인간 유전형-표현형 데이터베이스: 목적, 과제, 그리고 기회

박 경 진

서울아산병원 진단검사의학과

E-mail: unmar21@gmail.com

요약문

유전변이와 이 변이로부터 초래되는 결과 그리고 작용 기전에 대한 정보를 탐색하기 위해 유전형-표현형 데이터베이스를 이용할 수 있다. 데이터베이스는 종류와 중점분야, 그리고 운영 모드에 따라 다양하다. 염기서열분석, 오믹스, 표현형 발굴 기술의 발전에 따라 크고 복잡한 데이터셋이 생성됨에도 불구하고, 이 데이터셋을 개별적으로 이용하기보다는 데이터베이스를 통합시키는 방향을 통해 지속적 발전이 가능하다. 단일유전자변이 데이터에서 유전자 패널, 엑솜, 전장유전체 데이터로 초점이 이동하면서 데이터베이스 디자인, 변이의 병원성(pathogenicity) 평가, 데이터 제시 방식 및 사용과 관련된 새로운 과제 및 기회가 나타나고 있다.

Key Words: Genotype-phenotype database, Genotype, Phenotype, Database, Dataset, Variant pathogenicity

본 자료는 Human genotype-phenotype databases: aims, challenges and opportunities. *Nat Rev Genet.* 2015; 16(12): Pages 702-715 의 논문을 한글로 번역, 요약한 자료입니다.

목 차

1. 서론

'유전형-표현형 데이터베이스'라는 용어는 유전형 데이터(예를 들면 DNA 염기서열, 변이, 유전형)와 표현형 데이터(관찰된 특징), 그리고 이 두 데이터의 상관성에 대한 데이터셋을 기록하고 실행시키는 시스템을 총칭한다. 의학 영역에서 이러한 데이터베이스는 인간의 유전적 데이터와 표현형에 초점이 맞추어져 있다. 유전형-표현형 데이터베이스의 최종 목적은 유전변이의 기능적 중요성 및 병원성 확립을 위한 데이터와 지식을 제공하는 것에 있다.

수많은 양성 변이로부터 질환의 원인 유전자 변이를 구별하는 것이 중요하다. 변이의 질병 연관성을 잘못 해석하게 되면, 진단 및 질환 위험도를 부정확하게 평가할 수 있기 때문이다. 몇 개의 변이만이 큰 유효크기(effect size)를 가짐에도 불구하고 (측정 오류, 숨겨진 바이어스, 또는 다중 실험 등으로 인한) 위양성 가능성이 높기 때문에, 양성 변이로부터 원인유전자 변이를 명확하게 구별해내는 것은 어려운 일이다.

차세대염기서열분석법이 발전함에 따라, 지금은 전통적인 단일유전자 분석보다는 다중유전자 패널, 전장엑솜분석, 전장유전체분석 등을 이용하여 유전 요인과 질환의 관련성을 규명하고 있다. 하지만, 유전체학 기술은 양날의 검과 같은 효과가 있다: 이 방법을 이용하여 새로운 질환유전자 및 원인 변이를 발굴할 수도 있지만, 질환 관련성에 대한 판단 오류 역시 초래할 수 있기 때문이다. 예를 들면, 희귀질환에 대해 전장엑솜분석법을 시행하면 보통 30,000 – 100,000개의 변이를 발굴하게 되는데, 이 중 몇 개의 변이만이 질환의 원인이 된다. 이러한 변이 중 약 10,000개는 염기서열의 삽입 및 결실을 일으키거나 과오돌연변이 또는 무의미돌연변이, 짜깁기 돌연변이를 통해 단백질 기능의 변화를 초래할 수 있다. 건강한 일반인구집단에서 발견되는 정상 변이나 양성으로 예측되는 변이를 제외하면, 수백 개의 질환 원인변이가 남게 된다. 실제 인간의 정상 유전체에는 대략 100개의 기능상실변이와 20개의 불활성 유전자가 있는 것으로 추정된다. 그러므로 병원성 돌연변이를 발굴하는 것은 중요한 과제로 남아 있다. 이를 해결하기 위해 유전형-표현형 데이터베이스를 최대한 이용해볼 수 있다.

본 리뷰에서는 유전형-표현형 데이터베이스의 주 목적, 개발 현황, 해결 과제 등을 다루고자 한다. 또한, 필요성에도 불구하고 그 동안 간과되었던 요소들을 규명하고자 한다. 토픽의 범위가 광범위하기 때문에, 본 리뷰에서는 윤리적·법적·사회적 문제보다는, 실제적이고 기술적인 이슈 및 기회

에 초점을 맞추고자 한다.

2. 데이터베이스의 개요

인간 유전체의 규모, 복잡도, 다양성, 표현형의 범위, 그리고 데이터 활용의 이유 등을 고려할 때, 유전형-표현형 데이터베이스를 다양한 형태로 정리할 필요가 있다. 스프레드시트 또는 텍스트 형식으로 간단하게 데이터베이스를 생성할 수도 있지만, 데이터셋의 크기 및 복잡도가 지속적으로 증가하고 있으므로, 전산 및 데이터베이스 기술의 진보에 따라 데이터베이스를 개선해야 한다.

유전형-표현형 데이터베이스는 다음과 같은 요인에 따라 다양하게 분류할 수 있다: 첫째, 목적(보건 또는 연구); 둘째, 접근방식(개방형 또는 제한형); 셋째, 시스템(보관 시스템 또는 실시간 작동 시스템); 넷째, 데이터(일차 데이터 또는 데이터 집합); 다섯째, 데이터셋 수집 방식(중앙형 또는 분산형) 등의 요인이 있다. 또한, 데이터 형태 및 출처에 따라서도 데이터베이스를 분류할 수 있다. 서로 다른 형태의 유전형-표현형 데이터베이스 간에 중복되거나 상호보완적인 부분이 있을 수 있다. 본 리뷰에서는 데이터베이스의 상대적 규모에 따른 분류를 간단하게 제시하고자 한다. 현재 사용되고 있는 데이터베이스를 목적과 특성에 따라 분류하여 표 1에 제시하였다.

2.1 소규모 데이터베이스

유전형-표현형 데이터베이스는 단일 연구, 특정 질환 연구, 특정 환자군 또는 특정 유전자 연구 결과를 이용하여 생성될 수 있다. 전통적으로 인간 유전학 영역에서 진단 전략은 주로 단일 유전자, 소수의 유전자에 집중하여 변이를 발굴하고 locus specific mutation database(LSDB)에 제출하는 방식으로 진행되었다. 특정 유전자를 전문으로 하는 커뮤니티에서 변이 해석을 용이하게 하기 위해, 이 LSDB를 활용해 왔다. LSDB에는 소수의 질환 관련 유전자의 염기서열변이 정보가 포함하고 있다. 전문가 리뷰에 의해 질환 관련 유전변이와 해당 변이의 표현형 정보에 대한 완전하고 정확한 리스트를 수집하는 것이 중요하다. 표 1에 제시한 것처럼, LSDB에는 관련된 유전자가 함께 분류되어 제시되고 있다: 예를 들면, 판코니 빈혈(16개 유전자), 골형성 부전증(16개 유전자), 근위축성 측색 경화증(116개 유전자) 또는 면역결핍증(131개 유전자) 경우와 같다. 특정 국가나 특정 인종에서 중요한 유전 질환을 대상으로 하는 데이터베이스도 있다: 예를 들면, ETHNOS 데이터베이스가 그러하다. Leiden Open Variation Database와 Universal Mutation Database 같은 소프트웨어는 LSDB 큐레이션에 널리 활용될 수 있다.

기술적 완성도, 데이터 품질, 공용 데이터베이스가 표준화되지 않고 다양하게 존재한다. 이러한 데이터베이스는 생성에 이용한 소프트웨어의 내장된 특성을 고려하기보다는, 선호도에 따라 맞추어지는 경향이 있으며, 데이터베이스 디자인과 작동 방식, 다른 시스템과의 통합, 대용량 다운로드를 위한 표준도 거의 없다. 여러 곳에서 유전형-표현형 데이터를 디지털 방식으로 수집하였음에도 불구하고 아직도 소수의 사용자에 의해서만 데이터베이스가 사용되는 경향이 있으며, 보안이 충분치 않고, 지속성 문제로 인해 장기간 통합하기도 어렵다. 잘 관리가 되지 못한 데이터셋을 단편적으로 정렬하는 방식은, 좋은 데이터 활용을 어렵게 한다: 하지만, 전용 데이터베이스(private database)에는

연구 및 진단검사의학에 도움이 될 수 있는 데이터가 있기도 하다.

2.2 대규모 데이터베이스

다양한 유전형-표현형 데이터베이스 - 일차적으로 중앙 데이터베이스('central' database)- 가 표1에 요약되어 있다. 유전형-표현형 데이터베이스의 규모가 클수록 더 광범위한 돌연변이 정보를 제공하며, 이러한 데이터베이스로는 예를 들면, Human Gene Mutation Database나 ClinVar 등이 있다. 차세대유전체분석법을 통한 전장 유전체 데이터를 다루는 대용량 데이터베이스도 있으며, 예를 들면 DECIPHER 에는 array-CGH 및 전장엑솜분석법, 표현형 연구에서 얻은 정보가 요약되어 있다. 전장유전체연관분석 및 체성 돌연변이에 대한 데이터베이스도 있으며, 이 중 일부는 다운로드도 가능하다. 예를 들면, Cancer Genomics Hub는 2 페타바이트의 다운로드 가능한 데이터가 포함되어 있다.

통합 데이터베이스('Integration' database)는 다양한 소규모 데이터베이스를 통합하여 특정 토픽에 대해 더 광범위한 정보를 제공하기도 한다. 또한, 질환 유전자 및 관련된 임상 토픽에 대한 정보를 포괄적으로 제공하기 위한 지식 베이스[예를 들면, 인간 유전학 분야에서는, Online Mendelian Inheritance in Man (OMIM)이나 Orphanet]도 있다. 이러한 유형의 데이터베이스는 수집과 기획, 표준화 작업에 상당한 수작업이 요구된다.

중앙 데이터베이스와 통합 데이터베이스는 Locus Reference Genomic sequence 표준 등과 같은 표준을 사용하며, 웹 서비스를 제공하거나 강력한 데이터 검색 기능 및 다운로드 옵션을 제공하기도 한다. 또한, 데이터베이스 간에는 서로 광범위하게 연결되기도 하며, 내용의 일부에 대해서는 다른 데이터베이스에서 재현되기도 한다. 이러한 데이터베이스를 상용화 시스템으로 제공하는 경우는 드물며, 대부분 학문연구기금을 사용하여 데이터베이스를 유지 보수한다.

표 1. 유전형-표현형 관계에 초점을 맞춘 의학 데이터베이스 모음

데이터베이스	범위와 규모	표준	데이터 도입	데이터 접근 정책	참고 문헌
유전자 변이 데이터베이스(LSDB 또는 MDB)					
ClinVar	유전자 변이와 표현형 125,520개 변이	HGVS, HPO, MeSH, OMIM, RefSeq, SO	큐레이터, 사용자	공용	63
Human Gene Mutation Database (HGMD)	유전자 변이와 표현형 163,610개 변이	HGNC, HGVS	큐레이터	상용, 공용	96
Leiden Open Variation Databases (LOVD)	유전자 변이와 표현형 86개의 LOVD 기관, 248,807명에서 발견된 3,334,104개 변이 (2,400,084개는 고유 변이)	HGVS, Mutalyzer	큐레이터	공용, 컨소시움 공개	12

Universal Mutation Database (UMD)	유전자 변이와 표현형, 40개 데이터베이스의 90,383개 변이	HGVS	큐레이터	공용, 컨소시움 공개	13
Amyotrophic Lateral Sclerosis Online genetics Database (ALSoD)	LSDB, 근위축성 측삭경화증 관련 유전자 116개, 569개 변이	HGVS	큐레이터	공용	97
CFTR2	낭포성 섬유증 LSDB, 88,000명의 환자	HGNC, HGVS	큐레이터	공용	23
Fanconi Anemia Mutation Database	LSDB, 판코니 빈혈-BRCA 경로와 관련된 유전자 16개, ~3,000개 변이	HGNC, HGVS, LRG	큐레이터	공용	98
Osteogenesis Imperfecta Variant Database	LSDB, 불완전 골형성증과 관련된 유전자 16개, ~1,500개 변이	HGNC, HGVS, LRG	큐레이터	공용	99
IDbases	LSDB, 면역결핍과 관련된 유전자 131개, 7,292명의 환자 데이터	HGNC, HGVS, VariO	큐레이터	공용	100
MITOMAP	LSDB, 미토콘드리아 변이, 1,746개의 변이		큐레이터	공용	101
FINDbase	나라와 인종에 따른 변이 빈도에 대한 통합 정보, ~100 데이터베이스	HGNC, HGVS	큐레이터	공용	102
Array-CGH, 전장엑솜분석, 전장유전체분석(희귀질환)					
DECIPHER	유전자 변이와 표현형 진단과 발굴, 42,815건	HGNC, HGVS, HPO	사용자	공용, 컨소시움 공개, 연구자 공개, 유사환자 찾기,	38, 103
PhenomeCentral	유전자 변이와 표현형, Genomic "matchmaking" 600건	HPO, VCF	사용자	유사환자 찾기	65
PhenoDB	유전자 변이와 표현형 진단과 발굴, 3,300건	EoM, HPO, OMIM, PhenoDB	사용자	유사환자 찾기	104
GeneMatcher	유전자 찾기 (발굴) 668개 유전자	HGNC, Ensembl, EntrezGene, OMIM	사용자	유사환자 찾기	105
멘델성 질환과 다른 희귀질환 지식 베이스					
Online Mendelian Inheritance in Man (OMIM)	지식 베이스, 22,644 유전자 또는 질환	HGNC, HPO, ICD, OMIM, PhenoDB,	큐레이터	학술기관 공개	106

		SNoMED, UMLS			
Orphanet	지식 베이스 5,833 질환, 다른 희귀질환에 대한 방대한 데이터	HGNC, ICD, MedDRA, MeSH, OMIM, UMLS, Uniprot	큐레이터	학술기관 공개	73
Monarch Initiative Konwledge base	인간 및 모델동물 유전과 표현형, 36K 질환, 33K 표현형, 500K 유전형, 30K 유전형, 2M 표현형 연관성, >100 종	HPO, MPO	큐레이터	공용	107
암 유전체학과 변이					
CancerGenomics Hub	유전자 변이와 표현형 저장소, 82,140개 파일 (1870 Tb)	Sequence Read Archive Metadata XML	NCI 프로젝트, 큐레이터	공용, 접근 제한	108
Catalogue Of Somatic Mutations In Cancer (COSMIC)	변이와 유전형, 표현형 데이터 2,139,424 고유 변이	HGNC, CCDS	큐레이터	공용	109
DriverDB	변이, 유전형/표현형 데이터 6,079 데이터셋	HGNC	큐레이터	공용	110
전장 유전체연관분석 및 기타 연구를 위한 유전형-표현형 정보					
Database of Genotypes and Phenotypes (dbGAP)	유전형과 표현형 데이터 508개 연구	dbGAP, XML	큐레이터	공용, 접근제한	43
European Variation Archive (EVA)	여러 종의 유전변이 ~40개 연구, 35개 종, 150,000개 검체에서 발굴된 ~400 M 고유 유전자형	VCF, dbSNP	사용자, 큐레이터	공용	
European Genome–Phenome Archive (EGA)	유전형과 표현형 데이터 1,555개 데이터셋	VCF, FASTQ, BAM, EFO	사용자	공용, 접근제한	42
GWAS Catalog	유전형과 표현형 데이터 18,697개 연관성	dbSNP, HGNC	큐레이터	공용	111
GWAS Central	유전형과 표현형 데이터 >75 M개 연관성	dbSNP, HGNC, HPO, MeSH	큐레이터	공용, 접근제한	47
GWASdb	유전형과 표현형 데이터 272,918개 연관성	dbSNP, DO, HPO	큐레이터	공용	112
Human Genome Variation Database	유전형과 표현형 데이터 6개의 통합 데이터베이스의 ~100개 데이터셋	HGNC, HGVS, dbSNP	사용자	공용, 접근제한	113

약물유전체학					
PharmacoGenomics Database (PharmGKB)	약물유전학 지식 출처, 변이, 경로, 약물용량, 임상정보, 약물 상표 등에 대한 광범위한 데이터	dbSNP, HGNC, MeSH, SNoMED, UMLS	큐레이터	학술기관 공개	114

약어: CCDS, consensus CDS; CGH, comparative genome hybridization; dbSNP, Database of Single Nucleotide Polymorphism; DO, Disease Ontology; EFO, Experimental Factor; Ontology; EoM, Elements of Morphology; GWAS, genome-wide association study; HGNC, HUGO Gene Nomenclature Committee; HGVS, Human Genome Variation Society; HPO, Human Phenotype Ontology; ICD, International Classification of Diseases; LRG, Locus Reference Genomic; LSDB, Locus-specific database; MDB, variation (mutation) database; MedDRA, Medical Dictionary for Regulatory Activities; MeSH, Medical Subject Headings; MPO, Mammalian Phenotype Ontology; NCI, National Cancer Institute; NEMDB, national and ethnic variation (mutation) databases; NIH, National Institutes of Health; OMIM, Online Mendelian Inheritance in Man; SNoMED, Systematized Nomenclature of Medical Terms; SO, Sequence Ontology; UMLS, Unified Medical Language System; VariO, Variation Ontology; VCF, Variant Call Format; XML, Extensible Markup Language

3. 해결 과제 및 공통 목적

3.1 데이터셋 증가

단일 유전자 검사에서 차세대유전체분석 검사로 전환됨에 따라, 데이터 양이 급격하게 증가하게 되었고, 유전형-표현형 데이터베이스의 용량과 범위도 증가되었다. 이 같은 현상에서 두 가지의 결과가 초래되었다. 첫째, 과거에는 소수의 유전자에만 집중하던 분석가들이 이제는 다수의 유전자에 집중해야 한다는 것이다. 차세대유전체분석을 이용하게 되면서, 더 다양한 데이터 출처에서 더 많은 근거를 고려하여, 진단적 결론을 얻어야 한다. 이에 따라 전보다 공용 유전형-표현형 데이터베이스를 사용할 필요가 증가되었다. 둘째, 분석 오류의 위험을 최소화하기 위해 데이터 출처의 품질이 보장되어야 한다. 현재 사용 중인 데이터베이스에는 상당 부분 오류가 포함되어 있다. 예를 들면, 어떤 데이터베이스에 등록된 유전 변이의 27%까지, 원인 유전자로 잘못 분류되거나 불완전하게 분류되었다는 보고가 있다.

3.2 변이의 병원성 평가

잘 알려진 질환 유전자에서조차도, 변이의 병원성(pathogenicity)을 평가하는 것이 명확하지 않은 경우가 있다. 예를 들면, *TRIM63 (MURF1* 으로도 알려짐)의 경우이다. 한 연구에서는 302명의 환자에서 2개의 과오돌연변이와 1개의 결실이 발견된 반면, 299명의 대조군에서는 발견되지 않았다는 사실에 근거하여 *TRIM63*을 확장성 심근병증의 원인 유전자로 보고하였다. 하지만 또 다른 연구에 따르면, *TRIM63*에서 무의미 돌연변이가 있음에도 불구하고 심근병증의 증후가 없는 경우도 있었다. 또한, 마우스에서 *Murf1*의 결실로도 심장 병리 소견을 초래하지 않았다는 보고가 있으며, *Murf1*

과 *Murf2*를 동시에 제거(Knock-out)함으로써 심근병을 일으켰다는 보고도 있다. 아직까지 *TRIM63*이 멘델 유전병의 원인 유전자인지 또는 다른 유전자와 상호작용을 통해 심근병증에 기여하는지가 명확하지 않다.

유전변이의 병원성을 판정하는 작업은 간단하지 않다. 왜냐하면, 인과관계를 확립되기 위해서는 다양한 증거에 기반해야 하며 어떤 증거에 가중치를 둘 것인지는 다소 주관적일 수 있기 때문이다. 흔히 사용되는 방법은 몇 개의 범주(definitely pathogenic, probably pathogenic, uncertain, probably not pathogenic or of little clinical significance, not pathogenic or of no clinical significance)로 나누어 변이를 분류하는 것이다. 하지만, 이러한 분류에서 사용되는 '병원성'의 의미가 명확하지 않다. 변이의 병원성의 의미가 보편적이고 일관되게 사용될 뿐만 아니라, 병원성을 판정할 때 투과도 및 표현도 등의 요소까지 고려하여 정확한 임상적 표현형이 뒷받침되는 것이 이상적이다. 하지만, 이러한 작업은 좋은 품질의 관찰 데이터를 광범위하게 요구되므로 적용되기 쉽지 않다. 예외적으로 CFTR2 데이터베이스에는 *CFTR* 돌연변이 데이터와 더불어, 88,000명 이상의 낭포성 섬유증 환자에서 땀의 염소농도, 폐 기능, 췌장 기능, *Pseudomonas* 감염률에 대한 데이터가 포함되어 있기는 하다.

4. 현황 및 동향

4.1 데이터베이스 유형 및 배열

유전형-표현형 데이터베이스의 개수, 유형, 배열에 대한 현황 및 동향은 그림 1 과 같다. 인식 필요성, 제작자에 대한 보상, 데이터 공유에 대한 제한, 데이터 관리 전문가의 필요성 등과 같은 요인들은 연방 데이터베이스(federated database)의 필요성으로 귀결된다. 몇 개의 중앙 보관소를 생성하는 것만이 비용효과적이며, 실제적이라는 의견도 있다. 실제, 일차 데이터베이스(primary database)에 비해 중앙 데이터베이스(central database)가 데이터 공탁자에게 더 매력적인 접근법일 수 있다. 왜냐하면, 중앙 데이터베이스의 경우, 데이터 보관 서비스, 데이터 접근요청 관리, 보관, 심도 있는 데이터 분석, 출처 데이터와의 연결 등과 같은 인센티브를 제공하기 때문이다. 최근 Collaborative Cancer Cloud 같은 프로젝트에서는 데이터를 모두 클라우드로 이동시키므로, 데이터베이스의 실존이나 출처의 시작과 끝에 대한 논란이 제기되기도 한다.

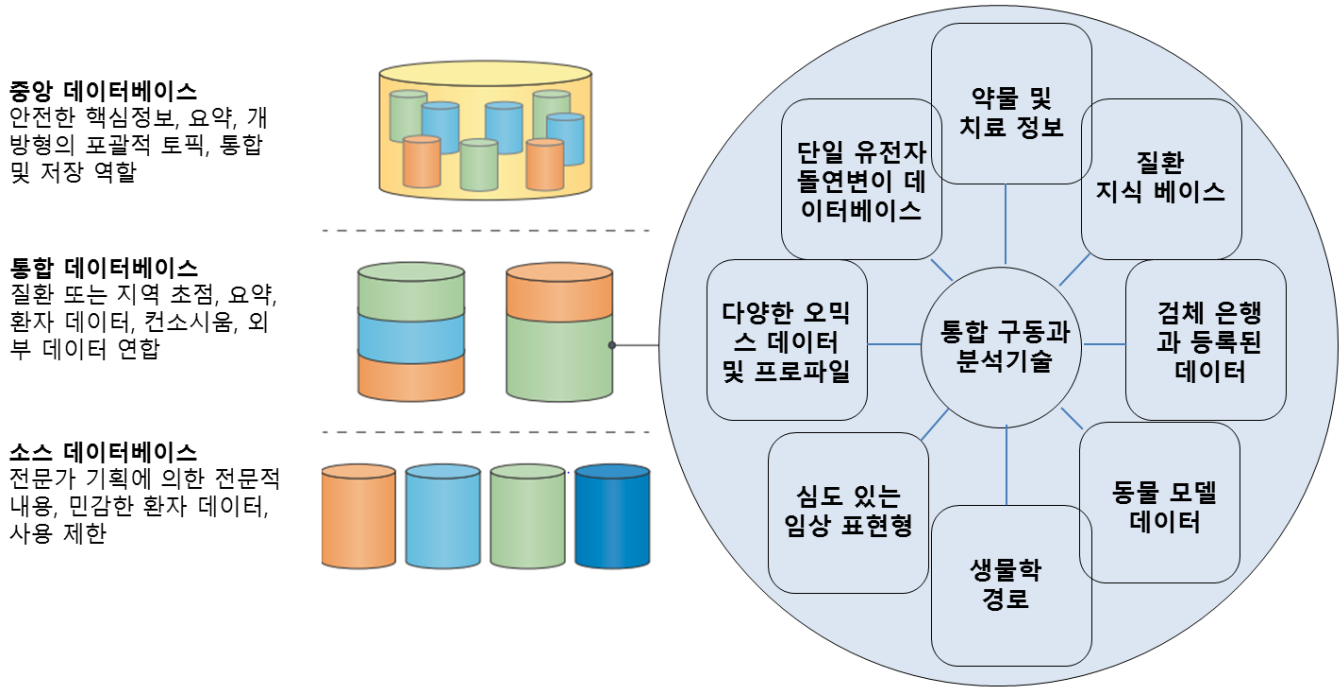

그림 1. 유전형-표현형 데이터베이스의 개요.

데이터베이스 유형에 따라, 소스 데이터베이스, 통합 데이터베이스, 중앙 데이터베이스로 구별함. 소스 데이터베이스는 개인 데이터를 포함하는 반면, 통합 데이터베이스는 다른 소스에서 나온 내용을 종합하여 특정 질환 및 프로젝트에 기반한 정보로 구성됨. 가장 높은 수준의 중앙 데이터베이스는 포괄적이고 보편적인 서비스를 제공함.

4.2 데이터 연결과 보고

앞서 기술한 바와 같이 중앙 데이터베이스의 장점에도 불구하고, 소규모 데이터베이스나 비-공용 데이터베이스들은 서로 연결되어 있지 않다. 또한, 유전형-표현형 상관관계를 보고할 때에도 데이터베이스로의 전송이 없어, 큐레이터가 직접 이 정보를 추출하고 통합하는 작업에 시간이 소요된다. 이러한 연결 부재는 데이터베이스의 포괄적 활용을 어렵게 만든다. 기금, 경쟁, 인센티브, 법적 제한 조치, 동의 등의 문제도 역시 데이터베이스 연결 작업을 어렵게 만든다. 저널이나 기금지원기관에서는 데이터 제출 파이프라인을 구축하는 작업에 투자할 수 있을 뿐만 아니라, 연구자들에게 이를 사용할 수 있도록 권장할 수도 있다. 실제 Human Mutation 같은 저널은 논문 투고 시에 데이터를 데이터베이스로 제출하도록 하고 있으며, 기금지원기관에서도 이와 관련한 필요조건(예를 들면, 미국 국립보건원의 데이터 공유 정책)을 만들고 있다.

유전형-표현형 데이터를 '언제 어떻게' 온라인 데이터베이스로 제출해야 하는지는 아직 명확하지 않다. 데이터 생성의 방법과 출처가 다양하다는 점을 고려하면, 데이터 사용이나 데이터베이스 배열 및 등록, 그리고 검체 은행에 대해 윤리적·법적 조치가 있을 수밖에 없다. 그러므로 다음과 같은 요소가 고려되어야 한다: 1) 데이터를 데이터베이스에 제출하는데 소요되는 시간과 능력 2) 데이터를 제출함으로써 얻는 보상과 위험의 균형 3) 사용 가능한 데이터베이스의 지속성 및 편리성 4) 데이터의 품질과 유용성 등이다. 이러한 고려사항을 모두 적용하려면 시간이 소요되므로, 우선은 차선을 생각해볼 수 있다. 예를 들면, Genetic Alliance 의 PEER 플랫폼, PatientsLikeMe, GenomeConnect

등과 같은 접근법으로, 환자의 유전형-표현형 데이터를 직접 생성하고 기록하는 것이다. 이러한 접근법은 환자에게 권한을 부여하고 환자 본인을 건강 관리의 중심에 두는 동향과도 잘 부합된다.

4.3 표현형 정보 및 변이의 병원성 판단 근거 향상

표 1에 기술된 데이터베이스에는 장점과 단점이 있다. 표준 데이터셋(예를 들어, 1000 genome과 Exome Variant Server) 생성에 사용한 차세대유전체분석기술과 대규모 질환 유전학 연구(예를 들면 Personal Genome Project, International Cancer Genome Consortium, The Cancer Genome Atlas 등) 덕분에, 유전형 정보는 대용량으로 빠르게 확장되고 있다. 반면, 표현형 정보는 규모 및 표준화 측면에서 상당히 뒤쳐지고 있다. 과거의 유전형-표현형 데이터베이스에는 질환명 정보만 있을 뿐, 표현형 정보가 제한적으로 있었다. 유전체 분석의 핵심은 전장 유전체 변이 데이터와 표현형 데이터 임상 분석이라고 할 수 있다. 예를 들면, 치료 가능한 유전 변이 중 보고된 변이의 10% 미만은 문헌 고찰을 통해 무증상 성인에서 '우연한 발견(incidental finding)'으로 보고될 수 있다. 전장엑솜분석법 또는 전장유전체분석법을 통해 발굴한 변이를 진단용으로 보고할 때, 변이의 병원성에 대한 기존 보고를 그대로 따르는 것은 주의해야 한다. 표현형 정보에 더 많은 주의를 기울인다면 변이의 임상적 해석의 정확성을 더 높일 수도 있다. 최근 CARE4RARE, DECIPHER, Genomes Management Application, PheWAS Catalog, Kaiser Permanente Research Program on Genes, Environment and Health 등의 프로젝트에서는 유전형 정보 증가에 상응하여 표현형 정보 수집의 필요성을 강조하고 있다. 이러한 움직임은 제한적 접근만을 허용하는 플랫폼에서도 마찬가지로 요구된다(예를 들면, European Genome-Phenome Archive, database of Genotypes and Phenotypes). 유전형과 표현형 데이터의 양과 질에 대한 격차는, 양질의 표현형 데이터를 생성하기 위해 향후 극복해야 할 과제의 정도를 반영한다.

유전형 데이터, 표현형 데이터 이외에도, 변이의 병원성 예측 데이터 및 방법, 해석 등이 데이터베이스에 포함되고 있다. 자료의 다양성 때문에, 이를 사용하기 위해서는 i2b2(Informatics for Integrating Biology and the Bedside)와 Observ-OM 같은 데이터베이스 기술이 이용되어야 한다.

4.4 여러 가지 데이터 제공 방식

단일 형태의 데이터가 아니라 연결과 중복이 있는 데이터의 경우, 데이터베이스를 생성하는 일이 간단하지 않다. 대부분 복잡한 데이터를 다양한 형태의 프로젝트에서 추출하기 때문에, 자체의 한계와 불확실성의 이슈가 있을 수밖에 없다. 따라서 서로 다른 전략들을 이용하여 주의 깊게 데이터를 통합하는 것이 필요하다.

데이터 이용 전략을 그림 2에 표시하였다. 한가지 방법은 데이터를 다른 형식으로 전환함으로써, 개인 식별의 위험을 제거하고 제시하는 것이다. 예를 들어 GWAS Central 에서처럼 메타데이터를 생성하거나 그래픽 형식으로 제시하거나, 또는 패턴을 사용할 수도 있다. 이러한 접근법은 통합 데이터베이스와 질환 특이적 지식 포탈의 핵심이며(그림 1), 이를 사용하여 LSDB, 등록, 검체 은행, 연구, 진단 정보 등을 통합할 수 있다.

또 다른 접근법은 흩어진 데이터를 수집하여 제공하는 방법이다. 데이터를 공유하지 않고,

흩어져 있는 데이터를 모아 분석하기 위해 예를 들면 DataSHIELD 와 같은 기술적 방법이 필요할 수 있다. 데이터를 직접 공유한다면, 보안을 강화하기 위해 여러 측면에서 데이터를 암호화할 수 있다. 구글 같은 일반 인터넷 검색 엔진의 사용에 대한 대안으로, 여러 곳에서 제한적 데이터를 수집하는 프로젝트가 있다. 이러한 노력은 SNPedia, MalaCards, WAVe, Café Variome Central, Kaviar, the European Variation Archive, 그리고 the Exome Aggregation Consortium 등에서 시도되고 있다.

그림 2. 데이터 제공 방식 (http://www.nature.com/nrg/journal/v16/n12/fig_tab/nrg3932_F2.html)

4.5 데이터 발굴에 대한 초점

지속적이고 포괄적인 유전형-표현형 '데이터 발굴 레이어(data discovery layer)' 가 등장하여, 관련 있는 데이터의 위치가 다양한 포털에서 확인되기 시작했다. 합법화된 조항에 기초한 동의가 이루어진다면 이 과정은 자동화되어 메타데이터로 제공된 후 사용자와 제공자의 컴퓨터 시스템간 상호작용을 촉진할 수도 있다. 탐색에 이어 데이터를 반환해주는 대신, 시스템에서 엄청난 양의 데이터셋을 조사하여 앞서 언급한 여러 가지 데이터 제공 방식을 통해 유용한 신호 패턴을 찾아주는 것이다.

다양한 데이터 발굴 방식이 있다(그림 3). 데이터 그 자체보다는 예를 들면, 유전형-표현형 관계를 아이콘으로 표시함으로써 탐색 과정을 대체하거나 카탈로그 방식으로 안전한 출처의 데이터를 탐색하거나, 탐색 파라미터에 따라 기 생성된 메타데이터 등을 발굴하는 것이다. 뿐만 아니라, 요약 또는 그래픽 형태의 데이터를 제공한다거나 데이터 소유자에게 연락할 수 있도록 다양한 옵션을 제공할 수도 있다.

유전형-표현형 데이터 발굴 시스템은 검체 은행에서 가장 명확하게 확립되어 있다. 카탈로그 형식으로 점차 통합된 형태로 제시하며, 일차 데이터를 어디에서 어떻게 요청할지에 대해 상세사항까지 안내해준다. 최근 몇 가지 프로젝트에 의해 일차 데이터 발굴이 촉진되었다: 예를 들면, International Rare Diseases Research Consortium (IRDiRC)와 Global Alliance for Genomics and Health (GA4GH)에 의해 주도되는 'MatchMaker' 교환을 통해 유전형과 표현형을 가진 희귀질환 환자를 발굴할 수 있게 되었다. 초반에는 공용 데이터를 이용하였으나, 점차 인터넷으로는 검색될 수 없는 민감한 정보까지 이용하게 되었다(GeneYenta를 이용하여 유사한 표현형을 매칭하여 유사한 희귀질환을 가진 환자를 검색할 수 있다; Café Variome을 통해 비-공용으로 사용되는 진단 및 연구 네트워크에서 변이와 표현형 데이터를 발굴할 수 있다; Beacon 프로젝트를 이용하여 여러 출처를 탐색하고 특정 유전변이 유무를 검색할 수 있다). 데이터 발굴이라는 개념은 검체 은행을 넘어 공중보건 및 역학 영역까지도 확장되고 있다.

데이터 발굴의 최종 목표는 데이터 생성자로 하여금 데이터가 언제 어떻게 이용될 것인지는 통제하면서도 최대한 정보를 공개할 수 있도록 하는 것이다. 이러한 옵션은, (협력, 기금, 인지 등을 통해) 데이터가 적절하게 이용되고 있는지에 대해 책임이 있는 데이터 관리인에게 특히 중요하다. 수 많은 데이터가 이용되지 않고 있음을 고려할 때, 데이터 발굴 기술의 개발은 전반적으로 데이터를 더 가시적으로 드러냄으로써, 데이터의 사용을 증가시킬 수 있다. 그렇지만 데이터 관리인이 데

이터베이스에서 데이터 공유 옵션보다는 데이터 발굴 인터페이스만을 우선적으로 고려한다면, 이러한 접근법은 데이터 공유를 제한시킬 수도 있다.

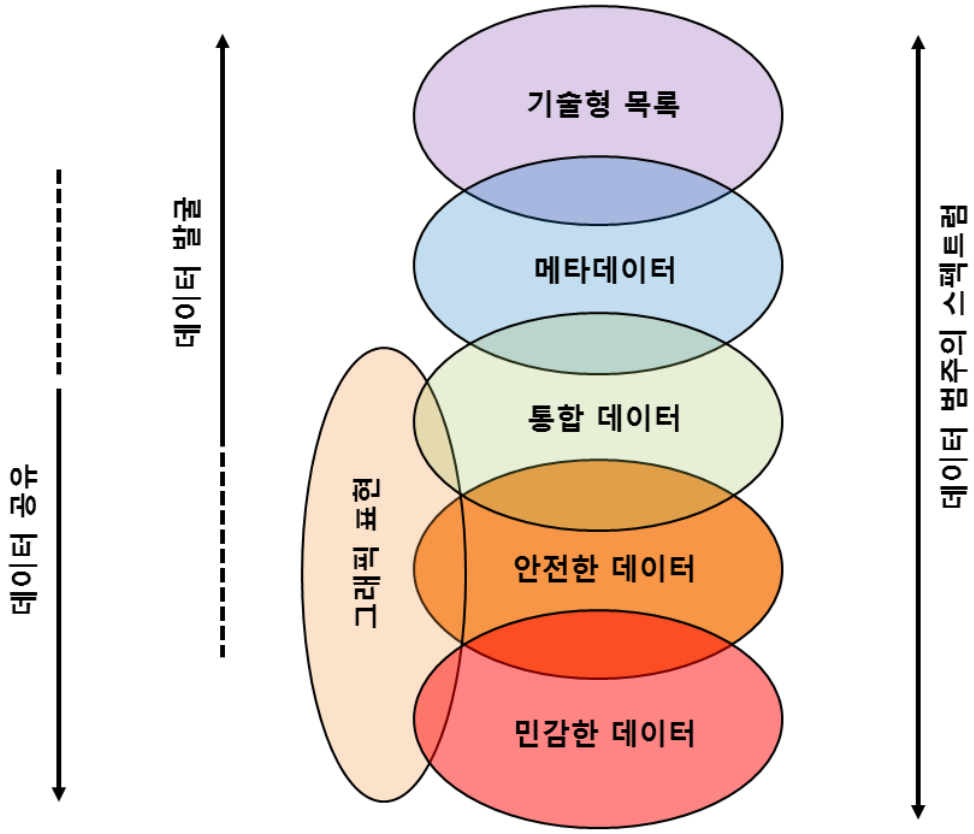

그림 3. 데이터 공유 및 데이터 발굴.

데이터베이스에는 일련의 정보가 포함됨. 민감한 데이터(빨간색)나 확산되기에 안전한 데이터(오렌지색) 등 일차 데이터는 다른 형태로 전환되기 쉬움. 요약 통계(통합 데이터, 녹색)나 그래픽으로 표현된 데이터(분홍색)는 편리하고 공유하기에 안전하며 지식을 더 잘 표현하는 데이터임. 이 범주의 데이터를 전달하는 것이 일반적인 데이터 공유를 의미함. 민감한 데이터를 예외로 하면, 데이터 발굴 접근법도 같은 범주의 데이터를 이용하는 것임. 그러한 접근법은 데이터 발굴의 범주를 넓혀 통합 데이터, 특히 메타데이터(하늘색)와 기술형 목록(보라색)에 더 중점을 두는 것임. 데이터 발굴과 공유 개념 사이에 겹침을 고려할 때, 데이터 유형, 인터페이스, 기능과 절차가 고안되어 정보의 확산에 보완적으로 이용되어야 함.

5. 커뮤니티의 최근 노력

1990년대에는 유전형-표현형 데이터베이스를 생성하거나 다루는 작업이 상당히 제한적이었으나, 2000년도에 들어 차세대유전체분석법이 도입됨으로써 데이터 관리에 많은 자원이 배정되어 현재 데이터베이스 영역은 꽃을 피우고 있다. 대규모 연구를 통해 유전형과 표현형 데이터를 생성하고 데이터를 수집하고 국제적 협력과 표준화 작업이 진행되고 있다. 희귀질환이라 할지라도 데이터베이스 생성 프로젝트가 활발하게 진행되고 있다(예를 들면, IRDiRC와 GA4GH). 더구나, ClinGen 네트워크도 ClinVar에 데이터를 기탁함으로써 이러한 이슈를 다루고 있으며, Canadian Forge/Care4Rare 프로그램과 DECIPHER 프로젝트 역시 진단에 도움이 되고자 환자 데이터를 수집하고 있다. 동시에 다양

한 컨소시움 전문가들도 희귀질환의 원인 변이를 발굴하기 위해 프로젝트를 운영하고 있다; 이러한 프로젝트는 European RD-Connect project 로부터 전산 측면에서 도움을 받고 있다.

데이터와 정보를 공유하지 않았던 과거에 비해, 최근에는 데이터 공유 방향으로 패러다임이 변하고 있다. 희귀질환 이외에도, 예를 들면 ICGC와 GERA study 등을 포함한 많은 컨소시움 프로그램들이 구성되어 대용량의 데이터를 공유하고 있다. 프로그램 참여자의 이해관계가 상당히 섞여 있음을 고려할 때, 한 명의 이해관계자를 규정하는 일에는 상당히 어려움이 따른다. 또한 프로젝트에 참여한 사람이 누구인지, 어떤 방법과 자원을 사용했는지에 대한 기록을 남기는 것도 쉽지 않다. 그러므로 다수의 자원 발굴 목록 예를 들면, IRDiRC, RD-Connect, Orphanet, GA4GH 같은 색인을 개발할 필요가 있다.

6. 미래 전망

유전형-표현형 연구는 관계자들간의 역동적 상호작용, 해결법의 확산, 혁신 등을 통해 빠르게 성장하고 있다. 그럼에도 불구하고, 현 이슈를 해결하기 위한 노력으로 이어지지 않고 있으며, 실제 데이터베이스 이용 없이, 상향식으로 진행되고 있다. 커뮤니티에 의해 보고된 몇 가지 미충족 필요를 그림 4에 제시하였다.

(1) 메타데이터: 유전형과 표현형 데이터의 모든 측면(예를 들면, 출처, 관리, 소유, 동의, 목적 등)에 대해 메타데이터의 깊이와 유형에 대해 규정해두는 것이 좋다. 데이터를 맥락화시키면 상당히 가치있는 정보가 되기 때문이다. 이를 위해 메타데이터의 표준화된 구조, 필요한 영역의 최소한의 셋, 메타데이터 생성과 공유 시스템을 필요로 한다.
(2) 데이터 형태 및 주석 달기: 이질적 데이터를 교환할 때, 데이터 전환 및 정확한 '주석 달기' 방법이 필수적이다.
(3) 데이터 발굴: 데이터 발굴에 초점을 둠으로써 데이터 공유를 촉진하기 위한 광범위한 노력이 증가되었다. 데이터 발굴이 강조됨으로써 포괄적·연합형의 발굴 '에코시스템'이 출현하게 되었다. 이 에코시스템은 데이터 공유를 위한 기반으로 생성되어야 한다.
(4) 동의: 형식의 다양성과 기타 실제적 이슈 등으로 인해 데이터 공유는 쉽지 않다. 표준화된 동의 문구의 핵심은 통합 연구의 여러 출처에서 검체 또는 데이터를 최대한 수집할 수 있도록 하는 것이다.
(5) 데이터 분석 및 보고: 데이터를 분석과 정도 관리는 진실성, 벤치마킹, 포괄성, 큐레이션 과정, 임상정보, 기존에 보고된 변이의 포함 여부 등과 같은 측면과 마찬가지로 필수적이다.
(6) 식별자: 환자 및 개인 기록, 데이터베이스에 대한 보편적 식별자가 필요하다. 이는 출판에서의 '디지털 객체 식별자(Digital object identifier)' 시스템 또는 '연구자 및 저자 고유 식별 번호(Open Researcher and Contributor ID)' 시스템과 유사하다.
(7) 응용 프로그램 인터페이스(application programming interface)와 온톨로지: 표준 응용 프로그램 인터페이스는 데이터베이스간 효율적인 연결과 상호작용에 필요하다. 희귀질환 커뮤니티에서는 Human Phenotype Ontology 같은 온톨로지가 널리 사용되고는 있으며, 종양이나 다빈도 복합질환에

서도 이와 같은 자원이 개발될 필요가 있다.

(8) 지속 가능성: 아무리 정교한 데이터베이스가 생성된다고 해도, 연구기금이 만료되면 데이터베이스 자원을 계속 유지하는 것은 어려운 일이다. 기금지원기관으로부터 장기간 기금을 받을 수 있을지, 상업화할지 결정해야 한다.

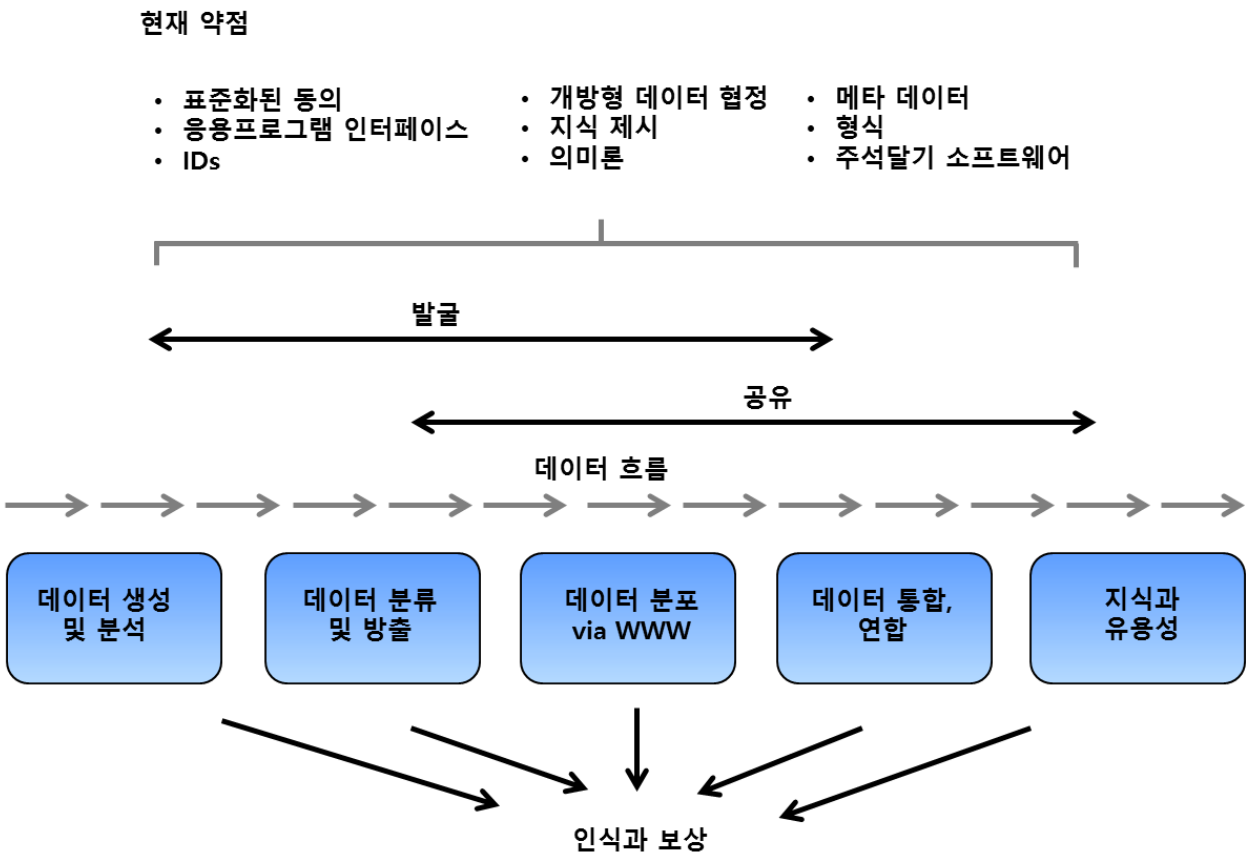

그림 4. 유전형-표현형 데이터베이스에서 데이터 처리.

유전형-표현형 데이터베이스의 개수, 기술적 정밀도, 연결성 등이 지속적으로 발전하면서 데이터 관리의 효율적인 방법이 제안되고 있음. 데이터 생성과 분석이 더 빠르고 정교해짐에 따라, 소스 데이터베이스 내에 정보를 큐레이션하는 효율적 방법이 요구됨. 이 데이터와 결과를 인터넷에서 쉽게 접근해서 사용할 필요가 있음. 이렇게 이용함으로써, 종종 데이터베이스 내에서 다른 내용과 통합되고 재조직되기도 함. 이 과정을 통해 더 나은 과학적 지식을 생성할 수 있으며, 이는 임상연구 및 진단과 관련된 중앙 데이터베이스에서 가장 잘 조직될 수 있음. 복잡한 데이터 흐름을 만드는 일은 여러 전산 기술자, 정보 큐레이터, 데이터 과학자 등의 노력을 요구하므로, 기여도는 지속적으로 기록되고 평가되고 보상되어야 함. 데이터 관리의 다양한 측면을 통해 데이터 발굴 및 공유를 최대화할 수는 있으나, 이것으로 충분하지는 않음.

7. 결론

불과 몇 년 전만해도 데이터셋은 간단하고 전산 시스템은 거의 이용되지 않았으나, 미래에는 전 인구의 유전체와 표현형 데이터를 시스템에 보관하고 유용하게 이용할 수 있을 것으로 전망된다. 이 데이터는 오믹스 결과, 환경적 요인, 경로 데이터 등의 추가적 정보와 통합되어 포괄적 데이터베이스의 필요성을 증가시킬 것이다.

본 리뷰에서 언급한 몇 가지 사항들은 실제 의학 발전에 충분하지는 않지만, 필요한 것들이다. 데이터를 잘 이용하기 위해서는, 데이터에서 임상적 지식을 추출하여 통합하고 공유할 수 있도록 알고리즘을 개발하는 것이 중요하다. 차세대 유전형-표현형 데이터베이스에는 고도의 통합 데이터와 그래픽, 그리고 인과관계에 대한 예측적 모델을 규명하는 알고리즘이 포함될 가능성이 높다. 신뢰성 있는 진단을 하기 위해, 기존 데이터 및 지식과 더불어, 이러한 데이터베이스를 이용하여 환자 및 연구 대상의 임상적·분자적 소견을 통합할 수 있을 것으로 기대된다.

상자 1 온톨로지(ontology)와 명명법

유전형-표현형 데이터를 관리하고 공유할 때, 정확한 방법으로 데이터를 구성하는 것이 중요하다. 이와 같은 이유로 온톨로지와 명명법 표준을 이용하도록 권고하고 있으며, 염기변이 명명을 위해 Human Genome Variation Society (HGVS) 표준과 더불어 Mutalyzer와 같은 방법을 이용하여 프로그램에 의한 정도관리법을 사용하고 있다. 이 외에도 차세대유전체분석법이 대두됨에 따라 파일 형식에 대한 표준으로 FASTQ, SAM/BAM, VCF 등과 같은 파일이 사용되고 있으며, 유전변이 보고에 대해서는 Locus Reference Genomic 표준 서열이 사용되고 있다. 메타데이터와 변이를 보고하기 위한(예를 들면 VarioML) 표준과 변이의 기전과 영향(예를 들면 VariO)에 대한 표준도 개발되었다

표현형 데이터의 경우 기술하고 수집하는 작업은 온톨로지 표준의 사용에 따라 다르다. 심도 깊은 표현형 분석을 시행하려면, 개별 표현 형질 이상에 대한 정확하고 포괄적인 분석이 요구된다. 온톨로지는 질환과 표현형에 대한 표준 용어를 제공할 뿐만 아니라, 표현형 분석을 위한 프로그램도 제공한다. 예를 들면, 생어 연구소의 DECIPHER, Deciphering Developmental Disorders 프로젝트, UK 100,000 Genome 프로젝트, NIH의 Undiagnosed Disease Program 등을 포함하여 유전체 검사법으로 검사한 환자의 표현형을 기록하고 분석하기 위해 Human Phenotype Ontology(HPO)를 사용하였다. 뿐만 아니라, 엑솜의 표현형 분석에서도 HPO를 사용하기도 한다. 종양과 복합질환에 대한 여러 개별 프로젝트에서 표현형 분석을 채택하고 있음에도 불구하고, 임상 연구의 전반적 범주에 개별 환자를 포함하는 것은 상당히 흔하다. 예를 들면, 병기는 종양의 범위와 전이의 유무 및 분포 등에 기반하여 종양 환자를 몇 가지 범주로 분류하기 위해 흔히 사용된다. 이런 접근법은 임상진료에 유용성이 높음에도 불구하고, 비슷한 병기를 가진 환자들이 종종 다른 예후를 보이고 있어, 질환의 스펙트럼이 현행 병기 시스템으로 검출되기 어렵다는 것을 시사한다. 이런 사항은 복합질환에서도 적용된다. 예를 들면, 현재의 정신 질환 진단 분류는 다양한 원인에서 기인하는 다양한 표현형을 보이는 환자들을 같은 집단으로 분류하고 있다. 정밀의학에 기반하여 질환을 분류하고 관리하기 위해서는, 표현형을 정확하게 분석하고 유전변이의 임상적 결과를 제대로 해석하는 것이 필요한데, 이를 위해 표현형 온톨로지를 포함하여 컴퓨터 자원을 개발하는 것에 상당한 노력이 요청된다.

박경진(2016). 인간 유전형-표현형 데이터베이스: 목적, 과제, 그리고 기회. BRIC View 2016-R16.
Available from http://www.ibric.org/myboard/read.php?Board=report&id=2575 (Sep 12, 2016)

Email: member@ibric.org

※ 본 콘텐츠는 invitrogen applied biosystems 의 후원으로 작성되었습니다.