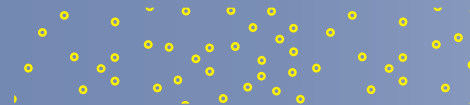




유전과 분자생물학



WEB2

유전체학

- W2.1 유전체지도 작성
- W2.2 유전체 염기서열 결정
- W2.3 유전체 프로젝트
- W2.4 유전체 주석달기와 데이터베이스
- W2.5 비교유전체학과 기능유전체학
- W2.6 유전체학의 응용

©William C. Ray, Director, Bioinformatics and Computational Biology Division, Biophysics Program, The Ohio State University

도입

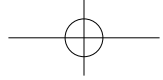
지난 30년 동안 생물학은 극적인 발전 속도를 보여 왔다. 1970년대 중반에 첫 번째 유전자의 동정을 시작으로, 연구자들은 1990년대 중반에 위 사진의 세균 종인 *Haemophilus influenzae*에서 처음으로 완전한 유전체 염기서열을 밝혔다(비슷한 기능의 유전자들이 같은 색으로 표시되어 있음). 21세기에 들어와 인간 유전체의 초안이 완성되었다. 20년 동안 단일 유전자 클로닝으로 백만 염기쌍의 서열을 결정하였고, 이후 5년 동안 30억 개 염기쌍의 서열을 분석하였다. 현재의 기술로는 몇 시간 안에 1000달러 정도의 비용만으로 200억 개의 염기서열을 분석할 수 있다. 16장에서는 생명공학에 대해 살펴보았고, 이번 장에서는 그러한 기술들이 전체 유전체 분석에 어떻게 응용되는지 알아본다. 유전체학에서는 고전유전학 및 분자유전학을 생명공학에 통합하여 유전체의 구조, 기능 및 연관성에 대해 연구한다.

W2.1 유전체지도 작성

학습 목표

1. 유전자지도와 물리지도를 비교하여 설명할 수 있다.
2. 제한지도, 염색체지도 작성 및 STS에 의한 지도 작성의 장단점을 설명할 수 있다.
3. 유전자지도와 물리지도가 어떻게 연관되는지 설명할 수 있다.

우리는 지도를 사용하여 위치를 찾고, 위치의 특성에 따라 세계 지도, 국가 지도 또는 도시의 지도 등 다양한 유형의 지도를 사용한다. 유전체학(genomics)에서는 유전자가 위치한 염색체, 염색체의 소구역, 그리고 최종적으로는 염색체 DNA 서열의 정확한 위치까지도 알 수 있게 해준다. DNA 서열 수준의 지도를 위해서는 유전체(genome) 전체의 서열을 아는 것이 필요한데, 아직까지는 기술적으로 불가능하다. 하지만 유전적 영역이 특정 표현형에 얼마나 영향을 주는지 알지 못한다면 유전체의 전체 서열은 무용지물이다. 이를 위해 유전자지도(genetic map)를 사용하게 된다. 이는 지리적 지도



(geographic map)에 비유할 수 있는데, 인간 유전체 서열 내에서 단일 유전자를 찾는 것은 세계 지도에서 집을 찾는 것과 같다.

유전체학은 전체 유전체를 분석하기 위한 많은 접근법을 사용한다

우리는 다른 종류의 데이터로부터 그리고 다른 종류의 분석을 위해 다른 종류의 유전체 지도를 사용한다. 기본적으로 유전자지도와 물리지도가 있는데, 이러한 유형의 지도는 **유전적 표지**(genetic marker)를 사용하여 개개인의 차이를 검출할 수 있게 한다. **유전자지도**(genetic map)는 재조합빈도에 따라 염색체상에 유전자 표지의 상대적 위치를 나타낸 추상적 지도이다(12장 참조). **물리지도**(physical map)는 유전체상에 유전자 표지의 정밀한 위치를 나타낸 것이고, 궁극적으로 전체 유전체의 DNA 염기서열을 나타낸 것이다. 유전체학 분야는 대규모 분석이 용이하도록 유전체의 다양한 유형의 지도나 그림 등을 총괄한다.

기술적으로 유전체에 대한 연구는 아니지만, 유전체와 RNA 발현 간의 관계, 그리고 유전체와 단백질의 발현 역시 연구할 수 있다. 유전체 연구는 진화적 유연관계, 질병에 대한 유전적 감수성 및 발생에 대한 관계를 밝혀냈다.

유전자지도는 유전적 표지 사이의 상대적 거리를 제공한다

연관지도(linkage map)라고도 불리는 유전자지도는 감수분열 동안 염색체 내 대립유전자의 연관관계가 바뀌는 재조합 과정을 사용한다(12장 참조). “유전적 표지”는 16장에서 기술된 PCR이나 제한핵산내부분해효소(restriction endonuclease)에 의해 확인할 수 있는 DNA 서열의 차이를 보이거나, 표현형의 차이로 검출되는 유전자를 말한다. 과거 유전자지도는 교배가 제어되는 환경에서 특정 유전적 표지 쌍에 대한 재조합 자손의 수로 작성되었다. 하지만 이러한 방법은 사람에게에는 적합하지 않았기 때문에, 가족 관계나 족보 등 이미 존재하고 있던 자료들을 활용하였고, 재조합빈도를 측정하기 위해 통계분석을 사용하였다.

유전자지도상의 거리는 cM(centimorgans, 센티모건) 단위로 표시되는데, 여기서 1 cM은 두 유전자좌(loci) 사이의 1% 재조합빈도(recombination frequency)를 나타낸다. 감수분열 동안 분리될 가능성이 1%밖에 안 된다는 점에서, 두 표지 간의 거리가 1 cM이라는 것은 매우 가까움을 의미한다. 초기에는 전체가 아닌 특정 염색체에 포함되는 질병감수성(disease susceptibility) 연관유전자만으로 사람의 유전자지도를 구성하였는데, 이것은 표현형으로는 검출할 수 없는 분자적 표지(molecular marker)가 나타난 후에야 바뀌게 되었

다. 이것은 많은 다형성 표지(polymorphic marker)를 제공할 뿐 아니라, 분자지도(molecular map)와 쉽게 통합이 가능하다. 사람의 유전자지도는 현재 매우 밀도가 높으며, 대략 1 cM마다 표지가 있다.

유전자지도는 염색체에 유전적 특성을 위치시키기 때문에 매우 중요하지만, 한계점은 있다. 먼저 교차(crossover)나 재조합의 빈도는 무작위적이지 않다는 점이다. 실제로, 인간, 생쥐 그리고 대부분의 생물에게서 재조합이 특히 많이 일어나는 곳들이 있다는 광범위한 자료들이 보고되었다. 게다가, 재조합빈도는 개인마다 다를 수 있고, 염색질 구조에 의해 영향을 받을 수도 있다. 이러한 사실들은 유전적 거리(cM, base-pair)와 실제 물리적 거리(bp) 사이의 관계가 유전체에 따라 다르다는 것을 의미한다. 그럼에도 불구하고 완전 유전체 염기서열과 밀집된 유전자지도는 함께 사용되어 유전체에 대한 보완적 시각을 제공할 수 있다.

물리지도는 유전적 표지 사이의 절대적 거리를 제공한다

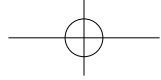
유전체 내 표지의 상대적 위치를 제공하는 유전자지도와는 다르게, 물리지도는 표지의 정확한 위치를 나타낸다. 최초의 물리지도는 분자생물학의 초기 도구였던 제한효소를 사용했다. 이후 기술이 정교해지면서 물리지도의 특성도 향상되었다.

물리지도의 궁극적 형태는 유전체의 완전한 DNA 서열에 많은 유전자 표지를 표시하는 것이다. 물리지도에서 표지 사이의 거리는 염기쌍[1000염기쌍(bp)=1킬로베이스(kb, kilobase)]으로 측정된다. DNA 단편의 염기서열이나 유전자 암호화 여부를 알지 못해도 물리적으로 지도 작성이 가능하다. 실제로, 많은 유전체 염기서열 분석 프로젝트가 물리지도를 작성하는 것으로 시작된다. (1) 제한핵산내부분해효소를 사용해 구성되는 제한효소지도(restriction map), (2) 염색체지도(chromosome map), 그리고 (3) 서열표지부위(sequence-tagged site, STS) 등의 일반적인 물리지도가 있다.

제한효소지도

제한효소지도(restriction map)는 일반적으로 50 kb 이상의 DNA 분자에는 적합하지 않다. 즉, 일부세포소기관 및 일부 바이러스 유전체에만 사용 가능하다. 먼저 물리지도는 단일 제한효소와 여러 제한효소의 조합으로 DNA를 절단하여 만들어진다(그림 W2.1). 제한효소로 절단된 DNA 절편들의 패턴을 분석하여 지도를 생성한다.

더 큰 유전체의 제한효소지도는 DNA를 먼저 작은 조각으로 자른 후 제작할 수 있다. 이러한 방법을 반복한 후 DNA 조각의 크기와 겹침에 따라 조각들을 다시 연속된 유전체 조각으로 모으는데, 이렇게 만들어진 하나의 연속된 조각을 **콘티그(contig)**라고 한다.



1. 여러 개의 복제된 DNA 절편이 제한 효소에 의해 절단 된다.

2. 효소 A, 효소 B, 효소 A와 B를 동시에 처리하여 생성된 절편들을 크기에 따라 분리하는 겔에 나란히 전기영동한다.

3. 동시에 잘려 만들어진 작은 것들이 각각에 의해 만들어진 큰 절편을 만들 수 있도록 모아서 절편들을 배열한다.

4. 물리지도가 만들어졌다.

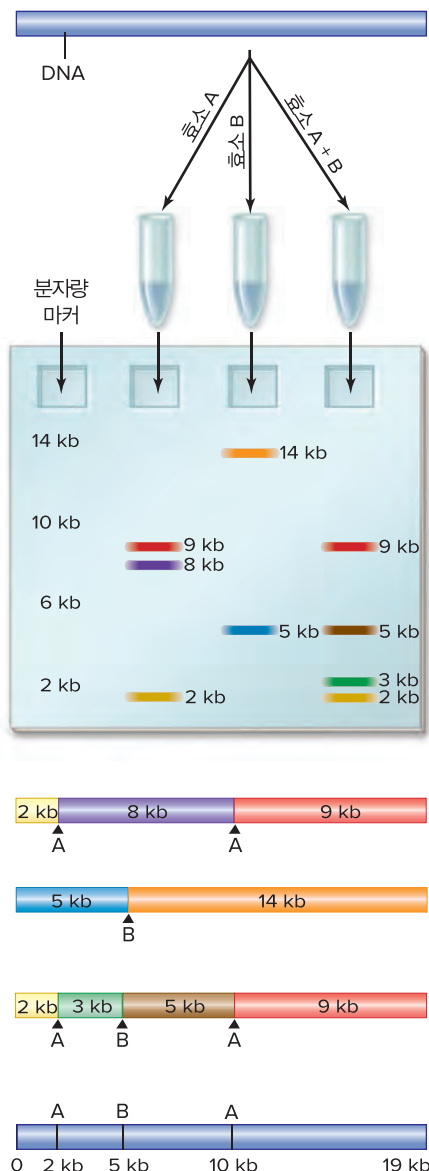
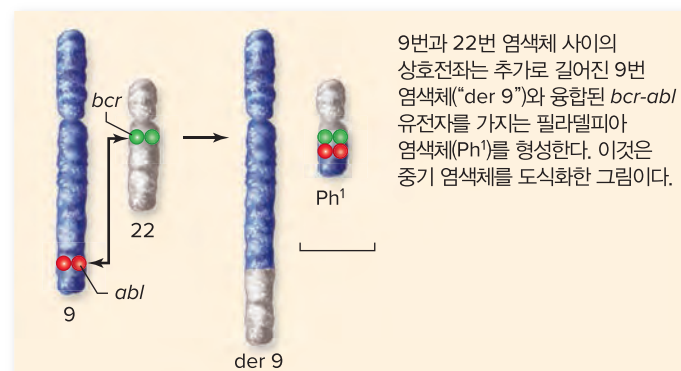


그림 W2.1 물리지도 작성에 이용되는 제한효소.

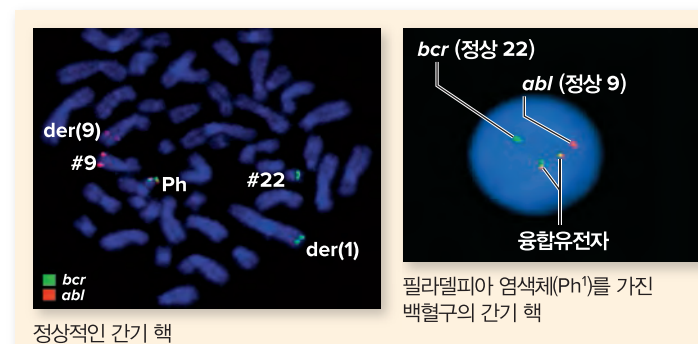
DNA를 2가지 다른 제한효소로 각각 처리한 것과 동시에 처리한 것의 절편을 전기영동을 통해 크기별로 분리한다. 위의 결과로 나타난 절편 크기를 비교하여 부위를 추론할 수 있다.

염색체 지도

세포학자들은 광학현미경을 이용한 염색체의 연구를 통해 여러 생물 변종의 염색체에서 재현성 있는 분염(banding) 패턴을 관찰하였다. 이 방법을 통한 분염 패턴에 따라 모든 염색체와 각 염색체의 소구역을 나눌 수 있었다. 여러 변종을 사용하여 전체 유전체의 세포학적 지도를 구축할 수 있는데, 각 염색체를 “짧은 팔(p arm)”과 “긴 팔(q arm)”로 나눈 후 분염 패턴에 따라 영역과 하위 영역으로 다시 한 번 나눈다. 이런 지도는 저해상도이다. 이러한 대규모 물리지도는 전체 유전체를 포함하지만 그 해상도는 낮은 점에서 전국 지도와



a.



b.

그림 W2.2 클론 DNA와 세포학적 지도의 연관을 위한 형광핵산혼성화(FISH)의 이용.

a. 사람 염색체의 핵형에서 보이는 9번과 22번 염색체 사이의 전좌. b. bcr(초록색)과 abl(빨간색) 탐침을 이용한 FISH. 노란색은 혼성화된 유전자(빨간색과 초록색 형광이 결합됨)를 나타낸다. abl 유전자와 융합된 bcr-abl 유전자는 모두 타이로신 키네이스를 암호화하지만 융합된 유전자는 항상 발현된다.

? 탐구문제 그림 W2.2b의 오른쪽에 보이는 빨간색, 초록색, 노란색 점의 패턴을 어떻게 설명할 수 있는가?

유사하다.

세포학적 지도는 만성 골수성 백혈병(chronic myelogenous leukemia)과 같은 사람의 질병과 관계된 염색체 이상을 찾는 데 이용될 수 있다. 이 질병에서는 9번과 22번 염색체에서 전좌(translocation)가 나타나며, 그로 인해 생성된 변성 타이로신 키네이스(tyrosine kinase)가 항상 활성화되어 백혈구 증식 및 이로 인한 백혈병을 일으키게 된다.

16장에서 기술된 형광핵산혼성화(fluorescent in situ hybridization, FISH) 역시 염색체 지도를 만드는 데 사용될 수 있다(그림 W2.2b). 초기에 FISH의 해상도는 약 1 Mb(megabase)였는데, 이것은 1 Mb보다 가까운 염색체 내 두 표지는 구분할 수 없음을 의미한다. 이러한 한계점 때문에 FISH 기법은 특정 염색체에 특정 DNA 서열의 존재 여부를 검출할 때만 사용 가능하다. 1990년대 기술의

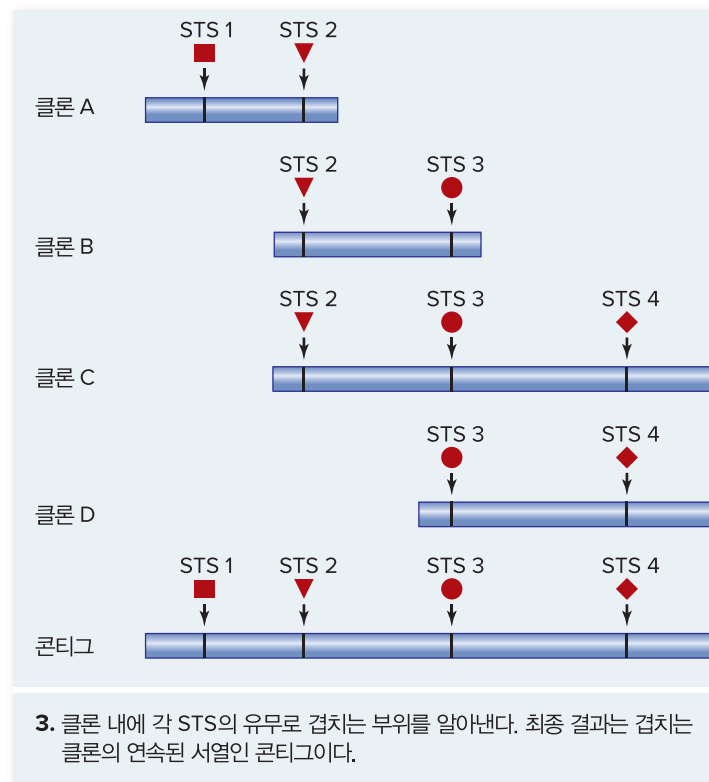
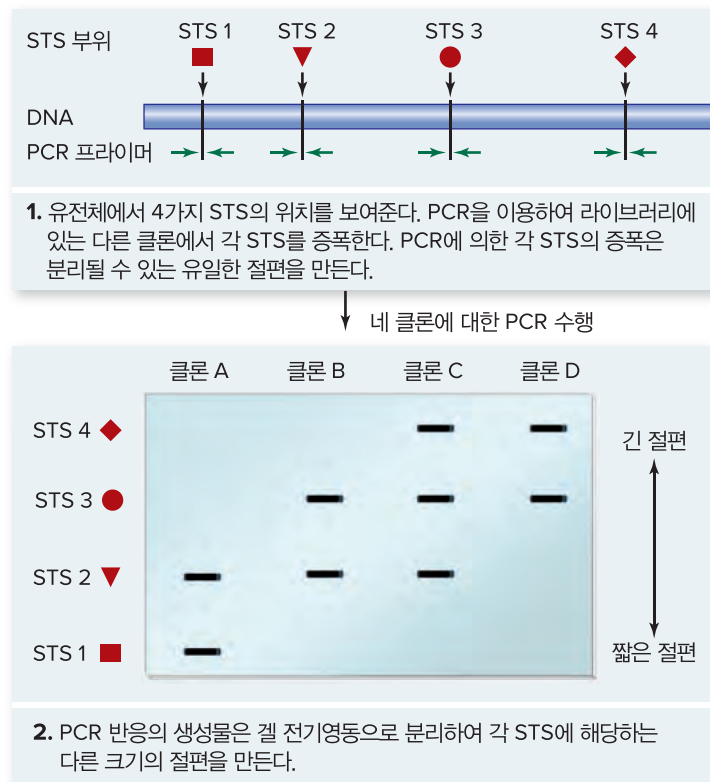
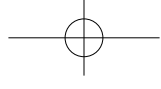


그림 W2.3 서열표지부위를 이용한 물리지도 작성.

사람의 유전체에서 서열표지부위(STS)라고 부르는 표지는 전체 유전체의 염기서열 결정을 위한 기초를 제공하기에 충분히 큰 물리지도 작성을 가능하게 해주었다. (1) 클론 DNA에 추가된 고유한 STS를 인지하는 프라이머(초록색 화살표)와 중합효소연쇄반응(PCR)을 통한 DNA 증폭. (2) PCR 생성물은 DNA 겔에서 크기에 따라 분리되고, 각 클론에 포함된 STS가 분리된다. (3) 클론 DNA 단편이 STS에 근거하여 정렬되어 콘티그를 만든다.

자료 분석 그림 W2.3의 패널 2에서 클론들의 작은 세트를 활용하여 물리지도를 구축할 수 있다. 물리지도를 구성할 수 있는 가장 적은 수의 클론 조합을 나열하시오.

발전으로 분열 중 간기와 중기의 염색체에도 사용하게 되었으며, FISH의 분해능을 25 kb로 개선하였다. 이러한 발전은 더욱 정밀한 지도 작성이 가능해졌음을 의미한다.

서열표지부위지도

제한효소지도는 상대적으로 작은 DNA 분자의 높은 처리량, 기술적 수월함, 그리고 고해상도를 가지는 반면, FISH 기법은 큰 DNA 분자의 낮은 해상도를 제공하고 기술적으로 까다로운 부분이 있다. 서열표지부위(sequence-tagged site, STS) 지도는 제한효소의 장점과 FISH의 장점이 조합된 매우 강력한 방법이다. STS 지도 작성을 통해 기술적 어려움이 없이 큰 DNA 분자의 고해상도 물리지도를 신속하게 구축할 수 있다.

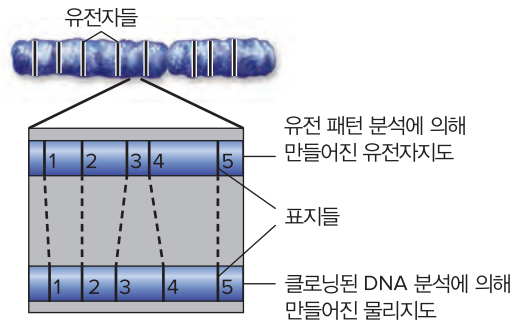
서열표지부위(STS)는 16장에 기술된 PCR에 의해 증폭될 수 있는 짧고 특이적인 유전체 DNA를 말한다. STS 지도 작성은 2개의 STS가 동일한 DNA 절편에 있는지 알아본다. 이것은 무작위로 절

단된 DNA 절편들을 큰 절편으로 중첩시키는 것도 가능하다. 만일 2개의 STS 표지가 하나의 DNA 절편에서 발견되었다면, 두 표지는 상대적으로 가깝다는 것을 의미한다. 하지만 다른 절편에서 두 STS 표지가 발견된다면, 더 멀리 떨어져 있을 수 있다(그림 W2.3). 유전자 지도 작성에서 표지가 재조합에 의해 분리되는 빈도에 근거한 것처럼, STS 지도 작성에서는 DNA를 절편으로 절단하면서 표지가 분리될 수 있는 빈도를 고려한다. 멀리 떨어진 표지는 서로 더 가까운 표지보다 자주 분리된다. 2개의 표지가 멀리 떨어져 있을수록 분리 빈도는 커지게 된다.

STS를 이용하면 중첩되는 부분을 식별하여 DNA 절편을 조립할 수 있다. 사람의 유전체에서는 STS의 밀도가 높고 DNA 클론에서 STS를 식별하기가 상대적으로 수월하였기 때문에, 연구자들은 1990년대 중반에 32억 bp 유전체에 대해 거대한 규모의 물리지도를 작성할 수 있었다(그림 W2.3). STS는 유전체 서열을 조립하기 위한 스캐폴드(scaffold)를 제공한다.

물리지도는 유전자지도와 연관될 수 있다

유전자지도를 통해 특정 표현형에 영향을 미치는 유전체의 부위는 식별할 수 있지만, 이 “유전자”의 특성에 대한 정보는 알 수 없다. 유전자지도에서 확인된 유전자의 서열을 알아내기 위해서는 물리지도와 유전자지도를 연관시키는 것이 필요하다.



유전자지도는 유전체 서열의 해상도 보다 좋지 않기 때문에 유전자를 발견하는 데 어려움이 있다. 사람의 유전체에서 1 cM 떨어져 있는 표지가 100만 bp 정도로 떨어져 있을 수 있기 때문이다.

현재 유전자지도는 주로 분자 표지를 사용하기 때문에 유전체 서열에서 그 위치를 쉽게 확인할 수 있다. 이와 유사하게, 클로닝된 유전자는 유전체 서열 내의 위치를 확인 할 수 있고, 유전자지도를 작성할 수 있다. 이를 통해 두 지도 간에 연관성이 자동적으로 만들어진다. 이론적으로 유전자지도에 작성된 유전자의 DNA 서열은 쉽게 찾을 수 있다. 하지만 실전에서는 유전적 거리와 물리적 거리의 관계가 유전체에 따라 변하기 때문에 쉽게 찾을 수 없다. 따라서 1 cM의 유전적 거리는 실제로는 유전체의 부위에 따라 염기쌍의 길이가 다르게 나타날 수 있다.

핵심 요약 W2.1

유전체지도에는 물리지도와 유전자지도가 있다. 물리지도는 제한효소지도와 염색체 분염의 세포유전학적 지도로 나뉜다. 유전자지도는 각 유전체의 고유한 STS(서열표지부위)와 같은 DNA 표지를 사용하고, 물리지도와 연관되어 있다. 유전자지도는 유전자와 같은 표지의 상대적 거리를 표시하는 반면, 물리지도는 그것들을 DNA 서열의 특정 유전자 위치에 표시한다. 유전자지도와 물리지도는 유전체 프로젝트의 염기서열 분석 단계를 돕기 위해 사용되었다.

- 유전자지도의 두 표지가 10 cM 떨어져 있고, 동일한 유전자지도의 다른 부분에 있는 두 표지는 12 cM 떨어져 있는데 어떻게 두 표지 세트가 같은 양의 DNA로 분리되어 있을까?

W2.2 유전체 염기서열 결정

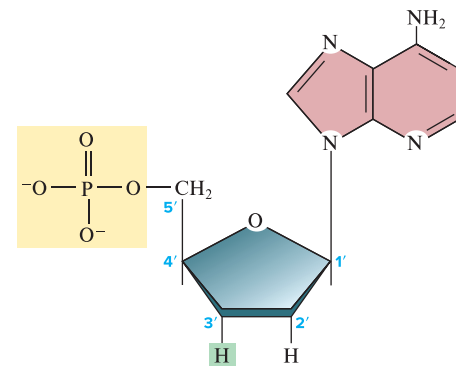
학습 목표

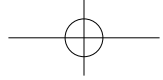
1. 디데옥시 종결 염기서열 결정법과 차세대 염기서열 결정법을 구별하여 설명할 수 있다.
2. 클론을 이용한 염기서열 결정법과 샷건 염기서열 결정법의 차이점을 설명할 수 있다.

궁극적인 물리지도는 전체 유전체의 염기서열을 결정하는 것이다. 대규모 및 고처리 DNA 염기서열 결정법은 소규모 실험실과 생명공학회사의 전체 유전체 염기서열 결정을 현실화하였다. 개념적으로, 모든 DNA 서열결정법(sequencing)은 시험관내(in vitro) 환경에서의 DNA 복제를 응용한다. 이전의 DNA 서열결정법은 복제를 중지 또는 일시정지시키는 변형된 뉴클레오타이드를 사용하였고, DNA의 단일가닥만을 복제한다. 보다 빠르고 자동화된 염기서열 결정 기술로 유전체학은 빠르게 발전할 수 있었다.

디데옥시 종결 염기서열 결정은 유전체 염기서열 결정에 여전히 중요하다

지난 10년 동안 DNA 서열분석에 사용된 기술은 크게 향상되었다. 그러나 과거의 DNA 염기서열 결정법도 여전히 유용하다. 1970년대 생거(Fred Sanger)는 DNA 복제를 모방하여 DNA 염기서열 결정법을 개발하였는데, 이 기술은 DNA 염기서열을 결정하는 반응에 디데옥시뉴클레오타이드(dideoxynucleotide)를 사용한다. 이 디데옥시뉴클레오타이드는 사슬종결자(chain terminator)로서의 역할을 하게 되는데, 복제 과정에 이 물질이 포함되면 반응이 멈추게 된다. 모든 DNA 뉴클레오타이드는 당의 2번 탄소에 수산기(hydroxyl group)가 없지만 디데옥시뉴클레오타이드는 3' OH도 가지지 않는다. 13장을





떠올려 보자. 추가되는 각 염기들은 DNA 가닥의 3' 방향으로 확장하게 되는데, 기존 뉴클레오타이드의 OH기와 첨가되는 뉴클레오타이드의 삼인산과 반응하게 된다(그림 13.12). 디데옥시뉴클레오타이드는 첨가될 수는 있지만 3' OH가 없기 때문에 다음 반응을 할 수 없게 되고 사슬의 확장이 종결된다.

디데옥시 종결 염기서열 결정법(dideoxy terminator sequencing)은 원래 각각 단일 디데옥시뉴클레오타이드와 4개의 데옥시뉴클레오타이드를 갖는 4개의 개별 반응을 실험자가 직접 수행하였다. 디데옥시뉴클레오타이드가 포함되면 특정 염기에서 합성이 종결된다. 따라서 각각의 반응은 특정 염기로 종결되는 절편의 세트를 생성하게 되고 4개의 반응을 함께 살펴보면 각 염기로 끝나는 절편들을 나타낼 수 있다. 이후 단일 염기 길이에 따라 절편들을 분리할 수 있는 고해상도 겔 전기영동으로 절편들을 분리시킨다. 최종적으로 DNA 염기서열을 맨 아래(가장 작은 절편)부터 맨 위(가장 긴 절편)까지 읽을 수 있는 조각들이 생성된다.

위 방법을 자동화하기 위해, 각각의 디데옥시뉴클레오타이드(dideoxynucleotide)를 서로 다른 형광염료로 표시시킨 후, 단일 DNA 합성반응에 4개를 모두 첨가한다. 위와 같이 디데옥시뉴클레오타이드의 무작위적인 첨가는 특정 염기로 종결되는 일련의 절편들을 생성하지만, 이 경우에는 어떤 염기로 종결되는지에 따라 서로 다른 색을 나타낸다. 염료로 표시되고 합성이 종결된 절편들은 전기영동에 의해 모세관(capillary tube)에서 분리되고 각 분자는 레이저를 통과할 때 형광을 발하게 된다. DNA 가닥을 종결시키는 뉴클레오타이드는 형광색을 기반으로 해석하고 컴퓨터에 자동으로 보고된다(그림 W2.4). 이 방법은 여전히 노동 집약적이지만, 자동화를 통해 수천 건의 염기서열 결정 반응을 병렬로 수행할 수 있게 되었다. 즉, 하루에 약 900 kb의 DNA 염기서열을 분석할 수 있게 되었다. 어떤 면에서 이 방법은 차세대 염기서열 결정 기술의 특징인 대규모 병렬 염기서열 결정법이 탄생되는 첫 번째 단계였다.

차세대 염기서열결정법에서는 처리량을 증가시키기 위해 대규모 병렬 기술을 사용한다

10여 년 전 차세대 염기서열 결정법(next-generation sequencing, NGS) 기술이 처음 등장하였다. 그 이후로 염기서열 결정법의 속도, 염기당 비용, 반응당 판독 가능 길이가 크게 개선되었다. 그 결과 지난 10년 동안 염기서열이 분석 및 재분석된 유전체의 수가 엄청나게 증가하였다. 이런 풍부한 데이터들은 앞으로 새로운 진화 관계를 나타내고, 개개인의 건강관리와 진단의학을 개선할 수 있게 할 것이다.

수많은 NGS 기술들이 개발되었지만 모두 공통된 특징이 있다. 첫

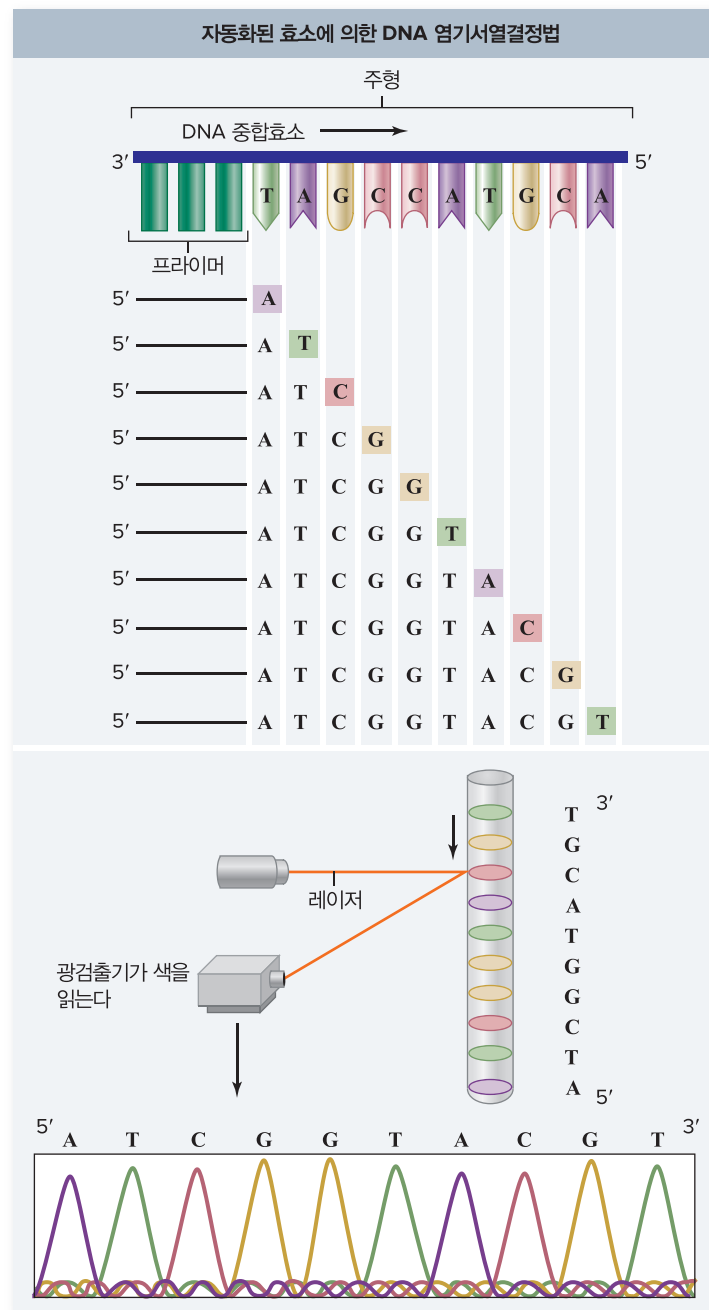
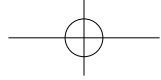


그림 W2.4 자동화된 디데옥시 DNA 염기서열 결정법.

분석될 DNA 서열은 프라이머가 부착된 DNA 중합효소에 대한 주형가닥으로서 상단에 나타내었다. 자동 염기서열 결정에서 각 ddNTP는 다른 색의 형광염료로 표시되어 단일 반응을 수행할 수 있다. 반응에 의해 생성된 절편을 그림으로 나타내었다. 이들이 모세관에서 전기영동될 때, 바닥에 있는 레이저에 의해 염료가 자극되고, 각 염료들은 서로 다른 색을 방출하고 광검출기에 의해 검출된다.

째, 디데옥시 종결 염기서열 결정법과 달리, 기존 클로닝에 의한 유전체 라이브러리의 구축 없이 염기서열 분석이 가능하다는 것이다(16장 참조). 둘째, 수천 개의 반응이 동시에 준비되고 분석되는 대신 수백만 건 이상의 염기서열 결정 반응이 동시에 수행될 수 있다. 따라서 이 기술에는 “대규모 병렬(massively parallel)”이라는 용어



가 사용된다. 셋째, 이전의 시간 소모적인 전기영동이 필요하지 않고, 염기서열 결정반응이 용액에서 동시에 일어나고 염기서열 결정 장비에서 직접 읽게 된다(그림 W2.5). 염기서열 결정 속도의 증가와 비용의 감소로 인해 일부 세균 유전체를 단 몇 시간 내에 분석할 수 있게 되었다.

차세대 염기서열 결정법으로 인해 비용이 크게 절감되고 훨씬 더 많은 양의 데이터가 생성되지만 단점 역시 존재한다. 이 기술을 통해

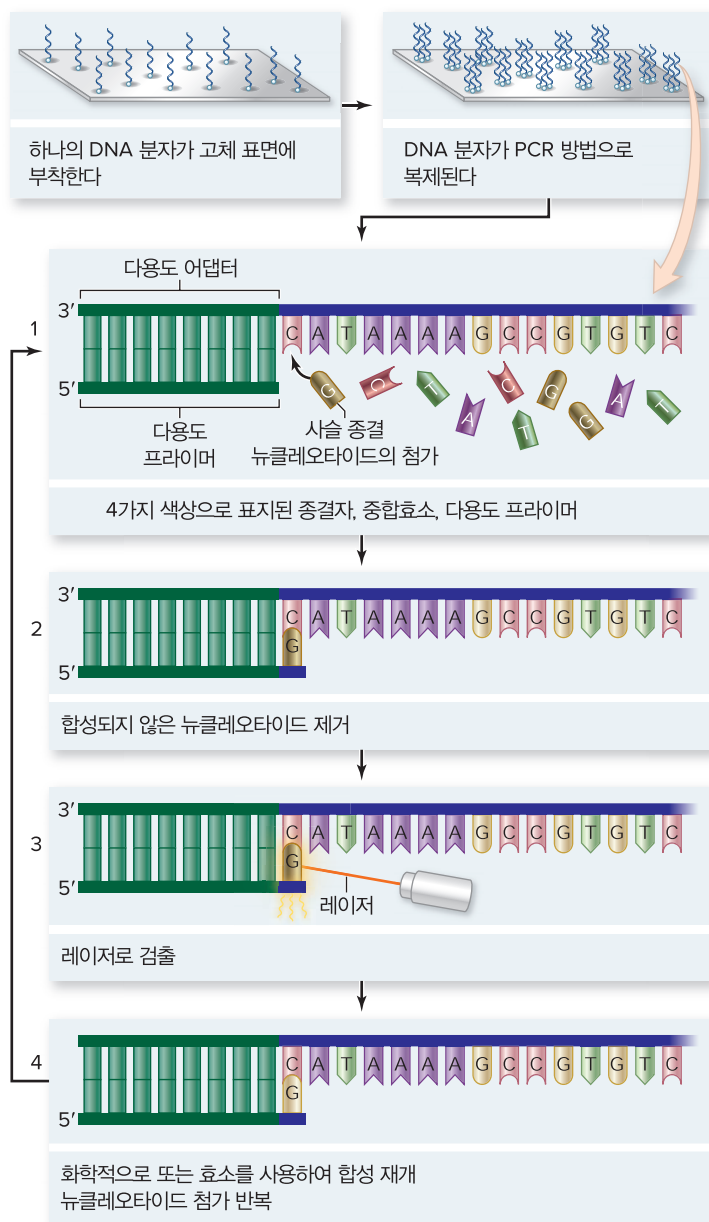


그림 W2.5 일루미나 차세대 염기서열 결정법.

일루미나(Illumina) 차세대 염기서열 결정법에서 주형 DNA 분자는 고체 표면에 부착되고 PCR에 의해 증폭되어 DNA 섬을 생성한다. 이후 각 4개의 형광 표지된 종결자 뉴클레오타이드가 첨가되고, 하나가 새로운 DNA에 도입된 후 합성이 중지된다. 첨가된 뉴클레오타이드의 형광색을 판독한 후, 뉴클레오타이드를 화학적으로 처리하여 다른 뉴클레오타이드가 첨가될 수 있게 한다. 이 단일 뉴클레오타이드 첨가 과정은 각 섬에서 DNA 서열을 읽기 위해 반복된다.

많은 분자들을 동시에 염기서열 결정을 할 수 있지만 분자당 정보가 적게 생성된다. 다시 말해서, 대부분의 차세대 염기서열 결정법의 염기서열 판독길이(read length)는 50~1000 bp로 짧은 경향이 있다. 따라서 정확한 유전체 조립(genome assembly)을 위한 컴퓨터 성능에 더욱 중점을 두게 되었다. 이로 인해 새로운 조립 알고리즘이 생겨났지만, 중복되는 DNA가 수없이 많아도 짧은 판독길이에 의해 어려움을 겪고 있다.

훨씬 더 긴 판독, 심지어 수만 bp까지 생성되는 최신 기술도 있다. 이러한 기술의 단점은 오류가 발생하기 쉽다는 것이다. 최근 염기서열 결정법은 길지만 덜 정확한 방법과, 짧지만 더 정확한 방법을 결합시키는 방향으로 발전되었다.

2014년에는 1000달러만으로도 사람의 유전체를 분석할 수 있게 되었다. 이것은 대략 10년 전의 사람 유전체 크기의 염기서열 결정 초기에 비해 거의 10,000배의 비용이 절감된 것이다.

염기서열이 결정된 유전체 절편들은 완전한 염기서열로 조립된다

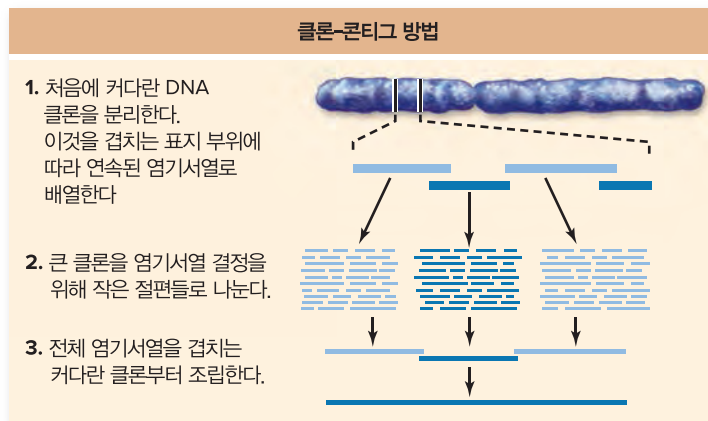
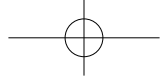
대규모 염기서열 프로젝트의 가장 복잡한 작업은 개별 염기서열 조각들을 완전한 유전체로 조립하는 것이다. 유전체 DNA는 먼저 겹치는 부분이 있는 조각으로 절단되고, 염기서열이 분석된 후에 다시 중첩되는 부분들을 일치시켜 재조립되어야 한다. 염기서열 결정이 되는 절편이 짧을수록, 그리고 유전체 DNA에 반복되는 부분이 많을수록 이 퍼즐을 맞추기가 더 어려워진다. 조립에는 (1) **클론-콘티그 조립**(염색체의 일부를 먼저 조립한 다음, 더 큰 조각을 결합)과 (2) **샷건 조립**(모든 조각을 한 단계씩이 아닌 한 번에 조립)의 2가지 전략이 이용될 수 있다.

클론-콘티그 조립

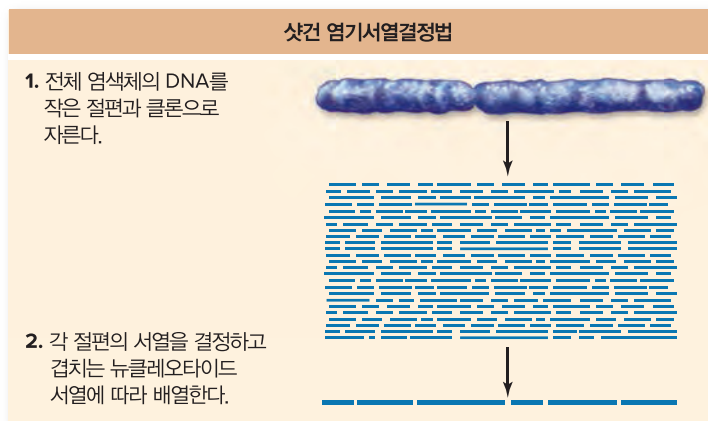
클론-콘티그 조립(clone-contig assembly, 그림 W2.6a)은 진핵생물과 같은 보다 복잡한 유전체의 DNA 염기서열 조각을 조립하는데 사용된다. 샷건 조립과는 달리, 이런 종류의 방법은 완전히 재조립된 유전체 염기서열을 구축하기 위해서 물리지도가 필요하다. 먼저 유전체를 약 1~1.5 Mb의 큰 절편으로 조각화한 후 이 절편을 STS 또는 제한효소부위와 같은 유전자 표지를 사용하여 **콘티그(contig)**라고 하는 더 크게 이어진 DNA 절편에 일치시킨다. 크기가 1~1.5 Mb인 DNA 절편들은 샷건 방법을 사용하여 염기서열들이 분석되고 조립될 수 있다.

샷건 조립

샷건 조립(shotgun assembly)의 장점은 유전자지도나 물리지도



a.



b.

그림 W2.6 염기서열 결정법과 유전체 조립법의 비교.

a. 클론-콘티그법은 STS에 의해 겹치는 부위를 조립한 커다란 클론을 사용한다. 한번 조립 되면 염기서열 결정을 위해 작은 절편으로 나누어질 수 있다. b. 샷건 염기서열 결정법에서는 전체 유전체를 작은 클론으로 조각내고 염기서열을 결정한다. 컴퓨터 알고리즘이 최종 DNA 서열을 겹치는 뉴클레오타이드 서열에 근거하여 조립한다.

에 의존하지 않는다는 점이다. 이는 유전체 염기서열에 대한 사전정보가 필요하지 않음을 의미한다. 최초로 전체 유전체의 염기서열이 분석된 세균인 *Haemophilus influenzae*는 샷건 방법을 사용하여 염기서열이 분석 및 조립되었다. 이 세균 유전체 DNA는 1.5~2 kb로 조각화되어 적절한 벡터(vector)로 클로닝되었다(16장 참조). 절편 말단의 염기서열을 분석하고, 중복된 부분을 기반으로 DNA 한 조각, 한 조각들을 분석하여 전체 1.8 Mb의 유전체를 만들 수 있었다. 이 방법으로 완전하게 조립된 유전체 서열을 얻기는 매우 어렵다. 따라서 샷건 방법으로 조립된 염기서열 조각들을 다양한 다른 방법을 사용하여 간격을 메우게 된다.

핵심 요약 W2.2

유전체의 크기가 매우 큰 생물도 있기 때문에, 염기서열 결정을 위해서는 많은 반응을 병렬로 실행할 수 있는 자동화된 분석기를 사용해야 한다. 전체 유전체 염기서열분석 및 조립을 위한 2가지 방식이 개발되었다. 하나는 물리지도에 의해 이미 정렬된 클론을 사용하는 방법(클론-콘티그 염기서열 결정 및 조립)이고, 다른 하나는 무작위 클론을 염기서열 결정을 하고 컴퓨터를 사용하여 최종 서열을 조립하는 방법(샷건 염기서열 결정과 조립)이다. 두 방식 모두 상당한 컴퓨터 성능을 필요로 한다.

■ 전체 유전체 염기서열 결정과 조립의 방식에 따른 장단점은 무엇일까?

W2.3 유전체 프로젝트

학습 목표

1. 인간 유전체 프로젝트에 대해 설명할 수 있다.
2. 인간 유전체 염기서열 결정이 왜 밀보다 더 간단하였는지 설명할 수 있다.
3. 1000 유전체 프로젝트의 잠재적 이점을 설명할 수 있다.

염기서열 결정 기술의 자동화로 대량의 서열 데이터를 얻게 되었고, 이를 통해 복잡한 문제를 연구하는 연구자들이 개별 유전자 분석에 국한되었던 연구방법의 한계를 넘어설 수 있게 되었다. 하지만 염기서열 분석 프로젝트는 유전체의 조직, 기능, 그들 간의 상호관계 등 어떠한 정보도 알 수 없는 단순히 나열된 염기서열만을 알게 해주었다. 이후 유전체 프로젝트의 데이터를 바탕으로 한 연구들을 통해 유전체에 대한 통찰력을 제공해 주고 있다.

인간 유전체 프로젝트를 통해 인간 유전체 대부분의 염기서열이 결정되었고 지도로 작성되었다

인간 유전체 프로젝트(Human Genome Project, HGP)는 공식적으로 1990년에 시작되었지만 1980년대 중반부터 그 뿌리를 추적할 수 있다. 인간 유전체의 유전자지도 및 물리지도의 제작은 궁극적으로 유전체의 실제 염기서열 분석 및 조립을 위한 노력이었다. 1995년에 인간 유전체의 94%가 평균 199 kb 간격의 표지로 이루어진 물리지도가 제작되었다. 1996년에는 약 1.6 cM마다 표지가 작성된 유전자지도가 발표되었다. 그 이후로 표지의 수와 지도의 해상도가 크

게 향상되는 동시에 염기서열 결정법 자동화의 발전으로 염기서열 결정 비용이 염기당 40센트로 낮아졌다. 인간 유전체 프로젝트는 클론-콘티그 및 샷건 방식을 기반으로 최종적으로 완성되었다.

1998년, 유전체 연구기관(Institute of Genomic Research)의 벤터(Craig Venter)는 자신의 민간 영리 회사가 HGP와 직접 경쟁하여 인간 유전체를 완성할 것이라고 선언하였다. 벤터는 그의 샷건 염기서열 결정법 및 조립 방식이 HGP의 클론-콘티그 방식보다 빠르고 저렴하다고 주장했다. 그의 주장이 사실이기는 했지만, 벤터의 회사 [현재 셀레라 제노믹스(Celera Genomics)]는 공개되어 사용 가능했던 HGP의 물리지도를 참고하였다. 벤터의 방식으로 거대하고 복잡한 유전체에 대해 성공했는지는 확실하지 않지만, 훨씬 작고 간단한 유전체인 *Haemophilus influenzae*에 대해서는 분명히 성공하였다. 2000년 6월, HGP에 공개적으로 자금을 지원한 회사와 벤터의 회사에 비밀리에 후원한 회사가 공동으로 인간 유전체의 초안을 발표하게 되지만, 몇몇 논문을 통해 많은 차이와 오류가 있음이 밝혀졌다. 하지만 누락된 서열의 많은 부분이 유전자를 포함할 가능성이 낮았기 때문에, 완전하지는 않지만 인간 유전체의 기능적인 부분은 모두 포함된 버전이라는 광범위한 합의가 있었다.

2003년 4월, 인간 유전체 염기서열 협력단(Human Genome Sequencing Consortium)은 인간 유전체의 염기서열 분석이 완료되었다고 발표하였다. 초안 유전체와 비교하여 완료된 유전체는 염기서열의 오류와 읽지 못한 부분이 적고, 더 없는 범위를 해독하였다. 이 유전체를 완성하는 데 13년이 소요되었고, 1991년 미국 달러의 가치를 기준으로 약 27억 달러의 비용이 지불되었다. 오늘날 인간 유전체는 고처리 염기서열 분석기를 사용하여 최소 2~3시간 안에 약 1000달러의 가격으로 분석할 수 있다.

자료 분석 인간 유전체에 약 30억 개의 염기쌍, 20,000개의 유전자 및 1%의 유전자 유전암호만 포함되어 있다면 각 유전자의 대략적인 길이는 얼마인가?

오랜 시간 동안 유전학자들은 인간 유전체의 유전자수가 약 100,000개라고 추정하였다. 인간 유전체 프로젝트에서 발견된 것 중 하나는 인간 유전체의 유전자 수가 약 20,000개라는 것이다. 이는 과일보다 약 1.5배 많고 쌀의 거의 절반에 해당한다(그림 W2.7). 인간, 생쥐 및 복어는 비슷한 수의 유전자를 가지고 있는데, 이는 생명체의 복잡성이 유전체의 유전자 수나 크기와 단순히 비례하는 것이 아님을 의미한다.

밀 유전체 프로젝트는 유전체 조립의 개선을 보여준다

밀(*Triticum aestivum*)은 전 세계 인구의 약 30%가 섭취한다. 밀 유전체학의 이해는 수확량을 늘릴 수 있는 교배정보를 확보하고, 가뭄과 질병에 대한 저항성을 향상시킬 수 있을 것이다. 또한 밀 종의 진화에 대한 통찰력을 제공할 수 있다. 인구 증가, 물 자원의 감소 그리고 기후변화를 고려할 때, 안정된 식량 공급을 유지하기 위해서는 현대유전학을 작물 개선에 적용하는 것이 중요해 질 것이다.

유전적으로 밀은 2번의 잡종화에 의해 만들어진 이질육배체(allohexaploid; WEB3장 참조)이다. 이것은 밀 유전체가 약 50억 bp의 DNA를 각각 함유하는 A, B, D로 지정된 3개의 2배체 조상을 포함하고 있다는 것을 의미한다. 현대 밀 유전체 크기는 16 Gbp로, 각 조상으로부터 유래한 7개씩, 총 21개의 염색체가 있다. 밀 유전체의 염기서열을 결정하고 조립하는데 (1) 너무 거대한 크기와 (2) 배수체 특성으로 인해 거의 동일한 서열이 반복되는 등, 2가지 주요 문제가 있다.

밀 유전체의 염기서열 결정이 처음 시도된 2012년에는 전체 유전체의 약 1/3인 5.4 Gbp만이 조립되었다. 2년 후에 발표된 두 번째 시도는 염색체 염기서열을 한 번에 분석하여 유전체의 약 2/3인 10.2 Gbp의 유전체를 조립하였다. 2017년 초 유전체의 78%를 조립하는 세 번째 시도가 이루어졌다. 이 3가지 모든 조립에 대한 문제점은 콘티그의 길이가 상대적으로 짧다는 것이다(W2.2절 참조). 2017년 후

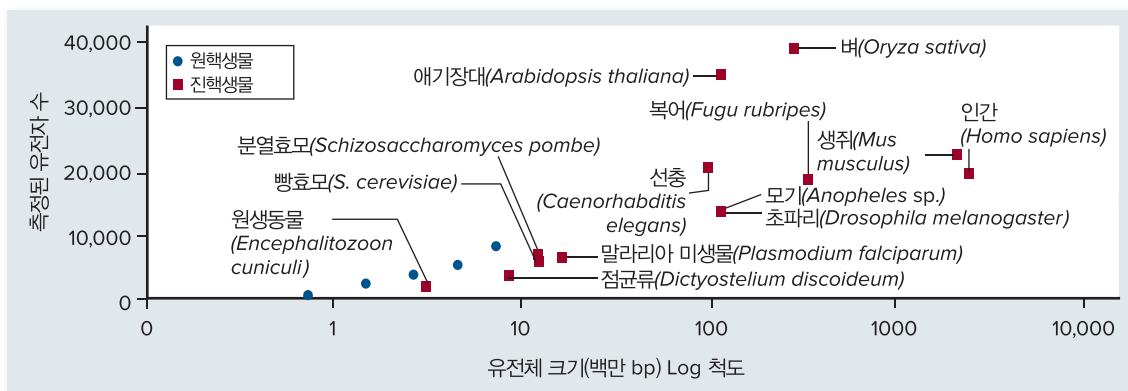
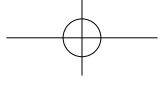


그림 W2.7 유전체의 크기와 복잡성.

유전체의 크기가 결정적인 요인은 아니지만, 일반적으로 진핵생물 유전체는 원핵생물보다 더 크고 많은 유전자를 갖는다. 생쥐의 유전체는 인간 유전체보다 크고 쌀 유전체는 인간과 생쥐보다 크다.



반, 이전의 노력보다 10배 이상 긴 콘티그로 구성된 15.3 Gbp의 거의 완전한 유전체 염기서열이 발표되었다. 이는 매우 긴 판독길이를 생성하지만 오류 빈도가 높은 기술과, 짧지만 더 정확한 판독이 가능한 최신 염기서열 결정법의 융합을 통해 달성할 수 있었다.

현대의 밀을 만들게 된 최근의 잡종화에 사용된 “D 유전체”인 *Aegilops tauschii*의 첫 번째 유전체 분석에도 비슷한 방식이 사용되었다. *Ae. tauschii* 유전체는 유전체의 84%를 구성하는 전이인자(transposable element)를 포함하여 고도로 흩어진 반복된 서열을 가지고 있다. *Ae. tauschii* 유전체는 이러한 반복적인 인자들 사이의 재조합으로 인해, 다른 곡식의 유전체보다 더 빠르게 대규모 재배열을 축적하고 있는 것으로 보인다.

1000 유전체 프로젝트는 유전적 다양성을 다룬다

인간 유전체 프로젝트는 인간 유전체의 표준서열(reference sequence)을 만들었다. 이 표준서열은 간극이 채워지고 정확도가 개선되었지만 결국 한 사람의 유전체만을 나타낸다. 즉, 한 사람의 유전체가 다른 누군가와 얼마나 다른지에 대한 궁금증은 해결할 수 없다. 1000 유전체 프로젝트(the 1000 Genomes Project)는 인간 유전자 내 공통 변이에 대한 포괄적인 목록을 만들기 위해 수행되었다. 이를 위해 26개의 다른 나라로부터 1000명 이상의 개개인 유전체 염기서열 결정이 이루어졌다.

이 연구는 인간 유전체 내 8000만 개의 유전자 변이를 확인하였다. 이 변이들을 기준으로 한다면, 우리의 유전체와 표준서열 사이에는 4~500만 개의 부위가 다를 것으로 예상된다. 변이의 99%는 SNP와 짧은 삽입 또는 결실(indels)로 구성된다. 구조적 변이로 불리는 큰 변이도 2100~2500개가 있는데, 여기에는 약 1000개의 큰 결실(large deletion), 160개의 복제수변이(copy-number variant) 및 약 1000개의 전이인자 삽입 등이 포함된다. 이러한 구조적 변이들은 크기가 크기 때문에 하나의 부위가 아닌 대규모 염기쌍에 영향을 미친다.

본 프로젝트에서 나타난 연구결과와 대다수 변이는 희귀하지만, 특정 유전체의 변이는 공통적으로 나타난다. 이것은 희귀한 변이가 집단에서 무작위로 일어나는 새로운 돌연변이임을 보여준다. 이 결과는 우리 인간 종의 진화 역사를 추론할 수 있게 해주는데, 인간은 15~20만 년 전에는 인구통계학적 역사를 공유하는 것으로 보인다. 약 1만 5천~2만 년 전에는 유럽, 아시아 및 아메리카의 일부 소집단들은 1000명 이하로 추정되는 집단 크기로 병목 현상이 있는 것으로 보인다. 이러한 인구 병목 현상은 급속한 인구증가에 의해 이루어졌고 현재 수많은 희귀한 변이 패턴을 만들어 냈다.

연구결과는 또한 유전자 기능을 조사하는 데 사용될 수 있다. 동형접합성 결실이 있는 유전자는 결실로 인한 명확한 표현형을 나타내지 않기 때문에 필수적이지 않음을 나타낸다. 이런 식으로 200개 이상의 유전자를 삭제할 수 있다. 이들 유전자 중 다수는 번역글로불린 영역을 갖는 단백질을 암호화하는 것으로 보이기 때문에 개인 간의 질병 감수성의 차이와 관련이 있을 것으로 추정한다.

핵심 요약 W2.3

인간 유전체 프로젝트는 완성하는 데 13년의 시간과 27억 달러의 비용이 필요했다. 밀 유전체는 훨씬 더 복잡하지만 염기서열 결성 기술의 발달로 더 빠르게 완성할 수 있었다. 이것은 점점 완성도 높아지는 조립체가 큰 역할을 하였다. 1000 유전체 프로젝트는 유전자 변이를 목록화한다. 보통 유전체는 표준서열과 400만 개의 부위에서 다르다. 이 연구결과는 인간이 20만 년 전에는 단일 집단으로 구성되었음을 암시한다.

- 밀과 같은 복잡한 유전체에 더 긴 염기서열 판독이 중요한 이유는 무엇일까?

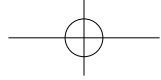
W2.4 유전체 주석달기와 데이터베이스

학습 목표

1. 유전체 주석달기가 왜 필요한지 설명할 수 있다.
2. 유전체 내 DNA 유형을 비교하고 대조할 수 있다.
3. 엔코드 프로젝트의 결론을 평가할 수 있다.

유전체 프로젝트의 유전체 염기서열 결정과 조립단계에서는 방대한 양의 데이터를 생성하게 된다. 여기에 유전체 서열 및 구조의 후속 분석에서 추가 데이터가 생성되어 이 데이터를 구성하고 저장하는 방법에 대한 문제가 나타난다. 서열 데이터 및 데이터의 분석 후 생성되는 정보는 검색이 가능하고, 공개된 다양한 데이터베이스에 저장되게 된다.

데이터 생산이 어려움에도 불구하고, 유전체 서열 자체는 본질적으로 의미가 없다. 유전체에 존재하는 유전자의 수, 종류 및 표현형 등이 더 중요한 의미를 갖는다. 유전체 서열에 이러한 정보를 배정하는 것을 유전체 주석달기(genome annotation)라고 한다.



유전체 주석달기는 유전체의 DNA 서열에 기능을 배정한다

주석달기(annotation)는 DNA 서열이 유전자로 식별되어 있는 데이터베이스에 DNA 서열을 첨부하여 분석하는 것이다. 이 분석에는 유전자, 유전자의 구조와 기능 그리고 궁극적으로 발현되는 RNA나 단백질에 대한 정보가 포함될 수 있다. 주석달기에는 유전자 기능이 어떻게 배정되었는지에 대한 정보도 포함될 수 있다. 가장 중요한 목표는 생물학적으로 관련된 정보를 유전체 DNA 서열에 추가하는 것이다.

유전체 주석달기는 전문가의 필요에 의해 만들어진 거대 자동화 과정이다. 주석달기 과정은 컴퓨터 알고리즘을 사용하여 유전자가 포함될 수 있는 DNA 영역을 찾는 것으로 시작한다. 알고리즘은 프로모터와 같은 조절영역을 포함하는 서열을 찾거나, 개시코돈으로 시작하여 아미노산으로 암호화되고 마지막으로 종결코돈으로 이어지는 부분을 찾을 수 있다. 이와 같은 염기서열을 **열린 번역틀(open reading frame, ORF)**이라고 한다. 또한 인트론/엑손 구조를 예측하기 위해 스플라이싱 신호를 검색하고, 최종적으로 잠재적 유전자로부터 생성될 수 있는 RNA의 서열 및 구조 그리고 아미노산 서열을 예측하는 것이 가능하다.

? **탐구문제** 유전체 서열이 주어진다면 어떻게 유전체에서 생성되는 단백질의 수를 예측할 수 있는가? 그 추정치가 정확하지 않은 이유는 무엇인가?

중 간의 기능추론: BLAST 알고리즘

잠재적 유전자나 생성물이 한번 확인된 유전자는 다른 종에서 유사한 서열을 찾아 기능을 유추할 수 있다. 이를 가능하게 하는 도구는 BLAST(Basic Local Alignment Search Tool)라는 검색 알고리즘이다. BLAST 검색 역시 자동화를 통해 주석달기 처리 속도를 높일 수 있다. 컴퓨터 네트워크를 이용하여 BLAST 서버에 염기서열을 입력하면 염기서열 데이터베이스에 이미 등록된 유사한 염기서열들을 얻을 수 있다. 다른 종의 다른 유전자 또는 유전자 산물과의 유의미하게 높은 유사성은 기능들이 진화과정 동안 보존되고 있음을 의미한다.

유전자 검색, 유전체 비교, 유전체 조립을 위해 컴퓨터 프로그램을 사용하는 것은 **생물정보학(bioinformatics)**이라고 하는 새로운 유전체 접근 방식 중 일부일 뿐이다.

발현 서열 표지를 이용한 전사된 유전자 식별

유전체에서 잠재적 유전자 서열을 확인하는 또 다른 방법으로는 다른 조직에서 생성된 mRNA를 분리하고 이를 이용해 cDNA를 합

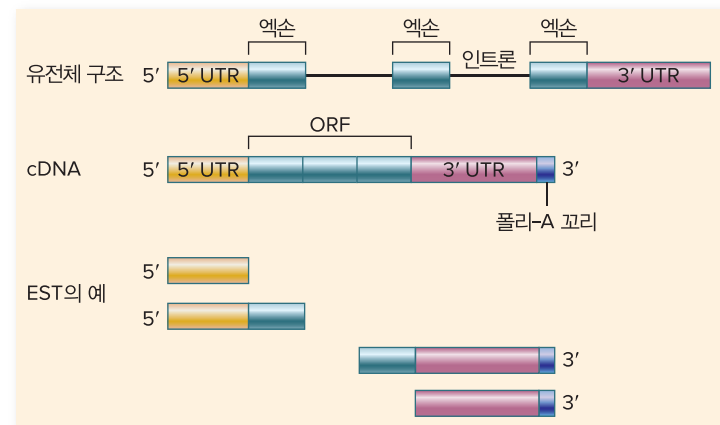


그림 W2.8 발현된 유전자 지도 작성을 할 수 있는 EST.

EST 말단의 염기서열 결정은 cDNA의 구조 예측과 해당하는 유전체 구조에 지도 작성 가능한 서열 데이터를 생성한다. 수많은 EST의 염기서열 결정을 통해 완전한 유전체 구조를 결정할 수 있다.

성한 후 염기서열 결정을 통해 유전체 서열에 지도 제작하는 것이다 (**그림 W2.8**). 이 과정은 가능한 많은 cDNA의 한쪽 또는 양쪽 말단의 염기서열을 결정하여 단순화할 수 있다. 이 역시 자동화할 수 있고, cDNA의 이러한 짧은 부위를 **발현 서열 표지(expressed sequence tags, EST)**라고 부른다. EST는 STS의 다른 형태로 물리 지도에 포함될 수 있다. EST를 이용한 BLAST 검색도 가능하며, 다른 종 EST와의 서열 유사성을 통해 암호화된 단백질 기능에 대한 예측이 가능하다.

EST는 사람의 여러 조직에서 93,000개의 cDNA를 식별하는 데 사용되었는데, 이 cDNA의 약 80%가 이전에는 알려지지 않았었다. 인간 유전체 내 약 20,000개의 유전자는 스플라이싱에 의해 93,000여 개의 cDNA를 만들게 된다(14장 참조).

유전체는 암호화와 비암호화 서열을 포함한다

단백질을 생산하는 DNA 염기서열 부위를 암호화 서열(coding sequence, CDS)이라고 부른다. 폴리펩타이드를 합성하는 이 부위를 제외한 모든 염기서열은 비암호화 서열(noncoding sequence)이 된다. 유전체 주석달기는 유전체에서 암호화서열의 분포를 식별하고 후속 실험들을 통해 이들 유전자 및 생성물의 기능을 밝혀낸다. 비암호화 DNA의 기능이 알려진 경우도 있다. 예를 들어, tRNA, rRNA 및 miRNA와 같은 기능성 RNA 분자를 생성하는 유전자의 기능적 특성이 잘 알려져 있다. 반면 아직 잘 알려지지 않은 비암호화 DNA도 많이 있다.

진핵생물의 비암호화 DNA

진핵생물 유전체의 주목할 만한 특징은 이들이 보유하고 있는 비

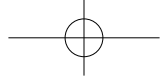


표 W2.1 인간 유전체에서 발견되는 DNA 서열의 종류

종류	설명
단백질 암호화 유전자	염색체에 흩어져 있는 2만 유전자가 번역되는 부분
인트론	각 인간 유전자의 다수를 차지하는 비암호화 DNA
단편 복제	복제된 유전체 부분
위유전자(비활성화 유전자)	유전자의 특성을 가지지만 기능이 없는 유전자 서열
구조 DNA	구성적 이질염색질, 중심립이나 말단소체 근처에 위치
단순 서열 반복	CGG와 같은 짧은 뉴클레오타이드가 수천 번 반복
전이인자	21%: 활성 전이인자인 긴 산재 전이인자(LINEs) 13%: 활성 전이인자인 짧은 산재 전이인자(SINEs) 8%: 각 말단에 긴 말단 반복서열(LTRs)을 가지는 레트로트랜스포존 3%: DNA 전이인자 화석
비암호화 RNA	하나 또는 그 이상의 성숙 mRNA와 상보적인 RNA를 암호화

암호화 DNA의 양이다. 인간 유전체 프로젝트는 특히 각 세포에는 약 2 m의 DNA가 포함되어 있지만, 약 2.5 cm만이 암호화 서열임을 밝혔다. 세포내 99%의 DNA가 비암호화 DNA인 것이다.

인간 유전체에서 단백질을 암호화하는 유전자는 비암호화 DNA 사이에 덩어리로 흩어져 있다. 7가지 유형의 주요 비암호화 DNA가 인간 유전체에 포함되어 있다(표 W2.1은 비암호화 DNA를 포함한 인간 유전체의 구성을 나타낸다).

유전자 내 비암호화 DNA: 15장에서 살펴본 것처럼 인간 유전자는 비암호화 DNA인 인트론과 산재된 여러 조각의 암호화 DNA인 엑손으로 구성된다. 인트론은 인간 유전체의 24%를 구성하고 엑손은 1.5% 이하만을 구성한다.

구조 DNA: 염색체의 일부 부위는 응축되고 단단하게 꼬여 세포주기에서 전사되지 않는 상태로 존재한다. 이를 **구성적 이질염색질** (constitutive heterochromatin)이라 부르는데, 이 부위는 동원체 부근이나 염색체 끝부분인 말단소체(telomere)에 위치하는 경향이 있다.

단순서열 반복. 단순서열 반복(simple sequence repeats, SSRs)은 염색체 여기저기에 흩어져 분포한다. SST은 CA 또는 CGG와 같은 1~6 뉴클레오타이드 서열이 수천 번 반복된 서열이다. SSR은 상동 재조합의 오류로부터 발생할 수 있으며, 인간 유전체의 3%를 구성한다.

부분복제: 염색체 내부에서 또는 비상동 염색체로 복제되어 이동하는 10,000~300,000 bp로 구성된 유전체 영역이다.

위유전자: 돌연변이에 의해 기능을 잃은 것으로 보이는 비활성 유전자를 일컫는다.

전이인자: 인간 유전체의 45%는 한 곳에서 다른 곳으로 이동할 수 있는 DNA 서열로 구성된다. 이들 중 일부는 이동에 관여하는 단백질을 암호화하지만 대부분은 그렇지 않다. 이 인자의 중요성은 이번

절에서 자세히 다룬다.

마이크로-RNA 유전자: 비암호화 DNA에는 유전자 발현을 제어하기 위한 기작이 숨겨져 있다. 한때 “정크(junk)”로 생각되었던 DNA가 21~23 nt의 길이로 전사되지만 번역은 되지 않는 마이크로-RNA(miRNA)를 암호화하는 것으로 밝혀졌다. 하나 이상의 성숙한 mRNA에 특이적인 약 1만 개의 miRNA가 확인되었다. 이들 miRNA는 번역을 조절함으로써 진핵생물의 복잡한 발생과정 중 일부를 조절한다.

긴 비암호화 RNA: 단백질로 번역되지 않지만 조절 역할을 하는 miRNA와 같은 많은 작은 RNA 외에도 수만 개의 더 긴 비암호화 RNA가 유전자 발현을 조절할 수 있다. 최근 발견된 숨겨진 조절 네트워크의 세계는 유전자 발현의 정밀한 제어에 새로운 수준의 복잡함을 보여준다. 긴 비암호화 RNA는 생리 및 발생에서 중요한 역할을 한다. 200개의 긴 비암호화 RNA의 기능이 밝혀졌지만 나머지는 아직 밝혀지지 않았다.

전이인자: 이동성 DNA

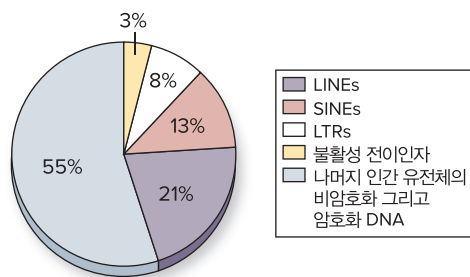
트랜스포존(transposon)이나 이동성 유전인자(mobile genetic elements)로도 불리는 **전이인자**(transposable element)는 유전체의 한 위치에서 다른 위치로 이동할 수 있는 DNA 서열을 지칭하는 용어이다. 전이인자는 다양한 경로로 이동한다. 트랜스포존이 복제되고, 복제된 DNA가 유전체의 새로운 위치로 이동하는데, 이 경우 전이인자의 복제수가 증가하게 된다. 다른 유형의 트랜스포존은 복제 없이 잘려 나와 유전체의 다른 부위에 삽입되는 것이다. 유전체 진화에서 전이인자의 역할은 WEB3장에서 다룬다.

인간 염색체는 네 종류의 전이인자를 가지며, 유전체의 약 21%는 **긴 산재전이인자**(long interspersed element, LINE)로 구성된다. 이

인자는 약 6000 bp 길이로 전이(transposition)에 필요한 모든 효소를 암호화한다. LINE은 전사된 LINE RNA의 cDNA 복제를 할 수 있는 역전사효소를 암호화하여 단백질로 번역되는 것이 아니라 유전체 내부로 재삽입이 가능한 이중가닥 절편을 생성한다. 이 인자들은 RNA 중간물을 이용하기 때문에 레트로트랜스포존(retrotransposon)이라고 부른다.

짧은 산재전이인자(short interspersed element, SINE)는 LINE과 비슷하지만 LINE의 전이 효소를 암호화하지 않기 때문에 LINE의 전이 장치 없이는 이동할 수 없다. 50만 개 이상의 Alu(서열 내부를 자르는 제한효소 이름에서 유래)라는 SINE 인자가 유전체의 LINE 내부에 중첩되어 있다. Alu SINE은 300 bp이고 인간 유전체의 10%를 차지한다. Alu SINE은 LINE의 효소를 이용하여 새로운 염색체 위치로 이동한다. Alu 서열은 암호화하는 유전자로 이동하여 심각한 돌연변이를 유발하기도 한다.

인간 유전체에서는 2가지 다른 종류의 전이인자도 발견된다. 첫째로, **긴 말단반복서열(long terminal repeat, LTR)**로 불리는 레트로트랜스포존은 인간 유전체의 8%를 차지하며, 전이 기작은 LINE과 약간 다르지만 LTR도 이중가닥이 되고 유전체 내로 재삽입될 수 있기 위하여 역전사효소를 사용한다. 둘째로, 유전체의 3%는 돌연변이로 인해 더 이상 이동할 수 없는 불활성 전이인자(dead transposon)로 구성된다.



؟ 탐구문제 DNA 서열이 기능을 하는지 여부를 어떻게 확인할 수 있는가?

엔코드 프로젝트는 인간 유전체의 모든 기능적 인자를 식별하고자 한다

유전체에서 비암호화 DNA의 존재는 수년간 알려져 왔지만 이 DNA의 생물학적 중요성은 밝혀지지 않았다. 기능을 알지 못했기 때문에 이 비암호화 DNA를 “정크 DNA(junk DNA)”라고 불렀지만 비암호화 DNA에 대한 추가 연구를 통해 “정크 DNA”라는 용어에 대해 재명명해야 한다는 의견이 있었다. 따라서 알려진 기능이 없는

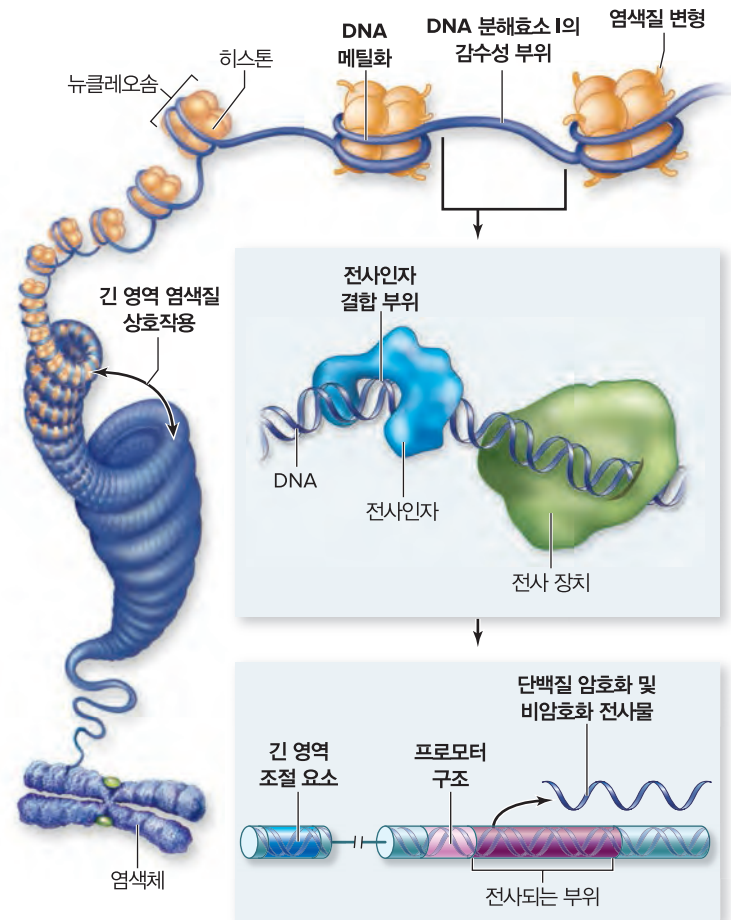


그림 W2.9 엔코드에 의해 정의된 기능적 DNA 요소.

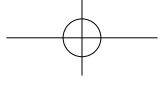
엔코드는 기능의 생화학적 정의를 사용한다. 생화학적 의미에는 DNA 서열의 메틸화 패턴, 염색질 내 히스톤의 변형, 전사 활동 영역으로 간주되는 DNA 분해효소 I에 대한 감수성, 전사인자 결합 부위, 프로모터, 암호화 및 비암호화 전사체의 생성, 인핸서와 같은 장거리 조절 요인, 그리고 장거리 염색질 상호작용 등이 포함된다.

DNA를 “비기능성 DNA(nonfunctional DNA)”라고 부르는 것이 적절할 것이다.

엔코드(Encyclopedia of DNA Elements, ENCODE) 프로젝트는 인간 유전체의 모든 기능적 요소를 식별하기 위한 공동 연구를 말한다. 이 연구를 통해 인간 유전체의 DNA 서열 중 80%가 기능적이라는 결론을 내렸다. 그중 62%는 실제로 전사되어 활성을 가지지만 그들의 기능까지는 알 수 없었다. 과학계에서는 이러한 수치가 실제로 생물학적 기능면에서 무엇을 의미하는지에 대해 많은 논쟁이 있었고, 엔코드 협력단(ENCODE consortium)은 이 추정치를 20~80% 사이로 수정하였다.

엔코드에서 내린 기능의 정의

기능적 요소에 대한 엔코드의 정의는 전사가 되거나, 독특하고 재현 가능한 생화학적 특성이 있는 단백질을 생산하는 DNA 서열을 말



한다(그림 W2.9). 여기에서 생화학적 특성이란 단백질이 DNA 서열에 결합할 때, 염색질이 구부러져 장거리 상호작용을 형성할 때, 또는 DNA 서열이 특정 패턴의 히스톤 변형을 나타내는 경우를 말한다. 서로 다른 세포 유형에서 이들 기능 사이의 유의미한 차이가 있을 수 있기 때문에, 엔코드는 기능적 요소에 대해 144개의 서로 다른 인간 세포주(cell line)의 유전체를 시험하여 80% 값을 도출하였다.

엔코드가 세포주 및 줄기 세포주를 배양된 상태로 사용하는 것에 대한 비판이 있었다. 이들 세포주는 생물의 세포 내 유전체의 상황과는 다를 수 있고, 조직 내 세포보다 전사가 더 활발히 일어날 수 있다. 비평가들은 또한 생물의 발생 단계 및 환경 조건에 따라 기능적 요소들 사이에 상당한 차이가 있을 가능성이 있다고 주장한다.

생물학적 기능

엔코드에서 사용하는 기능적 정의가 반드시 생물학적 기능과 동의어일 필요는 없다. 엔코드에 따르면 생화학적으로 기능을 하고 전사되는 DNA 염기서열이 생물을 위해 필요한 기능을 갖는 RNA를 생성하지 않을 수 있다. 또한 하나 또는 여러 개의 전사된 RNA의 생물학적 효과가 입증되었다고 해서 모든 전사된 RNA가 생물학적으로 효과를 갖는 것은 아니다. 기능적 요소의 정의를 생화학적 정의를 통해 본질적으로 확장시켜 보면 세포분열 동안 유전체의 100%가 복제되므로 유전체의 모든 부분이 기능적이라고도 생각할 수 있다.

기능의 선택된 효과

엔코드의 결론은 몇몇 진화학자들에게도 불씨를 제공했다. 많은 진화생물학자와 일반 생물학자들은 기능 정의에 있어 선택효과를 기술한다. 기능의 선택효과는 기능의 특성이, 정제된 선택에 의해 하나만 선택되어야 한다는 것을 뜻한다. 예를 들어, 포유류 폐의 선택기능은 가스교환이다. 선택압이 그 기능을 형성하게 했다는 것이다. 호흡에서 가슴의 상승 및 하강이 폐의 팽창과 수축을 야기하지만, 폐의 기능으로서 상승 및 하강 운동이 선택되지는 않았다.

유전자 서열이 특정 기능을 갖게 되는 경우, 상응하는 자연선택이 없다면 시간이 지나면서 무작위적인 돌연변이의 축적으로 결국 그 기능이 파괴될 것이다. 기능을 하는 DNA 서열이 파괴되지 않도록 보호하는 유일한 방법은 정제된 선택으로 기능을 파괴하는 돌연변이의 축적을 막는 것이다. 많은 생물학자들은 정제된 선택하에 DNA 서열만이 기능적인 것으로 간주되어야 한다고 주장한다. 이러한 서열은 특정 계통에서 보존될 가능성이 높고 연관된 종에서도 보존될 수 있다. 비교유전체학 연구는 이러한 사실에 기초하여 유전체의 5~15%만이 기능적임을 제안하고 있다.

인간 유전체에서 DNA 서열의 기능에 관해 상충되는 관점이 있으

며, 80%가 기능적이라는 엔코드의 계산은 논란의 여지가 분명히 있다. 하지만 엔코드의 노력에 의해 발전된 데이터 수집 기술과 방법은 유전체 내 DNA 서열의 기능에 대한 이해력을 크게 향상시킨다. 마찬가지로, 비암호화 DNA의 생화학적 기능에 대한 엔코드의 연구결과는 인간 유전체 내 확인된 기능적 요소의 생물학적 연관성을 찾기 위한 효율적인 지도를 제공한다.

핵심 요약 W2.4

유전체가 염기서열 결정 및 조립된 후 적절한 데이터베이스에 저장되면 유전체의 기능적 요소에 대한 유용한 정보가 주석달기로 표시된다. BLAST 프로그램을 사용한 검색을 포함하여 다양한 방식으로 유전적 요소들을 식별할 수 있다. 유전체에서 단백질을 만드는 정보를 암호화하는 부분은 DNA의 일부분이다. 남은 비암호화 DNA 부분도 기능을 한다. “기능”에 대한 정의가 모호하기 때문에, 유전체의 기능적 DNA의 구성 비율에 대한 결론은 논의가 더 필요하다.

- 인간 유전체내에서 수없이 반복되는 전이인자를 자연선택의 원리에 기초하여 어떻게 설명할 수 있을까?

W2.5 비교유전체학과 기능유전체학

학습 목표

1. 유전체의 특성을 이해하기 위한 비교유전체학의 필요성을 설명할 수 있다.
2. 유전체의 기능을 이해하기 위해 기능유전체학이 어떻게 사용되는지 설명할 수 있다.
3. 유전체, 전사체, 단백질체 사이의 관계를 설명할 수 있다.

비교유전체학(comparative genomics)은 한 유전체의 정보를 사용하여 다른 유전체에 대해 학습하는 것이다. 예를 들어, 한 생물의 유전체 내 유전자의 알려진 기능을 사용하여 연관 생물 유전체에서 유사한 유전자의 기능에 대해 추측할 수 있다. 인간의 암 유발에 관여하는 유전자의 60%가 초파리 유전체에 보존되어 있는 것으로 밝혀졌다. 따라서 이러한 유전자의 기능은 초파리를 통해 분석할 수 있고, 이것은 사람을 대상으로 연구하는 것보다 훨씬 수월하다. 비교유전체학은 또한 생물 간의 연관성, 다른 생물이 유사한 생물학적 기능

을 수행하는 방법, 종을 다르게 만드는 것, 심지어 다음 절에서 논의 될 합성생물학자가 관심 있어 할 기능세포를 만드는 데 필요한 최소한의 유전자의 수에 대한 정보를 보여준다.

기능유전체학은 유전자형과 표현형 사이의 연관성을 조사하는 유전체학의 연장선이다. 생물의 표현형은 유전자의 발현 패턴과 환경과의 상호작용에 의해 결정된다. 예를 들어, 건강한 세포와 질병에 걸린 세포의 유전자 발현 패턴을 비교함으로써, 질병 과정에 관여하는 유전자에 대해 알아볼 수 있다.

비교유전체학은 유전체의 보존된 영역을 나타낸다

인간 유전체 서열을 통해 알게 된 놀라운 사실 중 하나는 다른 생물들과 인간의 유전체가 매우 비슷하다는 것이었다. 초파리(*Drosophila*) 유전자의 절반 이상이 인간의 유전체에도 존재한다. 포유류에서는 그 유사도가 훨씬 더 큰데, 인간 유전체 중 생쥐에게 없는 유전자는 300개 정도밖에 되지 않는다. 진화적 궁금증에 대해

비교유전체학의 역할이 클 것으로 기대된다. 이미 많은 원핵생물 유전체의 비교를 통해 이전에 짐작했던 것보다 더 많은 수평적 유전자 전달(lateral gene transfer)이 일어난다는 것이 밝혀졌다.

신터니 기반 비교유전체학

비교유전체학은 신터니(synten)의 장점을 이용하여 거대 유전체 간의 비교를 가능하게 해준다. 신터니는 관련된 유전체에서 보존된 DNA 단편의 배열을 일컫는다. 물리지도 기술은 아직 서열이 알려지지 않은 유전체에서 신터니를 찾는 데 이용될 수 있어, 종 사이의 신터니 단편의 비교로 염기서열 결정이 안 된 유전체에 대해서도 정보를 얻을 수 있다.

이를 설명하기 위해 쌀과 옥수수, 보리 및 밀과 같은 가까운 곡물들을 살펴보자. 이들 중 쌀과 옥수수만이 완전 유전체 서열이 결정되었다. 비록 이런 식물들은 5000만 년 이전에 분기되었지만, 벼, 옥수수, 밀 그리고 다른 화본과 작물들의 염색체들은 폭 넓은 신터니를

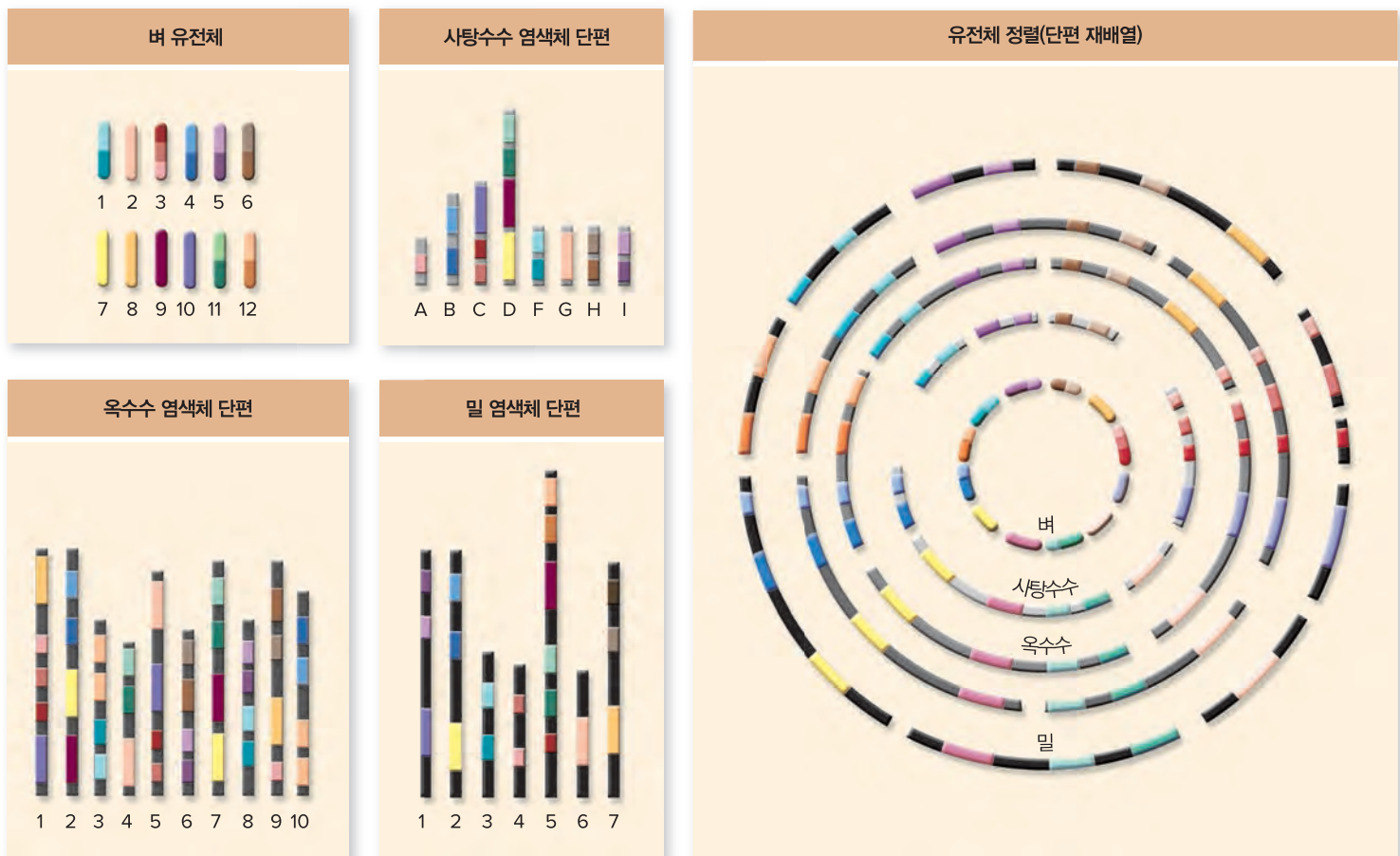


그림 W2.10 곡물 유전체들의 유사 염색체 절편의 재배열.

동일한 색은 서로 다른 종에서 보존되지만 재배열된 DNA 절편을 나타낸다. 주요 화본과 식물종의 각각의 염색체를 절편으로 쪼개고 그 절편의 재배열을 통해 연구자들은 벼, 사탕수수, 옥수수 그리고 밀의 유전체 구성요소가 크게 보존되어 있다는 것을 발견했다. 이것은 식물의 조상 유전체에 있는 절편들의 순서가 목초들이 진화해 오면서 재조합에 의해 재배열되어 왔음을 암시해 준다.

보여준다(그림 W2.10). 물리지도 및 유전자지도의 정보는 다양한 곡물의 염기서열 결정 프로젝트에 큰 도움이 된다.

현재 밀, 쌀, 보리 및 수수의 유전체 서열이 조립되었다. 이들 사이의 비교는 곡물의 핵심 유전자를 나타내는 것으로 보이는 12,607개의 유전자군(gene family)이 있음이 확인되었다. 이 연구결과는 질병 저항성, 작물 수확량 및 영양적 품질 관련된 유전자를 식별하는 데 유

용할 것이다.

이런 모든 유전체의 공통적인 특징 중 하나는 전이인자를 포함하는 반복성 DNA가 많은 양으로 존재한다는 것이다. 전체 유전체 중 60~80%가 전이인자로 보여진다. 물리지도 수준에서 보여지는 재배열들은 이러한 흠어지고 반복되는 인자들의 재조합을 통해 나타난 것으로 보여진다.

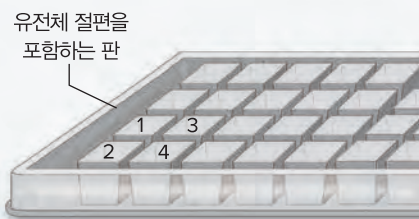
과학적 사고

가설: 꽃과 잎은 일정량의 동일한 유전자를 발현할 것이다.

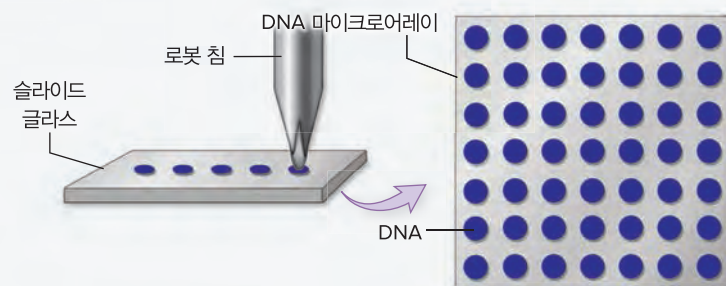
예상: 애기장대의 꽃과 잎으로부터 분리한 mRNA가 애기장대 유전체 마이크로어레이의 탐침으로 사용될 때, 2개의 서로 다른 탐침 세트는 공통의 서열과 각자의 독특한 서열 모두에 혼성화할 것이다.

실험:

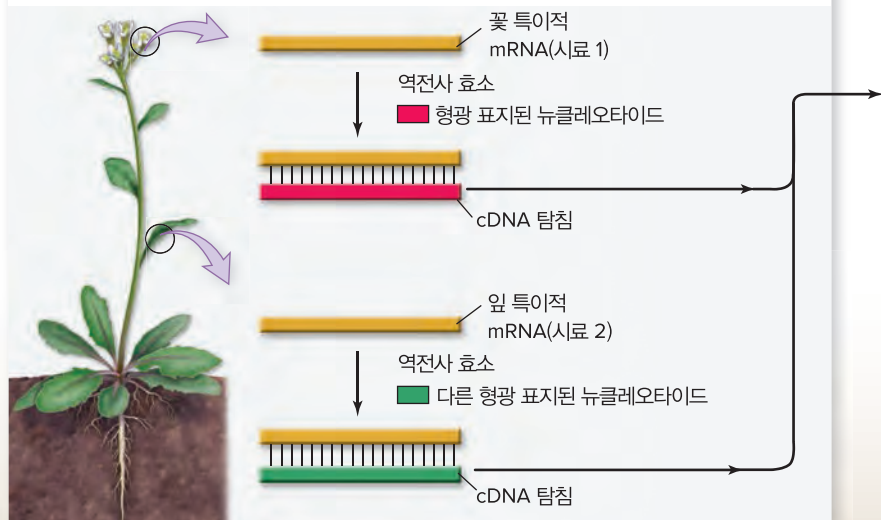
1. 애기장대 유전체로 마이크로어레이를 시작한다. PCR로 증폭된 애기장대 유전체 절편(1, 2, 3, 4.....)이 판(plate)의 각 구멍(well)에 담겨 있다.



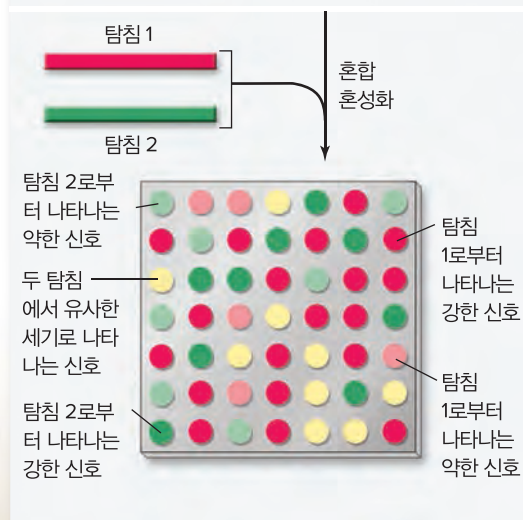
2. DNA는 슬라이드글라스 위로 부착된다.



3. 꽃과 잎으로부터 mRNA를 분리하여 cDNA를 만든 다음 형광으로 표지한다. 2개의 서로 다른 조직에서 mRNA 시료를 얻는다. 각 시료로부터 서로 다른 형광으로 표지된 핵산을 사용하여 탐침을 준비한다.



4. 표지된 cDNA를 가지고 마이크로어레이를 탐색한다. 2가지 탐침이 혼합되어 마이크로어레이와 함께 혼성화된다. 마이크로어레이 상의 형광 신호를 분석한다.

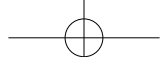


결과: 노란색 점은 꽃과 잎 모두에서 cDNA에 혼합된 염기서열을 나타낸다. 붉은색 점은 꽃에서만 발현되는 유전자를 나타낸다. 녹색 점은 잎에서만 발현되는 유전자를 나타낸다.

결론: 애기장대의 일부 유전자는 꽃과 잎 모두에서 발현되지만, 꽃에서만 발현되거나 잎에서만 발현되는 유전자들도 있다.

추가 실험: 꽃과 잎 모두에서 발현된 유전자가 하우스키핑 유전자인지 아니면 꽃과 잎에서만 발현되는 유전자인지 알아보기 위해 마이크로어레이를 어떻게 활용할 것인가?

그림 W2.11 마이크로어레이.



기능유전체학은 유전체 수준에서의 유전자 기능을 나타낸다

유전체가 어떻게 생물의 구조와 기능에 관여하는지 완전히 이해하려면 유전체의 산물인 RNA 분자 및 단백질에 대해 알아야 한다. 이러한 정보는 세포생물학, 생리학, 발생학 및 진화학을 이해하는 데 필수적이다. 기능유전체학(functional genomics)은 생물의 유전자형과 표현형 사이를 연결시킬 수 있다.

기능유전체학은 다양한 고처리 기술을 이용하여 유전체로부터 언제 생산물을 생산하는지, 어떻게 발생 과정 동안 변화하고 환경에 반응하는지 등에 대한 정보를 제공한다. 기능유전체학은 (1) 전사체(transcriptome): 유전체에 의해 생성되는 모든 RNA 분자에 대한 연구, (2) 단백질체(proteome): 유전체로부터 생성되는 모든 단백질에 대한 연구, (3) 단백질 간의 상호작용 및 상호작용에 의한 산물에 대한 연구 등으로 나눌 수 있다. 먼저 전사체를 연구하는 데 사용하는 몇 가지 기술을 소개한 후 단백질체를 알아본다.

DNA 마이크로어레이

전체 유전체 수준에서 유전자 발현의 분석을 위해 DNA 마이크로어레이(DNA microarray)를 수행한다. 마이크로어레이는 발생 과정 동안, 환경 자극에 반응하는 동안, 그리고 질병이 발생하는 동안 유전자의 발현 패턴이 어떻게 변화하는지에 대한 궁금증을 해결해 준다. 이를 위해 유전체의 정확한 주석달기가 필요하고 암호화 서열을 포함하는 마이크로어레이를 구성해야 한다(그림 W2.11).

마이크로어레이를 준비하기 위해 DNA 절편들은 유리나 실리콘 슬라이드의 특정 위치(an array, 하나의 정렬)로 로봇에 의해 집적된다. 각 DNA 절편은 특정 유전자 서열에 상응하게 되고, 따라서 전체 배열은 모든 암호화 전사물들이 고유한 특정 위치를 가지게 된다. 마이크로어레이의 사용법을 그림 W2.11에 나타내었다. 칩은 서로 다른 부위로부터 얻어진 표지된 mRNA와 함께 혼성화 실험에 사용될 수 있다. 이것은 여러 조건이나 상태에 따라 유전자의 활성화 및 비활성화에 대한 고차원적인 분석을 가능하게 해준다. 단일염기다형성(single-nucleotide polymorphism, SNP)의 전사물을 검출하기 위한 DNA 염기서열을 포함하는 마이크로어레이도 가능하다.

마이크로어레이도 몇 가지 한계점은 있다. 마이크로어레이의 적절한 DNA 배열을 위해 유전체의 서열을 미리 알고 있어야 한다. 또한 시료에서 희귀 다형성 전사물(rare polymorphic transcript)을 찾기 위한 마이크로어레이의 배열을 생성하기 위해서는 이러한 다형성에 대한 사전 지식이 필요하다. 이런 한계점 때문에 최근에는 전사체 연구를 위해 더욱 발전된 다른 방식을 사용한다.

유전자 발현의 순차적 분석

유전자 발현의 순차적 분석(serial analysis of gene expression, SAGE)은 암 세포 같은 시료에서 발현되는 모든 mRNA를 한 번에 조사할 수 있는 고처리 기술이다. 이를 위해서는 먼저 mRNA를 세포로부터 추출하고, 17장에서 기술한 바와 같이 역전사효소를 사용하여 cDNA로 전환시킨다. cDNA는 긴 사슬로 연결되는 작은 조각으로 절단된다. cDNA 조각들은 벡터로 클로닝되고 차세대 염기서열 결정법을 사용하여 결정된다. 잘려진 cDNA의 조각은 mRNA를 충분히 식별할 수 있는 정도의 길이이다. SAGE는 마이크로어레이에 비해 3가지 주요 이점이 있다. 첫째, 유전체에 대한 사전 지식이 필요하지 않다. 둘째, 차세대 염기서열 결정법에 의해 생성된 방대한 양의 데이터로 인해 매우 희귀한 mRNA 분자도 식별 가능하다. 셋째, 각 mRNA에 해당하는 cDNA를 정량하는 데 있어 마이크로어레이보다 매우 정확하다.

RNA 염기서열 결정

RNA 염기서열 결정법(RNA-seq, RNA sequencing)은 SAGE와 기능적으로 매우 유사하다. SAGE와 마찬가지로 RNA-seq은 시료로부터 모든 mRNA의 정보를 제공하지만, cDNA로부터 직접 염기서열 결정을 수행한다. 또한 차세대 염기서열 결정법을 사용하여 염기서열 결정을 한다. RNA-seq은 무엇을 필요로 하는지에 따라 변형할 수 있다. 예를 들어, RNA-seq은 인트론-엑손의 경계를 식별하고, 서열분석이 되지 않은 유전체의 단일염기다형성 지도를 제작하고, 대립유전자가 표현형에 어떻게 영향을 미치는지 등을 연구하는 데 사용될 수 있다. SAGE와 마찬가지로 RNA-seq의 문제점은 표현형에 더 많이 영향을 주는 것은 RNA보다 단백질이라는 것이다. RNA-seq이나 SAGE에 의해 분석된 RNA는 단백질 생성을 막는 전사 후 조절도 있다는 점을 고려하지 않는다. 세포 내 전사체의 분석은 잠재적으로 많은 양의 중요한 정보를 제공해 주지만, 특정 시간의 세포 내 전체적인 단백질의 분석이 중요할 수 있다. 이러한 연구를 단백질체학(proteomics)이라고 하고, 기능유전체학 분야 중 하나이다.

단백질체학은 유전체 내 암호화되는 단백질을 분류한다

많은 단백질들은 번역 후에 작용기나 다른 화학구조가 첨가되기 때문에 단백질체(proteome)에 대한 정보는 유전체 또는 전사체에 대한 정보보다 해석하기가 어렵다. 이러한 첨가는 세포 내 단백질의 활성화 및 위치에 영향을 준다. 또한 많은 단백질은 4차 구조에서만 활성을 갖고 서로 상호작용하여 세포 구조 및 기능을 조절한다. 뿐만 아니라, 단백질체학(proteomics)은 RNA의 대체 스플라이싱

(alternative splicing)으로 인해 단일 유전자가 여러 개의 단백질을 암호화할 수 있다는 점을 고려해야 한다(14장 참조). 유전체에 의해서도 다르게 생성되는 전사체와 단백질체 사이의 상관관계를 연구하기 위해서는 다른 종류의 세포와 조직에 대한 단백질체 분석이 필요하다.

단백질체를 연구하기 위해 많은 고처리 기술들이 개발되었다. 이러한 기술들은 연구 목적에 따라 다르게 사용된다. 단백질체를 분석하기 위해 점점 보편화되고 있는 기술 중 하나는 **질량분석(mass spectroscopy)**이다. 질량분석은 분자의 전하 대 질량비(charge to mass ratio)를 결정하는 방식으로 매우 작은 분자에 대해서는 구분이 어렵다. 오랫동안 단백질 분자의 작은 크기 때문에 활용이 어려웠으나, 정제와 이온화를 결합시킨 새로운 기술의 개발로 복잡한 단백질 혼합물의 질량분석이 가능해졌다(그림 W2.12). 인산화와 같은 번역 후 변형도 분석이 가능하다는 장점이 있다.

단백질체학은 앞 부분에서 설명한 마이크로어레이 기술도 사용한다. 칩에서 DNA를 집적하는 대신 특정 단백질을 인식할 수 있는 항체를 칩에 적용할 수 있다. 시료에서 분리된 단백질 혼합물을 항체 칩에 적용하면 16장에서 언급된 면역측정법(immunoassay)과 동일한 방식으로 항체가 특정 단백질에 결합한다. **생체표지자(biomarker)**로 알려진 특정 단백질을 사용하면 복잡한 시료를 빠르게 식별할 수 있다. 이 기술은 연구에서도 유용할 뿐 아니라 질병 진단 및 스크리닝에도 사용 가능하다.

생물정보학과 단백질체학

유전체 및 전사체 분석과 마찬가지로 고처리 단백질체 분석은 거대한 양의 데이터를 생성한다. 따라서 이런 거대한 데이터를 연구하고 해석하기 위한 방법이 필요하다. 컴퓨터 과학과 생물학을 연결 짓는 학문으로 생물정보학(bioinformatics)이라는 학문이 있다. 생물

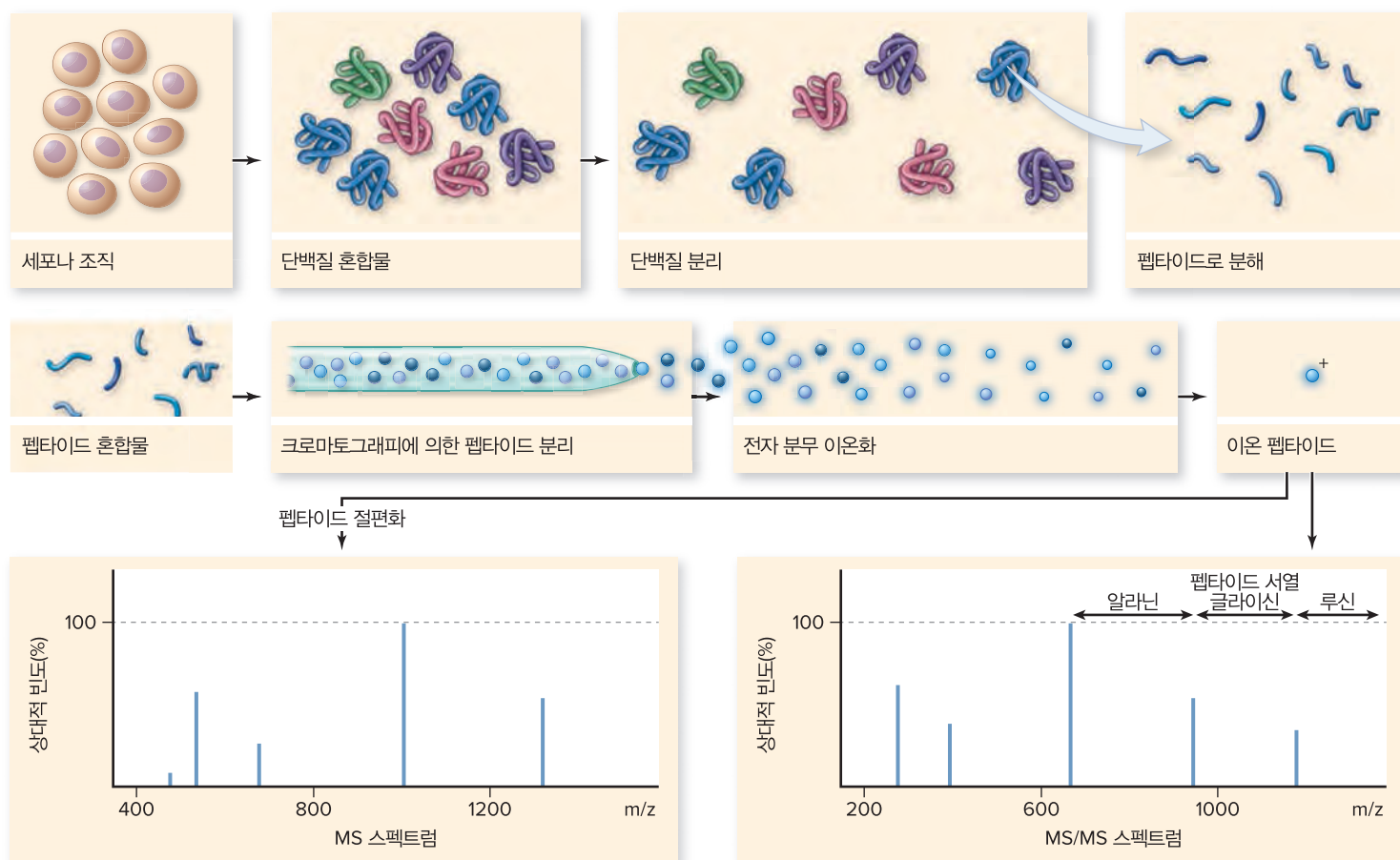
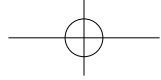


그림 W2.12 질량분석법을 사용한 단백질 분석.

단백질은 세포 또는 조직으로부터 추출되고 겔 또는 생화학적 기술을 사용하여 부분적으로 분리할 수 있다. 부분적으로 정제되거나 개별적인 단백질은 단백질분해효소에 의해 작은 펩타이드로 분해되고 이온화된다. 하전된 펩타이드는 분광계로부터 질량에 의해 분리될 수 있다. 컴퓨터 알고리즘은 단백질을 식별하기 위해 펩타이드 절편과 예상되는 분해 패턴을 일치시킨다. 각각의 펩타이드를 추가적으로 절단한 후 질량분석기로 아미노산 조성을 결정할 수 있다. MS 스펙트럼은 이온의 단위전하당 질량(mass to charge ratio)을, MS/MS 스펙트럼은 이중질량 이온화(tandem ionization)에 의해 이온이 생성될 때 이온의 단위전하당 질량을 사용하고, m/z는 이온의 질량을 이온의 전하로 나눈 값, 일반적으로 단위전하당 질량값을 사용한다.



정보학은 대규모 생물학 데이터를 분석하는 데 컴퓨터 프로그래밍, 수학 및 모델링을 적용하는 것이다. 단백질체학에 생물정보학을 적용시켜 고처리 프로파일링(profiling)을 통한 빠른 단백질 식별, 질량분석을 통한 유전체 주석달기, 단백질 구조 및 기능의 예측 등이 가능해졌다.

? **탐구문제** 단백질체학 실험을 통해 단백질의 아미노산 서열 정보를 얻을 수 있었다. 이를 이용하여 어떻게 유전체 데이터베이스의 유전자 기능 주석달기를 할 수 있는가?

단백질 구조 및 기능 예측

전통적인 단백질생화학(protein biochemistry)과는 달리 단백질체학은 수많은 단백질을 빠르게 식별하고 특성을 분석할 수 있는 새로운 컴퓨터 방법을 사용한다. 유전체학에서와 마찬가지로 규모의 문제인 것이다.

이상적으로 우리는 유전자 서열을 이용해 암호화되는 단백질의 구조와 기능을 알고자 한다. 현재로서는 이미 알려진 단백질 서열과의 비교를 통해서만 새로운 단백질의 구조와 기능을 알 수 있다. 유사한 아미노산 서열은 유사한 구조를 생성하기 때문에 단백질의 구조와 기능을 유추할 수 있는 것이다. 하지만 매우 다른 아미노산 서열로부터 유사한 구조가 생성되기도 한다. 이러한 경우에는 구조 정보를 바탕으로 유사한 기능을 유추할 수 있다.

구조를 바탕으로 기능을 알아내는 것은 매우 어려운 작업이다. 유사한 서열과 구조를 갖는 단백질은 보통 유사한 기능을 가지지만, 그렇지 않은 경우도 있다. 구조가 상당히 다르지만 유사한 기능을 하는 단백질도 많이 있다. 단백질의 진화 이력에 대한 정보가 있으면 도움이 된다. 공통조상으로부터 아직 기능이 분기되지 않은 단백질을 비교할 경우 공유되는 구조와 기능에 대한 예측을 더 정확히 할 수 있다.

최근까지도 아미노산 사슬이 2차 및 3차구조로 접히는 다양한 방법을 예측하는 것이 매우 어려웠다(예측되는 단백질의 구조를 **그림 W2.13**에 나타냄). 컴퓨터 알고리즘은 단백질 구조를 예측하기 위해 단백질 도메인의 특성을 갖는 아미노산 부분을 먼저 찾는다. 그 후 알고리즘은 단백질 영역의 짧은 절편을 구조가 알려진 단백질의 접힘과 비교한다. 컴퓨터 프로그램은 이러한 데이터를 사용하여 모든 절편들이 어떻게 특정 구조로 배열되는지에 대한 모델을 구축한다. 여러 방식의 독립적인 모의실험을 통해 동일한 구조가 나타날 때 정확할 가능성이 높다.

? **탐구문제** 유전체, 전사체, 그리고 단백질체의 관계는 무엇인가?



그림 W2.13 효소의 컴퓨터 모델.

위 그림은 인간 알도스 환원효소(human aldose reductase)를 포함한 알려진 단백질 구조를 포함하고 있다. 2차구조 모티프는 서로 다른 색상으로 나타냈다.

핵심 요약 W2.5

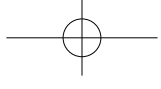
유전학자들은 서로 다른 유전체의 비교를 통해 종 사이의 연관관계뿐만 아니라 유전자와 단백질 사이의 구조적, 기능적 및 진화적 연관관계를 유추한다. 기능유전체학은 유전체에서 유전자의 기능을 이해하기 시작했던 도구이다. DNA 마이크로어레이, SAGE 및 RNA-seq을 사용하여 전사체에 대해 연구할 수 있지만 방식에 따른 장단점이 있다. 단백질체학은 특정 조건에서 유전체에 의해 생성되는 단백질의 분석을 말한다. 단백질 구조와 기능에 대한 예측은 가능하지만 분석이 복잡하고 고성능의 컴퓨터가 필요하다.

■ 한 종의 전사체 확립이 단백질체를 연구하는 데 왜 중요할까?

W2.6 유전체의 응용

학습 목표

1. 유전체 데이터의 사용으로부터 발생하는 윤리적, 사회적 문제를 설명할 수 있다.



유전체학은 생물학에 대한 근본적인 궁금증을 해결해 줄 수 있다. 예를 들어, 한 종을 다른 종과 근본적으로 다르게 만드는 것이 무엇인지, 유전체의 기능에 대한 이해를 통해 표현형을 결정하는 데 있어 유전체의 역할과, 서로 다른 종의 유전체의 유전적 연관관계를 알 수 있게 해준다. 이러한 유전체의 구조와 기능을 알게 되고 적용함으로써 여러 사회적, 윤리적 문제가 생기게 되었다.

유전체 데이터는 첫 번째 “합성” 세포를 만들어냈다

생명은 원자가 분자로, 분자가 거대분자로, 거대분자가 세포 구조로 조립되는 일련의 특성을 가진 것으로 생각된다. 이들 분자를 합성하고, 세포 구조를 형성하는 과정들은 유전체를 통해 제어된다. 이렇게 독자적으로 복제 가능한 세포를 만드는 데 필요한 최소한의 유전체 크기는 어느 정도일까?

살아 있는 생물 중에서 가장 작은 유전체를 찾아 생명을 유지하는 데 필요한 최소한의 유전체에 대한 정보를 얻을 수 있었다. 가장 작은 바이러스의 유전체는 5386 nt이지만, 많은 과학자들은 바이러스는 스스로를 독립적으로 복제할 수 없기 때문에 살아 있는 것으로 간주하지 않는다. 반독립적인 미토콘드리아의 유전체는 약 16,600 nt 길이이지만 미토콘드리아는 숙주세포와는 독립적으로 복제할 수 없다. 현재 알려진 가장 작은 자유생활(free-living)을 하는 세균은 *Mycoplasma genitalium*으로서 약 500,000 nt의 유전체를 가진다. 이는 살아 있는 세포의 최소 유전체가 16,000~500,000 nt 길이인 것을 의미한다.

2010년 제이크레이그벤터연구소(J. Craig Venter Institute)의 과학자들은 컴퓨터를 사용하여 *Mycoplasma mycoides*의 유전체를



그림 W2.14 합성된 *Mycoplasma mycoides* JCVI-syn1.0 세균의 주사전자현미경 사진.

설계하고, DNA를 화학적으로 합성하여, 효모 세포 내부의 염색체에 조립한 다음, 염색체를 세균 *Mycoplasma capricolum*으로 옮겼다고 발표했다. 세균의 원래 유전체는 그 과정에서 파괴되었고, 새로운 합성 유전체는 *Mycoplasma mycoides* JCVI-syn1.0이라는 새로운 세균 세포의 성장을 조절할 수 있게 되었다(**그림 W2.14**). 이러한 방식으로 최소 유전체에 대한 실험적 검증이 가능할 것이다.

유전체학은 질병을 진단하고 치료하는 데 도움을 줄 수 있다

유전체학(genomics)의 혁명으로 수백만 개의 새로운 유전자가 밝혀졌다. 유전체학은 인간의 건강을 향상시키기 위한 매우 커다란 잠재력을 지니고 있다. 단일 유전자에 생긴 돌연변이로 설명할 수 있는 유전질환(hereditary diseases)은 많지 않다. 하지만 유전체학은 유전적으로 영향을 받는 모든 특성들을 알 수 있는 잠재력이 있다.

단백질체학(proteomics)에 의해 신약 개발이 이루어질 수 있지만, 유전체학으로는 질병의 진단에 바로 효과가 있을 수 있다. 기술의 향상과 유전자의 발견이 유전 이상의 진단 가능성을 높여주고 있다. 개인을 판별할 수 있는 진단법이 이용되고 있는데, 그 실례로 유전체학 연구에서 발견된 STR(short tandem repeat, 짧은염기서열반복구간)이 2001년 발생한 미국 911 테러 사건의 시신 확인을 위해 법의학적 진단 도구(forensic diagnostic tool)로 사용된 바 있다.

911 테러로 인해 미국은 생물학적 무기에 대한 우려가 깊어졌다. 미국 우편을 통해 발송된 봉투에 탄저병 포자가 들어 있어 5명이 사망하고 17명이 감염되었다. 유전체 염기서열 결정을 통해 이 사건에 사용된 세균의 출처를 특정 지을 수 있었는데, FBI는 군주 간의 10 bp 차이만으로도 미 육군 감염질환연구센터(U.S. Army Medical Research Institute for Infectious Diseases)의 백신 연구 프로그램에서 사용된 탄저균이 들어 있는 유리병을 추적할 수 있었다. 결국 이 테러에서 공식적으로 기소된 사람은 없었지만, 의심을 받던 연구원은 자살했고, 또 다른 한 명은 무죄임이 밝혀졌다. 이 사건 이후 미국 질병통제예방센터(Centers for Disease Control and Prevention, CDC)는 생물 테러의 대상이 될 수 있는 세균과 바이러스를 분류해 왔다(**표 W2.2**).

인간 유전체의 지식은 복잡한 질병의 유전적 기반을 찾는 데도 도움이 되었다. 전장 유전체 연관분석(genome-wide association studies, GWAS)이라고 하는 연구는 SNP와 같은 유전자 표지(12장 참조), 특성, 질병상태 간의 상관관계를 찾는다. 이러한 연구는 키와 같은 특성과 유전적 요인이 있는 거의 모든 질병에 대해 진행되고 있다. 이러한 방식으로 꽤 많은 유전자가 확인되고 있지만, 대다수의 유전적 변이에 대해서는 아직까지 찾지 못했다.

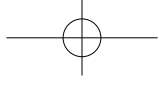


표 W2.2 유전체 연구를 위한 높은 우선 순위의 병원균		
병원균	질병	유전체*
대마마바이러스	천연두	완료
탄저균	탄저병	완료
역병균	역병, 페스트	완료
보툴리누스균	보툴리눔독소증	완료
들토끼병균	들토끼병	완료
필로바이러스	에볼라와 마르버그 출혈성열	둘 다 완료
아레나바이러스	라사열과 아르헨티나 출혈성열	둘 다 완료

* 바이러스와 세균은 많은 계통들이 존재한다. “완료”는 적어도 하나의 계통에서 염기서열 분석이 완료된 것을 의미한다. 그 실례로 탄저병의 플로리다 계통의 염기서열 분석이 완료된 것이 있다.

유전체학은 사회적, 윤리적 문제를 야기한다

유전체학은 우리의 유전자형과 표현형 사이의 관계를 이해하고 질병을 진단하고 치료할 수 있는 엄청난 기회를 제공한다. 하지만 이러한 기회와 함께 유전체 정보 사용에 대한 사회적, 윤리적 책임이 따른다. 유전자 정보 및 그 사용에 대한 소유권, 정보의 잠재적 오용, 모든 유전체 실험 수행 여부에 대한 윤리적, 사회적 결과에 대해서는 신중히 고려해야 한다.

1918 독감바이러스의 재현

2005년 5월, 1918년 스페인 독감 바이러스의 유전체 염기서열 결정 및 재조립에 대한 논문이 발표되었다. 1918년 이 독감 바이러스에 의해 2~5천만 명이 사망했다. 1990년대에는 애틀랜타에 있는 질병통제예방센터(CDC)와 워싱턴 D.C에 있는 군병리학연구소(Armed Forces Institute of Pathology)를 포함한 여러 연구팀이 해당 바이러스의 유전체 절편을 발견하여 염기서열을 결정한 후 전체 바이러스 유전체를 재구성하였다. 이 바이러스는 생쥐와 같은 다른 포유류에 시험하였을 때 본래 바이러스와 비슷한 사망률을 갖는 것으로 밝혀졌다. 물론 인간에게는 시험하지 않았다.

이 실험은 일부 과학자, 정치인, 공익단체 및 일반 대중들에게 심각한 걱정거리를 떠안게 하였다. 하지만 반대로 생각한다면, 바이러스의 유전체 및 바이러스 자체의 특성 연구를 통해 효과적인 백신을 만들 수 있고, 미래의 전염병을 예방할 수 있다. 이 실험의 비평가들은 바이러스의 재생산을 통한 불확실한 잠재적 이점의 가치가 불운하거나 부주의한 과학자가 고의적 또는 우연히 바이러스를 방출하게 될 경우보다 더 높지 않다고 주장한다. 많은 사람들은 또한 생물학적 테러리스트들이 무기화된 바이러스의 형태를 만들기 위해 이러한 정보를 사용하는 것에 대해 우려를 표명한다.

과학계의 광범위한 검토, CDC의 신중한 조사 및 국가자문위원회

(National Science Advisory Board for Biosecurity)의 감찰을 거쳐, 위험보다는 공식적인 연구의 이점이 더 큰 것으로 결정되었다. 물론 이러한 결정은 주관적이기는 하지만, 유전체학에 대한 더욱 큰 이해를 바탕으로 이렇게 논란의 여지가 많은 주제에 대해 더 합리적인 결정을 내릴 수 있다.

유전자 소유권 및 특허

여러분은 자신의 유전자 정보가 본인 소유라고 생각하겠지만, 2013년까지는 다른 사람이 유전자 서열에 대한 권리를 소유할 수 있었다. 1990년대에 미리어드 제네틱스(Myriad Genetics)라는 회사는 유방암의 일부 형태의 발달에 관여하는 것으로 알려진 2개의 유전자에 대해 특허를 냈다. *BRCA1* 및 *BRCA2* 유전자에 대해 특허를 출원함으로써 이 회사는 유전자의 특정 변이체의 존재를 시험하기 위해 염기서열을 사용할 권리를 독점했다. 이것은 유방암의 특정 유형에 대한 소인(predisposition)에 관한 정보를 알기 위한 매우 기초적인 시험이다. *BRCA1*과 *BRCA2*의 특정 변이를 가진 여성의 유방암 위험이 증가하고 암 진단의 심각성을 고려할 때, 많은 사람들은 다른 진단 의견을 받을 수 없다는 것이 시민의 자유를 침해한다고 생각했다.

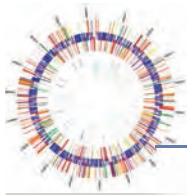
미국 시민자유연맹(American Civil Liberties Union, ACLU)이 주도하는 연방 법원에서의 오랜 분쟁 끝에 미리어드 제네틱스 회신에 대한 사건이 미국 대법원에 도달했다. 대법원은 미리어드 제네틱스가 유전자를 발명한 것이 아니기 때문에, 유전자 서열은 특허출원할 수 없다고 만장일치로 판결했다. 이 판결에서 법원은 염기서열로부터 파생된 것(예, cDNA 서열)은 그것들에 대한 사용으로 만들어진 것이기 때문에 특허를 받을 수 있다고 언급했다.

? **탐구문제** 바이오연료 생산에 사용되는 조류의 지질 수율을 증가시키는 합성 유전자의 특허 출원을 어떤 논거로 뒷받침할 수 있는가?

핵심 요약 W2.6

유전체 염기서열 결정법, 주석달기, 그리고 비교유전체학 및 기능유전체학은 질병에 관여하는 유전자의 발견으로 이어졌다. 이러한 유전자 중 일부는 질병을 더 잘 진단하거나 보다 정확하게 예측하는 데 사용될 수 있다. 유전체학은 또한 질병 발생 및 역학에 대해 조사할 수 있다. 유전체학은 사회적, 윤리적 책임을 동반한다. 규제 기관은 생물윤리학자, 정책 입안자 및 다양한 이해 관계자와 협의하여 실험 및 적절한 데이터 사용에 대해 알게 될 것이다.

- 사람에게 의해 서열이 구축된 유전체로 대체되어 만들어진 세포는 새로운 형태의 생명체가 될까?



핵심 개념 요약

W2.1 유전체지도 작성

■ 유전체학은 전체 유전체를 분석하기 위한 많은 접근법을 사용한다

유전체학은 유전자지도와 물리지도를 사용하여 표지라고 부르는 유전적 랜드마크를 유전체에 위치시킨다.

■ 유전자지도는 유전적 표지 사이의 상대적 거리를 제공한다

연관지도 작성은 유전자 재조합을 활용하여 유전체 내 유전자 표지의 상대적 위치를 결정한다.

■ 물리지도는 유전적 표지 사이의 절대적 거리를 제공한다

제한효소지도, 염색체 지도 및 STS 지도는 모두 표지를 유전체 서열에 위치시킨다(그림 W2.1~W2.3).

■ 물리지도는 유전자지도와 연관될 수 있다

유전자지도와 물리지도는 연관될 수 있다. 클로닝된 유전자는 유전자지도와 물리지도에 위치시킬 수 있다.

W2.2 유전체 염기서열 결정

■ 디데옥시 종결 염기서열 결정은 유전체 염기서열 결정에 여전히 중요하다

디데옥시 종결 염기서열 결정은 형광 뉴클레오타이드와 함께 DNA 합성을 종결시키며 수행된다(그림 W2.4).

■ 차세대 염기서열결정법에서는 처리량을 증가시키기 위해 대규모 병렬 기술을 사용한다

인간 유전체 프로젝트가 시작된 이래로 유전체 서열을 결정하는 데 필요한 비용과 시간이 약 10,000배 감소했다. 차세대 염기서열 분석기는 대량의 결과를 생성하기 위해 분리된 DNA로 한 번에 수천 개의 반응을 수행한다(그림 W2.5).

■ 염기서열이 결정된 유전체 절편들은 완전한 염기서열로 조립된다

샷건 염기서열 결정 및 조립 방식은 약간씩 중첩되는 DNA의 짧은 조각들로부터 유전체 서열을 재생성한다. 전체 유전체를 조립하기 위해 클론-콘티그 염기서열 결정과 조립방법은 더 큰 조각의 DNA와 타일링 방식을 사용하여 더 큰 중첩되는 절편을 구축한다(그림 W2.6).

W2.3 유전체 프로젝트

■ 인간 유전체 프로젝트를 통해 인간 유전체 대부분의 염기서열이 결정되었고 지도로 작성되었다

인간 유전체 서열을 조립하기 위한 경쟁은 빠른 기술의 발전과 완성을 가능하게 했다. 인간 유전체에는 예상보다 훨씬 적은 유전자가 포함되어 있었다(그림 W2.7).

■ 밀 유전체 프로젝트는 유전체 조립의 개선을 보여준다

밀 유전체는 4번에 걸쳐 점점 더 완성에 가까워졌다. 최신 버전에서는 더 길지만 오류가 많고, 짧지만 오류가 적은 두 염기서열 결정법이 결합되었다.

■ 1000 유전체 프로젝트는 유전적 다양성을 다룬다

1000 유전체 프로젝트는 1000명 이상의 개인의 유전체를 분석하고 8000만 개의 변이를 조사하였다. 평균적으로 유전체는 4~500만 개의 부위가 표준서열과 다르다. 표현형 변화 없이 최대 200개의 유전자를 삭제할 수 있다.

W2.4 유전체 주석달기와 데이터베이스

■ 유전체 주석달기는 유전체의 DNA 서열에 기능을 배정한다

단백질을 암호화하는 유전자는 DNA 서열에서 열린 번역틀로 식별된다. 알려지지 않은 유전자와 알려진 유전자를 비교하여 유전자의 기능을 유추할 수 있다.

■ 유전체는 암호화와 비암호화 서열을 포함한다

단백질을 암호화하는 DNA는 단일 복제 유전자, 단편 복제, 다중 유전자 패밀리 및 반복 집단을 포함한다. 진핵생물에서의 비암호화 DNA는 전체 DNA의 약 99%를 구성한다. 인간 유전체의 약 45%는 전이인자로 구성되어 있다(표 W2.1).

■ 엔코드 프로젝트는 인간 유전체의 모든 기능적 인자를 식별하고자 한다

엔코드 협력단은 인간 유전체의 20~80%가 기능적이라고 예측한다. 그들의 접근 방식에 대해 비평가들은 유전체 기능에 대한 진화적 선택을 고려하지 않는 기능의 정의에 대한 의문을 제기한다(그림 W2.9).

W2.5 비교유전체학과 기능유전체학

■ 비교유전체학은 유전체의 보존된 영역을 나타낸다

초파리 유전자의 절반 이상이 인간에 상응한다. 다른 종류의 곡물 유전체는 넓은 영역의 신터니를 가진다.

■ 기능유전체학은 유전체 수준에서의 유전자 기능을 나타낸다

기능유전체학은 유전자의 기능과 산물을 분석하기 위해 고처리 방식을 사용한다. DNA 마이크로어레이는 세포에서 모든 유전자의 발현을 한 번에 관찰할 수 있다(그림 W2.11). SAGE와 RNA-seq는 전사체를 직접적으로 연구한다.

■ 단백질체학은 유전체 내 암호화되는 단백질을 분류한다

단백질체학은 세포에 의해 생성되는 모든 단백질을 연구한다. 단백질체학은 번역 후 변형 및 대체 스플라이싱에 의해 복잡하다(그림 W2.12와 W2.13).

W2.6 유전체학의 응용

■ 유전체 데이터는 첫 번째 “합성” 세포를 만들어냈다

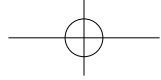
컴퓨터에서 설계되고 화학적으로 합성된 유전체를 효모에서 조립하고 세균에 첨가하여 새로운 세포를 만들어냈다(그림 W2.14).

■ 유전체학은 질병을 진단하고 치료하는 데 도움을 줄 수 있다

유전체학은 유전학에 영향을 받아 진단되는 질병의 숫자를 늘린다. 이것은 더욱 개인화된 약물로 이어질 수 있다.

■ 유전체학은 사회적, 윤리적 문제를 야기한다

사회단체와 생명 윤리학자들은 합성세포를 만들고 위험한 바이러스를 재현하는 것에 염려하고 있다. 개별 유전자로는 특허를 받을 수 없지만 유전자가 만들어내는 cDNA와 같은 합성 생성물의 경우는 가능하다.



연습문제

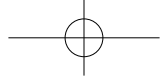
■ 1단계: 이해력 확인하기

- 유전자지도가 제공하는 것은 무엇인가?
 - 유전체 내 DNA 서열
 - 염색체상에서 유전자의 상대적 위치
 - 알려진 DNA 서열 내의 제한효소 절단 부위 위치
 - 염색체 분염 패턴
- STS란 무엇인가?
 - 지도 작성에 사용될 수 있는 DNA 내 고유한 서열
 - 지도 작성에 사용될 수 있는 DNA 내 반복서열
 - 유전자의 3' 영역의 지도 작성을 가능하게 하는 상위 요소
 - 클론-콘티그법에 의해 조립된 유전체 내 절편과 겹치는 DNA 서열
- 인간 유전체에 있는 유전자는 대략 몇 개인가?
 - 2500개
 - 10,000개
 - 20,000개
 - 100,000개
- 열린번역틀(open reading frame, ORF)을 구분하는 것은 무엇인가?
 - 종결코돈
 - 개시코돈
 - 단백질을 암호화하기에 충분한 길이의 DNA 서열
 - a, b, c 모두 맞다.
- BLAST 검색이란 무엇인가?
 - 전체 유전체 서열을 결정하는 동안 공통된 영역을 배열하기 위한 기작
 - 다른 종으로부터 유사한 유전자 서열을 찾는 것
 - DNA 라이브러리를 선별하는 방법
 - ORF를 찾는 방법
- 샷건 염기서열 결정법 및 유전체 조립에서 반복 DNA가 문제가 되는 이유는 무엇인가?
 - 반복 DNA의 염기서열 결정이 쉽지 않기 때문에
 - 염기서열 결정된 절편 사이의 고유성이 부족하기 때문에
 - 반복 DNA가 유전체의 GC 빈도를 높이기 때문에
 - 너무 비싸고 시간이 많이 걸려서
- 단백질체란 무엇인가?
 - 단백질을 암호화하는 모든 유전자
 - 유전체에 의해 암호화된 모든 단백질
 - 세포 내 존재하는 모든 단백질
 - 단백질의 아미노산 서열
- 전사체로부터 단백질체를 완벽하게 예측할 수 없는 이유는 무엇인가?
 - 유전체처럼 서열 분석을 할 수 없기 때문에
 - 전사체는 너무 역동적이어서 예측할 수 없기 때문에

- 모든 유전자가 전사되는 것은 아니기 때문에
- 많은 전사체가 대체 스플라이싱에 의해 다른 단백질로 생성되기 때문에

■ 2단계: 적용력 평가하기

- 유전체학을 통해 치료가 되지 않는 환자들의 공통되는 돌연변이가 존재한다는 것을 발견하게 된다면 어떠한 결론을 내릴 수 있는가?
 - 돌연변이가 치료에 반응하지 못하게 하는 원인이다.
 - 돌연변이를 좋은 진단 도구로 사용한다.
 - 해당 환자의 자손들이 같은 암에 걸릴 위험이 높다.
 - 환자는 모두 같은 유형의 암에 걸렸다.
- 엔코드는 생화학적 방식으로 기능을 정의한다. 다음 중 어떤 유전체학 방식이 인간 유전체의 기능에 대한 엔코드의 정확도를 높일 수 있는가?
 - 기능유전체학
 - 비교유전체학
 - 유전체 염기서열 결정 및 조립
 - a, b, c를 모두 선택하여 정확도를 증가시킴
- 아미노산을 화학적으로 연결하고 몇 가지 추가적인 아미노산을 첨가한 단백질을 합성하여, 단백질을 더 안정적으로 만들고 질병을 치료하는데 도움이 되는 단백질을 합성하였다. 이때 해당 단백질에 대한 특허 출원을 가능하게 하는 무엇인가?
 - 해당 단백질은 인간에게 유익하므로 나의 발명에 비용을 지불해야 한다.
 - 해당 단백질은 합성된 것이고 자연적인 형태와 다르다.
 - 이 단백질은 화학적으로 합성되었다.
 - 이 단백질은 제조비용이 비싸고 특허는 나의 생산비용을 상쇄시키는 데 도움이 될 것이다.
- 개구리(*Xenopus tropicalis*)의 유전체는 4배체이다. 이것이 유전체 염기서열 결정과 조립에 어떤 문제점을 야기할 수 있는가?
 - 유전체 조립을 어렵게 만드는 중복된 유전체 조각이 있을 수 있다.
 - 염기서열 결정을 어렵게 하는 LINE이 있을 수 있다.
 - 유전자지도와 물리지도를 만들 수 없다.
 - b와 c
- DNA 마이크로어레이로부터 얻을 수 있는 정보는 무엇인가?
 - 특정 유전자의 서열
 - 특정 조직 내 유전자의 존재
 - 유전자 발현 패턴
 - 유전체 간의 차이점
- 마이크로어레이 기술과 암에 관련하여 옳은 것은 무엇인가?
 - DNA 마이크로어레이는 암의 종류를 결정할 수 있다.
 - DNA 마이크로어레이는 암 치료에 대한 암의 반응을 측정할 수 있다.



c. DNA 마이크로어레이는 암이 전이할 것인지에 대한 예측을 하는 데 사용될 수 있다.

d. a, b, c 모두 맞다.

7. 비교단백질체학은 다른 조건의 조직이나 다른 세포의 단백질체를 비교할 수 있다. 단백질이 동일한 아미노산 서열을 갖더라도, 일부 암 세포에서 망막 아세포종 단백질은 종종 정상세포보다 약간 더 높은 분자 질량을 갖는 것으로 보인다. 이 차이점은 어떻게 설명할 수 있는가?

a. 번역 후 변형

b. 대체 스플라이싱

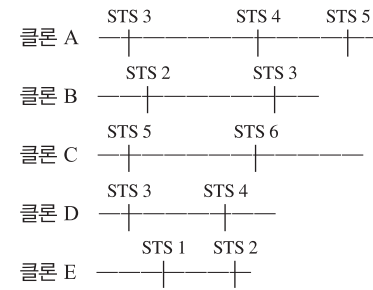
c. 다른 단백질과의 연관

d. 단백질을 암호화하는 유전자의 돌연변이

■ 3단계: 사고력 함양하기

1. 유전체 염기서열 사업의 초기 단계에서 BAC 라이브러리로부터 수많은 클론을 분리하고 STS를 이용하여 이런 클론 내부의 삽입된 DNA의 지

도를 작성하였을 때, STS를 사용하여 클론을 유전체의 연속된 서열(콘티그)로 배열하시오.



2. 유전체 연구는 전염병이 자연적으로 발생한 것인지 혹은 “의도적으로” 발생한 것인지 결정하는 데 사용될 수 있다. 탄저균 같은 질병을 의도적으로 발생시켰다는 의심을 해결하기 위해 유전체 연구자들이 무엇을 찾고 있었는지에 대해 설명하시오.