KISTI
ISSUE BRIEF

ISSN 2635-5728

과학기술인프라, **데이터로 세상을 바꾸다**

https://doi.org/10.22810/2023KIB058

2023. 06. 28. 제58호

국가 바이오 빅데이터 인프라의 미래 : 바이오 빅데이터 인프라 구축 동향 및 발전방향

이용호 · 이준학 · 강효진

국가 바이오 빅데이터 인프라는 국가 차원에서 바이오 데이터를 수집, 관리하고 이를 활용하기 위한 시스템과 기술로 구성된 총합으로서, 디지털 바이오 패러다임 하에서 질병 진단, 신약 개발 및 치료법 개발 등 연구혁신 및 바이오산업 성장에 있어서 필수적 요소라 할 수 있다. 국내에서는 생명연구자원 빅데이터 구축전략, 디지털바이오 혁신전략 등 보건의료 빅데이터 구축을 통한 정밀의료 실현을 목표로 데이터를 축적, 활용하는 다양한 정책을 추진하고 있다. 특히 다부처 국가생명연구자원 선진화 사업을 통해 국가바이오데이터스테이션을 중심으로 데이터를 통합 관리하고, 분석·활용환경을 구축하여 서비스를 제공하는 사업이 추진 중이다. 해외 주요 선도국가에서는 국가 주도의 바이오 비전 수립 및 후속 실행전략을 수립하여 국가 바이오 데이터 생산 및 서비스가 구축, 운영되고 있으며 정밀의료 이니셔티브 등의 국가 정책사업으로 맞춤형 진단 및 의료로 분석결과를 연계하고자 노력하고 있다. 선도국 사례를 볼 때 국내에서도 국가 기술경쟁력을 확보하기 위한 선도적 노력이 필요하며 이를 위한 바이오 빅데이터 인프라의 발전 방향 수립 및 추진이 필요하다. 첫째, 데이터 생산 부문에서는 대규모 바이오 데이터의 구축 및 국가 전략적 자산으로서의 관리 및 활용체계 수립이 필요하다. 세부 방안으로 민감정보 활용과 개인정보 보호 문제를 해결하기 위한 법적 기반과 기술적 보호 체계 조성이 요구된다. 둘째, 데이터 저장 및 유통 부문에서는 현재 추진하고 있는 국가 데이터 수집 및 공유체계를 더욱 강화하는 한편 수집 데이터 간 연결성을 강화하여 데이터 제공과 유통을 활성화하는 노력이 필요하다. 셋째, 연구 및 산업 주체별 활용 부문에서는 대규모 데이터 활용을 지원하기 위한 하드웨어 및 정보지원 인프라 구축이 필요하다. 구체적으로 대규모 데이터 분석을 위한 컴퓨팅 계산자원을 구축하여 제공하고, 디지털 바이오 패러다임에 대응하는 AI 모델 및 연구 협업 네트워크 제공과 활용 지원을 통해 바이오 데이터 기반 연구를 활성화할 필요가 있다.

CONTENTS

1. 국가 바이오 빅데이터 인프라 구축 배경

바이오 빅데이터 인프라 개요

- **(기본 개념) 바이오 빅데이터는 바이오 연구 수행을 통해 생산, 활용되는 모든 빅데이터를 의미하며 주로 다양한 생물체의 실물에서 도출되는 전체 정보를 뜻함**

  ※ 바이오 연구 소재(인체유래물, 동·식물, 미생물, 병원체 등 연구에 활용되는 실물소재)에서 산출되는 유전체, 전사체, 구조 데이터 등과 의료활동 및 시험을 통해 산출되는 임상·전임상 정보, 생활(라이프로그)정보를 포함하는 개념

  ※ 생명연구자원법상 바이오 연구 소재와 이로부터 도출된 데이터 및 관련 정보를 생명연구자원으로 규정하여 관리하고 있음

  - 바이오 빅데이터 인프라는 국가 차원에서 수집, 보유, 관리되는 바이오 빅데이터를 활용하는데 필요한 기술과 시스템으로 구성되는 인프라를 의미

  - 특히 다양한 분야의 바이오 빅데이터를 종합적으로 수집하고 분석하는데 필요한 시스템 및 공유·개방을 위한 데이터 서비스 등 제반 활용 기반을 포함함

- **(바이오 빅데이터의 특징) 기존 빅데이터 대비 바이오 빅데이터는 다음과 같은 특징을 지니고 있어 데이터 처리 및 분석을 위한 별도의 인프라 구축이 요구됨**

  - **(데이터 규모와 복잡도)** 타 분야 대비 크기 측면에서 대규모의 빅데이터를 다루며 다양한 데이터 형식과 복잡도를 가지고 있음

    ※ 인간 게놈 데이터는 30억개의 염기쌍, 약 2만 개의 유전자로 구성되어, 단위 샘플당 전장유전체 약 120GB, 전사체 10GB, 메타지놈 20GB가 생산[1]. 유전체, 단백질, 대사체 등 데이터 차원도 매우 높아 수백만~수십억개의 관측치 및 수천~수백만개의 특성(feature)을 가질 수 있음

    ※ 연구 목적과 방법에 따라 다르나 유전체 분석의 경우, 100명 이상의 샘플 데이터로 분석하는 것이 신뢰성을 보장하는 최소 규모로 알려져 있음. 일례로 2020년 Nature에 발표된 “Pan-Cancer Analysis of Whole Genomes” 연구에서는 2,658명의 참여자 데이터로 2.4PB 규모의 데이터 분석이 수행된 바 있음

  - **(데이터 유형의 다양성)** 바이오 빅데이터는 분석 대상에 따라 다양한 형태의 정보가 수집될 수 있음. 개인을 기준으로 볼 때 임상정보, 유전체·오믹스(단백체, 대사체, 전사체) 데이터, 생체 신호, 의료 영상 등 다양한 데이터가 생산되며 이러한 이종 데이터를 통합하여 분석해야 함. 따라서 데이터 분석에 앞서 데이터 표준화 및 통합 이슈를 선결해야 하는 경우가 많음

  - **(고도의 정보보호 수준)** 바이오 데이터는 개인정보를 포함하고 있기 때문에 정보 보호를 위한 데이터 보안이 필요하며 이를 해결하기 위해서 안전한 데이터 저장 및 처리 시스템과 엄격한 보안 프로토콜이 필요

1) 유전체 원본 시퀀싱 파일(FASTQ)과 맵핑(BAM) 파일을 합한 저장 용량

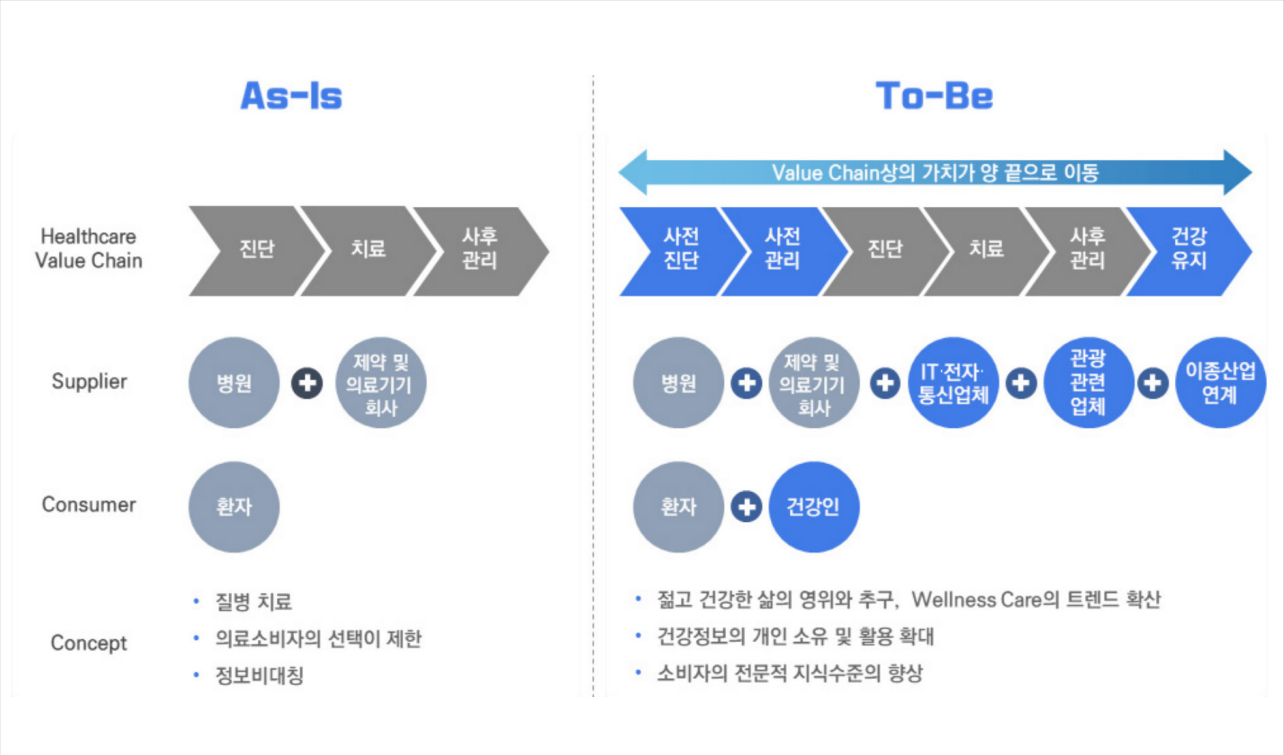
- (데이터 신뢰성) 의료분야에 활용되는 데이터이므로 데이터의 정확성과 신뢰성이 매우 중요함. 이로 인해 데이터 활용에 대한 규제와 윤리적인 문제를 함께 고려해야 함

바이오 빅데이터 인프라 구축의 필요성

- **(보건의료 패러다임의 전환) 고령화에 따른 의료비 증가 등 사회적 부담 경감을 위해 진단·치료 중심 의료에서 개인의 유전체 정보 기반 4P 의료로 패러다임 전환 추세**

※ 미래의 의료 패러다임은 4P 의학, 즉 개인의 유전자 정보나 생활 습관 등에 대한 빅데이터를 활용하여 질병이 발생하기 전에 예측하여 대응하는 예측(predictive) 의학, 유전자 조작이나 각종 기능의 보강을 통하여 원하지 않는 질병의 발생을 막는 예방(preventive)의학, 환자가 자신의 질병 치료에 주도적인 역할을 하는 참여(participatory) 의학 및 개별 환자에 특화된 맞춤(personalized) 의료 패러다임 중심으로 전환되는 추세

<그림 1> 의료 트렌드의 변화



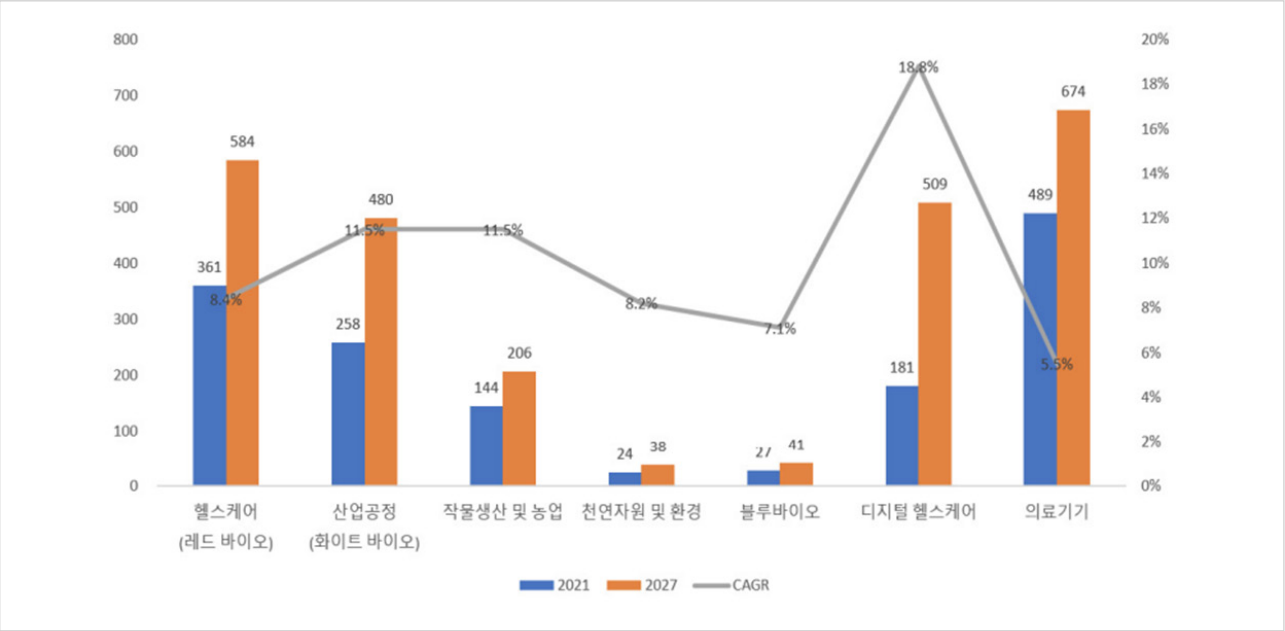
출처) 다부처 (2022)

- **(신성장 동력 육성) 국가 신성장 동력의 하나인 바이오 분야를 위한 정책적 육성 시책이 필요**

- 최근 부상하는 디지털 헬스케어 시장을 포함하여 전 산업이 성장하는데 바이오 빅데이터의 구축과 활용이 핵심 역할을 담당

※ 글로벌 헬스케어 시장 규모는 2020년 1.95조 달러~2.12조 달러로 연평균 3% 이상의 견조한 성장세를 유지

<그림 2> 글로벌 헬스케어 시장 전망(2021-2027)



출처) KIAT (2023)

- **(인프라 필요성) 개인의 환경적·유전적·생물학적 특성 등을 고려하여 질병을 세분화하고 이에 따른 맞춤형 질병 예측, 예방, 진단, 치료를 시행하기 위해서는 바이오 빅데이터 수집, 축적 및 분석을 위한 인프라 구축이 필수적임**

  - **(규모의 경제)** 빅데이터의 스노우볼 효과를 고려할 때, 개별 주체가 주도하는 산발적 구축은 선도국가 중심으로 견조하게 유지되고 있는 데이터 이니셔티브를 넘어서기 어려움

    ※ 바이오 빅데이터는 구축 시점부터 장기간 데이터의 추적이 시작되고 이들이 다시 축적되므로 시간이 흐를수록 데이터의 규모와 다양성이 크게 확대되는 규모의 경제효과 발생

  - **(공공재 특성)** 바이오 빅데이터 구축은 장기간에 걸쳐 대규모 투자가 필요한 영역으로 이윤극대화를 추구하는 민간에 맡겨두면 지속적 추진에 한계

2. 국내외 정책 및 투자 전략 동향

국내 정책 추진 현황

- **(국정과제 추진)** 정부는 '바이오·디지털헬스 글로벌 중심국가 도약'을 목표로 '보건의료 빅데이터 구축으로 정밀의료 실현'이라는 실천과제(25-4) 추진
  - '100만 명 규모의 임상·유전체 정보 기반 보건의료 빅데이터 구축'을 명시
- **(기본계획 및 전략수립)** 바이오 데이터 축적·활용의 중요성을 인식하여 바이오 헬스 산업 혁신전략('19), 제3차 국가생명연구자원 관리·활용 기본계획('20~'25), 생명연구자원 빅데이터 구축전략('20) 및 디지털 바이오혁신전략('22)을 통해 국가 바이오 빅데이터를 축적하여 활용토록 하는 정책을 추진
- **(인프라 구축)** 「생명연구자원 빅데이터 구축전략」('20)을 통해 국가 바이오 연구개발 데이터를 한 곳에서 연계·활용토록 하는 통합 체계*를 마련하였으며, 디지털바이오혁신전략('22)을 통해 국가바이오데이터스테이션 플랫폼 등을 5대 핵심 인프라로 육성

  * 국가바이오데이터스테이션(K-BDS, Korea-Bio Data Station): 부처·사업·연구자별로 흩어져 있는 바이오 연구데이터를 통합·수집·제공하는 플랫폼으로 정부에서 지원하는 바이오 분야 연구개발사업을 통해 생산되는 연구데이터(생화학분석, 이미지(영상), 임상 및 전임상, 오믹스, 분자구조, 표현형 정보)를 통합 수집해 고품질의 데이터가 연구현장에서 활용될 수 있는 기반을 마련

국내 바이오 빅데이터 인프라 구축 동향

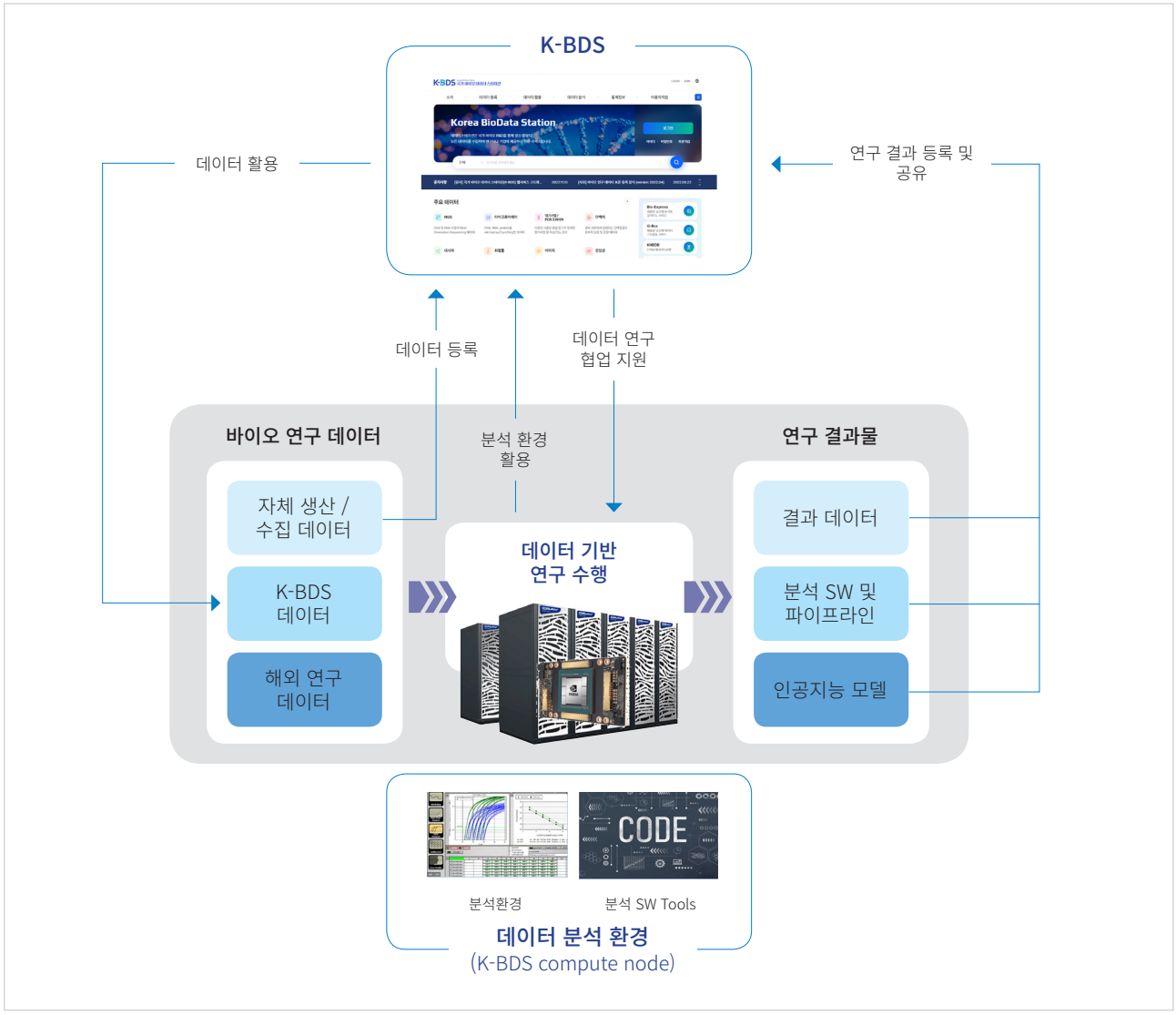
- **(거버넌스)** 국가 차원의 생명연구자원 생산·수집·활용 등 제반 체계는 제3차 국가생명연구자원 관리·활용 기본계획('20~'25)수립을 기점으로 부처별 개별 사업을 다부처사업인 국가생명연구자원 선진화 사업으로 통합하여 추진
- **(수집 및 제공)** 국립중앙인체자원은행(인체자원), 보건의료빅데이터 개방시스템(의료분야 공공데이터), 과기부(KOBIC)/보건복지부(CODA)/해양수산부(MAGIC)/농촌진흥청(NABIC)/국립생물자원관 등 부처별 정보센터에서 분야별 데이터 수집 및 제공
  - '21년 기준 부처별 정보센터 및 주요 데이터 보유 기관에서 오믹스(omics), 화합물, 천연물 등 약 9.12PB에 달하는 데이터를 확보하여 관리 중
  - 국가바이오데이터스테이션(K-BDS)에서 11개 부처·청 및 각 데이터센터와 연계하여 범부처 바이오연구데이터 통합·수집·관리·제공 서비스 추진('22년~)

    ※ '22년 기준 10PB 스토리지 구축 및 53종의 데이터베이스 표준등록 체계 완료

- **(분석·활용환경) K-BDS를 중심으로 바이오 연구데이터 분석을 위한 계산 자원 및 통합적 활용환경을 구축하여 제공**

  - 한국과학기술정보연구원은 K-BDS 분석·활용환경 구축의 총괄 기관으로서 '22년부터 2,432 CPU 코어, 2PB 스토리지 규모 인프라, 바이오 데이터 분석도구 및 DB를 탑재한 분석 환경을 구축하여 서비스

<그림 3> K-BDS 데이터 인프라(분석환경) 데이터 흐름도



출처) 자체 작성

해외 정책 추진 현황

- **(미국) 2015년부터 국민건강 및 질병 치료 개선, 의료비 절감 등을 위한 '정밀의료 이니셔티브(Precision medicine initiative, PMI)' 추진**

  - PMI 이니셔티브 발표로 정밀의료와 정밀의료 연구자원 구축 지원 본격화

- 4대 주요 프로젝트 중 대규모 코호트 구축, 개인맞춤형 암 치료법·예방법 개발, 데이터 공유 표준 제정과 함께 연구데이터 공유를 위한 research hub 플랫폼 개발을 추진하여 연구 결과와 데이터를 공유하고, 협업하도록 지원

※ 대규모 코호트 구축과 관련하여 2017년부터 2026년까지 100만 명 규모 데이터 및 생체자원 DB화를 목표로 사업을 추진하고 있으며 현재 51만 명 이상 참여, 36만 명 이상의 데이터 구축 완료(All of Us)

- **(영국) 2021년 7월 수립한 영국혁신전략(UK Innovation Strategy)에서 미래 영국의 경제를 변화시킬 분야로 생물정보학 및 유전체학(Bioinformatics and Genomics), 합성생물학(Engineering Biology) 등을 제시**

- 2021년 7월에 향후 10년 동안의 국가 바이오 비전인 영국생명과학비전(UK Life Sciences Vision)을 수립

- 일반인 50만 명 유전체 데이터 구축 및 25년간 추적조사를 목적으로 하는 바이오 데이터뱅크가 구축, 운영되고 있으며 '21년도 11월부터 20만명의 전장유전체분석 데이터를 승인된 연구자에게 제공 중

- '18년 10만 명 데이터가 구축 완료되어 목표를 달성하였으며, 500만 명의 Genomic Data 생산을 발표하고 프로젝트 추진 중(Genomics England)

- **(일본) 2010년 초부터 맞춤의료를 보건의료 분야 발전의 주요 전략으로 제시하고 맞춤의료 실현을 위해 바이오뱅크 구축 프로젝트 진행**

- 2030년 세계 최첨단 바이오경제사회 실현을 목표로 바이오전략 2020 수립을 통해 바이오와 디지털의 융합을 기반으로 바이오 데이터를 구축하고 이를 활용하는 계획 추진

- 제6기 과학기술·혁신기본계획('21~'25)을 통해 바이오데이터 수집·활용 지침 마련·추진 및 의료연구개발기구(AMED)를 설립하여 게놈·데이터 인프라 정비 및 데이터 활용 촉진을 추진

- **(중국) 21세기 초부터 바이오 데이터 구축 및 정밀의료 실현을 위한 정책을 수립하여 실행**

- 2030년까지 미국 정밀의료 프로젝트의 40배 규모인 600억 위안(약 10조 710억 원) 지원 계획

- 정밀의료 빅데이터의 자원통합·저장·이용·공유 플랫폼 구축 프로젝트, 질병 예방·진단·치료 방안의 정밀화 연구 프로젝트 진행

- **(EU) EU는 바이오경제를 유럽 그린 딜 구현과 지역적 다양성 강화를 위한 핵심 구성요소 중 하나로 인식하고 Horizon Europe(2021-2027) 프로그램을 통해 연구개발 추진**

- 2020년부터 백만명 이상의 유전체 정보와 관련 임상데이터를 수집하고 공유하기 위한 백만명 유전체 이니셔티브(1+Million GenomesInitiative, 1+MG)를 지원

- 정밀의료 빅데이터의 자원통합·저장·이용·공유 플랫폼 구축 프로젝트, 질병 예방·진단·치료 방안의 정밀화 연구 프로젝트 진행

<표 1> 주요 국가별 정책 추진 현황

	데이터 수집 규모/기간	사업목적	사업내용
미국, All of us	100만 명 / '17~'26년	장기적인 관점에서 다양한 국가의 연구자들이 사용가능한 대규모 코호트 및 데이터 플랫폼 구축을 통한 개인 맞춤 의료서비스 구현	국립보건원(NIH)이 미국 내 대학, 기업, 의료기관, 비영리기관 등과 협력하여 '17~'26년의 기간 동안 미국 전역의 자발적 참여자를 중심으로 구축
미국, Million Veteran Program (MVP)	100만 명 / '11~계속	유전자가 건강과 질병에 미치는 영향에 대한 연구를 통해 재향 군인과 일반인의 건강 개선	'11년 이후 VA에 등록된 재향 군인을 대상으로 자발적 참여자 100만 명의 설문조사 및 샘플을 수집하고 연구적 목적으로 유전자와 건강 및 질병의 연관성 연구에 활용하기 위한 사업
영국, UK Biobank	50만 명 / '06~'10년	만성질환 관련 위험요인들에 대한 종합적 추적과 자세한 평가를 통해 다양한 종류의 질병발생 원인 규명	NHS에 등록된 40-69세 일반성인을 대상으로 50만 명의 기본사항에 대한 설문조사, 신체검사 및 시료 등의 인체자원을 수집하여 만성질환의 유전 및 환경요인 연구를 위한 코호트 구축 (최소 30년 추적)
일본, Biobank Japan	약 27만 명/ (1차)'03~'08년 (2차)'13~'17년	질병의 상세한 원인 규명을 통한 새로운 약물 및 치료법 개발과 유전자의 분석을 통한 맞춤의료 실현	지역 병원으로부터 약 20만 명의 환자에 해당하는 임상 정보 및 인체시료를 수집·보유하고 데이터 서버에 정보를 통합하는 인프라를 구축하여 진단 및 신약개발을 위한 자원 제공하는 정밀의료 자원 구축 사업
중국, 정준의료 계획	100만 명 / '16~'20년	대규모 인체자원 수집을 통해 위암, 간암 등의 특정 암 연구에 주력한 유전체 인과 관계 규명과 기초연구 개발	'16~'30년의 기간 동안 1) 차세대 임상용 생명체학 기술연구, 2) 대규모 군층 (환자, 건강인) 연구, 3) 정밀의료 빅 데이터의 자원통합 저장·이용·공유 플랫폼 구축, 4) 질병예방·진단·치료 방안의 정밀화 연구, 5) 정밀의료 집적 응용시범 시스템 구축을 통한 중국형 정밀의료 육성 계획
핀란드, Finngen Research Project	50만 명 / '17~'23년	핀란드 국민의 유전자 게놈 데이터와 건강 정보 결합을 통한 의학 혁신을 창출과 헬스케어·제약 분야의 대학과 제약회사의 공공 협력 활성화를 통한 국민 맞춤의료 실현	'17년 시작된 핀란드 대규모 정밀의료 프로젝트이며 헬싱키 대학 주도로 9개의 대형 제약회사가 참여하여 50만명의 혈액을 분석. '17년부터 10개 권역별 공공병원에서 환자 동의 후 유전자 정보 표본 수집, 분석한 뒤 바이오뱅크로 전송
EU, 1+ Million Genomes	100만 명 / '21~'22년	유럽의 최소 100만 개의 유전자 염기 서열 수집과 EU 국가 간 협업 메커니즘을 구축하여 질병예방, 맞춤의료, 임상 연구를 위한 연구자원 제공	'20~'22년의 기간동안 유럽 국가의 100만 명의 유전체를 수집하고 분산·통합 인프라를 구축하여 유럽 전역에서 유전체 데이터를 활용할 수 있도록 1)국가들과 이해관계자들의 참여 모집 및 거버넌스 모델 정의, 2)구체적인 인프라·가이드라인 구축 및 시범 운영 진행, 3)공유, 확장 및 이니셔티브 지속 추진 사업을 진행하는 대규모 코호트 구축 사업

출처) 다부처 (2022), 이병욱 (2023)을 바탕으로 재작성

해외 바이오 빅데이터 인프라 구축 동향

- **실험 위주의 연구 패러다임 하에서 이루어지던 바이오 연구 소재 위주의 축적 및 관리 체계에서 컴퓨팅 인프라의 발전, 데이터의 폭발적 성장, AI 혁신 등 ICT기술의 발달로 데이터 기반 연구로의 패러다임에 대응하는 인프라 체계 전환이 이루어지고 있음**

- **(미국, NCBI)** 미국 NCBI(National Center for Biotechnology Information)는 국립보건원 (NIH) 산하 기관으로, 2017년부터 NIH가 모든 지원과제에서 생산되는 데이터를 NCBI에 의무 등록하는 정책을 시행함으로써 세계 최대 바이오 데이터 센터로의 지위를 공고히 유지 중

  ※ 주요 데이터로는 문헌, 임상정보, 발현체, 단백체, 유전체, 구조, 변이, 화합물 정보, 질병, 바이러스 등 35종의 바이오 데이터베이스가 있으며, 약 30억 건의 데이터를 보유

- **(EU, EBI)** EU의 대표적인 바이오 데이터 센터로 EBI(European Bioinformatics Institute)가 있으며 발현체, 단백체, 유전체, 대사체, 이미지, 구조, 화합물 정보, 상호작용 등 39종을 보유

  ※ 생명과학 분야의 데이터 저장, 관리, 활용을 위한 하드웨어 인프라로 273PB의 저장공간과 40,000 CPU 코어 이상의 인프라를 구축하고 ELIXIR라는 이름으로 데이터 저장, 관리, 활용 인프라 서비스 제공중

- **(중국, NGDC)** 2015년 NGDC(National Genome Data Center)를 설립하여 운영중이며 유전체, 발현체, 상호작용, 후성유전체, 유전변이, 마이크로바이옴 등 24종의 데이터를 보유

  ※ 총 55종의 데이터베이스를 보유하고 있으며 8,000 코어, 35PB 규모의 인프라 서비스 제공 중('21년 기준)

  ※ GSA(Genome Sequence Archive) 데이터베이스를 기준으로 현재까지 11PB 이상의 데이터가 축적되었으며, 2021년 한 해 2,049명의 사용자가 데이터를 제출하였고 연간 약 4PB의 데이터가 다운로드되는 등 서비스가 활용 중인 것으로 조사

- **(일본, DDBJ)** 일본의 대표적인 유전체 데이터 센터인 DDBJ(DNA Data Bank of Japan)는 일본 국립유전학 연구소(NIG, National Institute of Genetics) 산하기관으로 2021년 6월 기준 약 28억 건의 바이오 데이터 보유

  ※ NIG에서 47PB, 15,424 CPU 코어 규모의 슈퍼컴퓨팅 인프라를 구축하여 서비스 제공

3. 국가 바이오 빅데이터 인프라 발전 방향

바이오 데이터생태계 가치사슬

- **바이오 데이터생태계 가치사슬의 흐름은 ①바이오 데이터 생산 → ②바이오 데이터 정제·가공을 통한 저장 및 유통 → ③연구 및 산업 주체별 활용으로 이루어짐**

- **(활성화 요인) 각 단계를 거쳐 데이터 가치가 창출되도록 분야 특성에 적합한 활성화 요인(enabling factor)을 식별하여 제시하고 개선 및 발전방향을 제시함으로써 프로세스가 원활히 작동하도록 지원**

데이터 생산 부문 인프라 발전 방향

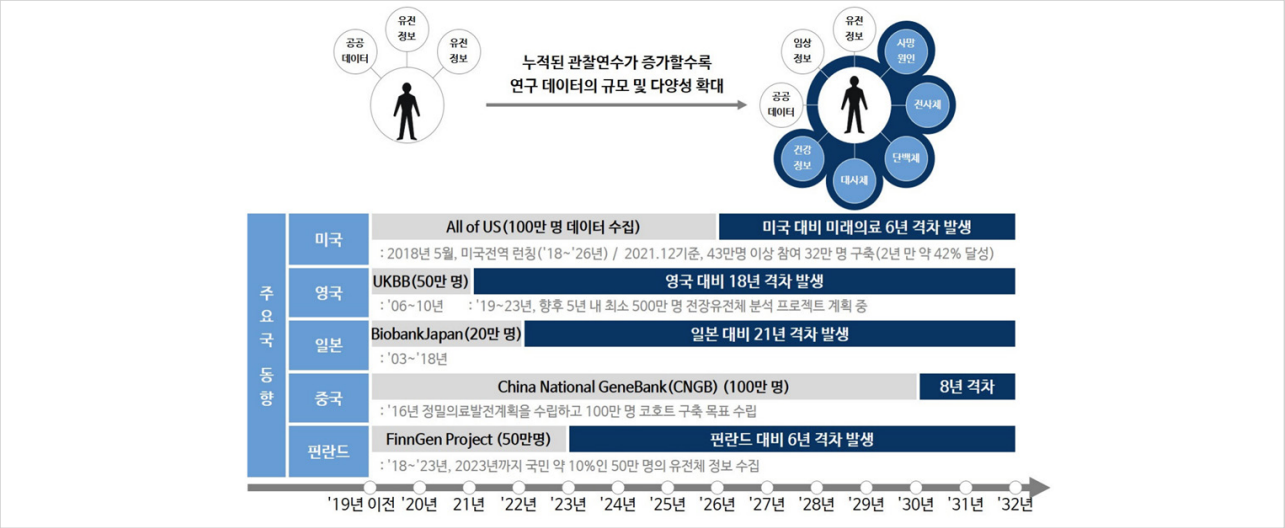
- **(기술패권 대응을 위한 바이오 빅데이터 구축 추진)** 선진국이 국가 전략자원으로서 바이오 빅데이터를 구축하는 현실을 감안하여 국내에서도 지속 확장성이 있는 국가전략자산으로 대규모 바이오 데이터 구축 필요

- 포스트게놈 다부처 유전체 사업(2014-2021)을 통해 개인별 맞춤의료 실현과 동·식물, 해양생물 등 다양한 유전정보를 수집하고 있고, 국가 바이오 빅데이터 시범사업(2020-2022)을 통해 2만5천명 규모의 한국인 전장유전체 및 임상데이터를 구축 중

- 정부 차원의 국가 바이오 빅데이터 생산 시범사업 완료 후 9,988억원 9년 사업으로 현재 예비타당성 조사가 진행되고 있음

- 데이터 기반 정밀의료 분야의 연구개발 경쟁력 강화를 위해 국가 간 격차를 감안한 조속 추진 필요

<그림 4> 주요국 바이오 빅데이터 구축 대비 한국과의 격차



출처: 다부처 (2022)

- 선도국 사례 및 난치 질환에의 활용을 고려하여 100만명 규모에 준하는 빅데이터 구축 필요

- 민감정보 활용에 대한 법적 제약, 개인정보보호, 보안 문제 해소가 필요하므로 제도적 기반과 함께 안전한 제공을 위한 기술적 기반 조성 필요

  - 통합 바이오 데이터의 수집·제공·활용이 원활하도록 개인 중심의 통합적 동의 기반으로 데이터를 수집하여 법적 장애요소 해소

  - 데이터 제공 주체의 포괄적 동의체계 하에 기술적으로 안전한 접근제어 시스템 등 데이터 보호 체계 구축

데이터 저장·유통 부문 인프라 발전 방향

- 10만명 이상의 대규모 데이터 축적이 가능한 공유 플랫폼이 현 시점에서 부재하므로 생산 단계 이후 이를 구축하고 공유·개방하는 별도의 플랫폼 필요
- BAM, FASTQ 등 원데이터 및 백업 데이터를 포함하면 10년 내 100PB~150PB 이상의 스토리지가 필요한 것으로 추산
- 정보 유통 시 개인 민감정보 보호를 위한 처리(암호화, 가명화, 비식별화) 방안 마련

대규모처리량 기술 등장과 데이터 제공 유통의 활성화

- 대규모의 바이오 데이터 처리가 가능한 대규모처리량 기술(high-throughput technology)이 등장하고 다중 오믹스 데이터가 축적되면서 데이터 간 연결성 강화는 데이터 제공·유통의 주요 활성화 요인으로 대두
- 멀티오믹스 데이터, 임상정보, 일상정보(라이프로그) 등 다양한 보건의료데이터 간 연계가 제공되는 유전체-오믹스-공공데이터 통합 서비스 체계 필요
- 분야 간 연계 외 국가 간 연계에 필요한 데이터 표준화 서비스 체계도 활성화의 주요 요인으로 고려
- 국가바이오데이터스테이션(K-BDS)을 통해 ’21년부터 53개 유형의 바이오데이터 통합 제공 체계가 마련된 바 있으므로 향후 다양한 국가 및 지역 바이오 데이터의 통합분석을 위한 연계 체계 및 기술적 실행기반 필요

연구 및 산업 주체별 활용을 위한 인프라 발전방향

- 연구 및 산업 주체별 활용 프로세스가 원활히 이루어지기 위해서는 대규모 데이터 활용을 원활하게 하도록 지원하는 요인 필요
- 인프라 관점에서 ①빅데이터 분석에 필요한 대규모 컴퓨팅 인프라 공급, ②빅데이터 분석을 위한 협업 인프라의 제공으로 구분

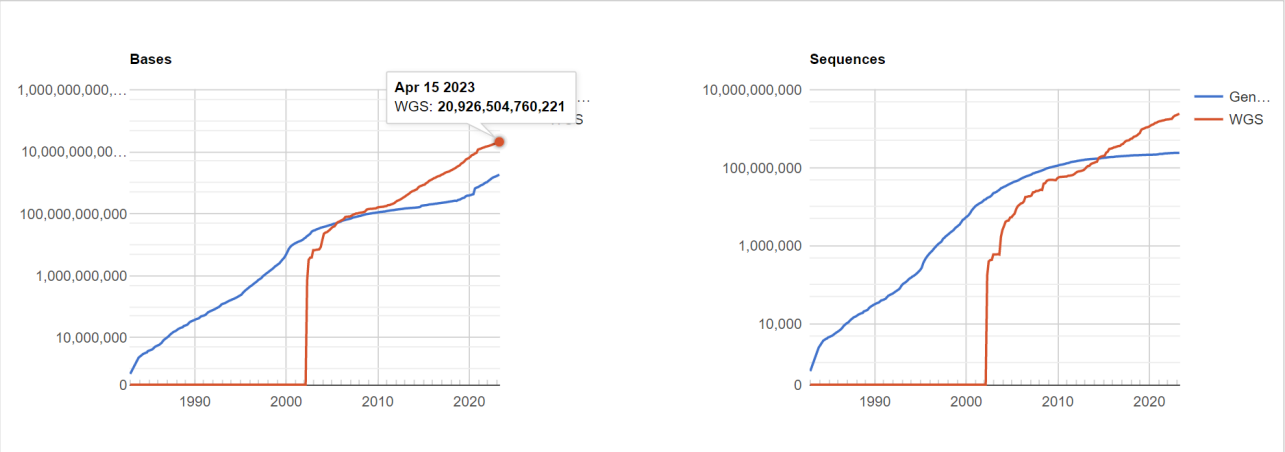
빅데이터 분석에 필요한 대규모 컴퓨팅 인프라 공급

- **(데이터 활용환경 강화)** 대규모 빅데이터 분석을 위한 국가 차원의 분석·활용환경 강화 및 제공 필요

  -「제3차 국가생명연구자원 관리·활용 기본계획」,「생명연구자원 빅데이터 구축전략」에 의거하여 국가 차원의 바이오 데이터 통합 수집·제공이 추진되었고, 추진과제로 데이터 활용환경 조성이 포함되어 있음

- '21년을 제외하고 국가 차원의 데이터 활용환경 구축은 국가 바이오 데이터 축적이 가시화되는 1단계 사업(~'23년) 종료 후 추진 예정
- 바이오 연구데이터의 기하급수적인 증가 추세 및 구축 소요기간을 감안할 때, 분석 활용환경의 선도적 구축 및 강화가 필요한 시점임

<그림 5> Genbank and WGS 데이터의 증가현황 (美 NIH 사례)



출처) https://www.ncbi.nlm.nih.gov/genbank/statistics/

- 향후 축적되는 국내 바이오 데이터를 추산했을 때('26년까지 14PB 규모), 필요 분석 인프라 규모[2] 로 최소 10,000코어 이상의 컴퓨팅 인프라가 필요한 것으로 조사

- **KISTI가 데이터 연구 활성화를 위해 국가 차원에서 구축, 운용 중인 초고성능컴퓨팅 연구지원체계가 있으나 과학기술 전 분야를 대상으로 하고 있고 신청자원 대비 자원 제공율이 65% 수준으로 자원 제공 여력이 부족한 상태임을 감안할 때 바이오 분야에 필요 계산자원을 온전히 제공하기 어려움**
  - 즉 바이오 분야에 특화된 전용 컴퓨팅 인프라가 필요하며 클라우드 기반의 연구환경 구축 및 데이터 특성에 맞는 분석·시뮬레이션을 수행할 수 있는 환경을 동적으로 구성하여 제공할 필요가 있음

디지털 바이오 패러다임 대응

- **바이오 분야 전반에 AI 등 데이터 기반 기술을 적용하는 디지털 바이오 패러다임*이 확산되고 있어 이에 대응하는 활용환경 강화 필요**

  * 디지털 바이오 패러다임: 첨단 디지털 기술과 바이오 간 융합 가속화로 새롭게 등장한 패러다임으로 데이터와 S/W로 연구하는 바이오 연구 패러다임

- **신약 개발을 위한 약물-단백질의 상호작용 예측, DNA서열분석, 단백질 구조 예측, 유전체 변이/발현 및 기작 예측 등 데이터 분석과 예측모델에 AI가 적극적으로 활용되는 추세**

2) 데이터 1PB당 요구되는 컴퓨팅 인프라 규모는 600코어로 산정(암 유전체 데이터센터, 영국 Wellcome Sanger 연구소의 인프라 규모를 기반으로 산정)

- **디지털 바이오 혁신전략('22) 등에 따르면 국가 차원에서 데이터 기반 연구 확산을 위한 기반 기술 확보 등 다양한 시책을 추진하고 있음**

- **인프라 제공 측면에서는 국가바이오데이터스테이션, 바이오소재 플랫폼 등을 통해 데이터 수집·관리·공유 체계 조성에 집중하고 있음**

- **향후 데이터 활용 활성화를 위한 환경 구축 및 AI 활용 플랫폼 강화 필요**

  - 바이오 분야 AI 활용 모델 구축에 있어서 학습 데이터 규모가 거대해짐에 따라 학습 및 검증에서 투입되는 컴퓨팅 인프라 또한 개인이 구축하거나 상용 클라우드를 활용하기 어려운 규모를 요구하고 있음

  - 최근에는 각종 분야별 특이 데이터로 사전에 미리 비지도 훈련을 적용하여 어느 정도의 예측능력을 갖춘 딥러닝 모델을 배포하고 있으며, 사용자는 미리 훈련된 모델을 자신의 데이터에 적용하여 학습시간을 줄이는 추세임

  - 따라서 인공지능 모델 개발을 위한 클라우드 플랫폼 및 개발된 사전학습 모델로부터 응용 AI 모델을 개발하고 서비스할 수 있는 분석 플랫폼 제공 필요

  ※ 레이블링, 어노테이션 등을 통한 AI학습 데이터 제공→고성능 클라우드 컴퓨팅 인프라를 기반으로 한 AI모델 학습 지원 또는 사전학습모델의 개발 및 배포 → 연구자 입력 데이터를 학습하여 응용 AI 모델을 개발하도록 지원하는 AI모델 지원 특화 플랫폼 필요

빅데이터 분석을 위한 협업 인프라의 제공

- **바이오 데이터 기반 연구는 그 특성상 협업이 매우 필수적인 분야로 협업 네트워크의 구성과 활용에 대한 지원이 매우 중요한 활성화 요인으로 작동**

- **분석 및 해석과 응용에 다양한 분야별 전문성이 필요함**

  - 바이오 데이터는 다양한 소스에서 수집되며, 다양한 형식으로 제공되는데 유전체 데이터, 의료 기록, 생물학 실험 결과 등과 같이 하나의 분석에 필요한 데이터가 서로 다른 형식과 구조로 제공되므로 데이터 엔지니어 등 전문가들이 협업해야 데이터 품질을 평가하고, 데이터 간 호환성을 확보하여 통합 데이터셋을 구축할 수 있음

  - 유전자와 질병 사이의 관계분석 단계에서도 유전체 전문가와 데이터 과학자가 협업해야 유전자와 질병 사이의 관계 분석이 용이하며 유전자 변이와 질병 사이의 연관성을 분석한 후에는, 임상 의사 및 신약개발 전문가와 협업을 수행해야 진단법 개발, 유전자 치료법 개발, 신약 개발로 연결할 수 있음

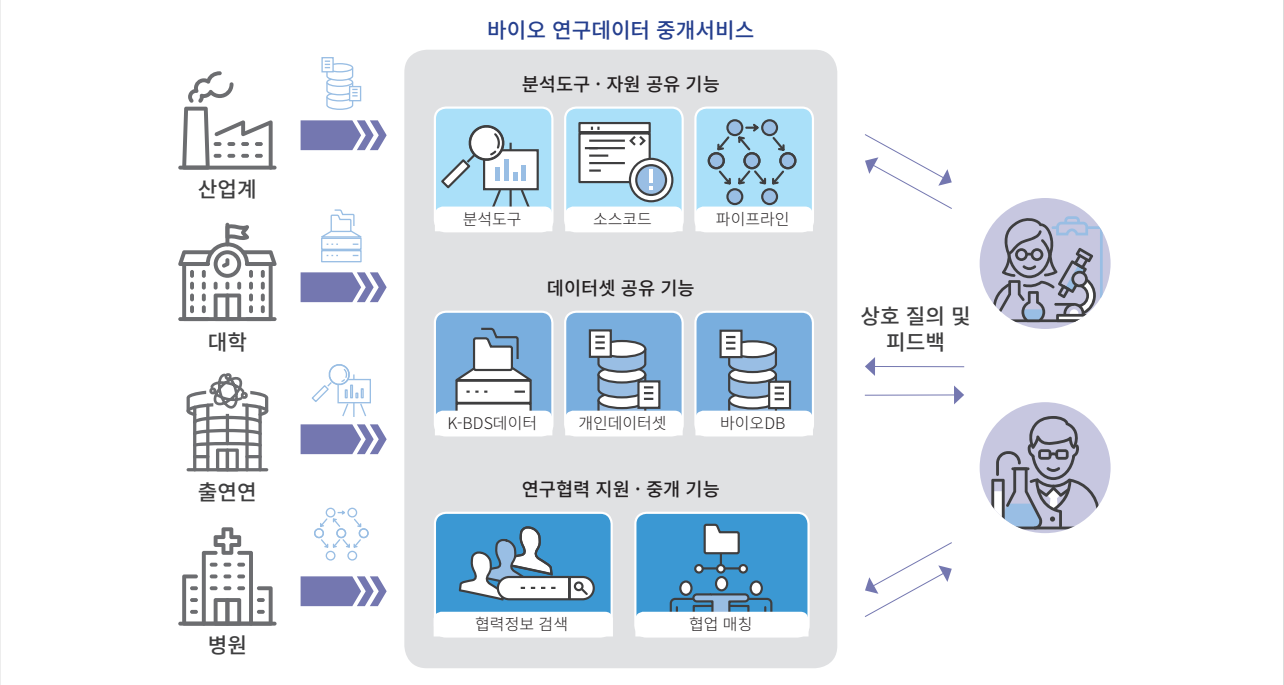
- **데이터 생산자, 전처리 및 분석 전문가·전문기관, 분야별 연구자 간 네트워크 구축 지원 및 수요 맞춤형 중개는 바이오 데이터 활용 활성화의 주요 요인으로 필요함**

  - 연구자간 협업을 위한 제한적 공유를 목적으로 하는 데이터 저장·정보 지원 인프라 공유체계 필요

  - 데이터 외 분석SW·소스코드·파이프라인 등 분석 관련 도구 공유와 토의를 위한 사용자 교류 및 지원 시스템 필요

  - 연구협업과 중개지원을 위해 동종 및 이종 분야 연구자·기관·기업 풀을 제공하고, 데이터 전처리·분석·응용(후속연구) 및 사업화 관련 공급-수요를 매칭하는 중개 서비스 필요

<그림 6> 바이오 연구협업을 위한 중개서비스 개념도



출처: 자체 작성

4. 시사점 및 제언

- 국가 바이오 데이터 기반 산업 및 연구경쟁력 강화를 위해서는 바이오 데이터 생태계의 가치사슬 흐름(생산→저장 및 유통→주체별 활용)에 대한 이해 및 단계별 활성화 방안이 필요
- 생산 단계에서 바이오 기술 패권 대응을 위한 100만명 규모의 자체 바이오 빅데이터 구축이 필요한 시점으로 구축한 데이터는 국가 전략 자산으로서의 역할을 고려한 제도적 기반과 기술적 서비스 기반 조성이 필요
- 데이터 공유와 연결성 강화는 바이오 데이터 제공과 유통의 주요 활성화 요인임. 이를 위한 통합 데이터 인프라 확대 및 AI 등 디지털 기술 활용 기반 강화가 중요
- 대규모 빅데이터 분석을 위한 컴퓨팅 인프라와 정보 인프라를 제공하여 연구 및 산업 주체가 원활하게 데이터를 활용할 수 있도록 해야 하며, 특히 데이터 생산자, 전문기관, 분야별 연구자 간의 연구협업 네트워크 구축 지원 및 데이터 공유체계를 통해 연구 협업을 촉진하는 것이 필요
- 지속적 인프라 정책 및 전략을 통해 바이오 데이터 생태계의 발전을 추진해야 하므로 이를 위한 산업, 학계, 정부 등의 다양한 이해 관계자들 간의 협력이 필요

참고문헌

- 과학기술정보통신부 (2022), 「바이오 대전환 시대 디지털바이오 혁신전략(안)」, 2022.12.
- 관계부처 합동 (2019), 「바이오헬스 산업 혁신전략」, 2019.5.
- 관계부처 합동 (2020), 「제3차 국가생명연구자원 관리·활용 기본계획('20~'25)」, 2020.5.
- 관계부처 합동 (2020), 「생명연구자원 빅데이터 구축전략」, 2020.7.
- 김준 (2023), 「유전체 빅데이터 기반의 맞춤 의료」, BioINpro Vol.109, 2023.
- 다부처 (2022), 「국가 통합 바이오 빅데이터 구축 사업」 기획보고서 , 2022.11.
- 생명(연) 워킹그룹 (2022), 「데이터 주도 과학 시대의 바이오 디지털 전환」, BioINpro vol.107, 2022.12.
- 이병욱 (2023), 「데이터 주도 과학 시대의 바이오 디지털 전환」, BioINpro Vol.109, 2023.3.
- KIAT (2023), 「산업기술환경예측–바이오·헬스」, 산업기술환경예측 보고서, 2023.2.
- KISTEP (2021), 2021년도 사업계획 적정성 재검토 보고서 「바이오 연구 데이터 활용 기반 조성사업」, 2021.12.
- KISTEP (2022), 2021년도 예비타당성조사 보고서 「국가 플래그십 초고성능컴퓨팅 인프라 고도화 사업」, 2022.10.

저 자

이 용 호
KISTI 국가슈퍼컴퓨팅본부
슈퍼컴퓨팅응용센터 바이오의료팀
책임연구원
T. 042-869-0923
E. stylee@kisti.re.kr

이 준 학
KISTI 국가슈퍼컴퓨팅본부
슈퍼컴퓨팅응용센터 바이오의료팀
책임연구원
T. 042-869-0714
E. juneh@kisti.re.kr

강 효 진
KISTI 국가슈퍼컴퓨팅본부
슈퍼컴퓨팅응용센터 바이오의료팀
책임연구원
T. 042-869-1049
E. hjkang@kisti.re.kr

KISTI ISSUE BRIEF 제58호

발 행 일 2023. 06. 28.

발 행 인 김재수

편 집 위 원 조민수, 서태설, 김한국, 고미현, 이상환, 최희석, 최장원, 곽영

발 행 처 34141 대전광역시 유성구 대학로 245
한국과학기술정보연구원 정책연구센터
https://www.kisti.re.kr

I S S N 2635-5728