

# BIO ECONOMY REPORT

June 2020, Issue 24

## 바이오 빅데이터 (Bio BigData) - 데이터가 생명을 살린다(Data Saves Lives)

## 바이오 빅데이터 (Bio BigData) - 데이터가 생명을 살린다(Data Saves Lives)

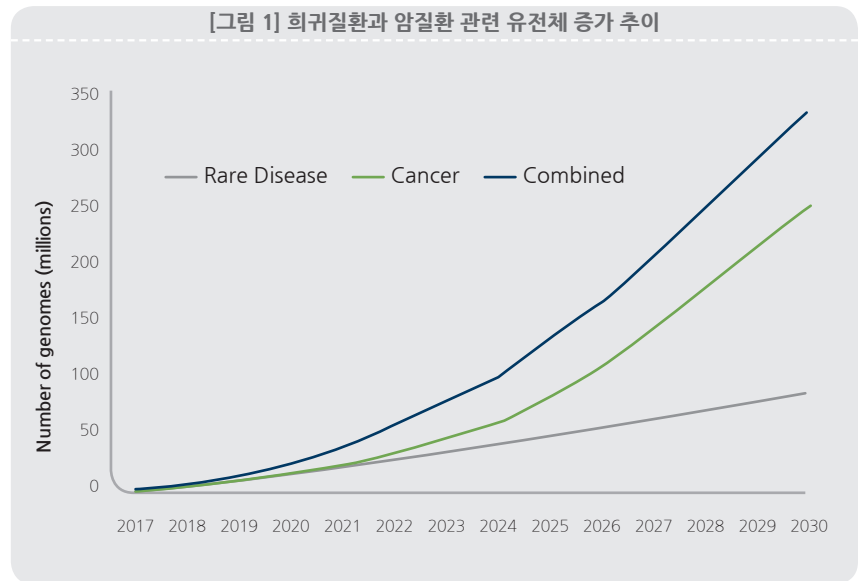
김태형 테라젠바이오 상무

### 개요

인간 게놈을 해독하는데 들어가는 비용은 2006년만 해도 약 1000억 원이 들어갔다. 차세대 염기 해독(NGS, Next Generation Sequencing) 기술 덕분에 현재는 게놈당 100만 원이면 시퀀싱이 가능해졌다. 또한 휴먼 게놈 프로젝트 당시만 해도 한 명의 게놈을 해독하는데 수년간의 시간이 들어갔지만 지금은 시간도 엄청 단축되어 48시간 정도면 한 사람의 유전체를 해독할 수 있는 시대가 되었다.

바이오 빅데이터 분야에 있어서 유전체뿐만 아니라 데이터 마이닝 분야에서도 새로운 비즈니스 기회를 창출하려고 하고 있다. 인간 게놈은 30억 개의 염기쌍으로 구성된 약 2만 개의 유전자로 구성된 3 기가바이트의 데이터이다. 서열 이외에도 구조/기능 해석, 질병 상관관계, 인구 집단내 차이 등의 가치 있는 데이터들을 어떻게 마이닝을 하느냐에 따라서 엄청난 보석을 얻을 수가 있는 무궁무진한 데이터이다. 이 데이터를 기반으로 다양한 기업 및 연구소들은 암, 유전질환 및 만성질환 환자들을 진단하고 치료하기 위해 다양한 오믹스 빅데이터 및 인공지능 기술을 적용하고 있다. 이러한 가치 있는 바이오 빅데이터들을 활용하고 공유하기 위한 많은 시도도 함께 이뤄지고 있다. 이번 호에서는 바이오 빅데이터가 적용된 사례 위주로 소개하고자 한다.

## 유전체 빅데이터 및 정밀의료 사업



출처: GA4GH

2017년까지만 해도 미국 국립생명공학정보센터(NCBI)에 등록되는 유전체 데이터가 약 2년마다 2배씩 증가 되고 있다고 했다. 이 기록은 최근에 그 주기가 더 짧아져 7개월밖에 되지 않을 정도로 유전체 데이터가 급속히 모이고 있다. [1]

여기에 가장 강력한 주체가 되어 가고 있는 단체로 GA4GH(Global alliance for genomic & health) 컨소시엄은 향후 몇 년 간 우선적으로 진행할 유전체와 임상 데이터를 수집하고 안전하게 공유하는 기술 및 표준화 개발 로드맵을 발표했다. GA4GH는 정밀 의료를 위한 표준화와 프레임워크를 만들기 위한 컨소시엄 형태의 프로젝트로 625개 기관 회원으로 구성된 큰 단체로 발전했다. 이들은 2025년까지 약 1억 명 이상의 환자 유래 유전체와 임상 데이터가 전 세계적으로 모일 거라고 예상하며 이를 위해 이 데이터들이 전 세계 연구자 및 기업들이 참여해 활용할 수 있게 이들 데이터를 안전하게 공유하는 시스템을 구축하겠다고 발표했다. [2]

이러한 바이오 빅데이터를 가장 먼저 생산하고 활용을 가장 먼저 실행하고 있는 국가는 영국이다. 2012년 캐머런 수상이 발표한 영국 10만 명 유전체 프로젝트의 구체적인 실행을 위해 영국 NHS는 공공기업인 지놈잉글랜드를 설립하고 이후 5년 만인 2018년 10만 명 유전체 해독을 모두 완료했다. 2019년 10월 1일부터는 전체 국민을 대상으로 'Data Saves Lives'를 위해 유전체 검사를 진행하기로 결정했다. 그리고 이렇게 모인 유전체/임상 데이터로 최대하고

부가가치를 만들기 위해 5개 유전체 기업과 13개 글로벌 제약사로부터 연구비/기술/노하우도 기부받고 각 영리 기업들이 이 데이터를 활용해 미래의 진단기술과 신약개발을 하도록 적극 독려하고 있다. 최근 영국 보건복지부 장관은 유전체의학을 통해 향후 5년 안에 500만 명 유전체를 확보한다는 계획도 추가로 발표했다. [3]



출처: NHS Health Education England

영국 NHS의 유전체 빅데이터 기반의 정밀의료 사업의 성과를 정리하면 다음과 같다.

- 유전성 질환, 암 질환 환자들을 위해 10만 명의 유전체 정보를 제공할 수 있게 되었다.
- 전 세계 연구자들과 산업체와 파트너십을 맺고 고부가가치를 창출할 수 있게 되었다.
- 10만 명 유전체 데이터를 기반으로 환자들을 치료하기 위한 다양한 미래의 진단 기술을 좀 더 신속하게 개발할 수 있게 되었다.
- 게놈 해독 센터로부터 데이터를 생산하고 분석하고 보관하고 배포하는 것에 있어 보안이 취약한 부분을 모두 파악했고 안전하게 사용하는 방법을 찾게 되었다.
- 10만 유전체 사업 경험을 바탕으로 2018년 10월 1일부터는 NHS는 세계 최초로 전 국민 대상으로 특정 적응증에는 WGS 루틴 검사를 수행하게 되었다.

여기에 미국, 중국 등의 국가들도 국민들을 대상으로 각자 100만 명에서 1억 명 정도의 유전체 빅데이터를 확보하는 것을 목표로 대규모 바이오 빅데이터 기반의 구축을 통한 정밀의료 사업들을 진행 중이고 이외에도 여러 나라들이 국가 주도하에 자신의 국가의 미래 헬스케어 분야의 큰 가치를 만들기 위해 유전체 빅데이터를 확보하고 있는 중이다.

특히 미국의 경우 NIH 4,000억 원 펀드를 만들어 670여 개 정밀의료 관련 연구를 지원하고 있으며 특히 GTEx 7천 명 이상의 정상인 조직 유래 오믹스 데이터(200TB), TCGA 1만 명 이상 암환자 유래 오믹스 데이터(4PB) 및 TOPMed 10만 명 이상의 다양한 만성질환 환자 전장 유전체 해독(WGS) 데이터(15PB)를 생산해 정밀의료 연구 및 임상 적용을 위해 연구자들에게 이미 배포하고 있다.

중국의 경우 전 세계에서 가장 빠르게 유전체 진단 규제정책을 국가 주도하에 빠르게 진행해 이를 기반으로 유전체 시장이 급속히 성장하고 있으며 2022년까지 3조 원 시장으로 성장을 예상하고 있다고 한다. 2020년 현재 전 세계 유전체 시장 규모가 약 5조 원 정도 규모인 것을 감안하면 매우 빠르게 그리고 크게 성장하고 있다.

한국도 조금 늦은 감은 있지만 국가 바이오 빅데이터 구축 사업을 통해 100만 명 유전체 정보 수집을 위한 사업을 추진계획을 발표했으며 바이오 빅데이터 사업 분야를 포함해 바이오 R&D 투자를 약 4조 원 이상으로 확대할 예정이라고 한국 정부 발표도 있었다. 이를 기반으로 국내에서도 바이오 빅데이터가 실제 임상에 적용될 수 있는 범위까지 확대되고 이를 많은 연구자와 임상 의사 및 바이오/제약 기업들이 경험해 곧 다가올 정밀 의료 시대를 준비해 나갈 수 있으면 하는 바람이다. [4]

## DTC 데이터 증가 및 디지털 바이오뱅크

2019년까지 DTC(direct-to-customer) 유전자 검사 서비스를 받은 사용자가 3000만 명을 돌파했다. DTC 유전자 검사 건수는 기하급수적으로 증가하고 있다. 전체 건수의 80% 이상을 차지하는 두 회사인 AncestryDNA는 1500만 명, 23andMe는 1000만 명 이상 서비스 건수를 각각 돌파했다. 이들 회사들은 자체적인 디지털 표현형 수집을 통해 코호트 기반의 유전체/임상 연구의 새로운 대안을 제안하고 있으며 유전정보, DNA족보, 다양한 질병 표현형, 그리고 디지털 표현형 정보까지 통합된 정보를 수집한 대규모 연구 결과를 연이어 발표하고 있다.

이러한 분위기에 힘입어 DNA랜드(DNA.Land)라는 새로운 디지털 바이오뱅크의 새로운 프레임워크가 나왔으며 DNA랜드는 여러 연구기관, 병원들이 컨소시엄을 통해 대규모의 자원을 필요로 하는 기존 바이오뱅크(Biobank) 프로젝트와는 완전히 다른 접근 방식으로써 DTC 유전자 서비스를 받은 개개인들로부터 자발적으로 유전체 데이터를 제공받아 이를 기반으로

유전체 데이터를 수집하는 디지털 바이오뱅크(Digital Biobank)이다. 현재(2018년 2월 10일 기준) DNA랜드의 인터넷 웹사이트(<https://dna.land>) 상에는 수십만 명의 사용자가 자신의 유전체 정보와 표현형 정보를 연구에 기여하기 위해 스스로 공유한 현황을 볼 수 있다. 미국 유방암 연합(National Breast Cancer Coalition, NBCC)에 참여한 사용자에게도 이러한 절차를 통해 다시 그들의 유전체를 공유하고 설문조사를 참여하도록 시도하였고 약 6개월 동안 약 1만 명 이상의 참여자들이 설문조사를 마쳐 매우 성공적으로 완료하기도 했다고 한다.

참가자들은 자신의 유전 정보(23andMe, AncestryDNA, FamilyTreeDNA 및 MyHeritage에서 서비스받은 파일)를 업로드하고 자신의 정보를 제공하는 범위를 지정할 수가 있는데 보통 업로드하고 24시간 이후에 DNA족보, 신체적 특징 및 웰니스 관련 등의 예측 결과 리포트를 제공해준다(여기서 만들어진 모든 프로그램의 소스들은 깃허브, <https://github.com/TeamErich/dnaland>를 통해 모두 공유하고 있다). 현재는 복잡한 규제를 피하기 위해 질병 관련 표현형 결과는 제공하지 않으며 단지 유방암 가족력과 같은 질병 표현형에 대해서만 표현형 정보를 수집하고 있다.

궁극적으로는 이러한 시도를 통해 유전정보, DNA족보, 다양한 질병 표현형, 그리고 디지털 표현형 데이터까지 통합된 바이오 빅데이터를 수집함으로써 기존의 대규모 코호트를 기반으로 한 바이오뱅크와는 또 다른 보완적인 시스템이 될 것으로 예상된다.

마지막으로 이러한 DTC 유전자 검사가 급속히 성장하고 특히 DNA랜드와 같은 디지털 바이오뱅크가 발전할 수 있게 된 가장 큰 이유는 미국 내 DTC 유전자 검사에 관한 규제의 완화일 것이다. 개인이 회사에 의뢰해 자신의 유전정보를 획득하는 데 있어 제약이 없는 환경이 조성되었고 이를 통해 이런 건설적이고 다양한 시도가 가능해진 것으로 보인다. 이런 점에 비추어볼 때 미국보다 상대적으로 엄격한 국내 DTC 유전자 검사 규제로 인해 유전체 산업이 발전 못하는 부분에 있어 개선이 필요해 보인다. [5]



최근 10년간 많은 스타트업들은 기존의 약물 데이터베이스에서 기존과 다른, 새로운 용도로 사용할 수 있는 약물을 효율적으로 찾기 위해 노력하고 있다. 공개된 유전체 데이터베이스를 가지고서 인공지능 기술을 활용해 데이터 마이닝하는 것이다. 즉, 공개된 유전자 발현 (전사체; Transcriptome) 데이터베이스를 활용해 정상 세포/조직과 비교하여 특정 질환과 관련된 세포/조직에서는 특정 유전자 발현 패턴이 어떻게 변화하는지를 분석하고, 특정 질환 세포/조직에서의 발현 패턴을 정상 세포/조직의 패턴으로 전환하기 위해 필요한 화합물을 대규모 화합물 라이브러리에서 스크리닝 하여 신약개발 성공률을 높여 가겠다는 전략인 것이다. 이미 이러한 유전자 발현 데이터는 유전체 기술의 급격한 발전으로 인해 최근 20년 사이에 엄청난 연구가 이루어졌고 그 데이터의 대부분이 모두 공개되어 있다.

이러한 접근은 처음부터 화합물을 만들지 않고도 치료제를 효율적으로 개발할 수 있다는 장점을 가지고 있어 노바티스(Novartis)와 글락소스미스클라인(GlaxoSmithKline)과 같은 여러 대형 글로벌 제약 회사들도 이러한 형태로 빅데이터 분석을 적용하여 기존 약물에 대한 새로운 용도를 밝혀내는 시도를 하고 있다. 전산 알고리즘과 공개된 데이터베이스를 통해 우리가 과거에는 발견하지 못했던 새로운 타겟과 적응증을 확인함으로써 약물 개발 비용을 절감할 수 있다는 부분에서 장점이 매우 크다고 할 수 있다. 특히 정상, 질환 세포/조직에서 다양한 약물이 적용되어 축적된 수십 만개의 전사체 프로파일링 데이터가 NCBI GEO(Gene Expression Omnibus), SRA(Sequence Read Archive), TCGA(The Cancer Genome Atlas)에 공개되어 있으며 전 세계 기업들과 연구자들은 이를 활용한 약물 개발을 위해 데이터 마이닝을 진행하고 있다.

더불어 약물 개발에 빅데이터 기술을 적용하는 것의 가장 큰 장점은 신약 개발 초기단계부터 차별화가 없어서 상품가치가 없는 후보 화합물들을 모두 제외시키거나 임상시험 시 실패 가능성이 큰 파이프라인들을 초기 임상시험 단계에서 빠르게 실패 유무를 확인함으로써 전체 신약 개발의 시간과 비용을 절감할 수 있다는 점이다. [7]

## 바이오빅데이터와 시기술 융합

사람의 유전체는 약 2만여 개의 단백질을 코딩하는 유전자로 구성되어 있으며 그 외에도 2만 5000개가 넘는 비단백질 코딩 유전자가 존재하고 있다. 이들은 다양한 조합을 통해 여러 전사체를 생성하며 유전자 발현에 관여하고 있으며 이를 통해서 세포 내 복잡한 생물학적 기능들을 담당하고 있다. 그래서 이들을 모두 이해하려면 세포가 자신의 유전체를 처리하는 것처럼 해석할 수 있는 분석능력을 갖춰야 하는데 이에 접근하는 데 있어서 인간의 인지능력 이상



의 분석 능력이 요구되며 이를 가능하게 하는 시스템이 필요로 한데 딥러닝 기술이 하나의 대안으로 떠오르고 있다.

특히 차세대 유전체 해독(Next Generation Sequencing, NGS) 기술의 발전과 함께 사람마다 약 100 기가 염기(gigabase)에 해당하는 수천~수십만 건의 유전체 빅데이터를 활용할 수 있게 되면서 이러한 데이터셋을 기반으로 머신러닝/딥러닝 기술을 적용하면 실험실에서 모델 생물을 대상으로 유전적 변이의 기능을 해석하는 것보다도 더 효율적인 것으로 보고 있다. 더 나아가 질병의 원인이 되는 변이를 발견하고 타겟 치료제를 스크리닝 하는 시도도 진행 중에 있다.

최근 CRISPR-Cas9과 같은 유전자 편집 기술이 등장하면서 유전체 정보를 자유자재로 읽고(decoding), 쓸(encoding) 수 있게 됨으로써 이를 기반으로 세포 내 현상 및 유전체 특정 영역의 기능을 해석하고 특정 질병과의 연관성을 분석하기 위한 다양한 데이터셋을 대량으로 얻을 수 있게 되었으며 기계학습/딥러닝을 통해 이를 분석하는 시도가 이루어지고 있다.

2014년에는 딥러닝 기술을 이용해 조직별로 다양한 엑손-인트론 스플라이싱 패턴을 예측하는 SPIDEX, 2015년에는 단백질이 결합하는 유전체 사이트를 예측하는 DeepBind와 알츠하이머 환자 임상시험 참여를 돕는 톨 등이 딥러닝으로 구현이 되었다. 2016년에는 현미경의 기능을 개선하는 분야를 비롯해 당뇨병성 망막증(diabetic retinopathy)을 스크리닝 하기 위해 적용되기도 하였다. 2017년에는 구글에서 개발한 딥베리언트(DeepVariant)는 차세대 유전체 해독(Next Generation Sequencing, NGS) 데이터에서 정확한 염기 서열의 콜링을 위해 적용되기 시작하였다. 2018년에는 EHR(Electronic Health Records)로 부터 환자의 예후를 예측하거나 단백질 구조를 예측하거나 영상 이미지를 통해 피부암을 진단하는 다양한 분야로 확장되어 딥러닝 기술이 적용되고 있다.

하지만 이들 딥러닝 기술이 학습될 당시의 성능이 실제로 임상현장에서 적용될 때와 똑같은 성능이 재현성 있게 작동될지에 대해서는 아직은 큰 의문을 가지고 있으며 의료기관의 임상 의들과 이를 심의하는 규제 기관에서는 주의 깊게 이들 기술을 평가하고 고려 중에 있다. 아직 딥러닝 기술의 예측 모델을 만드는 곳에서는 제대로 된 근거를 만들기 위한 노력을 더 해야 하는 숙제가 많이 남아 있기도 하다. [8]

## 바이오빅데이터 적용 사례

최근 몇 년 사이에 바이오 빅데이터를 적용해 환자들을 위해 질병 진단 및 치료에 활용된 사례들이 있어 소개하고자 한다.

### - 미국 유행성 이하선염

2016~2017년에 걸쳐 미국 보스턴 지역에서 유행성 이하선염 (동의를: 볼거리) 감염이 급증했다. 매사추세츠 주의 공중 보건국과 지역 대학 보건 서비스와 협력 아래 환자들 유래 유전체 데이터를 사용해 해당 연구를 진행했다. 브로드 연구소 연구팀은 역학 데이터와 201명의 감염된 환자들에게서 유행성 이하선염 전체 유전체 서열을 분석하였으며 전 세계 연구자들에 의해 시퀀싱 되고 공개된 유행성 이하선염 3646개의 GISAID 데이터베이스 내 존재하는 게놈 서열들을 사용했다. 이러한 자료 및 데이터 수집을 바탕으로 실행된 연구는 실시간으로 전염병의 확산을 이해하는 데 큰 도움이 되었으며, 연구자들은 발생원을 찾게 되어, 전파되는 경로 및 패턴을 해결함으로써 질병의 심각도 및 백신과 진단에 영향을 줄 수 있는 유전체의 변화를 확인할 수 있었다. 적절한 데이터의 수집과 활용으로 큰 전염병의 유행을 확인할 수 있었으며, 백신 개발 부분에 있어서도 크게 도움이 되었다. [9]

### - 미국 혈우병 환자 네트워크

미국 6개 주의 NIH와 CDC (Centers for Disease Control and Prevention)는 혈우병 감시 시스템을 후원하여, 해당 주에 거주 중인 혈우병 환자를 모두 식별하고 의료 기록에서 인구 통계 및 임상적인 특징에 대해 자세한 정보를 수집하였다. 해당 기록은 6년에 걸쳐 혈우병 환자 3000명으로부터 발생률 및 치료와 합병증의 결과를 연구할 수 있는 데 도움이 되는 데이터를 사용하게 된다. 그 결과, 1975년, 미국 혈우병 환자에게 종합적인 치료를 제공할 수 있는 네트워크가 구축되었으며, 표준화된 데이터 수집 양식을 사용하여 혈우병을 비롯하여 각종 다양한 혈액 장애에 관련해서도 100여 개의 국가에서 데이터를 수집하고 있었다. 이러한 데이터를 바탕으로, 가난해 최하위 의료 서비스를 받을 수밖에 없는 나라에서도 혈우병 치료 및 생존율이 향상되었음이 증명되기도 하였다. 이렇듯, 의료 데이터 수집 및 활용은 공공의 이득을 추구하며 진행되는 경우도 많다. 또한, 정부 차원에서의 지원이 필수적이며 관련 회사들이 협력하여, 데이터 수집뿐만 아니라 관련 신약 개발까지 나아간다면 더욱 공공의 이득이 활성화될 것이다. 그리고 환자의 데이터 수집과 더불어 개인 정보 보호와 보안은 헬스케어 분야에 있어서 시간이 갈수록 매우 중요해지고 있다. [10]

### - 메이요 클리닉(Mayo Clinic) & 헬릭스(Helix)

2020년 초 메이요 클리닉과 캘리포니아 기반의 유전체 회사인 헬릭스와의 파트너십이 시작

되었다. 올 1월 9일, 메이요 클리닉은 캘리포니아 기반의 유전체 분석회사인 헤릭스와 파트너십을 맺어 의사와 연구원이 건강 상태 모니터링을 목적으로 유전체 데이터 베이스 시스템을 개발하였다. 메이요 클리닉은 해당 파트너십을 맺고 10만 명의 동의된 환자를 모집하여 연구 데이터베이스를 개발하기 위해 유전자 검사를 실행한다. 연구에 활용될 환자 데이터는 10만 명의 환자 각각의 전체 유전체 엑솜(Exome) 맵핑을 통해 예방 조치를 취할 수 있는 초기에 사용될 수 있다. 해당 연구는 유방암 및 난소암, 유전성 고 콜레스테롤 혈증 및 유전성 결장 직장암의 한 형태인 린치 증후군과 같은 유전성 질환에 대해 사용될 것이며, 이를 통해 정밀 의료 및 진단을 더욱 향상할 것을 명시했다. [11]

#### - 노스쇼어 대학 병원(NorthShore University Health System)

2020년 1월 13일, 일리노이 주 기반의 노스쇼어 대학교는 임상 유전학 서비스를 1만 명의 환자에게 제공하기 위해 미국 유전체 분석회사인 Color와의 협력을 시작했다. 1만 명의 환자 테스트 결과에 대해 시퀀싱을 진행하였고, 이러한 검사는 환자의 유방, 결장 직장암 및 심장병 등 특정 질환 발생 여부를 확인하는 데 사용이 된다. 1년 여 간의 시범 프로그램 동안 검사가 무료로 진행되었지만, 그 이후로는 환자들은 175달러 정도를 지불할 예정이다. 노스쇼어에서 제공하는 테스트 결과는 환자 치료를 위한 의사결정에 사용될 수 있으며, 환자의 치료에 직접 참여하여 질병예측 및 사전 예방을 목표로 한다. [12]

#### - 저스틴 오디세이 (Justin's Odyssey)

1997년, 저스틴 배천 (Justin Vachon)은 희귀 신경 장애를 갖고 태어났다. 그가 3살이 되던 해, 저수초형성(Hypomyelination)인 펠리제우스-메르츠바하 병(Pelizaeus-Merzbacher disease)을 진단받았다. 그 후 그가 13살이 되었을 때 캐나다 이스턴 온타리오 아동병원 (Children's Hospital of Eastern Ontario)의 신경학자들은 현재까지의 알려진 임상적인 질병과 저스틴의 증상이 맞지 않는다는 것을 알았다. 저스틴의 장애를 정확하게 확인하기 위해 전체 엑솜 시퀀싱(Whole Exome Sequencing)을 수행하였고, 2017년 저스틴의 신경 장애와 반응하는 유전자인 USP 7 (Ubiquitin-specific-processing protease 7)를 확인할 수 있었고 해당 유전자를 매치메이커 익스체인지(Matchmaker Exchange)에 증상에 대한 세부 정보를 공유하여 연구자들이 관련된 다른 유전자에 대한 새로운 변화를 확인할 수 있게 되었다. 네덜란드에 거주하는 한 가족과 해당 유전자를 일치시킬 수 있었으며, 호주에서 유전자가 일치된 아이들과 저스틴이 동일한 정도의 증상을 가진 것을 확인한 후 그의 질병을 담당하는 2가지의 변종을 알아냈다. 매치메이커 익스체인지(Matchmaker Exchange)와 유전체 연구 연합 조직인 GA4GH (Global Alliance for Genomics and Health)의 노력으로 인프라를 구축하여, 희귀 질환 진단율을 높일 수 있었다. [13]

## 바이오빅데이터와 인공지능이 열어가는 미래의 모습 시나리오

가까운 미래에 현실화될 미래를 상상한 사례(출처: Wainberg, M., et al. Deep learning in biomedicine. Nat Biotechnol, 2018)가 있어 소개하고자 한다.

한 여성은 직장에 출근하는 중 자신의 휴대폰으로 약국에 들러 혈액과 소변 검사를 하라는 알람을 받는다. 이 알람은 그녀의 건강관리 기록, 의료영상 이미지, 유전체 데이터, 혈액 기반의 전사체, 대사체 및 그녀의 심박수, 혈압, 근력, 및 기타 정신적 상태 데이터들을 프로파일링 하는 인공지능 시스템을 통해 전달되었다. 이 권고안들은 엄청난 빅데이터를 트레이닝해서 나온 것이다. 즉, 공개된 수백만 개의 바이오 빅 데이터셋뿐만 아니라 자신과 가까운 혈연관계의 친척들 뿐만 아니라 비슷한 임상적 소견을 보이는 사람들과 수억 명에 해당하는 일반인들 데이터를 비교 분석을 통해 도출된 근거 기반의 권고안이라고 볼 수 있다.

이 여성은 약국에서 혈액검사에 기반한 전사체의 변화 및 소변검사를 통한 대사체의 변화를 통해 신경근 퇴행성 질환 가능성이 발생했음을 알게 되었다. 이러한 결과에도 이 여성은 그렇게 놀라지 않았다. 그 이유는 이 여성은 이미 오래전부터 데이터를 통해 이러한 일이 내년 언제쯤에는 일어나리라는 것을 예상하고 있었기 때문이다.

사실 자신의 어머니가 자신을 임신해 있을 때부터 신경근 퇴행성 질환 관련 병원성 유전 변이를 가지고 있다는 것을 알고 있었고 그 변이에 대해 편집을 해서 복구할 수 있었지만 인공지능 시스템이 이 변이로 인해 자신의 딸이 살면서 그 질병이 발병할 가능성에 대해서 아주 낮게 평가함으로써 그녀의 어머니는 유전적 치료를 선택하지 않고 자신의 딸을 낳기로 결정했었다.

아무튼 현재 자신의 건강 상태를 알게 된 이 여성은 자신의 유전체와 전사체를 포함한 그녀의 모든 데이터를 분석해 그녀에게 최적화된 개인 맞춤형 유전체 의학 서비스(치료제)를 제공받게 되었다. 인공지능 시스템의 도움을 받아 설계된 치료제에 대해 걱정하는 사람들이 아직 많을 수는 있지만 결론부터 말하자면 미래에는 시원한 바람이 부는 가을 저녁에 산책하는 것만큼 안전하다. 게다가 100번 적용하면 99번 정도가 높은 수준의 효능까지 보여주고 있다.

이 치료제의 인체 내 안전성을 예측하기 위해 인공지능 기술을 이용해 예측하게 되는데 영양류 또는 동물을 통해 전임상 시험을 통해 예측하는 것보다도 인공지능 시스템을 통해 안전성을 예측하는 것이 더 정확하다는 것도 증명되어 그 인공지능 시스템은 이미 규제 기관의 승인을 받은 상태이다.

다시 그 여성의 이야기를 마저 해보자면, 그 여성은 다음날 바로 약을 선택하고 또 근처 자신의 치료를 담당하는 카운슬러를 만나게 된다. 그리고 그 카운슬러와의 대화를 통해 치료법에 대해 이야기를 나눈다.

그리고는 1년 넘게 그녀 자신이 직접 약 처방 및 치료 결정에 관여하면서 그녀가 가지고 있는 스마트폰이나 웨어러블 장치 및 생체 정보를 측정할 수 있는 다양한 장치 등을 통해 걷는 속도나 팔을 흔드는 강도 등을 포함해 그녀가 하는 모든 신체 활동 및 모든 신체 지표를 계속해서 수집하고 기록하며 그녀는 2주마다 한 번씩 약 국을 들러 그녀의 전사체 및 대사체를 지속적으로 모니터링하게 되었다.

치료 후 1년 만에 그녀의 신경근 퇴화 증상이 완전히 멈추었다는 결론을 자신의 데이터를 통해 도출하게 되었으며 그녀 자신의 질병도 효율적으로 치료해냈을 뿐 아니라, 그녀의 모든 데이터는 본인 개인의 동의하에 비식별화 및 익명화를 거쳐 미래의 자신과 같이 고통받을 사람들에게 더 나은 치료를 제공할 수 있게 되기 위해 인공지능 시스템에 입력되었다.

이 시나리오가 현실화되는 시점이 지금 당장은 아니겠지만 미래 언젠가는 이런 방식과 비슷하게 구현될 것으로 예상된다. 그리고 지금부터도 바이오 빅데이터가 인공지능/딥러닝 기술과 만나 헬스케어 분야 많은 난제를 신속히 해결하면서 우리에게 가시적인 성과들을 보여주게 될 것으로 예상된다.

---

 <참고문헌>

1. NCBI Data and User Services, <https://www.nlm.nih.gov/about/2019CJ.html>
2. Genomics in healthcare: GA4GH looks to 2022, <https://www.biorxiv.org/content/10.1101/203554v1>
3. [바이오스펙테이터] 美 국민 유전체검사 시행의 의미, [http://www.biospectator.com/view/news\\_view.php?varAtcId=5872](http://www.biospectator.com/view/news_view.php?varAtcId=5872)
4. [바이오스펙테이터] "100만명 바이오 빅데이터 구축-연 4조 R&D 투자", [http://www.biospectator.com/view/news\\_view.php?varAtcId=7684](http://www.biospectator.com/view/news_view.php?varAtcId=7684)
5. DNA.Land is a framework to collect genomes and phenomes in the era of abundant genetic information, Nature Genetics 50, 160-165(2018), <https://www.nature.com/articles/s41588-017-0021-8>
6. [메디게이터] GSK와 23앤드미의 신약개발을 위한 개인 유전체 데이터 빅딜, <https://www.medigatenews.com/news/1059961358>
7. [메디게이터] 유전체 빅데이터 마이닝 기술 기반의 신약 개발, <http://www.medigatenews.com/news/2320049700>
8. Deep learning in biomedicine, Nat Biotechnol. 2018 Oct;36(9):829-838, <https://www.nature.com/articles/nbt.4233>
9. Combining genomics and epidemiology to track mumps virus transmission in the United States, PLOS Biology, 2020, <https://doi.org/10.1371/journal.pbio.3000611>
10. Public health surveillance and data collection: general principles and impact on hemophilia care, Hematology, 2012 <https://doi.org/10.1179/102453312X13336169156537>
11. Mayo Clinic to sequence 100K participants to build genomic database for improved care and research in collaboration with Helix, <https://newsnetwork.mayoclinic.org/discussion/mayo-clinic-to-sequence-100k-participants-to-build-genomic-database-for-improved-care-and-research-in-collaboration-with-helix/>
12. Through partnership with Color, 10K NorthShore patients will get whole genome sequencing. <https://medcitynews.com/2019/01/color-northshore/>
13. Data-sharing brings hope to rare disease patients around the globe, <https://www.ga4gh.org/news/data-sharing-brings-hope-to-rare-disease-patients-around-the-globe/>



June 2020, Issue 24

Writer

김태형 테라젠바이오, 상무  
e-mail : thkim@therabio.kr

BIO ECONOMY REPORT

발행 : 2020년 06월 | 발행인 : 서정선 | 발행처 : 한국바이오협회 한국바이오경제연구센터  
13488 경기도 성남시 분당구 대왕판교로 700 (삼평동, 코리아바이오파크) C동 1층, www.koreabio.or.kr  
\* 관련 문의 : 한국바이오협회 한국바이오경제연구센터 e-mail : Koreabio1@koreabio.org



한국바이오경제연구센터  
KOREA BIO-ECONOMY RESEARCH CENTER

Innovating Data Into Strategy & Business



9 772508 682002 11  
ISSN 2508-6820