BIO ECONOMY BRIEF

September 2022. Issue 163

생명과학 분야에서 빅데이터의 영향과 향후 방향

박봉현 책임연구원 **한국바이오협회 바이오경제연구센터**
조인산 대표 **에비드넷**

개요[1,2,3]

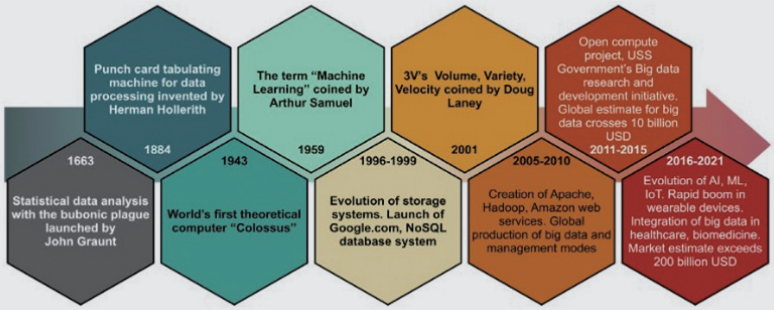
연구에 따르면 인터넷에 연결된 많은 장치가 정보를 추적하고 생성 및 저장함에 따라 매일 2,500조 바이트의 데이터가 생성되고 있다. 그 수는 인터넷 액세스 및 사용이 개선되고 전 세계적으로 확장됨에 따라 계속 증가할 것으로 예상된다. Markets and Markets에서 발간한 "Global Forecast to 2026"에 따르면 빅데이터 시장 규모는 '21년 1,626억 달러에서 '26년 2,734억 달러로 약 11%의 증가율을 보이며 성장할 것이다. 생명공학 분야는 기술혁신으로 인해 다양한 소스에서 정보를 수집하는 것이 점점 더 가능해짐에 따라 의료산업에 침투하고 있는 빅데이터 분야의 최전선에 있으며 데이터는 생명공학에 적용할 수 있는 가장 중요한 영역 중 하나가 되었다.

빅데이터란[4,5]

'빅데이터'는 이름에서 알 수 있듯이 기존의 소프트웨어나 인터넷 기반 플랫폼으로는 관리할 수 없는 대용량 데이터를 의미한다. 빅데이터에 대한 정의는 여러 가지가 있지만 널리 사용되는 정의는 Doug Laney가 정의한 3V의 특징을 가지고 있다. 엄청난 양의 데이터를 포함하는 용량(Volume), 실시간으로 생성되는 속도(Velocity), 다양한 형태로 제공되는 다양성(Variety)이 그것이다.

역사적으로 빅데이터의 진화는 개인·직장생활의 모든 측면을 현저하게 변화시켰다. 현대시대에 빅데이터의 탄생은 1663년 런던 전염병 동안 John Graunt의 데이터 분석에 대한 최초의 보고서 중 일부에 기인한다. 여러 하드웨어 처리 및 저장 장치의 혁신과 발명이 생성된 데이터를 관리하기 위한 길을 열어 주었다. 그 후 컴퓨터 언어의 등장과 함께 데이터베이스 관리시스템은 정보기술과 빅 테크 분야의 세계적인 발전에 기여하였다.

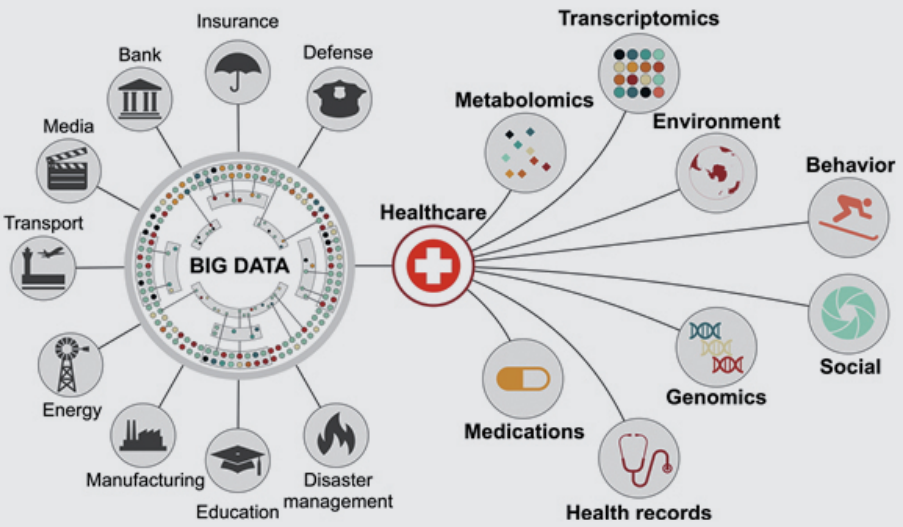
[그림 1] 빅데이터의 진화 타임라인



빅데이터에 의한 생명과학 연구의 발전[5,6,7,8]

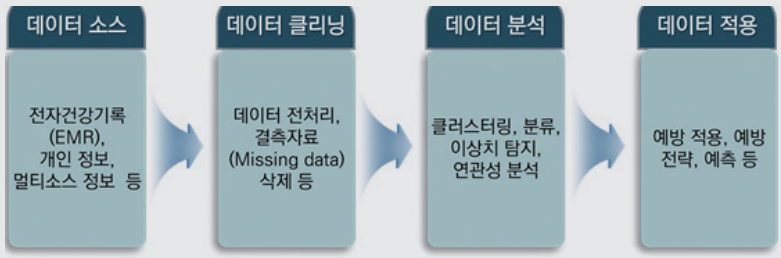
현재 하루에 생성되는 연구 데이터의 양은 이전에 10년 동안 생성된 양과 비슷할 것으로 추정되고 있고 특히 헬스케어 산업은 빅데이터의 도래로 많은 영향을 받았다. 다양한 조직에서 분석 도구, 인공 지능(AI) 및 머신러닝(ML) 기술을 사용하여 데이터 기반 통찰력을 도출하여 의료 비용을 절감하고, 수익흐름을 개선하고 개인화된 의학을 개발하며 사전 예방적으로 환자 진료를 관리하고 있다.

[그림 2] 빅데이터가 기업과 의료산업에 미치는 영향



의료 및 의료연구는 사용 가능한 풍부한 데이터 리소스에서 파생된 데이터 내에 숨겨진 연관성이나 패턴을 찾아 질병을 개선하는데 초점을 맞추고 있다. 기존의 의료데이터 분석은 예측력이 제한적이였으나 데이터 마이닝을 통해 다양한 차원 또는 관점에서 데이터를 분석하고 추출된 정보를 사용하여 예측모델을 구축한다. 이러한 과정은 개별환자의 요구에 맞는 맞춤형 의료를 제공하고 질병의 진단 및 치료를 지원하는 자동화된 분석의 개발로 발전하고 있다.

[그림 3] 의료데이터 분석 과정



빅데이터는 또한 다양한 병원체에 대한 집단 내의 역학을 더 잘 이해하는데 기여하였다. 면역감시센터는 정기적으로 빅데이터를 처리하여 집단 내에서 풍토병이 될 위험이 높은 병원체를 식별하며 마찬가지로 전체 게놈 시퀀싱 및 전체 엑솜 시퀀싱 라이브러리와 같은 게놈 라이브러리의 빅데이터는 데이터 기반 생물의학 발전을 가속화하는데 핵심적인 역할을 수행한다.

유전체학에서 빅데이터의 적용은 치료제 개발을 빠르게 변화시키고 있다. 단일세포 게놈 및 액체생검에서 종양 DNA의 전사체 시퀀싱을 포함한 유전체학의 최근 발전과 메타지노믹스는 이미 의학에 현저한 영향을 미치고 있다. 과거에는 세포나 동물에서 수행된 실험에서 질병과 관련된 경로 및 지식에 따라 약물이 개발되었지만 이 결과는 꽤 자주 인간에게까지 적용이 되지 않았다. 유전적 변이가 있기 때문에 종의 차이와 생물학적 및 기술적 변수를 고려하여 추론하기 위해 빅데이터가 필요하며 유전체 빅데이터를 기반으로 환자에게 개별화되거나 개별화될 수 있는 치료방법의 개발을 도와주고 있다.

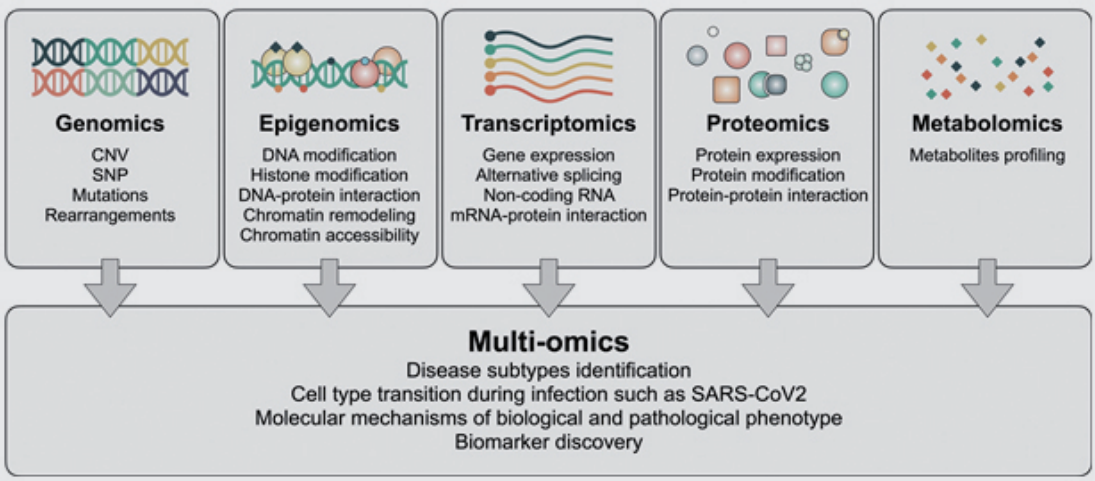
인간게놈 서열은 유전적인 지시가 인간의 생물학적 기능으로 어떻게 이어지는지 이해하기 위한 기초이다. 환자집단의 유전적 프로파일을 평가하고 질병의 위험 및 치료적 접근 및 반응과 관련된 패턴을 식별하기 위해 국제 이니셔티브가 착수되었다. The Cancer Genome Atlas(TCGA)는 여러 질병 부위의 표본을 배열하고 암과 관련된 유전적 변형에 대한 심층적

인 이해를 얻기 위해 미국 국립 보건원과 국립 인간 게놈 연구소에서 시작한 프로젝트이다. 국제 HapMap 프로젝트는 인간 게놈 변이의 패턴을 식별하고 약물 또는 환경요인에 대한 반응을 포함하여 건강 및 질병에 대한 영향을 결정하기 위한 프로젝트로 미국, 유럽, 중국 및 일본이 참여하였다.

빅데이터에 활용된 기술[5]

오믹스 기술은 정교한 생체정보 분석을 필요로 하는 빅데이터를 생성하며 많은 세포와 조직에 걸쳐 많은 세포 구성요소를 상세하게 연구할 수 있게 했다. 고품질의 오믹스 데이터 가용성은 건강 및 질병의 생물학적 조건의 여러 계층을 이해하는데 크게 기여하였으며 멀티오믹스를 통해 여러분야에 걸쳐 데이터를 통합함으로써 생명과학 연구가 가속화되었다. 다중 오믹스 접근방식의 데이터 통합은 생물학적 조건의 여러 데이터 레이어(게놈, 전사체, 후성 유전체 등)의 결과를 결합하여 다각적 판독을 제공하고 실제로 여러 그룹에서 다중 오믹스 데이터를 성공적으로 사용하여 생물학적 표현형을 프로파일링하고 다양한 질병 메커니즘에 대한 이해를 얻었다.

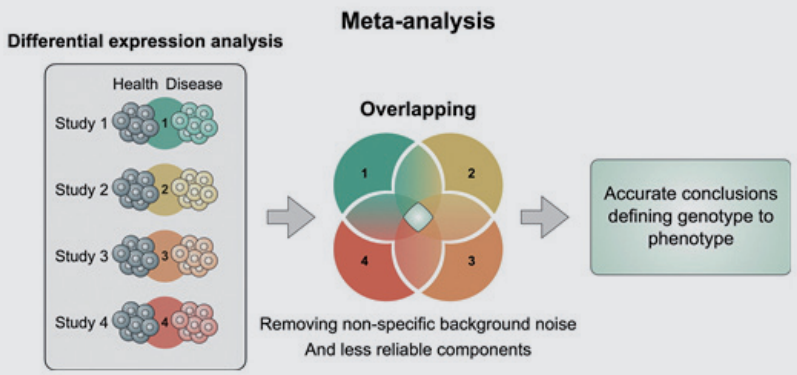
[그림 4] 멀티오믹스의 데이터 소스 및 응용



메타분석은 비특이적이거나 또는 신뢰성이 낮은 구성요소를 제거하고 동일한 데이터 유형의 데이터 세트에서 공유되는 가장 빈번한 신호를 증폭시키는 것을 목표로 한다. 이는 데이터 세트의 품질이 이미 낮은 경우 다중 오믹스 접근방식을 사용하면 잘못된 결과를 발생할 수 있는

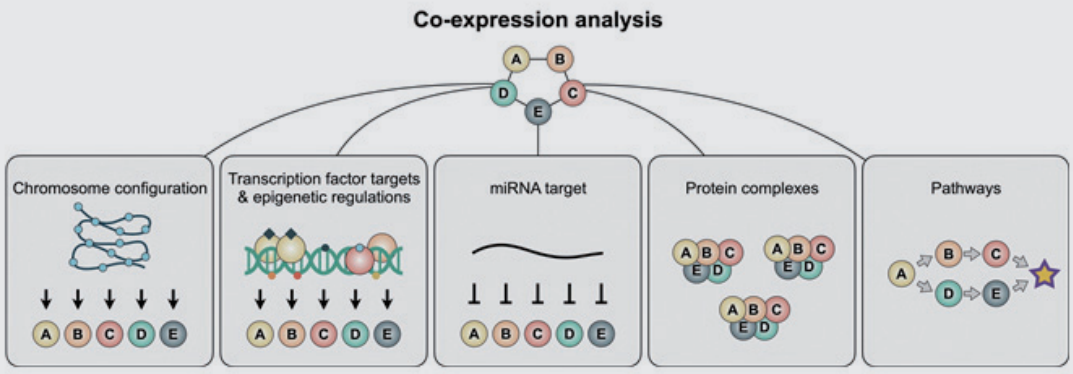
위험을 감소시킬 수 있다. 예를 들어, 질병과 관련된 데이터 세트가 생성되면 각 데이터 유형에 대한 메타분석을 결합한 다음 통합하는 것이 다중 오믹스 분석을 수행하는 가장 정확한 방법일 수 있다.

[그림 5] 메타분석 개념



세포 기능에는 메커니즘이 작동하는 데 필요한 RNA 및 단백질 구성 요소를 인코딩하는 유전자 네트워크가 필요하지만 어떤 유전자가 어떤 기능과 관련되어 있는지 결정하는 것은 어려울 수 있다. 공동발현 분석은 유사한 특성을 공유하는 유전자 클러스터를 식별하는 데 사용할 수 있으며 잠재적으로 세포기능을 제어하는 동일한 네트워크에 관련되어 있음을 나타낸다. 공동발현 분석 접근방식을 통해 전사 인자와 같은 조절 요소 및 메틸화와 같은 후성 유전학적 변형을 식별할 수 있다.

[그림 6] 공동발현 분석의 생물학적 해석 및 결과



빅데이터 기술 활용 및 연구 현황[9,10,11]

에비드넷은 의료 빅데이터를 바탕으로 질병현황, 약물 처방 빈도, 수술 빈도, 검사 빈도 등의 메타데이터 분석 및 패턴정보를 제공하는 EVIX-INSIGHT™ 플랫폼 개발을 완료하였다.

신테카바이오는 전 세계 다양한 인종, 수천 명의 유전체시퀀싱 데이터를 마하 슈퍼컴퓨팅 기술로 분석을 수행하고 있으며 자체 기술인 Adiscan 엔진으로 3가지의 DB인 대립유전자깊이 정보, 유전형질정보, 반수체정보를 생성한다. 유전체 빅데이터를 생성하고 운영하는 시스템은 암 약물선별 및 희귀질환 진단과 같은 병원 정밀의료를 위해 활용되어 질병 연관성 검증에 역할을 하고 있다.

테라젠바이오는 첨단 유전체 분석 기술을 바탕으로 맞춤형 진단 및 솔루션과 차세대 염기서열 분석(NGS) 임상 검사, 의료 빅데이터 등의 서비스를 제공하고 있으며 유전자 분석 기반의 암 위험도 예측, 약물 기전 파악, 맞춤형 항암제 선별 등이 가능한 알고리즘을 개발해 특허를 취득하였다.

쓰리빌리언은 인공지능 유전변이 해석 시스템을 활용하여 10만 개의 유전변이에서 병원성 변이를 판별하고 동시에 환자가 가지고 있는 증상이 알려진 7,000여 개의 유전질병과 상관성이 있는지 검정하는 과정을 통해 최종 진단하는 서비스를 제공하고 있다. 이외에도 국내 여러 기업들이 빅데이터 기술을 활용한 플랫폼 개발이나 진단·분석 서비스 등을 제공하고 있다.

향후 방향[5,12]

빅데이터라는 용어는 최근 몇 년 동안 전 세계적으로 매우 인기가 있고 산업계든 학계든 거의 모든 연구 분야에서 빅데이터를 생성하고 분석하고 있다. 이러한 현상은 매우 두드러져 '데이터 과학'이라는 새로운 과학분야의 탄생으로 이어졌으며 데이터 관리 및 분석을 포함한 다양한 측면을 다루고 있다.

빅데이터의 광범위한 사용을 촉진하기 위해 극복해야 하는 주요 장애물은 대규모 데이터 세트의 복잡한 특성을 작업하고 서로 다른 데이터 유형을 통합하는 가장 효율적인 방법을 식별하는 것이다. 멀티오믹스, 메타분석, 공동발현 분석 모두 대규모 복잡한 데이터에서 정보를 추출하는 것을 기반으로 한다. 또한 데이터 간의 이질성, 데이터 이해 관계자 간의 갈등, 데이터 소유권, 데이터 개인정보 보호 및 무결성 등 생물의학 연구의 빅데이터 분야에서 많은 시급한 과제가 존재하고 있다.

생물의학 연구에서 빅데이터를 개선하기 위한 몇 가지 실행 가능한 접근으로는 1) 데이터 형평성을 촉진하고 정크데이터의 중복성과 과부하를 최소화하기 위해 기업, 산업 및 데이터 생성 팀 간의 협업 촉진 2) 컴퓨터 엔지니어, 데이터과학자, 생물의학 과학자 및 임상의사를 초빙하여 학제 간 팀, 부서 및 센터 구축 3) 약물 파이프라인에 적용할 수 있도록 규제기관의 AI-ML 모듈 및 방법 승인 4) TCGA와 같은 범용적 고가치 데이터 세트 축적에 대한 자금 지원 5) 높은 처리량 컴퓨팅 및 머신러닝 접근에서 얻은 결과의 투명성 및 재현성 발전 등의 조치가 포함될 수 있다.

〈 참고자료 〉

1. Big Data Is Changing The Way People Live Their Lives, Forbes, 2018.05.16.
2. Big Data Market worth $273.4 billion by 2026, Markets and markets, 2022.07.02.
3. How Big Data Is Used in Biotech, The George Washington University, 2021.11.15.
4. What makes Big Data, Big Data?Exploring the ontological characteristics of 26 datasets,Big data & society, 2016
5. Big data: Historic advances and emerging trends in biomedical research, current Research in Biotechnology, 2022.02.18.
6. Big data analytics for preventive medicine, Neural Computing and Applications, 2019.03.16
7. https://www.cancer.gov/
8. The International HapMap Project, Nature, 2003
9. 헬스케어 데이터 공공 플랫폼의 활성화를 위한 통합적 전략 연구, 과학기술정책연구원, 2021
10. 신테카바이오 홈페이지
11. UKBB '20만 전장 게놈' 데이터 공개...유전체 분석 우리는?, 히트뉴스, 2021.11.29
12. Big data in healthcare: management, analysis and future prospects, Journal of Big Data volume, 2019.06.19

Writer	Reviewer
박봉현 한국바이오협회 바이오경제연구센터, 책임연구원	**조인산** 에비드넷, 대표

BIO ECONOMY BRIEF

발행 : **2022년 9월** | 발행인 : **오기환** | 발행처 : **한국바이오협회 한국바이오경제연구센터**

13488 경기도 성남시 분당구 대왕판교로 700 (삼평동, 코리아바이오파크) C동 1층, www.koreabio.org

* 관련 문의 : 한국바이오협회 한국바이오경제연구센터 e-mail : kberc@koreabio.org

한국바이오경제연구센터
KOREA BIO-ECONOMY RESEARCH CENTER
Innovating Data Into Strategy & Business

65

9 772508 681005

ISSN 2508-6812