COMPUTATIONAL TOOLS

Human genotype–phenotype databases: aims, challenges and opportunities

Anthony J. Brookes[1,2] *and Peter N. Robinson*[3–5]

Abstract | Genotype–phenotype databases provide information about genetic variation, its consequences and its mechanisms of action for research and health care purposes. Existing databases vary greatly in type, areas of focus and modes of operation. Despite ever larger and more intricate datasets — made possible by advances in DNA sequencing, omics methods and phenotyping technologies — steady progress is being made towards integrating these databases rather than using them as separate entities. The consequential shift in focus from single-gene variants towards large gene panels, exomes, whole genomes and myriad observable characteristics creates new challenges and opportunities in database design, interpretation of variant pathogenicity and modes of data representation and use.

The term ‘genotype–phenotype database’ covers a wide range of online and institutional implementations of systems used for recording and making available datasets that include genetic data (for example, DNA sequences, variants and genotypes), phenotype data (observable characteristics of an individual) and correlations between the two. In biomedicine, such databases are principally focused on human genetic data and the resulting normal or disease phenotypes. Genotype–phenotype databases have one fundamental goal: to provide access to sufficient data and knowledge to enable the functional and pathogenic significance of genetic variants to be reliably established and documented[1].

It is critical to distinguish disease-causing alleles from the abundance of neutral variants that co-occur in normal and disease-affected individuals. Incorrectly assigning pathogenicity to variants can lead to inaccurate genetic diagnoses or disease risk assessments in other individuals harbouring these variants. Achieving this distinction is challenging given that relatively few alleles have sufficiently large effect sizes to unambiguously stand out from background noise (that is, false positives due to measurement errors, hidden biases or multiple testing).

With the advent of next-generation sequencing (NGS), traditional single-gene analyses using Sanger sequencing are increasingly being superseded in both research and diagnostic settings as relationships between an individual’s genetic make-up (genotype) and disease (phenotype) can now be explored using gene panels[2], whole-exome sequencing (WES)[3], clinically focused WES[4] and even whole-genome sequencing[5]. However, the power of genomics technologies can be a double-edged sword: although high-throughput methods help find novel disease genes and pathogenic variants, they also generate many misleading pathogenicity assignments. For instance, a typical WES in a rare disease medical context will uncover 30,000–100,000 variants relative to today’s reference genome[6], of which only one or a few may be causative. Approximately 10,000 of these variants will have putative molecular consequences by inserting or deleting genic sequences, or by causing missense or nonsense amino acid changes or alterations of conserved splice site residues[7,8]. After eliminating all those that are common in the overall, healthy population or otherwise predicted to be non-pathogenic, several hundred possible disease-causing candidates remain. Indeed, all ‘normal’ human genomes are estimated to contain ~100 loss-of-function alleles, with ~20 genes completely inactivated in each individual[9]. Thus, identifying true pathogenic variants among the many alleles that are putatively pathogenic remains a major challenge, and one that can best be tackled by maximizing the reliability, usefulness and availability of data and their annotations, which are recorded in genotype–phenotype databases.

In this Review, we consider the main objectives, state of progress, challenges and initiatives that presently characterize genotype–phenotype databases. We

[1]Department of Genetics, University of Leicester, Leicester LE1 7RH, UK.
[2]Data to Knowledge for Practice Facility, Cardiovascular Research Centre, Glenfield Hospital, Leicester LE1 9HN, UK.
[3]Institute for Medical Genetics and Human Genetics, and the Berlin Brandenburg Center for Regenerative Therapies, Charité Universitätsmedizin Berlin, 13353 Berlin, Germany.
[4]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany.
[5]Institute for Bioinformatics, Department of Mathematics and Computer Science, Free University of Berlin, 14195 Berlin, Germany.
Correspondence to A.J.B.
e-mail: ajb97@leicester.ac.uk

doi:10.1038/nrg3932
Published online
10 November 2015

also identify important missing elements that need to be developed to maximize the field's potential. Given the vastness of the topic, we have limited the focus to practical and technical issues and opportunities rather than ethical, legal and social matters, which cut across a broader swathe of biomedical practice.

Overview of the current database landscape

Given the immense size, complexity and variability of human genomes, the large range of possible normal and disease phenotypes, and the myriad settings in which one might wish to capture, organize and utilize human genotype–phenotype data, there is a need for many different types of genotype–phenotype databases. These databases are sometimes created using simple spreadsheets or text file approaches, but the rapidly increasing size and complexity of relevant datasets is generally being matched by progress in computing and database technologies.

Categorization of genotype–phenotype databases can be broadly based on many factors, such as the following: distinctions between health care and research settings; open versus controlled access or limited access systems; archival systems as opposed to those that handle real-time content (including Big data projects); primary data through to aggregated forms of data; and the spectrum between centralized (single-site data depositories) and federated (multiple interconnected data sites) ways of bringing datasets together. Furthermore, databases can be classified on the basis of the type of stored data and the source of this information (for example, disease-related mutations collected from the literature, primary through to fully processed and interpreted data from NGS-based investigations, and manually curated data combined from various sources). Evidently, there is substantial overlap and interplay between these different forms of genotype–phenotype databases, and when and where each option might best be deployed. In this Review, we provide a simplified categorization on the basis of relative magnitude. TABLE 1 provides a selection of current databases organized according to scope and purpose.

Small-scale databases. Many genotype–phenotype databases are small-scale efforts that support single research projects, specific disease areas, defined patient registries or specific sets of genes. Traditionally, diagnostic strategies in human genetics have focused on single genes or small sets of genes, leading to variant discoveries that have typically been submitted to locus-specific mutation databases (LSDBs; see TABLE 1 for examples). These databases aim to serve international communities of diagnosticians who specialize in the gene or genes in question by facilitating the interpretation of variants[10]. In other words, LSDBs contain sequence variation information about one or a few specific genes, usually pertinent to one or a few diseases. The aim is to collect complete and accurate lists of disease-related genetic variations and their phenotypic effects based on expert curation of published and unpublished data to produce these datasets. Increasingly, LSDBs for a number of related genes are grouped together, such as the resources listed in TABLE 1 for Fanconi anaemia (16 genes), osteogenesis imperfecta (16 genes), amyotrophic lateral sclerosis (116 genes) or immunodeficiency (131 genes). Other databases concentrate on genetic diseases of particular importance in a given country or ethnic group[11]; for example, the collection of ETHNOS databases. The Leiden Open Variation Database[12] and the Universal Mutation Database[13] software platforms are widely used for LSDB creation, as they were designed from the outset to be generic solutions for this domain.

The technical proficiency and data quality of small-scale, public databases varies. Beyond the inbuilt properties of the adopted database software they are created from, these databases tend to be tailored to local preferences and may use few standards in their design and operation, with little or no integration with other systems or options for bulk download. As such, despite contributing to the global digital collection of data relating genotypes to phenotypes, such databases tend to each be visited by only a limited number of users and may not be secured or integrated long-term due to sustainability challenges. This piecemeal arrangement of mixed quality datasets obviously limits the scope for good data exploitation; however, even greater amounts of poorly exploited data reside in non-public genotype–phenotype databases that support research and diagnostic laboratories in non-profit or commercial environments.

Large-scale databases. Numerous larger genotype–phenotype databases — primarily 'central' databases — have been created; a selection of the most notable of these is summarized in TABLE 1. Some larger genotype–phenotype databases seek to provide comprehensive coverage of germline variations in single genes across the majority of known Mendelian diseases. Principal examples include the commercial Human Gene Mutation Database, which obtains its core data from publications, and the public domain-based ClinVar, which has various input sources and links out to variants, diseases and other metadata. Other large databases have been created to manage genome-wide genetic data from NGS studies. For example, DECIPHER provides summary information derived from genomic screening investigations such as Array-CGH and exome sequencing, together with phenotypic descriptions of affected individuals. Other databases cover genome-wide association studies (GWASs) or somatic variations in cancer. Some of these repositories allow users to download not only information about individual variants but also files containing primary information such as whole-exome or whole-genome DNA sequences. For instance, the Cancer Genomics Hub currently contains almost two petabytes of downloadable data.

A special subset of databases could be called 'integration' databases because they unify other resources. Their objective is to bring together (via direct data compilation, hotlinks to remote data or virtualization approaches[14]) the content from various smaller databases and/or interrelated specialized databases to provide a single site with a more complete picture of a particular topic.

Genotype
In biology, genotype refers to the genetic makeup of an organism with reference to either a single nucleotide, a larger genetic locus or the entire genome. In the current context, genotype refers to a genetic sequence variant being assessed for potential causality of a disease, as well as its status as heterozygous, homozygous or hemizygous.

Phenotype
In biology, phenotype refers to the observable characteristics of an organism, but in medicine, the word is usually used to describe clinically relevant abnormalities, including signs, symptoms and abnormal findings of laboratory analyses, imaging studies, physiological examinations, as well as behavioural anomalies.

Variants
Genetic variants describe any deviations from a normal or reference sequence. For example, a substitution of one nucleotide for another at a certain chromosomal position, an insertion or deletion of one or more nucleotides, a chromosomal microdeletion encompassing several million nucleotides or a trisomy of an entire chromosome.

Pathogenicity
The tendency of a genetic variant in a person's genome to produce disease. The term is most often used in the context of cancer or inherited disease, when a genetic variant has a substantial deleterious effect on the function of the gene product that leads to, or substantially contributes to, the development of disease.

Effect sizes
The percentages of genetic variance explained by a specific locus, ranging from less than 1% for many common traits up to 100% for some Mendelian diseases.

Table 1 | **A selection of databases with a focus on genotype–phenotype relationships in human medicine***

Database	Scope and scale	Standards	Data entry	Data access policies	Refs
Gene variation database (LSDB or MDB)					
ClinVar	• Genetic variants and phenotypes • 125,520 variants	HGVS, HPO, MeSH, OMIM, RefSeq, SO	Curators and users	P	63
Human Gene Mutation Database (HGMD)	• Genetic variants and phenotypes • 163,610 variants	HGNC, HGVS	Curators	Com, P	96
Leiden Open Variation Databases (LOVD)	• Genetic variants and phenotypes • 3,334,104 variants (2,400,084 unique) in 248,807 individuals in 86 LOVD installations	HGVS, Mutalyzer	Curators	P, Cs	12
Universal Mutation Database (UMD)	• Genetic variants and phenotypes • 90,383 variants in 40 databases	HGVS	Curators	P, Cs	13
Amyotrophic Lateral Sclerosis Online genetics Database (ALSoD)	• LSDB • 116 genes associated with amyotrophic lateral sclerosis • 569 variations	HGNC	Curators	P	97
CFTR2	• Cystic fibrosis LSDB • 88,000 patients	HGNC, HGVS	Curators	P	23
Fanconi Anemia Mutation Database	• LSDB • 16 genes associated with Fanconi anaemia–BRCA pathway • ~3,000 variations	HGNC, HGVS, LRG	Curators	P	98
Osteogenesis Imperfecta Variant Database	• LSDB • 16 genes associated with osteogenesis imperfecta • ~1,500 variations	HGNC, HGVS, LRG	Curators	P	99
IDbases	• LSDB • 131 genes associated with immunodeficiency • Data for 7,292 patients	HGNC, HGVS, VariO	Curators	P	100
MITOMAP	• LSDB • Mitochondrial DNA variation • 1,746 variants		Curators	P	101
FINDbase	• Aggregated information on national and ethnic variation frequencies • ~100 NEMDBs	HGNC, HGVS	Curators	P	102
Array-CGH, WES, WGS (rare disease)					
DECIPHER	• Genetic and phenotypic data • Diagnostics and discovery. • 42,815 cases	HGVS, HGNC, HPO	Users	P, Cs, RG, MM	38, 103
PhenomeCentral	• Genetic and phenotypic data • Genomic “matchmaking” • 600 cases	HPO, VCF	Users	MM	65
PhenoDB	• Genetic and phenotypic data • Diagnostics and discovery. • 3,300 cases	EoM, HPO, OMIM, PhenoDB	Users	MM	104
GeneMatcher	• Gene matching (discovery) • 668 genes	HGNC, Ensembl, Entrez Gene, OMIM	Users	MM	105
Mendelian and other rare disease knowledge bases					
Online Mendelian Inheritance in Man (OMIM)	• Knowledge base • 22,644 entries (genes or diseases)	HGNC, HPO, ICD, OMIM, PhenoDB, SNoMED, UMLS	Curators	Ac	106
Orphanet	• Knowledge base • 5,833 disease entries, copious data on other rare disease topics	HGNC, ICD, MedDRA, MeSH, OMIM, UMLS, Uniprot	Curators	Ac	73
Monarch Initiative knowledge base	• Human and model organism genetics and phenotypes • 36K diseases, 33K phenotypes, 500K genotypes, 30K genes, 2M curated phenotype associations, >100 species	HPO, MPO	Curators	P	107

Table 1 (cont.) | **A selection of databases with a focus on genotype–phenotype relationships in human medicine***

Database	Scope and scale	Standards	Data entry	Data access policies	Refs
Cancer genomics and variations					
Cancer Genomics Hub	• Genetic and phenotypic data repository • 82,140 files (1870 Tb)	Sequence Read Archive Metadata XML	NCI projects, curators	P, CA	108
Catalogue Of Somatic Mutations In Cancer (COSMIC)	• Variation and genetic and phenotypic data • 2,139,424 unique variants	HGNC, CCDS	Curators	P	109
DriverDB	• Variation and genetic/phenotypic data • 6,079 datasets	HGNC	Curators	P	110
Genotype–phenotype information for GWAS and other studies					
Database of Genotypes and Phenotypes (dbGAP)	• Genetic and phenotypic data • 508 studies	dbGAP, XML	Curators (focus on NIH-funded projects)	P, CA	43
European Variation Archive (EVA)	• All types of genetic variants from any species • ~40 studies, representing 35 species, describing ~400 million unique alleles from more than 150,000 samples	VCF, dbSNP	Users, curators	P	
European Genome–Phenome Archive (EGA)	• Genetic and phenotypic data • 1,555 datasets	VCF, FASTQ, BAM, EFO	Users	P, CA	42
GWAS Catalog	• Genetic and phenotypic data • 18,697 associations	dbSNP, HGNC	Curators	P	111
GWAS Central	• Genetic and phenotypic data • >75 million associations	dbSNP, HGNC, HPO, MeSH	Curators	P, CA	47
GWASdb	• Genetic and phenotypic data • 272,918 associations	dbSNP, DO, HPO	Curators	P	112
Human Genome Variation Database	• Genetic and phenotypic data • ~100 datasets in 6 integrated databases	HGNC, HGVS, dbSNP	Users	P, CA	113
Pharmacogenomics					
PharmacoGenomics Database (PharmGKB)	• Pharmacogenomics knowledge resource • Extensive data on variants, pathways, dosing, clinical annotations, drug labels	dbSNP, HGNC, MeSH, SNoMED, UMLS	Curators, NLP	Ac	114

Ac, open to academics, but commercial entities require license; CA, controlled access; CCDS, Consensus CDS; CGH, comparative genome hybridization; Com, commercial; Cs, restricted to consortium members; dbSNP, Database of Single Nucleotide Polymorphism; DO, Disease Ontology[87]; EFO, Experimental Factor Ontology; EoM, Elements of Morphology[115]; 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; MM, MatchMaking (finding similar patients); 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; P, public; RG, access restricted to a specific research group; SNoMED, Systematized Nomenclature of Medical Terms; SO, Sequence Ontology; UMLS, Unified Medical Language System; VariO, Variation Ontology; VCF, Variant Call Format; WES, whole-exome sequencing; WGS, whole-genome sequencing; XML, Extensible Markup Language. *Further details are available at the websites of the databases. Owing to space constraints, many important databases had to be omitted. Databases are divided into major categories, but it is recognized that many of the databases can be considered to belong to multiple categories. Scope refers to the major focus of the database. If applicable, information is given about the number of items currently contained in the database. Standards indicates the major terminologies or ontologies that the database uses to annotate and organize data. Data entry indicates whether the data in the database are primarily entered by curators or by users of the database, and if natural language processing (NLP) is employed to gather data. Note that for data access policies, databases with more than one access level tend to provide more information as data access becomes stricter.

There are also generalist knowledge bases, of which two prominent examples for human genetics and rare disease would be Online Mendelian Inheritance in Man (OMIM) and Orphanet, respectively. These types of database invest substantial manual effort to collect, curate, harmonize and inter-relate large amounts of primary information to create broad and powerful compilations of knowledge about disease genes and related clinical or research topics.

Central and integration databases tend to make greater use of standards (in some cases also precipitating the creation and adoption of standards, such as Locus Reference Genomic sequences[15,16]), provide web services, and offer powerful data search, display and download options. As such, they can and often do have extensive inter-database connections, and may replicate at least a minimal level of their content (sufficient for crosslinking) between one another. Rarely are they commercial systems and, similar to their smaller counterparts, rely almost exclusively on academic funding for their survival and growth.

Key challenges and common goals

Expanding datasets. The rapidly increasing abundance of genetic and phenotypic data, driven by the transition from single-gene testing to NGS gene panel, exome and genome sequencing, has resulted in a concomitant increase in the size, scope and sophistication of genotype–phenotype databases. This growth produces two further consequences. First, analysts who previously specialized in one or a few genes will now be confronted by variants from thousands of genes. In other words, only one or a few genes used to be tested in genetic diagnostics (for example, *NF1* in a patient with possible neurofibromatosis type 1) and any observed likely disruptive variant was taken as sufficient evidence to confirm the diagnosis. Now, with NGS, far more evidence from diverse data sources must be considered before drawing any diagnostic conclusions. Thus, practitioners will need to make much greater use of public genotype–phenotype databases than ever before. Second, the data sources used must be of extremely high quality to minimize the risk of erroneous analyses. Currently, gene- and disease-specific databases only sometimes meet this requirement because they vary widely in comprehensiveness, depth of clinical data, coverage of published and unpublished pathogenic variants, and may contain a substantial proportion of erroneous information[17]. For example, up to 27% of literature-cited disease-causing variants in some databases are incorrect or incomplete, being merely either common polymorphisms or sequencing errors, or lacking good evidence of pathogenicity[18].

Assigning pathogenicity. For all but the most highly studied variants in a relatively small list of well-understood disease genes, uncertainty exists as to whether they have a causative role in a particular disease. For instance, a recent study identified tripartite motif containing 63, E3 ubiquitin protein ligase (*TRIM63*; also known as *MURF1*) as a novel disease gene for hypertrophic cardiomyopathy on the basis of two missense variants and one deletion allele detected in 302 patients but in none of the 229 control individuals[19]. However, other studies reported a nonsense variation (p.Q247*) in *TRIM63* in individuals with no signs of cardiomyopathy[20]. Deletion of *Murf1* in mice did not result in heart pathology, but knocking out both *Murf1* and *Murf2* led to extreme cardiac hypertrophy[21]. Hence, it remains uncertain whether *TRIM63* is a Mendelian disease gene or whether it merely contributes to hypertrophic cardiomyopathy in conjunction with a variant in another gene.

As the above example illustrates, the task of assigning pathogenicity to a genetic variant is far from straightforward, primarily because establishing causality involves having to apply differential and often subjective weighting to multiple lines of evidence, and making choices over which analytical tools to use. Common practice in research and clinical diagnostics seeks to address this challenge by reducing pathogenicity down to a simple classification system with just a few categories (such as 'definitely pathogenic', 'probably pathogenic', 'uncertain', 'probably not pathogenic or of little clinical significance', 'not pathogenic or of no clinical significance')[22]. However, the actual concept of pathogenicity underlying these classifications is often not well defined. Ideally, the meaning of pathogenicity would be not only universally understood and consistently used but also broken down and quantified at the level of its underlying components — penetrance, expressivity and the precise functional or clinical phenotype under consideration — with each component being given in the context of a certain age range, gender, population and environment. However, elaborating this type of stratified medicine approach for each and every subgroup of interest (which could be a single person or thousands of individuals) will require an extensive amount of high-quality observational data to establish and interpret signal patterns reliably. In the past, this scale and quality of data simply has not existed for most disorders, and still rarely does today. A notable exception is the CFTR2 database[23], which contains data on more than 88,000 patients, thus enabling this database to provide information on sweat chloride, lung function, pancreatic status and *Pseudomonas* spp. infection rates associated with many cystic fibrosis transmembrane conductance regulator (*CFTR*) mutations. As such, genotype–phenotype databases previously needed only simple computing technologies, including very basic data fields relating to pathogenicity, and did not capture the process of pathogenicity interpretation or the employed evidence base. Going forward, this approach will have to change, especially if we wish to deliver truly personalized medicine, which will require mechanistic in addition to probabilistic modelling, and hence even more sophisticated sources of input information and tools for the recording of results.

Current progress and trends

Database types and arrangements. The ideal number, type and arrangement of genotype–phenotype databases that will ultimately be required cannot easily be predicted. Instead, we can merely observe the current reality and note some common themes and trends (FIG. 1). Factors such as the need for recognition and reward for creators of data and databases, valid limitations on data sharing, a desire to encourage widespread innovation, and the value of having domain-specific experts close to data curation and management processes, all argue in favour of federated databases. A counter argument would be that it is both more cost-efficient and practical in terms of standards-based inter-system compatibility to only build a few centralized repositories. In practice, real-world primary databases and data generation projects are deciding for themselves what subsets of their content are to be submitted into centralized depositories, whereas those centralized databases are making themselves more attractive to data depositors by offering incentives such as private data hosting services, managing data access requests, archiving, deeper and better data analysis, and full accreditation with links back to the sources on all displays of the submitted data.

Multiple testing
The process of using bioinformatics analysis to assess potential pathogenicity of a variant is often formulated as a statistical hypothesis test. As tens of thousands of such tests may be performed in the analysis of diagnostic next-generation sequencing data, adjustments of the *P* values resulting from assessments of individual variations are required to avoid numerous false positive results, a procedure known as multiple testing correction.

Whole-exome sequencing
(WES). A sequencing technique that seeks to selectively enrich and assay only the sequences belonging to the ~1.5% of the human genome consisting of the exons of protein-coding genes (called the exome) because the majority of causative variations identified in Mendelian diseases to date have been located in or very close to these exons.

Big data
This term is used to describe collections of data that are characterized by features such as being large in size, complex and heterogeneous in type, rapidly produced or frequently changing, and of uncertain veracity, such that analysis requires high-performance computing resources and sophisticated algorithms. In biomedicine, especially high-throughput omics data such as whole-genome sequencing, as well as ever increasing amounts of clinical data available in electronic health care records, are often regarded as big data.

Standards
In the present context, a formal set of specifications about the format and contents of data records of variants or diseases that are to be exchanged between databases.

Metadata
Metadata, literally 'data about data', refers to information that accompanies other data and explains their context or provenance.

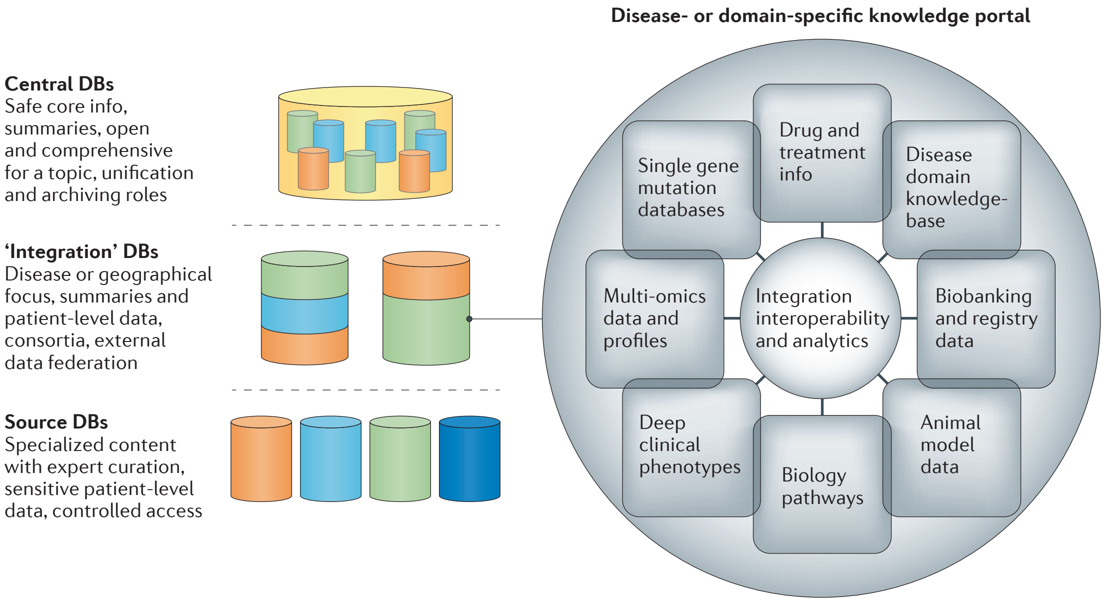
Disease- or domain-specific knowledge portal
Central DBs
Safe core info, summaries, open and comprehensive for a topic, unification and archiving roles
'Integration' DBs
Disease or geographical focus, summaries and patient-level data, consortia, external data federation
Source DBs
Specialized content with expert curation, sensitive patient-level data, controlled access
Single gene mutation databases
Drug and treatment info
Disease domain knowledge-base
Multi-omics data and profiles
Integration interoperability and analytics
Biobanking and registry data
Deep clinical phenotypes
Biology pathways
Animal model data

Figure 1 | **Emerging landscape of genotype–phenotype databases.** Based on the types of databases now in existence, the field can be described as having a three-level architecture: source databases (DBs), 'integration' databases and central databases. Source databases hold highly sensitive and private data of a specialized nature, whereas integration databases (knowledge portals) aggregate some depth of content from source databases and other resources to support specific disease areas or regional and project-based activities. Top-level central databases sit above the source and integration databases and provide a (semi-) comprehensive overview and universal services. The orange, blue and green database sections shown in this diagram represent different types or categories of data. The optimum number of each type of database, and the degree of federation and communication across and between the layers, will presumably find its own optimum as the field further evolves. This process could be accelerated by creating interoperability methods and standards, consensus principles by which data can be linked, discovered or shared across databases, robust consent and legal frameworks under which they interoperate, and models for sustainable funding.

Some recent initiatives, such as the Collaborative Cancer Cloud even seek to move all of the above onto the cloud, and hence the question of where a database physically resides, or debates about where one resource starts and another federated one ends, are becoming increasingly meaningless.

Data linking and reporting. Despite the impressive sophistication of central databases, there is often no direct connection between these resources and the far larger number of smaller databases or non-public datasets. Similarly, the process of reporting genotype–phenotype relationships in traditional publications does not normally involve the transfer of those findings into any database, and so curators must subsequently spend time extracting and integrating this information into structured repositories. These disconnects make it difficult for any database to present a complete view, even though the technical solutions to this issue (for example, web services, data linking and core data transfer) are obvious and relatively simple to implement. Clearly, other factors such as funding, competences, incentives, legal restrictions and informed consent, are holding up progress in this area. Arguably, journals and funding agencies are best placed to address these issues, as both can direct resources to getting pipelines built for data submission, and apply pressures to make researchers use them. Some journals have now begun to move in this direction (notably *Human Mutation*[24], which requires published variant data to be submitted to a database, and more recently *CSH Molecular Case Studies* for phenotypic abnormalities) while funding agencies are beginning to formulate rules and requirements that will work best if supported by real sanctions (such as the National Institutes of Health Data Sharing Policy). However, the pace of change must be managed carefully, as it would make no sense to insist on digital deposition of data if the necessary journal procedures and adequately governed databases are not first established and placed on a sustainable footing.

The above considerations raise the question of how and when genotype–phenotype data are, could or should be entered into an online database. Given the diversity of methods and sources for data generation, the many ethical, legal and other restrictions on the use of such data, and the array of databases, registries and biobanks that might want to receive the data, there is no straightforward answer. The following factors therefore become key considerations: whether the data generator and owner have the time and ability to submit to a database; the conditional balance of risks and rewards for doing so; the perceived strengths, longevity and convenience of the available databases; and the quality and utility of the data compared to the expectations of those databases. The

Array-CGH
Array-comparative genomic hybridization (CGH) enables the gain or loss of genetic material to be detected in the range of as little as 40 kilobases up to entire chromosomes. Array-CGH has become a standard diagnostic tool for the identification of copy number variants.

Web services
Databases, data processing or analytical functions that can be accessed by another computer program over the worldwide web.

Penetrance
The proportion of persons who carry a pathogenic germline variation and also show signs of a disease irrespective of the clinical severity.

Expressivity
The degree of clinical expression and severity of a disease in individuals who have inherited a given germline variation.

Stratified medicine
An approach to patient care that subdivides patients into groups that are defined on the basis of expected risk of developing disease or the expected response to a certain treatment.

Personalized medicine
This concept is synonymous with individualized medicine, and is used in varying ways to convey the idea of health or medical care being in some way tailored and optimized for a person. This typically means going beyond shaping care for groups of similar patients to the ultimate of uniquely customizing interventions for each separate individual.

Probabilistic modelling
A class of computational algorithms that describe data observed from a system in a way that takes uncertainty and noise associated with the model into account. It is one method for making predictions about disease onset or severity on the basis of genetic and other data.

Federation
A software strategy that allows data from disparate databases and other sources to be aggregated ad hoc as a virtual database that can be used for analysis. In the present context, federation involves connecting genotype–phenotype databases across networks to allow combined searches for information about variations or diseases.

Cloud
Remote servers that are accessed via the Internet and provide data storage and analysis resources.

Informed consent
An agreement on the part of a patient to take part in a clinical study and allow the results of the study to be used in some way, such as for additional research or health care activities or for sharing with others in a publication or database. Consent can only reasonably be given after the subject is informed and given the opportunity to discuss the purpose of the research and any potential harms and benefits.

Biobanks
Collections of biological (often medically relevant) specimens such as blood, saliva or tissue, associated with data annotations that describe the subjects from whom the specimens were obtained, such as age, gender, environmental exposures, phenotypic features, molecular test results or clinical diagnosis. Biobanks are used by researchers to obtain sets of data and specimens from subjects with the same diagnosis or with similar characteristics to undertake research investigations.

Registry
A registry comprises a collection of information about individuals affected by a specific disease or who share other similarities. Many registries collect information about individuals over time or are used to track information regarding the response of patients to treatments. A registry may, but does not necessarily, include genetic information.

complex equations that these factors raise are slowly and steadily being solved, such that more and better data are being organized and entered into databases each year; however, many further improvements are needed. As a workaround or parallel track, there has been considerable interest in recording and making available genotype–phenotype data collected directly from patients; for example, the PEER platform of the Genetic Alliance[25], PatientsLikeMe[26] and GenomeConnect[27]. Such an approach fits well with the global move towards patient empowerment and putting individuals at the centre of their own health care and health preservation[28,29].

Improving phenotypic information and pathogenicity evidence. The data that have so far been gathered into the types of databases listed in TABLE 1 have strengths and weaknesses. The genotype (that is, DNA sequence-related) content is already large in scale and growing rapidly, primarily as a result of the power of NGS technologies deployed for the generation of reference datasets (for example, 1000 genomes[30] and the Exome Variant Server of the NHLBI Exome Sequencing Project[31]) and owing to large studies of the genetics of disease (for example, the Personal Genome Project[32], the International Cancer Genome Consortium[33] and The Cancer Genome Atlas[34]), for which data access is often open or relatively straightforward to obtain. By contrast, the phenotype dimension is lagging behind in terms of the scale and the granularity of the information collected, and its degree of standardization (and hence its wider utility). Many older genotype–phenotype databases contain and provide only limited phenotype information, perhaps just the headline disease name at best. Critically, especially with genome-wide variation data, a careful clinical analysis of phenotype data is an essential component of genomic analysis[35]. For instance, less than 10% of published variants in genes associated with medically actionable genetic conditions were judged to be sufficiently supported by the medical literature to be reported as an incidental finding in adults otherwise not known to have the conditions in question[36]. This finding highlights the fact that, when considering variants identified by WES or whole-genome sequencing in diagnostic settings, practitioners should not rely on previously published claims that variants are putatively disease-causing, because such pathogenicity assertions are often erroneous. More attention to phenotypic information and other metadata in variant databases would help to improve the accuracy of clinical interpretation of WES data. Fortunately, more recent initiatives (for example, CARE4RARE[37], DECIPHER[38], the GEnomes Management Application[39], the PheWAS Catalog[40] and the Kaiser Permanente Research Program on Genes, Environment and Health[41]) are emphasizing the need for phenotype content to be as rich as the genotype data. This move is also true of platforms that manage the controlled access of detailed research studies (for example, the European Genome-Phenome Archive[42] and the database of Genotypes and Phenotypes[43]). Perhaps the disparity between the amount and quality of genotype and phenotype data reflects the degree of practical, financial, ethical, legal and organizational challenges that must be overcome to produce good phenotypic data on large numbers of individuals. If so, an unfortunate corollary might be that researchers that do take on this effort may subsequently be less willing or able to share their data widely. One way to minimize this problem would be for health care data (for example, clinical phenotypes, patient histories, medications and outcomes) to be processed and managed routinely in ways that make them more available for feeding into research programmes[44].

In addition to genotype and phenotype data, databases are in a few cases beginning to include the actual evidence used to infer variant pathogenicity, as well as the methods used to process and interpret such evidence. Because of the diversity of this evidence, its capture would ideally be facilitated by the use of innately flexible database technologies (such as i2b2 (Informatics for Integrating Biology and the Bedside)[45] and Observ-OM[46]), which are not constrained to holding only certain types of data structured in a specific way.

Alternative modes of data provision. Databases seek to gather and hold data, and enable users to search for and thereby access data of interest to them. This goal is simple to achieve for a solitary database that handles a single type of data record, but far from straightforward if the ultimate goal is a single genotype–phenotype databasing 'universe'; that is, a comprehensive system in which all information is interconnected, and data quality and data duplicates are fully apparent. Most data are large and complex, widely dispersed in different repositories or projects of different designs, with privacy concerns imposing limitations and uncertainties over which users can access which particular records in which situations. These challenges mean that progress towards completely unifying and optimally sharing individual-level, personal data (which includes substantial portions of one's genome sequence), will need to proceed carefully, with many different strategies being debated and applied in various settings.

As a reaction to these challenges, complementary strategies are being explored that can make data more immediately useful (FIG. 2). One approach involves converting data to other representations, and enhancing it in various ways, to counter the risk of subject identification. An obvious example would be the aggregation of individual-level genotype data into variation frequency information, split by population group and/or disease. More refined approaches would entail the creation and use of detailed metadata, and the use of graphical displays of signals and patterns in complex datasets (for example, GWAS Central[47]). Such approaches are a key facet of integration databases and disease-area-specific knowledge portals (FIG. 1), enabling them to combine LSDB, registry, biobank, research and diagnostics information, across species, and with functionalities serving many different audiences across research, health care and beyond.

Another approach is to enable many dispersed collections of data to be jointly exploited. Technically sophisticated ways to achieve this involve setting up mechanisms (for example, DataSHIELD[48]) that enable remote pooled data analysis, thereby eliminating the need to share data directly. Related approaches that do entail direct sharing of data can employ multi-party data encryption for added security[49]. More straightforward approaches involve bringing together a common, limited depth of data from many sites to provide a single, comprehensive search, access and analysis environment. Such projects are urgently needed as a better alternative to the use of generic Internet search engines, such as Google or Google Science, for accessing specific genotype–phenotype data subsets. Several efforts in this direction have been launched, not least SNPedia[50], MalaCards[51], WAVe[52], Café Variome Central, Kaviar[53], the European Variation Archive and the Exome Aggregation Consortium.

A focus on data discovery. A novel strategy for making databases more useful changes the traditional paradigm that data provision automatically follows from a successful search of a database by an approved user. In other words, searching through a database becomes a goal in and of itself, enabling the existence of data of interest to be discovered, irrespective of whether or how these data might subsequently be accessed.

Steadily, a comprehensive genotype–phenotype 'data discovery layer' is beginning to emerge, in which the location of all relevant data is established using one of various portals. This process could become highly automated, particularly if consent agreements can be based on more regularized clauses (including some that give explicit approval for data discovery as a particular form of data sharing), and then be represented as computer-readable data-use metadata to facilitate interactions between the computer systems of discovery users and discovery providers. Instead of returning data after a discovery search, a system might instead employ alternative data representation options, such as those mentioned above. As a result, immense combined datasets from many databases (even sensitive content) could be immediately examined for informative signal patterns.

Various different data discovery approaches are currently being explored (FIG. 3). Examples include the following: the search step itself can be enhanced with, or replaced by, visual modes of data interrogation (for example, via genotype–phenotype relationships marked as icons on a genome browser track); the content being searched could be a safer derivative of the data, such as descriptions of the data resource (catalogue approaches), study metadata or aggregated data (pre-prepared or generated dynamically according to the search parameters), rather than the data itself; and various options could be provided after a successful search, such as provision of metadata or data only in summary or graphical form, facilitation of data request procedures, or providing contact details of the data owners.

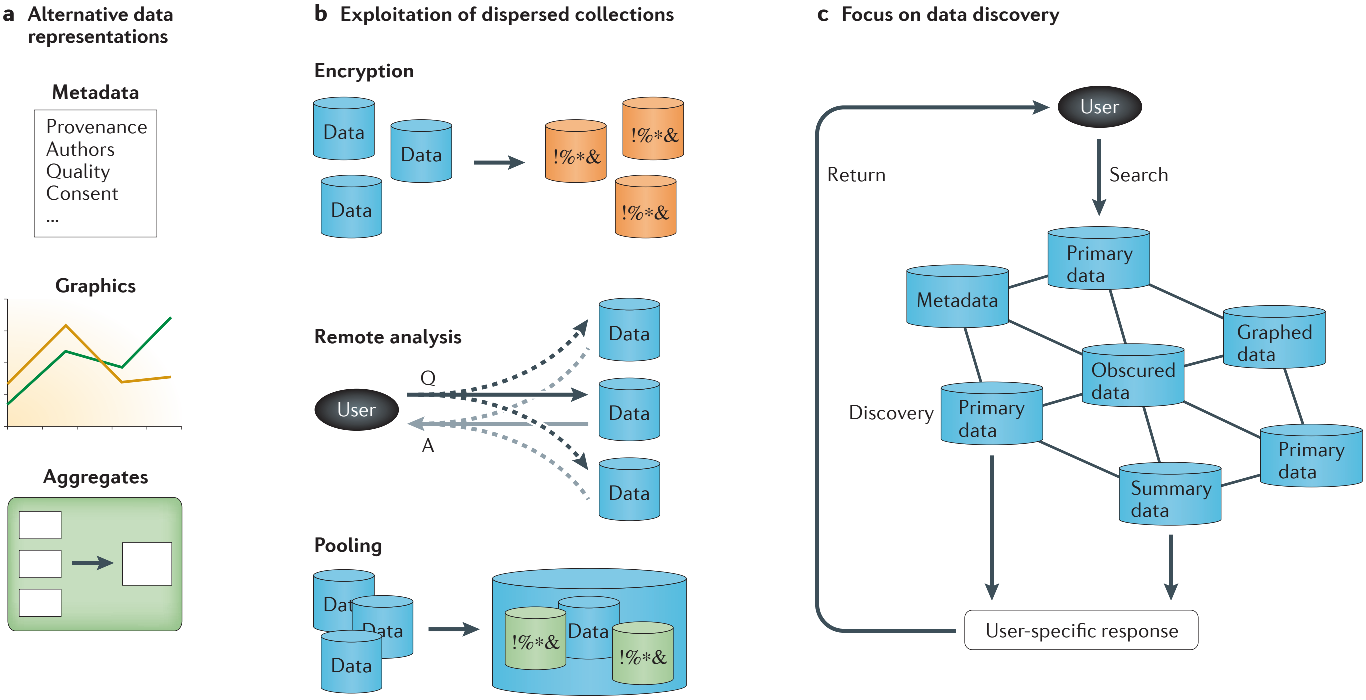

Figure 2 | **Modes of data provision.** In addition to the sharing of data to maximize benefits, other approaches can be followed in parallel, or instead, as and when suitable. **a** | For instance, other aspects or derivatives of the data, such as metadata, graphical representations or aggregates, which may be less sensitive, can be shared. **b** | Additional approaches include data encryption to further enhance data protection, and remote data querying, whereby the data are not transferred to the user (although may be pooled at a trusted broker). **c** | Data discovery spans a range of options that complement data sharing, ranging from completely risk-free querying of safe data components leading to simple confirmation that data of interest exist, through to enabling full datasets to be queried and some data elements to then be provided (which *in extremis* is full data sharing).

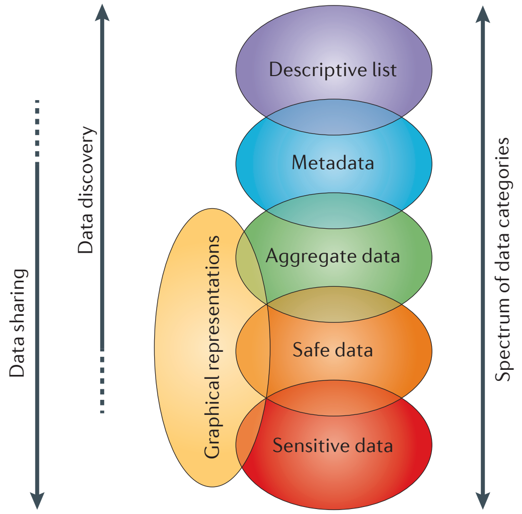

Figure 3 | **Data sharing and data discovery.** A database may contain a continuum of types of information, as illustrated. Most focus is usually given to the primary data, which may in part or in whole be sensitive in nature (such as identifiable sensitive data (red oval)) or content that is generally safe to further distribute (orange oval). But these data can easily be converted to other forms that may often be more convenient to consume, be even safer to share and can potentially better portray the knowledge in the data. This type of data includes summary statistics (aggregate data (green oval)) or graphical representations (yellow oval). Passing any of these categories of data on to others is what one normally means by the term 'data sharing'. With the possible exception of sensitive data in some settings, data discovery approaches could also make good use of these same categories of data. Such an approach would enhance the scope of data discovery, which has more typically concentrated on aggregate data, and especially metadata (light blue oval) and resource descriptions (purple oval). Given this overlap between the concepts of data discovery and data sharing, and the underlying data types they can use, interfaces, functions and procedures should therefore be devised that synergistically facilitate both of these complementary approaches to the dissemination and exploitation of information.

International Rare Diseases Research Consortium
(IRDiRC). This consortium comprises rare disease researchers and funding organizations and promotes the goal of developing 200 new therapies for rare diseases and a means to diagnose most rare diseases by the year 2020.

Global Alliance for Genomics and Health
(GA4GH). This alliance comprises more than 200 institutions working in health care, research, disease advocacy, life science and information technology with the goal of creating a common framework of harmonized approaches to enable the responsible, voluntary, and secure sharing of genomic and clinical data.

Human Variome Project
An umbrella organization that intends to help coordinate efforts to integrate the collection, curation, interpretation and sharing of information on variation in the human genome into routine clinical practice and research.

Genotype–phenotype data discovery systems to date are most apparent in the biobanking world, where catalogues are springing up frequently and progressively being unified, as well as other efforts that facilitate searching of aggregate data via visual displays (for example, GWAS Central[47]) or via metadata (for example, the European Variation Archive[42]) followed by providing details on where or how the primary data can be requested. More recently, the idea of enabling discovery on primary data has been promoted in initiatives such as the following: MatchMaker Exchange5[4], an initiative led by the International Rare Diseases Research Consortium (IRDiRC)[55] and Global Alliance for Genomics and Health (GA4GH)[56] to discover matched pairs of patients with rare diseases based on similar genotypes and phenotypes, initially using public domain data but eventually via sensitive information that cannot be openly exposed on the Internet; GeneYenta[57], which finds similar patients with rare diseases via phenotype level matching; Cafe Variome[58], which discovers variant and phenotype data in non-public diagnostic and research laboratory networks; and the Beacon project, which discovers the existence of specific genetic variant records in diverse sources. Additionally, data discovery concepts are spreading to other areas, such as public health and epidemiology[59] and even as the basis for a UK-wide discovery catalogue of research data from higher education institutes and data centres[60].

One final notable consequence of data discovery is that it enables those that generate and/or own data, which is typically held within individual level databases, to maximally expose this information for discovery while keeping far greater control over how and when it is used than that offered by more conventional data search and access approaches (such as having a 'terms of use' statement on an open access website). This option is particularly important for data custodians, who may be responsible for ensuring that data are used appropriately or stand to benefit if data are used by others (via, for example, collaboration, grants and recognition). Given that much data might otherwise remain hidden, the emergence of data discovery technologies is likely to increase data visibility overall, leading to more and better collaborative uses of data. Sometimes, however, this approach might limit data sharing, if particularly cautious data custodians elect to place only data discovery interfaces rather than data-sharing options on top of their databases.

Emerging community efforts

In the 1990s, funding or recognition for those involved in generating, managing and/or contributing to genotype–phenotype databases was limited; however, progressively, as NGS took hold in the first decade of the new millennium, more resources were allocated to data management. In recent years the field has truly blossomed. Large-scale studies are generating genotype and phenotype data on substantial numbers of individuals, data collection efforts have been ramped up, with database creators sourcing literature, clinical and research datasets, and international coordination and standardization work is taking place. Even just considering rare diseases, the list of community initiatives is long. For example, two global consortia (IRDiRC[55] representing and assisting funders, and GA4GH[56] developing precompetitive standards across this field and others) are building on the work of the Human Variome Project[61]. Moreover, the ClinGen[62] network of many co-funded North American laboratories is tackling issues covered in this Review while shepherding data into the ClinVar[63] database. In another example, the Canadian Forge/Care4Rare[64] programme and the DECIPHER[38] project (via the PhenomeCentral[65] and DECIPHER databases, respectively) are gathering data about many patient cases to help with diagnosis (a mission now bolstered by Genomics England targeting 100,000 patients with rare diseases or cancer). Simultaneously, many and various consortia involving research and health care

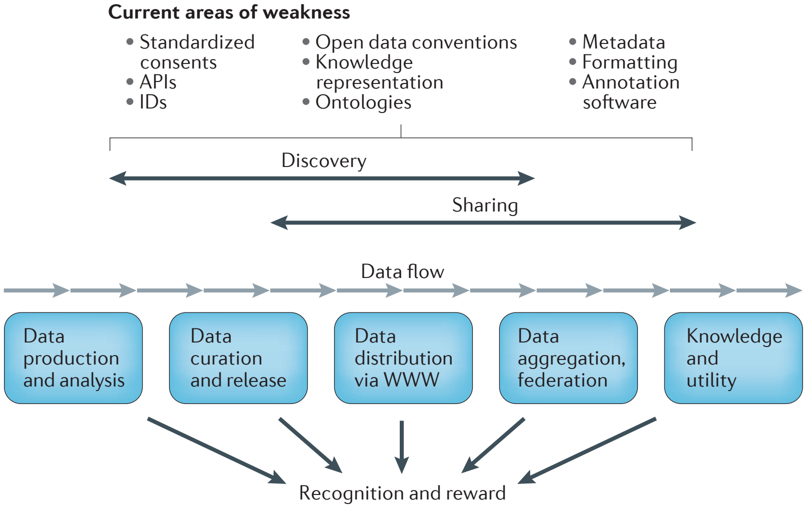

Figure 4 | **Data handling in genotype–phenotype databases.** Ongoing progress in the number, technical sophistication and inter-connectedness of genotype–phenotype databases is opening the way to increasingly effective and complex ways of managing and exploiting the data held within these depositories. Data production and analysis in research and health care settings is becoming faster and more elaborate, which demands ever better ways to curate (organize, describe, annotate, quality check and deposit) the resulting information within source databases. These data and results, along with associated metadata and governance information, then need to be made available for discovery, access and use by others over the Internet. Such uses will often involve reorganizing and federating data from a number of resources via reformatting, aggregation and merging with other content within integration databases. Ultimately, this process produces improved scientific knowledge that enables predictions to be made and hence utility to be derived, and this is best orchestrated and disseminated by central databases interested in translational research and/or clinical diagnostics. Making this complex data flow work well requires the efforts of many computer technologists, information curators and data scientists, the contributions of whom need to be consistently recorded, assessed and rewarded by approaches that go beyond the traditional measures of publication impact factor and H-scores. These various aspects of local through to international data management contribute to the maximization of data discovery and sharing activities, but they are not sufficient. Various additional elements need to be devised and improved, as discussed in the main text. APIs, application programming interfaces.

professionals are running projects to discover and characterize causative variations and genes in specific classes of rare diseases; such projects are being aided in the computational aspects of this task by groups such as the European RD-Connect project[66].

Previously, the field was overwhelmingly characterized by 'data silos', meaning that databases were isolated and did not share information on variants or variant interpretations. Indeed, the business model of commercial genetic testing entities still commonly involves proprietary and unshared data. For example, Myriad Genetics had a long-term monopoly on *BRCA1* and *BRCA2* genetic testing, and thus accrued large amounts of data on variants in the BRCA genes, giving the company a competitive advantage in the evaluation of rare *BRCA1* and *BRCA2* variants for which the medical significance has not been documented in the medical literature or public databases[67]. However, the human genetics community is rapidly shifting to a new paradigm of publicly sharing variant and variant interpretation data, as exemplified in projects such as ClinGen[62] and the GA4GH BRCA Challenge. Moreover, commercial databases are increasingly responding to customers' wishes to share data by allowing direct data deposition to public databases[68].

Beyond rare diseases, there are too many large community and consortium programmes available to list them comprehensively in this Review, but notable for their ambitious size and success to date are the global International Cancer Genome Consortium[69] (ICGC), whose aim is the characterization of germline and somatic variations in cancer, and the GERA[70] study, which explores adult health and ageing.

Strikingly, given the mixed interests of the range of participants in today's large programmes, it is becoming increasing difficult to categorize any one stakeholder as being from the research community, the health care sector, a commercial entity, a charity, an organized patient group or the general public. To help with these new constellations and interactions, there have been calls for a stakeholders' charter to reassure and set expectations, with at least two frameworks for this now created: the GA4GH framework[71] and the RD-Connect charter[72]. It is also increasingly difficult to keep track of who is engaged in which initiatives, or which tools, resources and skills are available from which sources. As such, there is a need for one or more resource discovery catalogues, which are currently being developed; for example, by IRDiRC, RD-Connect[66], Orphanet[73] and GA4GH[56].

Future perspective

The field of genotype–phenotype collation is maturing rapidly via dynamic interactions between many interested parties combined with innovation, testing and dissemination of solutions. With such productive, natural evolution taking place, it is difficult to identify major deficiencies that are not already being tackled. Nevertheless, the community is not served well by having too many disconnected efforts addressing current issues, nor by having top-down imposition of concepts devised without input from those actually building and using the databases. The larger initiatives now under way (such as GA4GH[56]) comprise diverse groups that are working together to devise standards guided by a deep understanding of the problem domain, organized and disseminated professionally, with a very open, inclusive and consultative approach. Several particularly important areas of unmet need have been identified via such community discussions, as elaborated below and illustrated in FIG. 4.

Metadata. Substantial progress could be made if the database community could agree as to what minimum amount of data (for example, pathogenic DNA variants and the main associated phenotypes) should be made available by publicly funded research projects or upon journal publication. Once made available, data would be immensely more valuable if accompanied by information that contextualizes the data. Therefore, it would be good to emphasize what depth and types of metadata

Stakeholder
In the present context, a person or organization with an interest or role in medical databases, including patients and families, physicians, researchers, public and private research institutions, and funding agencies.

Box 1 | Ontologies and nomenclature

When managing and sharing genotype–phenotype data with the goal of having them used effectively, it is important that these data be composed from sufficiently precise words used in the correct way. As such, many groups are working on ontologies and nomenclature standards, the widespread use of which needs to be encouraged. For instance, a specific variant in the prothrombin gene associated with hypercoagulability is variously called the 'prothrombin 20210 mutation', the 'prothrombin variant', 'prothrombin G20210A', and many other names[75]. Although such designations are recognized by specialists, they impede computational data consumption and make it difficult for non-specialists to interpret the variant. These issues motivated work that culminated in the Human Genome Variation Society (HGVS) standards for sequence variation nomenclature[76–78] (see Further information), along with computational quality control of variation nomenclature using tools such as Mutalyzer[79,80] and open-source code libraries[81]. Many other standards are equally gaining importance as genomic medicine moves from single genes towards panels, exomes and genomes. These standards include file formats for exchanging low-level next-generation sequencing (NGS) data as FASTQ[82] or SAM/BAM[83] files, the Variant Call Format[84] (VCF), which is nearly universally used for exchanging variants revealed by NGS, and the Locus Reference Genomic standard sequences for gene variant reporting, which provides a unique and stable single file reference DNA sequence along with all relevant transcript and protein sequences needed for describing gene variants[15,16]. Additional standards have been developed for reporting variants with metadata (for example, VarioML)[85] as well as for describing variation effects and mechanisms (for example, VariO)[86].

Particular challenges emerge in the case of phenotype data, as such data spans an almost infinite spectrum of possible observations about an individual. Describing, collating and computing on this semi-subjective data continuum are therefore highly dependent on the existence and use of good ontology standards, especially when interpreting genome-wide screening data[35]. Deep phenotyping can be defined as the precise and comprehensive analysis of phenotypic abnormalities in which the individual components of the phenotype are observed and described. Ontologies provide not only standard terminologies for diseases[87] and phenotypic features[74,88] but also come with supporting computational tools for deep phenotyping that allow sophisticated search and analysis routines. For instance, the Human Phenotype Ontology (HPO) is used by many groups to record and analyse the phenotypic features of patients being investigated by genome screening methodologies, including the Sanger Institute's DECIPHER and Deciphering Developmental Disorders projects[38,89], the rare disease section of the UK 100,000 Genomes project, and the Undiagnosed Diseases Program of the National Institutes for Health. Several groups are using HPO terms to enable phenotype-driven analysis of exomes[90–94]. Although numerous individual projects in oncology and complex disease follow deep-phenotyping approaches, it is still relatively common to assign individual patients to broad overall categories for clinical studies. For instance, staging is commonly used in oncology to classify patients into a small number of categories based primarily on the extent of the original tumour and the presence and distribution of metastases. Although this approach is of great utility for clinical management, similarly staged patients often have varied clinical outcomes, suggesting the existence of a spectrum of disease states that are not optimally captured by current staging systems. Similar remarks apply to many complex diseases. For example, current psychiatric diagnostic classifications group together patients that present with a heterogeneous range of phenotypes that likely result from heterogeneous aetiologies[95]. Substantial efforts will be required to develop computational resources, including phenotype ontologies, to adequately capture phenotypes and enable the full interpretation of the clinical consequences of genetic variation that may lead to precision medicine-based stratification and clinical management of disease.

would be needed to ideally cover all aspects of the genotype and phenotype data provenance, governance, quality, ownership, consent, technical backdrop and purpose. This emphasis requires a standard structure for metadata, a minimal set of required fields and systems for metadata production and sharing.

Data formatting and annotation. As more heterogeneous data are exchanged, it will become essential to have reliable and accurate tools for data conversion and annotation support, including format inter-conversion, sequence annotation and phenotype term cross-mapping.

Data discovery. Widespread efforts to promote responsible data sharing are increasingly being enhanced by a focus on data discovery. The latter needs to be emphasized further so that a comprehensive and federated discovery 'ecosystem' emerges. This ecosystem needs to be generated in such a way that it is part and parcel of global infrastructures for data sharing.

Consent agreements. The diversity in formats and many other practical issues make data sharing difficult. For instance, consent agreements can be so diverse that not all samples in a database or biobank are consented for a certain type of study. A core set of standardized consent clauses relating to sharing and discovery would go a long way towards maximizing our ability to aggregate data or samples from multiple sources for integrated studies.

Data analysis and reporting. Similarly, standards for analysing and reporting data quality are essential, covering items such as veracity, benchmarking, comprehensiveness, curation processes, depth of clinical data, and coverage of published and unpublished disease-causing variants.

Identifiers. There is an urgent need for globally accepted identifiers for patients (who may be involved in multiple studies), individual records (as these may

Phenotype term cross-mapping
A computational link between equivalent or related terms in two or more different phenotype ontologies. For instance, the Medical Dictionary for Regulatory Activities (MedDRA) term Platyspondylia (10068629) is mapped to the Human Phenotype Ontology term Platyspondyly (HP:0000926).

ORCID
ORCID provides a persistent digital identifier (for example, orcid.org/0000-0002-0736-9199) for each researcher that can be used to streamline workflows such as manuscript and grant submission and to unambiguously identify researchers in databases.

APIs
(Application programming interfaces). A specification of a software component in terms of functionalities, formats and data types. In the current context, an API is a framework that allows exchange and processing of data and contents between different websites and databases.

Ontologies
Ontologies are computational resources that combine catalogues of the relevant entities of a domain (a conceptualization) with a description of the interrelationships among those entities (a specification).

become replicated upon sharing and then recirculate and be mistaken as being independent observations) and databases. This would be analogous to the DOI system for publications and the ORCID system for researchers, both of which provide IDs that are unique, permanent and readable by humans and computers.

APIs and ontologies. Standard APIs are needed for efficient connection and interaction between databases for these systems to be able to exchange information, and agreements on semantics are needed to specify the meaning of the data. Although some ontologies such as the Human Phenotype Ontology[74] are now widely used in the rare disease community for this purpose, comparable resources need to be developed for oncology and common, complex diseases (BOX 1).

Long-term sustainability. Even though many academic groups develop sophisticated databases and algorithms for human biomedicine, it is often difficult to sustain such resources once initial grants have expired. It remains to be decided whether funding agencies will provide long-term funding for medical databases similar to the funding received by PubMed or some National Center for Biotechnology Information (NCBI) resources such as GenBank, or whether commercialization and for-profit solutions will be required to maintain the most important resources in the future.

Conclusions

Whereas merely a few decades ago datasets were simple and computer systems had little power, in the future it is likely that complete genomes of whole populations with extensive sets of phenotype data will need to be stored and made useful via interoperable systems working across projects and locations. These data may need to be integrated further with additional types of information, such as multi-omics results, environmental contexts, pathway data and model organism comparators, highlighting the increasing need for comprehensive database efforts to record and make widely available the ever-growing datasets.

Some of the resulting needs and requirements are described in this Review, and these should be regarded as necessary but not sufficient for real medical progress. To make best use of data, it will be essential to develop algorithms and applications that share and integrate knowledge derived from the data for the benefit of research and patient care. The next generation of genotype–phenotype databases is therefore likely to include highly aggregated and integrated data, graphical representations to allow humans to explore its complexity, and algorithms that reveal predictive, personalized mathematical models of causality. These databases will thereby integrate deep clinical and molecular findings of patients or research subjects with as much relevant existing data and knowledge as possible to enable reliable and useful diagnoses and discoveries to be made.

1. Johnston, J. J. & Biesecker, L. G. Databases of genomic variation and phenotypes: existing resources and future needs. *Hum. Mol. Genet.* **22**, R27–R31 (2013).
2. Rehm, H. L. Disease-targeted sequencing: a cornerstone in the clinic. *Nat. Rev. Genet.* **14**, 295–300 (2013).
3. Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).
4. Zemojtel, T. *et al.* Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci. Transl. Med.* **6**, 252ra123 (2014).
5. Saunders, C. J. *et al.* Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci. Transl. Med.* **4**, 154ra135 (2012).
6. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
7. Li, M. X. *et al.* Predicting Mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet.* **9**, e1003143 (2013).
8. Pelak, K. *et al.* The characterization of twenty sequenced human genomes. *PLoS Genet.* **6**, e1001111 (2010).
9. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
10. Horaitis, O. *et al.* A database of locus-specific databases. *Nat. Genet.* **39**, 425 (2007).
11. Patrinos, G. P. *et al.* Human Variome Project country nodes: documenting genetic information within a country. *Hum. Mutat.* **33**, 1513–1519 (2012).
12. Fokkema, I. F. *et al.* LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat.* **32**, 557–563 (2011).
13. Beroud, C., Collod-Beroud, G., Boileau, C., Soussi, T. & Junien, C. UMD (Universal Mutation Database): a generic software to build and analyze locus-specific databases. *Hum. Mutat.* **15**, 86–94 (2000).
References 12 and 13 describe the two most highly used software platforms for creating LSDBs and for annotating, analysing and displaying DNA variations in genes.
14. Polvi, A. *et al.* The Finnish disease heritage database (FinDis) update — a database for the genes mutated in the Finnish disease heritage brought to the next-generation sequencing era. *Hum. Mutat.* **34**, 1458–1466 (2013).
15. Dalgleish, R. *et al.* Locus Reference Genomic sequences: an improved basis for describing human DNA variants. *Genome Med.* **2**, 24 (2010).
16. MacArthur, J. A. *et al.* Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Res.* **42**, D873–D878 (2014).
17. Gout, A. M. *et al.* Analysis of published *PKD1* gene sequence variants. *Nat. Genet.* **39**, 427–428 (2007).
18. Bell, C. J. *et al.* Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci. Transl. Med.* **3**, 65ra4 (2011).
19. Chen, S. N. *et al.* Human molecular genetic and functional studies identify *TRIM63*, encoding muscle RING finger protein 1, as a novel gene for human hypertrophic cardiomyopathy. *Circ. Res.* **111**, 907–919 (2012).
20. Ploski, R. *et al.* Does p.Q247X in *TRIM63* cause human hypertrophic cardiomyopathy? *Circ. Res.* **114**, e2–e5 (2014).
21. Witt, C. C. *et al.* Cooperative control of striated muscle mass and metabolism by MuRF1 and MuRF2. *EMBO J.* **27**, 350–360 (2008).
22. Plon, S. E. *et al.* Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum. Mutat.* **29**, 1282–1291 (2008).
23. Sosnay, P. R. *et al.* Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat. Genet.* **45**, 1160–1167 (2013).
24. den Dunnen, J., Cutting, G. R. & Paalman, M. H. Mandatory variant submission — our experiences. *Hum. Mutat.* **33**, 1 (2012).
25. Terry, S. F. Disease advocacy organizations catalyze translational research. *Front. Genet.* **4**, 101 (2013).
26. Wicks, P. *et al.* Sharing health data for better outcomes on PatientsLikeMe. *J. Med. Internet. Res.* **12**, e19 (2010).
27. Kirkpatrick, B. E. *et al.* GenomeConnect: matchmaking between patients, clinical laboratories and researchers to improve genomic knowledge. *Hum. Mutat.* **36**, 974–978 (2015).
28. McAllister, M. & Dearing, A. Patient reported outcomes and patient empowerment in clinical genetics services. *Clin. Genet.* **88**, 114–121 (2015).
29. The Lancet Editorial. Patient empowerment — who empowers whom? *Lancet* **379**, 1677 (2012).
30. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
31. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
32. Ball, M. P. *et al.* A public resource facilitating clinical use of genomes. *Proc. Natl Acad. Sci. USA* **109**, 11920–11927 (2012).
33. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal — a one-stop shop for cancer genomics data. *Database (Oxford)* **2011**, bar026 (2011).
The ICGC data portal provides tools for visualizing, querying and downloading an immense amount of data from the ICGC, with an innovative approach to federating data and annotations across numerous participating centres.
34. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
35. Petrovski, S. & Goldstein, D. B. Phenomics and the interpretation of personal genomes. *Sci. Transl. Med.* **6**, 254fs35 (2014).
36. Dorschner, M. O. *et al.* Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am. J. Hum. Genet.* **93**, 631–640 (2013).
This report convincingly demonstrates that there is often insufficient evidence for pathogenicity of many variants reported in databases or in the medical literature.
37. Dyment, D. A. *et al.* Whole-exome sequencing broadens the phenotypic spectrum of rare pediatric epilepsy: a retrospective study. *Clin. Genet.* **88**, 34–40 (2015).

38. Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
DECIPHER is the largest publicly available database of genotypic and phenotypic data of mainly undiagnosed patients with rare diseases. DECIPHER has a large suite of tools to facilitate the interpretation of candidate variants.
39. Gonzalez, M. A. *et al.* GEnomes Management Application (GEM.app): a new software tool for large-scale collaborative genome analysis. *Hum. Mutat.* **34**, 842–846 (2013).
40. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
41. Schaefer, C. & RPGEH GO Project Collaboration. The Kaiser Permanente Research Program on Genes, Environment and Health: a resource for genetic epidemiology in adult health and aging. *Clin. Med. Res.* **9**, 177–178 (2011).
42. Lappalainen, I. *et al.* The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* **47**, 692–695 (2015).
43. Tryka, K. A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.* **42**, D975–D979 (2014).
44. Kohane, I. S. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* **12**, 417–428 (2011).
45. Murphy, S. N. *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc.* **17**, 124–130 (2010).
46. Adamusiak, T. *et al.* Observ-OM and Observ-TAB: universal syntax solutions for the integration, search, and exchange of phenotype and genotype information. *Hum. Mutat.* **33**, 867–873 (2012).
47. Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C. & Brookes, A. J. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur. J. Hum. Genet.* **22**, 949–952 (2014).
The largest publicly available compilation of summary-level findings from genetic association studies. Together with references 50 and 111–113, this provides alternative ways of searching and visualizing GWAS data.
48. Gaye, A. *et al.* DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int. J. Epidemiol.* **43**, 1929–1944 (2014).
49. Karr, A. *et al.* Secure, privacy-preserving analysis of distributed databases. *Technometrics* **49**, 335–345 (2007).
50. Cariaso, M. & Lennon, G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res.* **40**, D1308–D1312 (2012).
51. Rappaport, N. *et al.* MalaCards: an integrated compendium for diseases and their annotation. *Database (Oxford)* **2013**, bat018 (2013).
52. Lopes, P., Dalgleish, R. & Oliveira, J. L. WAVe: web analysis of the variome. *Hum. Mutat.* **32**, 729–734 (2011).
53. Glusman, G., Caballero, J., Mauldin, D. E., Hood, L. & Roach, J. C. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* **27**, 3216–3217 (2011).
54. Philippakis, A. A. *et al.* The matchmaker exchange: a platform for rare disease gene discovery. *Hum. Mutat.* **36**, 915–921 (2015).
55. Manolio, T. A. *et al.* Global implementation of genomic medicine: we are not alone. *Sci. Transl. Med.* **7**, 290ps13 (2015).
56. Hayden, E. C. Geneticists push for global data-sharing. *Nature* **498**, 16–17 (2013).
A report on the founding of the GA4GH.
57. Gottlieb, M. M. *et al.* GeneYenta: a phenotype-based rare disease case matching tool based on online dating algorithms for the acceleration of exome interpretation. *Hum. Mutat.* **36**, 432–438 (2015).
58. Lancaster, O. *et al.* Cafe Variome: general-purpose software for making genotype–phenotype data discoverable in restricted or open access contexts. *Hum. Mutat.* **36**, 957–964 (2015).
59. Wellcome Trust. Enhancing discoverability of public health and epidemiology research data. *Wellcome Trust* [online], http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTP054675.htm (2014).
60. Digital Curation Centre (DCC). Jisc Research Data Registry and Discovery Service *DCC* http://www.dcc.ac.uk/projects/research-data-registry-pilot (2014).
61. Cotton, R. G. *et al.* The Human Variome Project. *Science* **322**, 861–862 (2008).
62. Rehm, H. L. *et al.* ClinGen — the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
ClinGen is a National Institutes of Health-funded resource, building an authoritative knowledge base that promotes evidence-based clinical annotation and interpretation of genomic variants. ClinVar (reference 61), an active partner of the ClinGen project, is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence.
63. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
64. Beaulieu, C. L. *et al.* FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *Am. J. Hum. Genet.* **94**, 809–817 (2014).
65. Buske, O. J. PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. *Hum. Mutat.* **36**, 931–940 (2015).
66. Thompson, R. *et al.* RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. *J. Gen. Intern. Med.* **29**, S780–S787 (2014).
67. Conley, J. M., Cook-Deegan, R. & Lázaro-Muñoz, G. Myriad after *Myriad*: the proprietary data dilemma. *N. C. J. Law Technol.* **15**, 597–637 (2014).
68. Riggs, E. R., Jackson, L., Miller, D. T. & Van Vooren, S. Phenotypic information in genomic variant databases enhances clinical care and research: the International Standards for Cytogenomic Arrays Consortium experience. *Hum. Mutat.* **33**, 787–796 (2012).
69. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
70. Banda, Y. *et al.* Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. *Genetics* **200**, 1285–1295 (2015).
71. Knoppers, B. M. Framework for responsible sharing of genomic and health-related data. *HUGO J.* **8**, 3 (2014).
A report on an ethical framework for responsible data sharing developed in conjunction with a wide spectrum of the bioethics, genomics and clinical communities, under the auspices of the GA4GH.
72. Mascalzoni, D. *et al.* International Charter of principles for sharing bio-specimens and data. *Eur. J. Hum. Genet.* **23**, 721–728 (2015).
73. Rath, A. *et al.* Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum. Mutat.* **33**, 803–808 (2012).
Orphanet is a portal for rare disease and orphan drugs that provides an inventory of rare diseases and a classification system that serves as a model for updating international terminologies such as the International Classification of Diseases.
74. Köhler, S. *et al.* The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, D966–D974 (2014).
A report on the Human Phenotype Ontology, a widely used standard for annotating and analysing phenotypic abnormalities in diagnostic and translational research settings.
75. Varga, E. A. & Moll, S. Cardiology patient pages. Prothrombin 20210 mutation (factor II mutation). *Circulation* **110**, e15–e18 (2004).
76. Beaudet, A. L. & Tsui, L. C. A suggested nomenclature for designating mutations. *Hum. Mutat.* **2**, 245–248 (1993).
77. den Dunnen, J. T. & Antonarakis, S. E. Nomenclature for the description of human sequence variations. *Hum. Genet.* **109**, 121–124 (2001).
An initial description of the Human Genome Variation Society's nomenclature standard for naming sequence variants.
78. Taschner, P. E. & den Dunnen, J. T. Describing structural changes by extending HGVS sequence variation nomenclature. *Hum. Mutat.* **32**, 507–511 (2011).
79. Laros, J. F., Blavier, A., den Dunnen, J. T. & Taschner, P. E. A formalized description of the standard human variant nomenclature in Extended Backus-Naur Form. *BMC Bioinformatics* **12**, S5 (2011).
80. Wildeman, M., van Ophuizen, E., den Dunnen, J. T. & Taschner, P. E. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum. Mutat.* **29**, 6–13 (2008).
81. Hart, R. K. *et al.* A Python package for parsing, validating, mapping and formatting sequence variants using HGVS nomenclature. *Bioinformatics* **31**, 268–270 (2015).
82. Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771 (2010).
83. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
84. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
85. Byrne, M. *et al.* VarioML framework for comprehensive variation data representation and exchange. *BMC Bioinformatics* **13**, 254 (2012).
86. Vihinen, M. Variation Ontology for annotation of variation effects and mechanisms. *Genome Res.* **24**, 356–364 (2014).
87. Kibbe, W. A. *et al.* Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* **43**, D1071–D1078 (2015).
88. Groza, T. *et al.* The Human Phenotype Ontology: semantic unification of common and rare disease. *Am. J. Hum. Genet.* **97**, 111–124 (2015).
89. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).
90. Sifrim, A. *et al.* eXtasy: variant prioritization by genomic data fusion. *Nat. Methods* **10**, 1083–1084 (2013).
91. Robinson, P. N. *et al.* Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.* **24**, 340–348 (2014).
92. Singleton, M. V. *et al.* Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am. J. Hum. Genet.* **94**, 599–610 (2014).
93. Javed, A., Agrawal, S. & Ng, P. C. Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat. Methods* **11**, 935–937 (2014).
94. Westbury, S. K. *et al.* Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Med.* **7**, 36 (2015).
95. Adam, D. Mental health: on the spectrum. *Nature* **496**, 416–418 (2013).
96. Stenson, P. D. *et al.* Human Gene Mutation Database: towards a comprehensive central mutation database. *J. Med. Genet.* **45**, 124–126 (2008).
97. Abel, O., Powell, J. F., Andersen, P. M. & Al-Chalabi, A. ALSoD: a user-friendly online bioinformatics tool for amyotrophic lateral sclerosis genetics. *Hum. Mutat.* **33**, 1345–1351 (2012).
98. Chandrasekharappa, S. C. *et al.* Massively parallel sequencing, aCGH, and RNA-seq technologies provide a comprehensive molecular diagnosis of Fanconi anemia. *Blood* **121**, e138–e148 (2013).
99. Dalgleish, R. The human type I collagen mutation database. *Nucleic Acids Res.* **25**, 181–187 (1997).
100. Piirila, H., Valiaho, J. & Vihinen, M. Immunodeficiency mutation databases (IDbases). *Hum. Mutat.* **27**, 1200–1208 (2006).
101. Ruiz-Pesini, E. *et al.* An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res.* **35**, D823–D828 (2007).
102. Papadopoulos, P. *et al.* Developments in FINDbase worldwide database for clinically relevant genomic variation allele frequencies. *Nucleic Acids Res.* **42**, D1020–D1026 (2014).
103. Bragin, E. *et al.* DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res.* **42**, D993–D1000 (2014).
104. Hamosh, A. *et al.* PhenoDB: a new web-based tool for the collection, storage, and analysis of phenotypic features. *Hum. Mutat.* **34**, 566–571 (2013).

105. Sobreira, N., Schiettecatte, F., Valle, D. & Hamosh, A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum. Mutat.* **36**, 928–930 (2015).

106. Amberger, J., Bocchini, C. & Hamosh, A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Hum. Mutat.* **32**, 564–567 (2011).
OMIM is one of the oldest and most important knowledge bases in human medicine, going back to work initiated in the early 1960s by Victor McKusick. In addition to 12 book editions of Mendelian Inheritance in Man (MIM) between 1966 and 1998, the online version (OMIM) has been available since 1987.

107. Mungall, C.J. *et al.* Use of model organism and disease databases to support matchmaking for human disease gene discovery. *Hum. Mutat.* **36**, 979–984 (2015).

108. Wilks, C. *et al.* The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database (Oxford)* **2014**, bau093 (2014).

109. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).

110. Cheng, W. C. *et al.* DriverDB: an exome sequencing database for cancer driver gene identification. *Nucleic Acids Res.* **42**, D1048–D1054 (2014).

111. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).

112. Li, M. J. *et al.* GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **40**, D1047–D1054 (2012).

113. Koike, A., Nishida, N., Inoue, I., Tsuji, S. & Tokunaga, K. Genome-wide association database developed in the Japanese Integrated Database Project. *J. Hum. Genet.* **54**, 543–546 (2009).

114. Whirl-Carrillo, M. *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92**, 414–417 (2012).

115. Carey, J. C., Allanson, J. E., Hennekam, R. C. & Biesecker, L. G. Standard terminology for phenotypic variations: The elements of morphology project, its current progress, and future directions. *Hum. Mutat.* **33**, 781–786 (2012).

Acknowledgements

Preparation of this article was facilitated by funding from the European Union Seventh Framework Programme (FP7/2007-2013; 'BioShaRE' grant no. 261433, 'SYBIL' grant no. 602300, 'EMIF' IMI-JU grant no. 115372), the National Institutes of Health Office of the Director (grant no. 5R24OD011883), and the Bundesministerium für Bildung und Forschung (BMBF; project no. 0313911). The authors also acknowledge the many key insights provided by attendees of an IRDiRC workshop dedicated to this topic, and expert suggestions made by colleague R. Dalgleish (University of Leicester, UK).

Competing interests statement

The authors declare no competing interests.

DATABASES

Amyotrophic Lateral Sclerosis Online Genetics Database: http://alsod.iop.kcl.ac.uk
Cancer Genomics Hub: https://cghub.ucsc.edu
Catalogue of Somatic Mutations in Cancer: http://cancer.sanger.ac.uk/cancergenome/projects/cosmic
CFTR2 database: http://www.cftr2.org
ClinVar: http://www.ncbi.nlm.nih.gov/clinvar
Database of Genotypes and Phenotypes: http://www.ncbi.nlm.nih.gov/gap
DECIPHER: https://decipher.sanger.ac.uk
DriverDB: http://driverdb.ym.edu.tw/DriverDB
ETHNOS databases: http://www.findbase.org/ethnos-databases
European Genome-Phenome Archive: https://www.ebi.ac.uk/ega
European Variation Archive: http://www.ebi.ac.uk/eva
Fanconi Anemia Mutation Database: http://www.rockefeller.edu/fanconi
FINDbase: http://www.findbase.org
GeneMatcher: https://genematcher.org
GWAS Catalog: http://www.genome.gov/gwastudies
GWAS Central: http://www.gwascentral.org
GWASdb: http://jjwanglab.org/gwasdb
Human Gene Mutation Database: http://www.hgmd.cf.ac.uk/ac/index.php
Human Genome Variation Database: http://gwas.biosciencedbc.jp
IDbases: http://structure.bmc.lu.se/idbase
Leiden Open Variation Database: http://www.lovd.nl
MITOMAP: http://www.mitomap.org
Online Mendelian Inheritance in Man: http://www.omim.org
Orphanet: http://www.orpha.net
Osteogenesis Imperfecta Variant Database: https://oi.gene.le.ac.uk
PharmacoGenomics Database: http://www.pharmgkb.org
PhenoDB: https://phenodb.org
PhenomeCentral: https://phenomecentral.org
PheWAS Catalog: https://phewas.mc.vanderbilt.edu
Universal Mutation Database: http://www.umd.be

FURTHER INFORMATION

1000 genomes: http://www.1000genomes.org
Beacon project: https://beacon-network.org
Café Variome Central: http://central.cafevariome.org
CARE4RARE: http://care4rare.ca
ClinGen: https://www.clinicalgenome.org
Collaborative Cancer Cloud: https://communities.intel.com/community/itpeernetwork/healthcare/blog/2015/08/17/intel-ohsu-announce-collaborative-cancer-cloud-at-intel-developers-forum
DataSHIELD: http://www.datashield.ac.uk/
Exome Aggregation Consortium: http://exac.broadinstitute.org
Exome Variant Server of the NHLBI Exome Sequencing Project: http://evs.gs.washington.edu/EVS
GA4GH BRCA Challenge: https://genomicsandhealth.org/work-products-demonstration-projects/brca-challenge-0
GeneYenta: https://geneyenta.com
GenomeConnect: https://www.clinicalgenome.org/genomeconnect
GEnomes Management Application: https://genomics.med.miami.edu/gem-app
Global Alliance for Genomics and Health: http://genomicsandhealth.org
Human Genome Variation Society Nomenclature: http://www.hgvs.org/mutnomen
Human Phenotype Ontology: http://www.human-phenotype-ontology.org
Human Variome Project: http://www.humanvariomeproject.org
International Cancer Genome Consortium: https://icgc.org
International Rare Diseases Research Consortium: http://www.irdirc.org
Kaiser Permanente Research Program on Genes, Environment and Health: http://www.dor.kaiser.org/external/DORExternal/rpgeh/index.aspx
Kaviar: http://db.systemsbiology.net/kaviar
Locus Reference Genomic: http://www.lrg-sequence.org
MalaCards: http://www.malacards.org
MatchMaker Exchange: http://www.matchmakerexchange.org
Monarch Initiative: http://monarchinitiative.org
Mutalyzer: https://mutalyzer.nl
National Institutes of Health Data Sharing Policy: http://grants.nih.gov/grants/sharing.htm
PatientsLikeMe: https://www.patientslikeme.com
PEER platform: http://www.geneticalliance.org/programs/biotrust/peer
Personal Genome Project: http://www.personalgenomes.org
RD-Connect: http://rd connect.eu
SNPedia: http://www.snpedia.com
The Cancer Genome Atlas: http://cancergenome.nih.gov
VariO: http://variationontology.org
VarioML: http://www.varioml.org
WAVe: http://bioinformatics.ua.pt/WAVe

ALL LINKS ARE ACTIVE IN THE ONLINE PDF