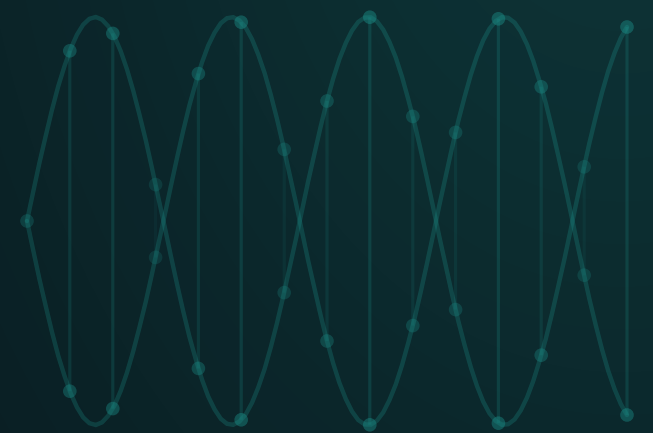


레슨: 유전자와 rsID, 그리고 목록이 실제로 재는 것



생명과학

유전학

만 16세+

입문

16세 이상을 위한, 바로 쓸 수 있는 유전학 레슨입니다. 학교·칼리지·대학교 어디서나 무료로 쓸 수 있습니다. 모두가 한 권씩 가진 책 한 권 — 한 장(章)이 유전자이고, rsID는 사람마다 달라지는 글자 하나의 이름입니다. 그다음 학생들이 실제 목록에서 그것을 세어 보고, 그 숫자가 생물학이 아니라 연구 관심도를 재고 있다는 것을 발견합니다.

자유롭게 읽고, 쓰고, 고쳐 쓰실 수 있습니다. 계정도, 구매도, 수집되는 정보도 없습니다.

전체 레슨과 그림, 갱신본: <https://www.mygenelog.com/ko/updates/lesson-genes-and-rsids-as-code>

이것은 무엇인가. 바로 쓸 수 있는 유전학 레슨입니다. 학교, 칼리지, 대학교 어디서나 자유롭게 쓰고 고칠 수 있습니다. 저희에게서 살 것도, 계정도, 누구에게서 수집하는 자료도 없습니다.

- 수준: 대략 16세부터 대학생까지. 유전학 사전 지식은 필요 없습니다. 소스 파일이라는 개념을 알면 도움이 되지만 필수는 아닙니다 — 비유는 처음부터 설명합니다.
- 시간: 45~60분. 세어 보는 활동 자체는 15분 정도입니다.
- 준비물: 학생 한 명 또는 두 명당 브라우저 하나, 그리고 숫자를 적을 칠판. 유전체도, 검체도, 어떤 종류의 검사도 필요 없습니다.
- 다루는 것: 유전자란 무엇이고 rsID란 무엇인지, 각각 몇 개나 되는지, 왜 목록이 곧 실체가 아닌지, 그리고 순전히 생물학처럼 보이는 데이터에 연구 관심도가 어떻게 지문을 남기는지.

다섯 줄로 끝내면

살아 있는 모든 사람이 똑같은 책을 한 권씩 가지고 있습니다.

- 책이 여러분의 DNA입니다. 30억 글자, 23권으로 묶여 있습니다.
- 한 장(章)이 유전자입니다. 무언가 하나를 만드는 법이 적힌 부분이고, 약 20,000장이 있습니다.
- 내 책과 당신 책은 거의 똑같습니다. 아주 가끔 글자 하나가 다릅니다. 내 책에는 C인데 당신 책에는 T인 식입니다.
- 과학자들이 그렇게 달라지는 자리마다 영구적인 번호를 붙였습니다. 그 번호가 rsID이고, rs1815739 같은 것입니다.
- 그러니 rsID는 자리의 이름이지 거기에 내가 가진 글자가 아닙니다. 내가 어떤 글자를 가졌는지는 따로 물어야 하고, 그것을 유전형이라고 합니다.

이 수업에 나오는 용어는 이게 전부입니다. 아래는 저 다섯 줄 중 하나를 더 자세히 보는 것이거나, 학생들이

그것으로 실제로 해 보는 활동입니다.

기억할 한 문장. 유전자는 한 장이고, rsID는 그 안의 글자 하나에 붙인 번호입니다. 한 장 안에 그런 글자가 여럿 있습니다 — 그래서 ”나는 BRCA1 유전자가 있다”는 말은 그 자체로는 성립하지 않습니다. BRCA1은 누구에게나 있습니다. 다른 것은 그 사람의 책에 어떤 글자가 들어 있는가입니다.

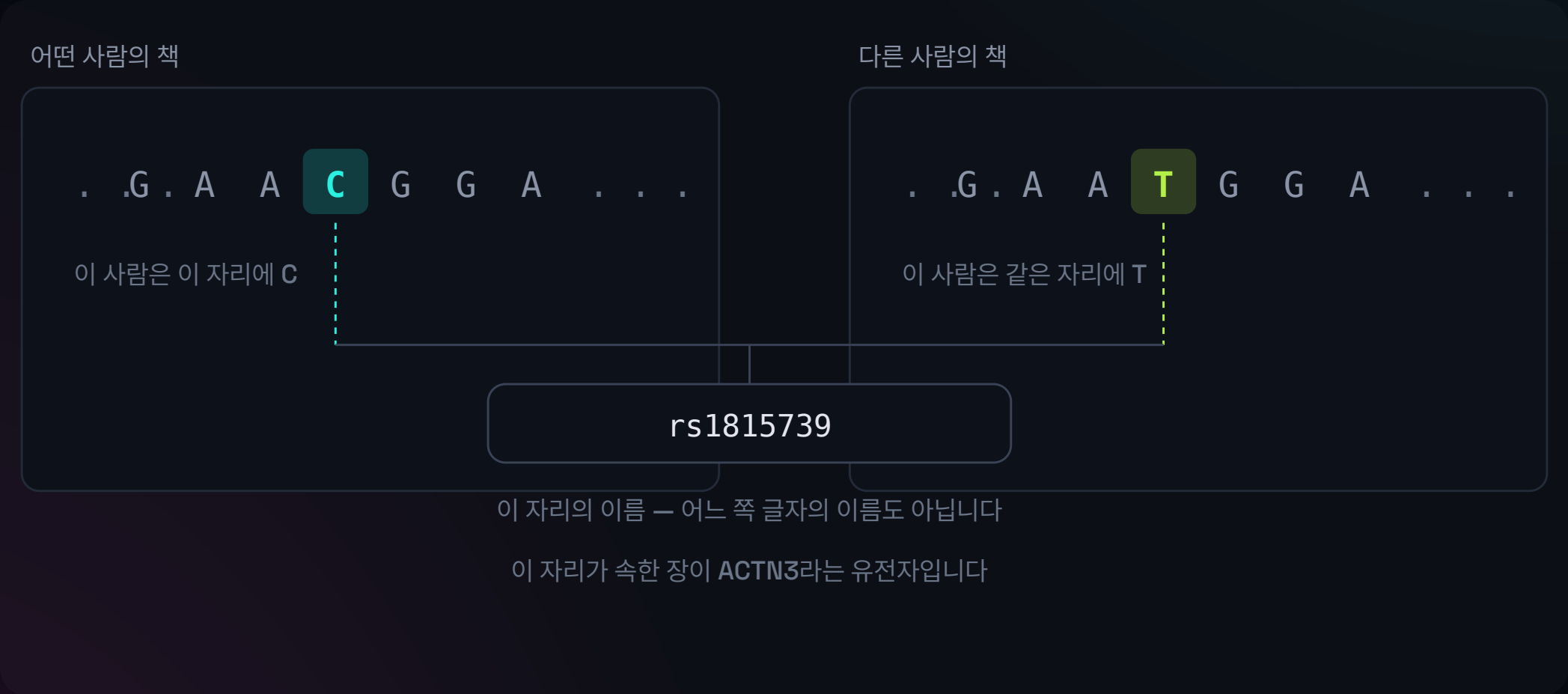
서로 다른 두 개의 ”둘” — 헷갈리면 안 됩니다

이 주제 전체에서 가장 흔한 혼동입니다. 그림을 보고 나서 다투는 것보다 보기 전에 2분 쓰는 편이 낫습니다. 지금부터 ”둘”이라는 말이 서로 다른 두 가지를 가리킵니다.

- 두 사람. 당신의 책과 나의 책은 거의 똑같고 가끔 글자 하나가 다릅니다. rsID는 그것을 위한 것입니다. 사람들 사이에서 달라지는 자리에 이름을 붙입니다.
- 한 사람 안의 두 벌. 여러분에게는 이 책이 두 벌 있습니다. 어머니에게서 한 벌, 아버지에게서 한 벌 물려받았기 때문입니다. 그래서 어느 한 자리를 펼치면 글자가 하나가 아니라 둘입니다. 그 짝이 유전형이고, 검사 결과가 글자 하나가 아니라 C/C, C/T, T/T처럼 적히는 이유입니다.

아래 그림은 첫 번째 경우입니다. 서로 다른 두 사람, 글자 하나 차이. 두 번째 경우는 머리에 담아 두십시오. 결과에 글자가 두 개 들어 있는 이유가 그것입니다.

같은 이야기를 그림으로



축척이 아닌 개념도입니다. 두 책은 글자 하나만 빼고 전부 같습니다. rs1815739는 그 자리의 이름이며, 거기에 누가 어떤 글자를 가졌는지는 말해 주지 않습니다. 내가 어떤 글자를 가졌는지가 유전형이고 별개의 질문입니다.

책에서	유전체에서	무엇인가
글자, 그리고 그것이 인쇄된 종이	DNA	A, T, G, C 네 글자가 사다리처럼 짝지어 꼬여 있는 것. 그 꼬임이 그 유명한 이중 나선입니다. 글이 아니라 종이입니다.
묶여 있는 23권	염색체	2미터의 DNA를 눈에 보이지도 않는 세포 안에 들어갈 만큼 단단히 감아 둔 것. 23쌍을 가집니다.
무언가 하나를 만드는 법이 적힌 한 장(章)	유전자	하나의 산물, 보통 단백질 하나의 설명서를 담은 구간. 약 20,000개 있습니다.
내 책과 당신 책에서 다르게 읽히는 글자	SNP	사람마다 달라지는 한 자리. 거의 모든 글자는 누구에게나 같고, 그렇지 않은 드문 자리가 이것입니다.
바로 그 자리에 붙인 목록 번호	rsID	모두가 어디서든 같은 자리를 가리키도록 붙인 영구적인 이름. rs1815739는 지구상의 모든 논문에서 같은 자리입니다. 그 자리의 이름이지, 거기에 당신이 어떤 글자를 가졌는지의 이름이 아닙니다.

포함 관계는 한 방향입니다. 한 장이 그런 자리를 여럿 담고, 한 자리는 많아야 한 장 안에 있습니다 — 아예 어느 장에도 속하지 않는 경우가 더 많습니다. 이 책의 대부분은 장이 아니기 때문입니다.

비유가 깨지는 지점, 그리고 그것이 중요한 이유. 책은 앞에서 뒤로 읽고 모든 독자에게 같은 말을 합니다. 이 책은 그렇지 않습니다. 세포마다 다른 장을 읽습니다 — 근육 세포와 간 세포는 같은 책을 읽고 서로 다른 부분을 씁니다. 한 장의 본문 대부분은 설명서가 쓰이기 전에 잘려 나갑니다. 그리고 한 장이 어느 부분을 남기느냐에 따라 여러 산물을 만들 수 있습니다. 이 비유를 "장과 글자" 너머로 밀고 가면 사실이 아닌 것을 알려주기 시작합니다.

프로그래밍을 해 보셨다면 더 딱 맞는 설명이 있습니다

유전자는 소스 파일이고, rsID는 그 안에서 갈래마다 달라지는 한 줄의 이름입니다. 같은 주의가 더 강하게 적용됩니다. 코드는 쓰인 대로 위에서 아래로 실행되지만 DNA는 그렇지 않습니다.

각각 몇 개나 되나

이 숫자들은 소리 내어 말해 볼 값어치가 있습니다. 대부분이 예상하는 숫자가 아니기 때문입니다.

대상	사람 한 명 기준 대략
염색체	23쌍, 46개

대상	사람 한 명 기준 대략
DNA 글자	한 벌당 약 30억 개
세포 하나의 DNA를 펼친 길이	약 2미터
단백질을 만드는 유전자	약 20,000개
전사되지만 단백질을 만들지 않는 유전자	그보다 수만 개 더
지금까지 기술된 rsID(전 인류 통틀어)	수억 개

놀라운 것은 20,000입니다. 유전체가 해독되기 전의 진지한 추정치는 100,000개 이상이었습니다. 복잡한 생물에는 설명서가 많이 필요하리라는 논리였습니다. 나온 숫자는 훨씬 적었고, 그 해답은 설명서가 재사용된다는 것이었습니다. 유전자 하나가 여러 산물을 낼 수 있고, 유전체의 상당 부분은 나머지가 언제 실행될지를 정하는 조절 영역입니다.

해 봅시다: 유전자 안의 rsID를 세어 보기

여기서부터 두 단어가 정의이기를 그만두고 측정할 수 있는 것이 됩니다. 아래 내용은 모두 이 사이트의 무료 페이지만 씁니다. 제출하는 것도, 계정도 필요 없습니다.

시작하기 전에 — 지켜야 할 한 가지. 누구도 자기 유전 정보를 이 수업에 가져오지 않고, 무엇을 가지고 있는지 묻지 않습니다. 여기에는 검사도 없고 교실 안 누구에 대한 결과도 없습니다. 이 활동은 공개된 목록에 관한 것이지 여러분에 관한 것이 아닙니다. 학생이 자기 DNA에 대해 물으면, 교실은 그것을 알아볼 자리가 아니며 의사나 유전상담사가 맞는 자리라고 정직하게 답해 주십시오.

1. 무엇이든 열기 전에, 먼저 예측합니다

검색을 시작하기 전에 학급이 합의한 두 숫자를 칠판에 적습니다.

- 교실에서 대부분이 이름을 들어 봤을 BRCA1 안에는 rsID가 몇 개나 있을 것 같습니까?
- 대부분이 들어 보지 못했을 PITX2 안에는 몇 개일 것 같습니까?

거의 모든 학급이 유명한 유전자 쪽이 더 많을 것이라고 예측합니다. 그 예측을 적어 두십시오. 예측이 빛나가는 것이 이 수업이고, 먼저 적어 두어야만 작동합니다.

2. 이 목록이 실제로 담고 있는 것을 세어 봅니다

이 사이트에서 각 유전자 기호를 검색해 나오는 변이 페이지를 세어 봅니다. 이 글을 쓰는 시점에 목록은 2,076개 유전자에 걸쳐 2,976개 변이를 담고 있고, 숫자는 이렇습니다.

유전자	알려진 이유	여기 있는 rsID	그 항목들이 다루는 것
PITX2	심장 박동	15	심방세동 — 열다섯 개 전부
FTO	체중	10	체질량과 관련 지표
ABO	혈액형	9	여러 형질
APOE	알츠하이머 위험	4	콜레스테롤과 알츠하이머병

유전자	알려진 이유	여기 있는 rsID	그 항목들이 다루는 것
TAS2R38	쓴맛	3	PTC 쓴맛 감지
MTHFR	엽산 처리	2	엽산 대사
CFTR	낭포성 섬유증	2	낭포성 섬유증
ACTN3	근섬유 유형	1	근섬유 유형
TP53	암 연구에서 가장 많이 연구된 유전자	1	기저세포암
BRCA1	유방암·난소암 위험	1	초경 연령

넘어가기 전에 마지막 열을 읽어 보십시오. 숫자보다 그쪽이 더 많은 일을 하고 있습니다.

3. 대조: 이 숫자가 무엇을 재고 있는지 묻습니다

예측을 결과 옆에 놓습니다. TP53은 암 생물학 전체에서 가장 많이 연구된 유전자인데 여기서는 항목이 하나입니다. PITX2는 열다섯 개입니다. 이제 이것을 단순한 웹 서핑이 아니라 과학 수업으로 만드는 질문입니다. 그러면 PITX2가 TP53보다 사람마다 더 많이 다른 것입니까?

아닙니다. 이 숫자는 이 목록이 어떻게 만들어졌는지를 켓니다. 여기 있는 거의 전부가 전장유전체 연관분석에서 왔습니다. 어떤 형질이 있는 사람과 없는 사람 사이에서 달라지는 자리를 유전체 전체에서 훑는 방법입니다. 심방세동은 그 방식으로 여러 번 연구되었고 PITX2는 매번 나오며, 연구마다 항목을 남깁니다. PITX2의 열다섯 줄은 전부 같은 형질이 반복해서 발견된 것입니다.

아래 두 줄이 어떤 설명보다 이 점을 날카롭게 보여 줍니다.

TP53은 암 연구에서 가장 중요한 유전자라고 할 만한데, 여기 있는 단 하나의 항목은 기저세포암에 관한 것입니다. 오류가 아닙니다. TP53은 주로 한 사람이 살아가는 동안 종양 안에서 생기는 돌연변이를 통해 문제가 되는데, 그것은 유전되지 않으며 이런 종류의 연구에는 아예 잡히지 않습니다.

BRCA1은 교실의 모두가 들어 본 유전자인데, 여기 있는 하나뿐인 항목은 암과 아무 관련이 없습니다. 초경 연령에 관한 것입니다. 암을 일으키는 BRCA1 변화는 드물고, 하나하나가 강력하며, 가계를 시퀀싱해서 찾습니다 — 이 목록을 만든 방법과는 완전히 다른 방법입니다. 그래서 그 유명한 암 유전자가 아무도 연결 짓지 않는 형질 아래 실려 있습니다. 그 방법이 볼 수 있는 변이가 그것뿐이었기 때문입니다.

항목의 개수, 나아가 항목의 주제까지도 어떤 질문을 어떤 도구로 던졌는지를 반영합니다. 그 유전자가 얼마나 변이가 많은지도, 무엇에 중요한지도 아닙니다. 이 하나를 가져가는 학생은 앞으로 만날 모든 데이터셋에 적용되는 것을 배운 것입니다.

이 목록에 없는 것

규모를 정직하게 비교해 봅시다. 이 사이트는 약 2,000개 유전자에 걸쳐 약 3,000개 변이를 담고 있습니다. 단백질을 만드는 유전자는 약 20,000개이고, 기술된 rsID는 수억 개입니다. 그러니 이 목록은 유전자의 10분의 1 정도와, 자리의 1퍼센트에도 한참 못 미치는 비율을 덮고 있습니다.

작다는 것보다 나쁜 것은 고르지 않게 작다는 점입니다. 이 목록이 딛고 선 연구들은 유럽계 조상을 가진 사람들에게 불균형하게 많이 수행되었습니다. 그래서 다른 집단에서 중요한 자리일수록 빠져 있을 가능성이

큽니다 — 존재하지 않아서가 아니라 찾아본 연구가 적어서입니다. 이것은 원자료의 한계이며, 어떤 목록에 어떤 변이가 없다는 것이 그 사람에게 없다는 근거가 결코 되지 못하는 이유입니다.

아직 열려 있는 논쟁

그렇다면 유전자당 변이 개수를 얼마나 믿어야 할까요?

한쪽 입장은, 이 숫자들은 연구 설계가 남긴 인공물이며 유전자의 성질처럼 제시되어서는 안 된다는 것입니다. 다른 입장은, 관심이 무작위가 아니라는 것입니다. 서로 다른 형질을 다룬 독립적인 연구들에서 계속 나오는 유전자는 실제로 넓게 중요한 일을 하고 있을 가능성이 있고, 그 숫자에는 편향에 묻힌 진짜 신호가 들어 있다는 것입니다.

양쪽 모두 문헌에서 주장되고 있고 결론은 나 있지 않습니다. 학생들은 이 분야가 가진 것과 같은 근거를 그대로 손에 짚 수 있습니다. 결론이 아니라 원래의 숫자를 보여 주는 이유가 그것입니다.

질문

몸풀기

1. 유전자가 무엇이고 rsID가 무엇인지 각각 한 문장으로 말해 보십시오.
2. 하나의 rsID가 서로 다른 두 유전자에 속할 수 있습니까? 하나의 유전자가 두 개의 rsID를 담을 수 있습니까?
3. 사람은 단백질을 만드는 유전자를 대략 몇 개 가지고 있으며, 처음 추정치는 왜 다섯 배나 높았습니까?
4. 이중 나선인 것은 무엇입니까 — 유전자입니까, DNA입니까, 염색체입니까?

본문제

1. 이 목록에서 TP53은 한 번, PITX2는 열다섯 번 나옵니다. 그 패턴을 만들어 낼 수 있는 서로 다른 설명 두 가지를 대고, 둘을 가려낼 근거가 무엇일지 말해 보십시오.
2. BRCA1의 유일한 항목은 암이 아니라 초경 연령입니다. 암 위험으로 가장 유명한 유전자가 어떻게 다른 항목 아래 실리게 되는지 설명하고, 그것이 이 목록을 만든 방법에 대해 무엇을 말해 주는지 쓰십시오.
3. 친구가 "내 DNA 검사에서 BRCA1에 아무것도 안 나왔으니 괜찮아"라고 말합니다. 이 수업의 규모 비교를 이용해 그 문장의 무엇이 잘못되었는지 설명하십시오.
4. "유전자는 한 장이고 rsID는 글자 하나에 붙인 번호다"라는 비유가 왜 쓸모 있는지 말하고, 그 비유가 틀리게 말하는 것을 하나 구체적으로 대십시오.
5. 어떤 목록이 주로 한 집단의 연구로 만들어졌다면, 다른 집단에서 더 중요한 유전자의 숫자는 어떻게 됩니까? 그렇게 생기는 차이는 생물학에 관한 사실입니까, 연구에 관한 사실입니까?

더 나아가기

1. 유전자당 변이 개수가 생물학을 반영하는지 관심도를 반영하는지 검증할 방법을 설계해 보십시오. 무엇을 측정하겠으며, 어떤 결과가 나오면 납득하겠습니까?
2. 2미터의 DNA가 눈에 보이지도 않는 세포 안에 들어갑니다. 압축 배율을 어렵하고, 유전자 하나를 읽어야 하는 세포에게 그 압축이 어떤 문제를 만드는지 말해 보십시오.
3. rsID는 위치로 부여되고, 위치는 인간 유전체의 서로 다른 참조 버전을 기준으로 기술될 수 있습니다. 두 연구실이 서로 다른 참조 버전의 결과를 비교할 때 무엇이 잘못될 수 있으며, 어떻게 방지하겠습니까?

교사용 메모 — 각 질문이 향하는 곳

1~4는 포함 관계를 확인하고, 물리적 구조(DNA, 나선, 염색체)와 정보 단위 (유전자)와 이름표(rsID)를 갈라 놓습니다. 4번은 교실에서 가장 흔한 혼동을 잡아냅니다. 학생들은 으레 유전자가 나선이라고 말합니다. 나선은 유전자가 적혀 있는 재료입니다.

5와 6이 이 수업의 중심입니다. 5번은 어떤 기제든 두 가지면 받아 주십시오 — 연구 수, 연구 종류, 유전자 길이, 그 구간의 변이 정도. 다만 뒷부분을 강하게 밀어 주십시오. 둘을 가려내는 근거는 유전자당 발표된 연구 수이고 그것은 셀 수 있습니다. 이 질문이 의견이 아니라 경험적 문제가 된다는 것을 학생들이 알아채야 합니다.

6번에는 짚고 넘어갈 답이 있습니다. 이 목록은 흔한 변이를 다룬 연구들로 만들어졌고, 암을 일으키는 BRCA1 변화는 가게 시퀀싱으로 찾는 드문 변이입니다. 흔한 차이만 볼 수 있는 방법은 유명한 유전자라도 그 유전자가 마침 지닌 흔한 차이 아래에 실을 수밖에 없습니다. 여기까지 스스로 도달한 학생은 도구가 무엇인지를 이해한 것입니다.

7번은 안전에 관한 질문이며, 묻지 않아도 가장 먼저 나올 질문입니다. 답은 소비자용 검사가 선별된 작은 패널만 본다는 것이고, 패널에 없다는 것은 그 사람에게 없다는 뜻이 아니라는 것입니다. 이것이 누군가의 실제 결과를 해석하는 쪽으로 흘러가지 않게 하십시오.

8번은 모형과 그 한계를 동시에 붙들 수 있는 학생에게 보상이 됩니다. 좋은 답: 이 책은 앞에서 뒤로 읽지 않는다, 세포마다 같은 책의 다른 장을 읽는다, 한 장의 본문 대부분은 산물이 만들어지기 전에 제거된다, 유전자 하나가 여러 산물을 만들 수 있다.

9번은 그 데이터가 누구에게서 수집되었는지에 관한 것입니다. 학생들은 목록이 자연을 그대로 비추는 중립적인 거울이라고 흔히 가정합니다. 목록은 무엇에 연구비가 갔고 무엇이 연구되었는지의 기록입니다.

10~12는 연구 문제를 원하는 학생을 위한 것입니다. 12번에는 구체적인 실제 답이 있습니다. 좌표는 유전체 빌드가 바뀌면 이동하는데 rsID는 안정적으로 유지되도록 되어 있고, 빌드가 어긋난 비교는 실무에서 실제로 오류의 원인이 됩니다.

더 해 보기

여기 나온 모든 유전자와 자리에는 이 사이트에 페이지가 있고, 무료이며 출처가 적혀 있습니다. 이어서 볼 만한 곳입니다.

- rs1815739 — 그림에 나온 ACTN3의 그 자리. 근육 단백질을 바꾸는 글자 하나입니다.
- 낭포성 섬유증 — 유전자 하나, 기술된 변이 2천 가지 이상, 그리고 어떤 변이를 지녔는지로 고르는 약.
- 유전자 지식 그래프 — 이 목록이 오늘 담고 있는 유전자와 변이와 연결의 개수, 그리고 각 연결을 어떻게 세는지.
- 이 사이트가 딴고 선 논문들 — 이 페이지들 뒤에 있는 모든 연구와, 이름이 다 적힌 저자들.

자유롭게 쓰십시오

어느 교실에서든, 어느 수준에서든, 상업적 이용을 포함해 자유롭게 쓰고 복사하고 고칠 수 있습니다. 출처 표기는 환영하지만 필수가 아닙니다. 이 사이트는 학생에게서 아무것도 수집하지 않습니다. 계정도, 업로드도, 유전 정보도, 언제까지나.