

GenBank 게놈 마이닝 방법

공개 유전체 데이터에서 원하는 유전자를 가진 균을 찾는 절차 · 2026-08-22

GenBank란

전 세계 연구자가 밝혀낸 유전자 서열을 모아 놓은 공개 데이터베이스입니다. 미국 NCBI가 운영하며 무료입니다.

여기서 자주 헷갈리는 것이 하나 있습니다.

	균주은행 (KCTC 등)	GenBank
보관하는 것	살아 있는 균 그 자체	유전자 서열 정보(데이터)
받는 것	앰플에 든 균	컴퓨터로 내려받는 텍스트
비용	균주당 수만 원	무료

이 둘을 함께 쓰면 연구 방식이 바뀝니다.

왜 이 방법을 쓰는가

균 100개를 주문해 하나씩 확인하는 것은 느리고 비쌉니다. 순서를 뒤집습니다.

① GenBank에서 목표 유전자 서열을 가져온다

← 무료

② 공개된 세균 유전체 수십만 건을 그 서열로 검색한다

← 무료

③ 유전자를 가진 균 목록이 나온다

④ 그중 안전하고 분양 가능한 것만 추린다

⑤ 그 몇 개만 실제로 주문해 실험한다

← 여기서 처음 비용 발생

①~④는 컴퓨터만 있으면 됩니다. 실험실도, 예산도, 균도 필요 없습니다.

0단계 · 씨앗 서열 확보

검색은 "이것과 비슷한 것을 찾아라"입니다. 씨앗이 틀리면 이후 전부가 무의미합니다.

⚠ 가장 중요한 원칙

NCBI에는 이름만 그럴듯하고 실제로는 검증되지 않은 서열이 매우 많습니다. 자동으로 붙은 이름이기 때문입니다.

논문으로 기능이 실증된 서열만 씨앗으로 쓰세요.

판별 방법은 두 가지입니다.

1. **SwissProt를 쓴다** — 사람이 직접 검토한 것만 모아 놓은 구역입니다. 검색어 끝에 붙이면 됩니다.

목표유전자이름[Protein Name] AND srcdb_swiss-prot[Properties]

2. **연결된 논문을 확인한다** — 서열 페이지의 Reference 항목에 논문이 있고, 그 논문이 효소 활성을 직접 측정했는지 봅니다. "Direct Submission"만 있으면 검증되지 않은 것입니다.

계열이 여럿이면 전부 넣기

같은 기능을 하는 효소라도 구조가 전혀 다른 계열이 존재할 수 있습니다. 한 계열만 씨앗으로 쓰면 나머지는 통째로 놓칩니다. 대상 효소에 몇 개 계열이 있는지 먼저 확인하고, 계열마다 씨앗을 하나씩 확보하세요.

음성 대조군도 함께

목표 유전자와 친척이지만 기능이 다른 서열도 같이 모읍니다. 나중에 계통수를 그릴 때 "우리 후보가 진짜 목표 쪽에 붙는가, 아니면 친척 쪽으로 붙는가"를 판정하는 기준선이 됩니다. 실험의 음성 대조군과 같은 역할입니다.

1단계 · 검색

같은 씨앗이라도 도구에 따라 찾는 범위가 크게 다릅니다. 낮은 단계부터 올리세요.

강도	도구	특징	코딩
1차	BLASTp (웹)	비슷한 것만. 확실하지만 좁음	불필요
1차 보완	tBLASTn (웹)	아직 주석이 안 붙은 유전체에서도 유전자를 찾아냄	불필요
2차	PSI-BLAST / jackhmmer	반복 검색으로 먼 친척까지 확장	불필요
3차	HMMER 프로파일	여러 씨앗을 한 덩어리로 묶어 검색. 훨씬 민감	약간
4차	Foldseek	서열이 달라도 3차원 모양이 같으면 찾아냄	불필요

Foldseek을 꼭 시도하세요

서열은 진화하며 많이 바뀌지만 단백질의 3차원 모양은 훨씬 오래 보존됩니다. AlphaFold가 예측해 둔 방대한 구조를 대상으로 "모양이 닮은 것"을 찾습니다. BLAST가 절대 못 찾는 것을 찾아냅니다. 신규성이 필요한

연구에 특히 유용합니다.

검색 대상

데이터베이스	내용
nr	NCBI 전체 단백질
RefSeq genomes	정제된 세균 유전체
BV-BRC	세균 유전체 특화
MGNify / IMG-M	메타게놈 — 아직 배양하지 못한 균의 유전자

마지막이 미개척지입니다. 다만 실물 균이 없어 유전자를 합성해 다른 균에 넣는 방식으로만 검증됩니다. 2차 트랙으로 두세요.

2단계 · 거르기

검색하면 수천 건이 쏟아집니다. 대부분 목표와 무관합니다.

왜 그런가

효소는 대개 거대한 "집안"에 속합니다. 같은 집안 안에는 하는 일의 원리는 같지만 만드는 물질이 전혀 다른 효소들이 잔뜩 들어 있습니다. 이름과 도메인이 비슷해 보여도 기능은 다릅니다.

"비슷하게 생겼다"는 근거가 되지 못합니다.

거르는 순서

- ① 기본 컷 — E-value뿐 아니라 정렬 커버리지(씨앗의 몇 %를 덮는가)를 반드시 보세요. 유사도만 높고 일부만 겹치는 것은 대개 가짜입니다.
- ② 도메인 확인 — Pfam·InterPro로 확인합니다. 단, 이것은 최소 조건일 뿐 증거가 아닙니다.
- ③ 계통수 — 0단계에서 모은 음성 대조군과 함께 나무를 그립니다. 후보가 진짜 목표 무리 안쪽에 들어가는지 봅니다.

추천 도구: EFI-EST / EFI-GNT

미국 Enzyme Function Initiative의 무료 웹 도구입니다. 코딩 없이 다음을 해 줍니다.

- **EFI-EST**: 후보 수천 개를 유사도 네트워크로 그려 줍니다. 자연스럽게 무리가 나뉘어, 알려진 목표 효소가 들어간 무리와 음성 대조군이 들어간 무리가 눈으로 구분됩니다.
- **EFI-GNT**: 각 후보의 유전자 이웃을 자동으로 그려 줍니다. (3단계에 직결)

3단계 · 유전자 이웃 — 여기가 승부처입니다

목표 유전자 하나만 가진 균은 대개 실제로 그 물질을 만들지 못합니다.

물질을 진짜 만들려면 재료를 공급하는 유전자들이 필요하고, 세균은 그런 유전자들을 바로 옆에 세트로 붙여 둡니다. 이것을 오페론이라 합니다.

... 목표유전자 — 재료공급유전자1 — 재료공급유전자2 ...

조립 기계 하나만 덩그러니 있는 공장과, 기계 옆에 부품 창고까지 갖춘 공장의 차이입니다. 후자만 실제로 생산합니다.

"비슷한 유전자"를 찾지 마시고 "세트를 통째로 가진 균"을 찾으세요. 가짜 후보가 극적으로 줄어듭니다.

도구 (전부 무료·웹)

도구	용도
EFI-GNT	이웃 유전자 자동 분석 (추천)
webFlaGs	주변 유전자 시각화, 사용 간단
NCBI Genome Data Viewer	개별 확인
clinker	여러 균의 유전자 배열 비교 그림

한 가지 더 — 유전자섬 신호

목표 유전자 세트의 양옆에 transposase(이동성 요소)가 붙어 있다면, 그 유전자가 균의 원래 것이 아니라 밖에서 들어와 끼어든 조각일 가능성이 큽니다.

이 신호는 중요합니다. 같은 종인데도 균주마다 있고 없고가 갈리는 이유를 설명하고, 동시에 다른 균에도 옮겨 갔을 수 있다는 뜻이 되어 탐색 범위를 넓혀 줍니다.

4단계 · 실물 균주와 연결

남은 후보로 표를 채웁니다.

균주명	유전체 번호	유전자 세트	위험균	분양 가능?	특허 선점
		완전/부분	1급/2급	KCTC-ATCC 번호	유/무

우선순위: ① 비병원성 ② 국내 은행 분양 가능 ③ 특허 미선점 ④ 유전자 세트 완전

여기까지 하면 근거가 명확한 주문 목록이 나옵니다. 이것이 목표였습니다.

최종 판정은 실험으로

컴퓨터는 후보를 좁힐 뿐입니다. 증명은 실험이 합니다.

간이 검출법은 대개 특이성이 낮아, 목표 물질이 아닌 것도 양성으로 나옵니다. 효소 특이 분해 시험 → 구조 분석(NMR 등) 까지 가야 확정됩니다. 선행 논문을 읽을 때도 "이 논문이 구조 확인까지 했는가"를 보면 신뢰도가 바로 갈립니다.

도구 요약 (전부 무료)

목적	도구	코딩
서열 확보	NCBI Protein / Datasets	불필요
기본 검색	BLASTp, tBLASTn	불필요
민감 검색	PSI-BLAST, jackhmmer	불필요
구조 검색	Foldseek, AlphaFold DB	불필요
네트워크·이웃	EFI-EST / EFI-GNT	불필요
이웃 시각화	webFlaGs, clinker	약간
계통수	MEGA, IQ-TREE, iTOL	약간
대량 처리	BLAST+, HMMER, MMseqs2	필요

대량 자동 수집을 하실 때만 NCBI API 키가 필요합니다. 웹으로 하는 작업은 키 없이 됩니다.

오늘 30분 안에 해 볼 것

1. NCBI Protein 접속 → 관심 유전자 이름으로 검색
 2. SwissProt 항목을 하나 열어 연결된 논문 확인
 3. FASTA 복사 → **BLASTp 실행** (DB: nr)
 4. 결과에서 예상 밖의 균이 나오는지 훑어보기
- 4번에서 낯선 균 이름이 보이면, 그것이 후보 목록의 첫 줄입니다.