Biomedicine Hub

Biomed Hub 2017;2(suppl 1):481682 (DOI: 10.1159/000481682)

Received: September 19, 2017
Accepted: September 20, 2017
Published online: November 21, 2017

Article

From the Data on Many, Precision Medicine for "One": The Case for Widespread Genomic Data Sharing

Serena Scollen[a, d] Angela Page[c, d] Julia Wilson[b, d]
on behalf of the Global Alliance for Genomics and Health

[a]ELIXIR Hub, Cambridge, and [b]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, UK; [c]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [d]Global Alliance for Genomics and Health, Toronto, ON, Canada

Keywords
Precision medicine · Genome sequencing · Genomic medicine · FAIR · Big data · Rare disease · Cancer · Global Alliance for Genomics and Health · ELIXIR

Abstract
Within the decade, genome sequencing promises to become a routine part of healthcare around the globe. Many millions of genomes linked to health records will soon be available for researchers and clinicians to make use of to advance precision medicine. To realise the full impact of genomic medicine, genomic and clinical data must be interoperable across traditional geographic, jurisdictional, sectoral, and domain boundaries. Extremely large and diverse data sets are needed to provide a context for interpretation of genetic sequences. No single country or institution can achieve the necessary scale and diversity alone. Data must be shared within an internationally federated, learning health system.

Introduction

Many countries around the globe have recently launched national genomic medicine initiatives that aim to both transform the way people are cared for and enable new medical research. By combining genomic sequencing data with health records, researchers and clinicians will have a vast resource that can be interpreted to improve patient outcomes and also be used to investigate the causes, diagnoses, and treatments of a disease.

Serena Scollen
ELIXIR Hub
South Building, Wellcome Genome Campus
Hinxton, Cambridge CB10 1SD (UK)
E-Mail serena.scollen@elixir-europe.org

This precision medicine "movement" is gaining momentum thanks to the rapid decline in the price of high-throughput genome sequencing, which is slated to reach the storied USD 100 range within the decade. This milestone will change the face of healthcare, with every patient and many healthy citizens knowing each A, C, T, and G that makes him or her unique. But for the precision medicine model to be truly successful for each individual, their collective data must be interoperable across traditional geographic, jurisdictional, sectoral, and domain boundaries. That's because a single genetic sequence is less meaningful outside the context of the broader population. For common complex disorders it is crucial to study hundreds of thousands of individuals – the greater the scale, the greater the understanding. For rare disorders it is often crucial to work on an international scale, identifying patients with similar conditions from all over the globe.

The genomic community has always been a leader in "FAIR" (Findable, Accessible, Interoperable, Reusable) principles, making well-annotated data available before publication [1]. The challenge is to preserve this great tradition within a framework that safeguards patients and study participants. Ideally, the clinical and research communities would have access to enormous FAIR data sets consisting of genome sequences and non-identifiable, relevant health information from many millions of people across a range of ethnic backgrounds. No single country or institution – be it academic, governmental, or commercial – is capable of achieving that scale and diversity alone. It is therefore critical – for both the health and knowledge economy agendas of precision medicine initiatives around the globe – that data are shared within an internationally federated, learning health system.

ELIXIR is an initiative that unites leading life science organisations around Europe in managing and safeguarding the growing volume of data generated by publicly funded research. It coordinates, integrates, and sustains bioinformatic resources across the region and enables users in academia and industry to access services that are vital for their research.

As a distributed infrastructure, the services run within ELIXIR are operated by ELIXIR Nodes, which comprise over 180 leading universities and centres of excellence across Europe [2]. The services run within ELIXIR include databases (both deposition databases and knowledge bases); analysis tools and software; interoperability services and expertise in FAIR data; computer services; and training. While ELIXIR's services cover life science data broadly, the model, remit, and learnings are particularly relevant to the genomics community.

ELIXIR is a continental example of what is needed across the globe: a federated research infrastructure of interoperable data sets. The Global Alliance for Genomics and Health (GA4GH) is beginning to enable that international ecosystem by creating the standards and tools required to share genomic and health data responsibly across national and other boundaries [3]. Founded in 2013, GA4GH includes more than 500 organisational members and more than 1,000 individual members, all committed to responsible genomic and health-related data sharing.

GA4GH takes a human rights approach to sharing data, stating in its *Framework for the Responsible Sharing of Genomic and Health-Related Data* that data sharing should be conducted "with a view towards minimizing harms and maximizing benefits to not just those who contribute their data, but also to society and healthcare systems as a whole" [4]. There is a fundamental human right to benefit from the fruits of research and care-based genomic sequencing. Both patients and the public deserve a greater level of education around the value of sharing and linking genomic data with health records.

GA4GH is guided by the explicit needs of the genomic medicine community and has convened the leading organisations in research, healthcare, and industry from around the globe to help guide its work. Groups such as ICGCmed, the Australian Genomics Health Alliance, and Genomics England have committed to the organisation and suggested real-world issues to tackle as GA4GH "Driver Projects," which work with the community's tech-

nical experts to create the policies, standards and scalable tools needed to make data interoperable across geographic, institutional, and sectoral boundaries within 5 years' time.

Already, GA4GH has developed several tools to enable data sharing, including the Beacon API, which allows institutions to make the contents of their data sets discoverable to any user of the Web. More than 25 organisations and institutions have implemented beacons since the project's launch in 2014, including 7 at institutions connected through ELIXIR. In total, there are now more than 60 beacons that make over 200 data sets discoverable (in sum, data on more than 100,000 individuals). This project has helped to challenge researchers' attitudes to sharing data whilst developing a secure platform to do so. The next step beyond simplistic discovery is to help researchers to more quickly identify data sets that they would want to access for reuse, for example, by making some level of metadata available.

Genome Sequencing in the Research Setting

To date, most genome sequences have been collected in the research setting. Since the completion of the Human Genome Project in 2001 (after a 10-year effort), genomic sequencing has skyrocketed, with estimates ranging from 100 million to as many as 2 billion human genomes expected to be sequenced by 2025 [5]. From its earliest days, genomic research has been a collaborative and international effort. The Wellcome Trust Sanger Institute has been a leader in some of the most significant initiatives in genome sequencing over the last 25 years. Building from the original genomic data sharing projects such as the International HapMap Project, the 1,000 Genomes Project, and UK10K, Sanger has made thousands of genomes from patients and healthy individuals available to the public.

The Sanger Institute makes most of its data open source and available, primarily through dedicated data services led by the European Bioinformatics Institute (EBI), one of the largest Nodes in ELIXIR. The EGA, which is jointly run by the EBI and the Centre for Genomic Regulation (Centre de Regulació Genòmica [CRG]), provides a service for the permanent archiving and distribution of personally identifiable genetic and phenotypic data that results from biomedical research projects. CRG archived data from individuals who consented to specific use by bona fide researchers. The organisation has strict protocols for how the information is managed, stored, and distributed. Dedicated repositories such as EGA and the federation thereof are two models that must and will coexist in this domain, and expertise from the former is increasingly being used to enable the latter. Today, established cohorts are being retrofitted toward FAIR-ness – making existing data even more accessible while the community simultaneously moves future research forward at even greater speed. A good example is the UK Biobank, which recently shared anonymised health information on more than 500,000 individuals via the European Genome-phenome Archive (EGA) [6].

The rare disease community provides a pioneering example of the promise of an internationally federated ecosystem for sharing that enables researchers from around the globe to discover and access data from any institution in the world. In rare disease diagnosis, the challenge is not identifying candidate genes for a pathologic phenotype, but rather confirming such candidates as culpable. Outside of lengthy, resource-intensive functional studies, the only way to do so is by identifying another individual with the same phenotype and candidate gene; this requires extremely large data sets – haystacks worthy of a search for the proverbial needle.

ELIXIR has prioritised rare disease as a life science area in order to bring together experts to develop specialised standards, services, and workshops. Ongoing studies include building a portal through which authorised researchers can access rare disease data from repositories and catalogues around Europe; building a portfolio of data resources and analytic tools is

critical for this community. The goal is to create a federated infrastructure that enables researchers to discover, access, and analyse data in rare disease repositories across Europe. It is doing this in partnership with other European infrastructures and projects, namely, RD-Connect and BBMRI. The Deciphering Developmental Disorders study, led by scientists at the Sanger Institute, brought together genomic data from 14,000 families around the UK – the largest-ever genetic study of children with previously undiagnosed rare developmental disorders. It has provided diagnoses for around 40% of the nearly 4,000 families analysed to date in the programme and identified several previously undocumented syndromes. Data from the study are stored in the DECIPHER database, which helps the clinical community share and compare human genome variants and phenotypes.

With nearly 2,000 new rare disease diagnoses per year in UK children alone, it is clear that even larger and more comprehensive data sets will be necessary to fully address the needs of rare disease patients and catalyse additional research. This requires rare disease data holders around the globe to share data across institutional and national boundaries. Several efforts to link disparate rare disease databases (including the Sanger Institute's DECIPHER) have emerged over the past 4 years, and a service supported by GA4GH and the International Rare Diseases Research Consortium (IRDiRC) called Matchmaker Exchange attempts to federate all of these into a single resource for rare disease gene matching [7]. These efforts have already led to diagnoses that were previously intractable using traditional approaches.

Genomic data sharing will be invaluable for improving cancer prevention, diagnosis, and treatment. Analysis of somatic genomes has been used to identify sequence variants and mutations critical for the development of human cancers. The Cancer Genome Project, The Cancer Genome Atlas (TCGA), and the International Cancer Genome Consortium (ICGC) have collectively documented the genomic changes present in different types and subtypes of cancer. Databases linking these data on molecular changes in particular cancers to drug sensitivity information and patient outcomes will ultimately realise greater clinical utility. For example, information on somatic mutations can be located in the COSMIC database, and the NCI Genomic Data Commons provides the cancer research community with a unified data repository allowing data sharing across cancer genomic studies. Databases that document the genetic heterogeneity and complexity of human cancer will realise their potential but will undoubtedly require international collaboration and interoperability.

Project GENIE (Genomics Evidence Neoplasia Information Exchange) of the American Association for Cancer Research is an international data-sharing project for precision oncology. Similar to Matchmaker Exchange, Project GENIE is creating a federated registry of clinical-grade cancer genomic data and outcomes from tens of thousands of cancer patients treated around the globe. The goal is to enable clinical data to inform a more robust cancer research endeavour and, eventually, to improve prevention, diagnosis, and treatment for future patients. Project GENIE brings together 8 clinical partners and includes genomic data from nearly 19,000 de-identified patient records.

Genome Sequencing in Healthcare

Building on the learnings from research-level efforts like Project GENIE and others mentioned above, national governments around the globe are beginning to see the value that genomic and clinical data contained in their citizens' health records can have for improved patient outcomes. Pioneering initiatives such as Genomics England's 100,000 Genomes Project and the Australian Genomics Health Alliance (Australian Genomics) are integrating genome sequencing into routine healthcare for the long-term benefit of human health. Both

are focusing initially on rare diseases and cancer. The 100,000 Genomes Project will ultimately sequence whole genomes from around 70,000 individuals – patients in the UK NHS system with a rare disease, plus their families and patients with cancer. The UK Chief Medical Officer annual report *Generation Genome* released in 2017 defines how the NHS should build on the 100,000 Genomes Project to deliver the promise of precision medicine [8].

Australian Genomics has published health economic data demonstrating the value for childhood syndrome patients seen at the Royal Children's Hospital of Melbourne, where the use of whole-exome sequencing increased the diagnostic rate 5-fold over traditional diagnostic tests while decreasing the average cost per patient by 75% [9]. Additionally, 4 times more patients saw improvements in their medical care.

The US National Institutes of Health's All of Us research program will gather new and existing genomic data into a cohort of 1 million or more individuals to advance precision medicine. The project is effectively aiming to build a formal national "learning health system" by leveraging healthcare and other data to create a research platform that feeds learnings back into medicine.

While some countries, such as the USA, are quickly trying to find solutions to their electronic health record systems, others, such as Estonia, Finland, and Iceland, already have easy-to-access health record data built in. This has permitted the genomic sequencing of large portions of their populations. For instance, the deCODE project, launched over two decades ago, has collected genomic and health-related data from more than 160,000 volunteers – more than half of the Icelandic adult population. Additional projects are underway in South America, Asia, Africa, and elsewhere, ranging from population cohorts to disease-focused precision medicine initiatives; many have implemented beacons to make their data discoverable to the global community.

GA4GH is working with these groups, which all support a culture of sharing best practices in order to improve their own outcomes and not reinvent the wheel with the launch of each new initiative. Although the many regulatory and ethical issues of implementing genomics in healthcare must be addressed within individual national contexts, each faces similar challenges. By working together, these groups believe they can transform healthcare for the benefit of both human health and their nations as a whole.

By the same token, aggregating data across nations is the only way to create a reference that is truly representative of human diversity. Currently, 81% of the genomes sequenced around the globe come from people of European descent – or, put another way, data on individuals from a country such as India, which represents 20% of the human population, contributes only 0.2% to genomic sequencing overall [10]. The Asian Genome Project and Global Gene Corp are two efforts to overcome this challenge by opening up data from the most populated continent on the planet.

Genome Sequencing and Changing Attitudes to Data Sharing in Pharma

To overcome the huge attrition in failed drug development programmes, one strategy pharmaceutical companies have employed is to identify and prioritise targets with compelling human genetic evidence associated with the therapeutic area of interest, either in drug development or drug repositioning. From the analysis of drug development pipelines, targets with human genetic data are more likely to succeed [11, 12]. There are many examples of success, both prospective and retrospective. For instance, ivacaftor was developed to treat cystic fibrosis, and a specific genotype (CTFR G551D) was identified by family-based sequencing and linkage studies. Denosumab was developed to treat osteoporosis, and a non-coding allele (TNFSF11) has been discovered by genome-wide association studies [13]. However, compared

to the advances in the genomics field in the last decade, these examples should be increasing at a higher rate than they are. One way to achieve this is through precompetitive data sharing and increased access to genomic data from around the globe.

Companies need access to genetic evidence including large data sets for a wide range of therapeutic areas of interest (e.g., cardiovascular diseases, diabetes, autoimmune diseases, etc.). Beyond that, they then need to identify the causal genes/variants underlying the associations and the causal mechanisms and biological pathways, all within a time frame that can impact a drug development programme. This could be 10 years from discovery to market, with genetic evidence ideally feeding in as early as possible. In addition, it is key for companies to be able to select the right patients for the right clinical trial, and genome sequencing can aid in this.

There are still huge unmet challenges, which are essentially the same as those for the wider genetics community. The infrastructure currently does not exist to generate, access, and analyse data all at scale, or to identify and reaccess patients and biospecimens efficiently. No one company can achieve this alone. Companies are realising that to move the field forward they will need to work precompetitively and share data where possible. Efforts are underway to do this via the Innovative Medicines Initiative or other public-private initiatives. For example, Open Targets sets out to transform drug discovery by permitting the systematic identification and prioritisation of targets. This is a collaboration between Biogen, GSK, EMBL-EBI, and Sanger. The benefit to the scientific community is that they are committed to data sharing through publications and databases.

International Data Sharing Is Necessary for the Full Realisation of Precision Medicine

Within the decade, genome sequencing promises to become a routine part of healthcare around the globe, with groups like H3ABioNet building the capacity for researchers in all geographies to participate. Many millions of genomes will be available and will collectively yield an invaluable resource that is only as good as the sum of its parts. The industry of precision medicine that will inevitably emerge as a result of this shift is fundamentally dependent on effective, responsible data sharing. Initiatives such as ELIXIR and GA4GH are enabling the infrastructure and the culture that are required for such sharing to be successful.

Disclosure Statement

The authors have no conflicts of interest to declare. The founding sponsors had no role in the writing of the manuscript and in the decision to publish.

References

1 Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.
2 ELIXIR: The ELIXIR Nodes. https://www.elixir-europe.org/about/elixir-nodes (accessed July 28, 2017).

3 Global Alliance for Genomics and Health, Page A, Baker D, Bobrow M, Boycott K, Burn J, Chanock S, et al: GENOMICS. A federated ecosystem for sharing genomic, clinical data. Science 2016;352:1278–1280.
4 Global Alliance for Genomics and Health: Framework for Responsible Sharing of Genomic and Health-Related Data. Toronto, GA4GH, 2014. https://genomicsandhealth.org/framework (accessed July 28, 2017).
5 Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE: Big data: astronomical or genomical? PLoS Biol 2015;13:e1002195.
6 EMBL-EBI: UK Biobank partners with the EGA. July 19, 2017. https://www.ebi.ac.uk/about/news/press-releases/ukbiobank-genetic-data-ega (accessed July 28, 2017).
7 Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al: The Matchmaker Exchange: a platform for rare disease gene discovery. Hum Mutat 2015;36:915–921.
8 Davies SC: Annual Report of the Chief Medical Officer 2016. London, Department of Health, Generation Genome, 2017. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/631043/CMO_annual_report_generation_genome.pdf (accessed July 28, 2017).
9 Stark Z, Schofield D, Alam K, Wilson W, Mupfeki N, Macciocca I, Shrestha R, White SM, Gaff C: Prospective comparison of the cost-effectiveness of clinical whole-exome sequencing with that of usual care overwhelmingly supports early use and reimbursement. Genet Med 2017;19:867–874.
10 Popejoy AB, Fullerton SM: Genomics is failing on diversity. Nature 2016;538:161–164.
11 Cook D, Brown D, Alexander R, March R, Morgan P, Satterthwaite G, Pangalos MN: Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. Nat Rev Drug Discov 2014;13:419–431.
12 Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, Whittaker JC, Sanseau P: The support of human genetic evidence for approved drug indications. Nat Genet 2015;47:856–860.
13 Plenge RM, Scolnick EM, Altshuler D: Validating therapeutic targets through human genetics. Nat Rev Drug Discov 2013;12:581–594.