가진 ANI95 단위 6개에서 하나씩이다; 좌위 아틀라스에는 이런 공개 좌위가 게놈 11개에 13개 올라 있고, 같은 단위의 추가 게놈 5개에 있는 좌위 7개는 총조사에 포함되지 않는다), 좌위마다 RT, −1, +1, +2, +3, +4의 여섯 슬롯을 본다. RecD 유사 +4 슬롯은 주석 슬롯으로 평가하되 카세트 구성 요소로는 세지 않는다(보충 데이터 S1, CENSUS_SLOTS genome_origin).

4.8 긴 읽기 선별, 키메라 기준선, 읽기 수준 대조

우리는 공개 시퀀싱 런에서 ESOL 카세트의 읽기를 네 종류의 분석으로 찾았다. 긴 읽기 메타게놈 런의 선별, 긴 읽기 전사체 런의 선별, 짧은 읽기 전사체 런과 짝을 이루는 메타게놈 런의 재스캔, 그리고 이 선별들이 내놓은 읽기에 대한 읽기 수준 대조이다. 긴 읽기 관측은 동결 표본에 카세트 서열 읽기가 존재한다는 것이며, 조립된 콘티그가 존재한다는 뜻이 아니다. 읽기의 숙주 속은 판정하지 않았다. 이 관측의 독립 단위는 시료가 아니라 수계이며, 시료의 상당수를 한 연구가 제공한다.

동결 표본에는 처리한 런 265개가 있으며, 담수 층에서 263개, 퇴적물 층에서 2개이고 읽기는 202.9 Gbp이다. 후보 런 799개에는 도달하지 못했다.

4.9 동반 출현 총조사, 쌍-앵커 스캔, 구성 요소 재출현, 회문, 외군 계통수

동반 출현 총조사(기탁한 표의 파일 이름과 열 이름에서는 ‘co-travel’이라 부른다)는 개별 좌위, 게놈, 속 각각에서 ESOL 카세트의 구성 요소가 K 계열 RT와 함께 얼마나 자주 발견되는지, 그리고 그 연관이 어느 방향으로 이어지는지를 묻는다. 이는 기존 검출기 표를 사후에 다시 집계하는 기술적 총조사이며, 기존 표가 포괄하지 않은 게놈의 보충 스캔을 제외하면 새 좌위를 검출하지 않는다.

쌍-앵커 스캔은 같은 가닥의 −1/+3형 유전자 쌍과 그 구성원 사이에 있는 RVT_1 유전자에 대한 집계를 다시 실행한 것으로, 검출기, 범위, RT 정의, 귀무 모형을 미리 고정했으며, 새 검출기가 내놓는 결과를 봉인 전에 한 집계와 일치하는지 여부와 관계없이 보고한다(보충 표 S10, 시트 F2).

전체 집합은 Bacteroidota 문의 GTDB R232 종 대표 23,901개였다.

RT가 없는 −1과 +3 유전자 쌍이 공개 Bacteroidota 게놈에서 되풀이되는지 묻기 위해, 분석은 어떤 쌍도 세기 전에 검출기, 모집단, 귀무, 결정 규칙을 고정했다.

4.10 사전 지정과 봉인

분석을 실행하기 전에 규칙 파일을 작성했다. 규칙 파일은 SHA-256 해시와 시각을 기록해 봉인했다. 이 타임스탬프는 내부 기록이며 시각에 대한 제3자 증거가 아니고, 결과를 본 뒤에 한 변경은 그렇게 표시했다. 프로토콜 기록은 이 연구의 규칙 파일을 모은 것이다. 각 규칙 파일은 그 파일이 규정하는 분석 전에 해시로 봉인했고, 모은 기록 자체는 나중에 봉인했으며, 그 날짜와 SHA-256 다이제스트는 보충 정보 안내서에 제시한다.

4.11 소프트웨어와 공개 카탈로그

서열 정렬에는 MAFFT [24], 트리 추론에는 IQ-TREE [25]를 썼고, 공분산 모형은 Infernal [26]로 만들어 탐색했으며 Rfam 모형 15개 [27]도 함께 탐색했다. 공통 구조는 ViennaRNA [28]의 RNAalifold로 구했다. 예측 구조는 Foldseek [29]으로 탐색했고, geNomad [30]는 기본 문턱값으로 돌렸으며, 분류기는 DefenseFinder [31]와 PADLOC [32] 이었다. GEM 단백질의 유전자 좌표는 Prodigal [33]에서 얻었고, 군집은 GTDB 탐색에는 DIAMOND [34], GEM에는 MMseqs2 [23]로 만들었다. 우리가 탐색한 공개 카탈로그에는 GEM 카탈로그 [35], Logan [36], SPIRE [37]가 있었다. ANI는 skani [38]로 계산했고, 코돈 모형 ω는 PAML 4의 codeml [39]로, 부위별 선택은 HyPhy [40] (FEL [41])로, 재조합 절단점은 GARD [42]로 추정했다. 긴 읽기는 minimap2 [43]와 BLAST+ [44]로 다시 점검했다. 분류는 GTDB R232 판 [17,45]을 따랐고, 카세트 MAG는 GTDB-Tk [46]로 분류했다.

선언

데이터 가용성. 이 연구에서 분석한 공개 서열 데이터는 보충 표 S9에 접근번호로 명시했고, 모든 그림과 보충 표의 바탕이 되는 좌위 표, 구성요소 표, 주석 표는 보충 데이터 S4 및 보충 데이터 S5에 담아 제공한다. 시드 조립체 16개(카세트를 가진 MAG)의 뉴클레오타이드 서열은 Zenodo(DOI 10.5281/zenodo.23346665)에 기탁했으며, 그 밖의 모든 좌위는 보충 데이터 S4 및 보충 데이터 S5에서 공개 접근번호와 좌표로 식별한다. 검출기 모형, 판정 문턱값, 씨앗 목록, 구축 명령, 도구 버전은 보충 표 S13과 보충 데이터 S3에 제시한다. 보충 표는 Excel 파일로, 보충 노트는 PDF 파일로 제출물과 함께 제공하고, 보충 데이터는 Zenodo 기록으로만 배포하며, Zenodo 기록에는 표와 노트도 들어 있다.

보충 자료. 보충정보는 항목 안내로 시작하며, 위의 절들이 인용하지 않는 항목은 여기에 이름을 적는다. 보충 표 S14에는 공개 시퀀싱 런의 긴 읽기 선별, K 계열의 담수 농축, 속의 서식지 기록, 연못 시계열이 담겨 있다. 보충 표 S15에는 문헌 선례 지도와 선점 문제를 다룬 방식에 대한 설명이 담겨 있다. 보충 표 S16에는 이 연구가 하지 않는 진술이 나열되어 있다. 보충 표 S17에는 카세트를 가진 게놈의 출처 시료 기록이 담겨 있다. 보충 표 S18에는 ESOL의 총조사 그룹에 적용한 알려진 유형 관문과 그 결과가 실려 있다. 보충 표 S19에는 RT 오픈 리딩 프레임을 기준으로 한 위치별 ESOL 카세트의 구성이 실려 있으며, RecD 유사 +4 유전자는 보조 유전자로, 카세트 밖의 이웃 유전자 두 개는 위치를 가늠하도록 함께 나타냈다. 보충 노트 S2에는 열린 질문의 자료집이 담겨 있다. 확장 데이터 그림 11에는 카세트의 배열 좌표, 간격 특징, xseA 유사 유전자에서 RecD 유사 유전자까지의 구간이 나타나 있다. 확장 데이터 그림 12에는 사후(post hoc) K 계열 집합이 여러 단계통군으로 갈라지는 기준 RT 도메인 계통수가 나타나 있다. 확장 데이터 그림 13에는 문턱값을 바꿔 가며 수행한 +1/+2 쌍 스캔, 같은 검출기로 알려진 구성원을 회수한 결과, 방향 없이 닫힌 갈래의 사전 지정된 규칙 결과가 나타나 있다. 확장 데이터 그림 14에는 탐색 파이프라인의 민감도가 나타나 있다. 확장 데이터 그림 15에는 탐색 경로와 알려진 유형 관문이 나타나 있다.

라이선스. 데이터는 Creative Commons Attribution 4.0 라이선스로 공개한다.

AI 사용. AI3 Discovery 의 과학적 발견 하네스를 AI 도구(Claude 와 GLM 모델)와 함께 사용했으며, 저자가 전 과정을 감독했고 모든 내용은 저자의 책임이다.

실험 연구. 이 연구에서 보고하는 습식 실험 결과는 없다. 예측을 검증할 습식 실험이 진행 중이며, 그 설계는 보충 노트 S1에 제시했다.

이해상충. 저자는 이해상충이 없음을 밝힌다.

연구비. 이 연구는 특정 재정 지원을 받지 않았으며 자체 재원으로 수행했다.

저자 기여. 표철민이 연구를 구상하고 분석과 초안 작성을 지휘했으며, 내용에 대한 책임을 진다.

참고문헌

1. Toro N, Nisa-Martínez R. Comprehensive phylogenetic analysis of bacterial reverse transcriptases. PLoS One 2014;9:e114083. https://doi.org/10.1371/journal.pone.0114083
2. Zimmerly S, Wu L. An Unexplored Diversity of Reverse Transcriptases in Bacteria. Microbiol Spectr 2015;3:MDNA3-0058-2014. https://doi.org/10.1128/microbiolspec.MDNA3-0058-2014
3. González-Delgado A, Mestre MR, Martínez-Abarca F, Toro N. Prokaryotic reverse transcriptases: from retroelements to specialized defense systems. FEMS Microbiol Rev 2021;45:fuab025. https://doi.org/10.1093/femsre/fuab025
4. Mestre MR, Gao LA, Shah SA, López-Beltrán A, González-Delgado A, Martínez-Abarca F, et al. UG/Abi: a highly diverse family of prokaryotic reverse transcriptases associated with defense functions. Nucleic Acids Res 2022;50:6084-6101. https://doi.org/10.1093/nar/gkac467
5. Toro N, Martínez-Abarca F, Mestre MR, González-Delgado A. Multiple origins of reverse transcriptases linked to CRISPR-Cas systems. RNA Biol 2019;16:1486-1493. https://doi.org/10.1080/15476286.2019.1639310
6. Millman A, Bernheim A, Stokar-Avihail A, Fedorenko T, Voichek M, Leavitt A, et al. Bacterial Retrons Function In Anti-Phage Defense. Cell 2020;183:1551-1561.e12. https://doi.org/10.1016/j.cell.2020.09.065
7. Mestre MR, González-Delgado A, Gutiérrez-Rus LI, Martínez-Abarca F, Toro N. Systematic prediction of genes functionally associated with bacterial retrons and classification of the encoded tripartite systems. Nucleic Acids Res 2020;48:12632-12647. https://doi.org/10.1093/nar/gkaa1149
8. Wang M, Yoneyama K, Žedaveinytė R, Ishikawa J, Tang S, Le HC, et al. Coordinated RNA- and protein-templated synthesis of double-stranded DNA by a dual reverse transcriptase immune system. Cell 2026;189:4997-5011.e11. https://doi.org/10.1016/j.cell.2026.07.012
9. Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, et al. Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. Nucleic Acids Res 2018;46:11-24. https://doi.org/10.1093/nar/gkx1150
10. Silas S, Mohr G, Sidote DJ, Markham LM, Sanchez-Amat A, Bhaya D, et al. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. Science 2016;351:aad4234. https://doi.org/10.1126/science.aad4234
11. Sánchez-Nieto E, Wilkinson ME, Millán V, Wenardy C, Cabrerizo D, Martínez-Abarca F, et al. Host exonuclease SbcB and a phage-encoded SSB-like protein control activation of the DRT10 reverse transcriptase defense system. bioRxiv (preprint) 2026. https://doi.org/10.64898/2026.04.29.721553
12. Tang S, Ramirez JL, Mestre MR, Zhang DJ, Wang M, Wiegand T, et al. Antiviral reverse transcriptases reveal the evolutionary origin of telomerase. bioRxiv (preprint) 2025. https://doi.org/10.1101/2025.10.16.682844
13. Ramirez JL, He Q, Wiegand T, Lampe GD, Wang J, Tang S, et al. Mechanism of tandem-repeat DNA synthesis by an antiviral reverse transcriptase. bioRxiv (preprint) 2026. https://doi.org/10.64898/2026.06.29.735271

14. Chen S, Zhang S, Zhang Y, Fu P, Mao T, Zhan Y, et al. Defense-associated reverse transcriptases synthesize DNA in a telomerase-like manner. bioRxiv (preprint) 2026. https://doi.org/10.64898/2026.10.08.757509
15. van den Berg DF, Costa AR, Esser JQ, Muralidharan A, van den Bossche H, Brouns SJJ. Discovery of phage defense systems through component modularity networks. bioRxiv (preprint) 2025. https://doi.org/10.1101/2025.09.30.679545
16. Toro N. Landscape of retron diversity across the SPIRE microbial metagenome resource reveals candidate novel type XI-like lineages. mSystems 2026;e00732-26. https://doi.org/10.1128/msystems.00732-26
17. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res 2022;50:D785-D794. https://doi.org/10.1093/nar/gkab776
18. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123-1130. https://doi.org/10.1126/science.ade2574
19. DeWeirdt PC, Mahoney EM, Laub MT. DefensePredictor: A machine learning model to discover prokaryotic immune systems. Science 2026;392:eadv7924. https://doi.org/10.1126/science.adv7924
20. Yoon PH, Athukoralage JS, Ameisen E, Kauderer-Abrams E, Perry NT, Durrant MG. Autonomous AI agents discover reverse transcriptases with tandem repeat arrays. bioRxiv (preprint) 2026. https://doi.org/10.64898/2026.10.05.755809
21. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol 2011;7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195
22. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res 2021;49:D412-D419. https://doi.org/10.1093/nar/gkaa913
23. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 2017;35:1026-1028. https://doi.org/10.1038/nbt.3988
24. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30:772-780. https://doi.org/10.1093/molbev/mst010
25. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 2020;37:1530-1534. https://doi.org/10.1093/molbev/msaa015
26. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 2013;29:2933-2935. https://doi.org/10.1093/bioinformatics/btt509
27. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res 2021;49:D192-D200. https://doi.org/10.1093/nar/gkaa1047
28. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol 2011;6:26. https://doi.org/10.1186/1748-7188-6-26
29. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol 2024;42:243-246. https://doi.org/10.1038/s41587-023-01773-0
30. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. Nat Biotechnol 2024;42:1303-1312. https://doi.org/10.1038/s41587-023-01953-y
31. Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat Commun 2022;13:2561. https://doi.org/10.1038/s41467-022-30269-9
32. Payne LJ, Todeschini TC, Wu Y, Perry BJ, Ronson CW, Fineran PC, et al. Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. Nucleic Acids Res 2021;49:10868-10878. https://doi.org/10.1093/nar/gkab883

33. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010;11:119. https://doi.org/10.1186/1471-2105-11-119
34. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 2021;18:366-368. https://doi.org/10.1038/s41592-021-01101-x
35. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. Nat Biotechnol 2021;39:499-509. https://doi.org/10.1038/s41587-020-0718-6
36. Chikhi R, Lemane T, Loll-Krippleber R, Montoliu-Nerin M, Raffestin B, Camargo AP, et al. Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity. bioRxiv preprint (version 2) 2025. https://doi.org/10.1101/2024.07.30.605881
37. Schmidt TSB, Fullam A, Ferretti P, Orakov A, Maistrenko OM, Ruscheweyh HJ, et al. SPIRE: a Searchable, Planetary-scale mIcrobiome REsource. Nucleic Acids Res 2024;52:D777-D783. https://doi.org/10.1093/nar/gkad943
38. Shaw J, Yu YW. Fast and robust metagenomic sequence comparison through sparse chaining with skani. Nat Methods 2023;20:1661-1665. https://doi.org/10.1038/s41592-023-02018-3
39. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 2007;24:1586-1591. https://doi.org/10.1093/molbev/msm088
40. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. Mol Biol Evol 2020;37:295-299. https://doi.org/10.1093/molbev/msz197
41. Kosakovsky Pond SL, Frost SDW. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 2005;22:1208-1222. https://doi.org/10.1093/molbev/msi105
42. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 2006;23:1891-1901. https://doi.org/10.1093/molbev/msl051
43. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34:3094-3100. https://doi.org/10.1093/bioinformatics/bty191
44. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009;10:421. https://doi.org/10.1186/1471-2105-10-421
45. Parks DH, Chaumeil PA, Mussig AJ, Rinke C, Chuvochina M, Hugenholtz P. GTDB release 10: a complete and systematic taxonomy for 715 230 bacterial and 17 245 archaeal genomes. Nucleic Acids Res 2026;54:D743-D754. https://doi.org/10.1093/nar/gkaf1040
46. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics 2022;38:5315-5316. https://doi.org/10.1093/bioinformatics/btac672