A Review of Recent Advances in Translational Bioinformatics: Bridges from Biology to Medicine

J. Vamathevan, E. Birney
European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

Summary

Objectives: To highlight and provide insights into key developments in translational bioinformatics between 2014 and 2016.

Methods: This review describes some of the most influential bioinformatics papers and resources that have been published between 2014 and 2016 as well as the national genome sequencing initiatives that utilize these resources to routinely embed genomic medicine into healthcare. Also discussed are some applications of the secondary use of patient data followed by a comprehensive view of the open challenges and emergent technologies.

Results: Although data generation can be performed routinely, analyses and data integration methods still require active research and standardization to improve streamlining of clinical interpretation. The secondary use of patient data has resulted in the development of novel algorithms and has enabled a refined understanding of cellular and phenotypic mechanisms. New data storage and data sharing approaches are required to enable diverse biomedical communities to contribute to genomic discovery.

Conclusion: The translation of genomics data into actionable knowledge for use in healthcare is transforming the clinical landscape in an unprecedented way. Exciting and innovative models that bridge the gap between clinical and academic research are set to open up the field of translational bioinformatics for rapid growth in a digital era.

Keywords

Translational bioinformatics; genomic medicine; genome sequencing; precision medicine

Yearb Med Inform 2017:178-87
http://dx.doi.org/10.15265/IY-2017-017
Published online August 18, 2017

Introduction

There has been a remarkable shift over the last decade in the costs of generating molecular measurements – most notably DNA sequencing but also transcriptomes, proteomes, and metabolomes [1, 2]. Most impacted by this change is the area of genomic medicine where it is now possible to move from generating reference or population level data to producing data from individuals. Concurrently, there have also been advances in the development of new algorithms and tools to integrate and interpret this data.

Bioinformatics is classically defined as the storage, analysis, and interpretation of biological data [3D]. In the 2016 edition of the IMIA Yearbook of Medical Informatics, Russ Altman summarized the evolution of the term Translational BioInformatics (TBI) from 2004, when the term 'biomedical informatics' was used. This term has come to describe the creation of informatics methods that may include the biological world (including DNA, RNA, proteins, small molecules, cells), and the clinical world (including patients, diagnoses, signs, symptoms) [4], while TBI is defined as the translation of basic capabilities and discoveries provided by informatics methods into clinically useful tools.

Previous reviews have classified translational bioinformatics into four themes: (i) linkage of Electronic Health Records (EHRs) to biobanks for genomic discovery, (ii) adoption of genomics and pharmacogenomics into routine clinical care, (iii) use of genomics in drug discovery and drug repositioning, and (iv) personal genomic testing [5, 6]. Over the last three years the translational bioinformatics global landscape has substantially changed and it is both interesting and exciting to see these themes have merged with genomic medicine very much at the fore of national policies and investment. The National Institute of Health (NIH) makes the difference between genomic and precision medicine by defining the broad term genomic medicine as "an emerging medical discipline that involves using genomic information about individuals as part of their clinical care (e.g., for diagnostic or therapeutic decision-making) and the health outcomes and policy implications of that clinical use" and precision medicine as the approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle.

This review outlines some of the most influential research publications and resources that have been developed between 2014 and 2016 and the national initiatives that have started to capitalize on these developments to routinely embed genomic medicine into healthcare. The field is broad and thus this review can neither cover all the aspects of TBI nor all the significant publications and outputs.

The translation of genomic data into clinically actionable knowledge is one of the key challenges of TBI. Even rare monogenic disorders can be influenced by a large number of different genes and biological pathways, as well as by environmental factors that are difficult to assess. Patients will also vary in how they present symptoms and in disease severity. Common medical problems such as heart disease, diabetes, and obesity do not have a single genetic cause—they are likely associated with the effects of multiple genes in combination with lifestyle and environmental factors. In addition, complex or multifactorial disorders do not have a clear-cut pattern of inheritance. This makes it difficult to determine a person's risk of inheriting or passing

on these disorders. Complex disorders are also difficult to study and treat because the specific genetic and environmental factors leading to most of these disorders have not yet been identified.

The genetic developments of the monogenic disorder cystic fibrosis exemplify some of the challenges of applying genomic information to the clinic. Although the genetic basis of cystic fibrosis was well established to be mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene in 1989 [7], it has taken many years to understand and categorize the 2000 variants within the CFTR gene into roughly six different functional classes and then stratify patients for appropriate treatments [8]. The mutations observed have a wide variety of effects on the CFTR protein: some disrupt the function of the chloride channel at the apical surface of epithelial cells; some affect the intracellular processing of CFTR reducing expression levels; and others alter the transcription of CFTR. The successful development of the drug ivacaftor in 2012, a potentiator of CFTR function, was the first targeted therapy for patients with cystic fibrosis caused by specific genotypes [6]. In 2015, the approval of the combination therapy of a potentiator and a corrector (ivacaftor and lumacaftor) offered tailored treatments to people with cystic fibrosis caused by the most common CFTR mutation, Phe508del [9]. This treatment is a powerful example of precision medicine.

One of the key challenges in the application of genomic medicine in healthcare is the necessity to protect patient privacy and security of data whilst making genomics diagnoses and discoveries. What makes this difficult are the wide differences in the way research and clinical practice are performed. Scientific research has in the last century been a global effort with English as the practicing language, similar systems for sharing results, open access to data via publications or other means, as well as funding via grants or awards. Healthcare practice, on the other hand, is primarily a national endeavor practiced under complex national legislation in the native language(s) of the country and is often contractually-funded. Hospital information systems also vary within and from country to country, with most of patient-specific data being private to patient and associated clinicians. Thus, when it comes to the application of bioinformatics methods into clinical studies a whole new paradigm needs to be created in order to be successful.

Advances in Genomic Medicine Discovery

Much has been done to advance and develop the capabilities provided by informatics methods for genomic medicine discovery. These range from the development of various sequencing technologies to new models for data storage and data sharing. Below, we review some of these advances and their contribution to the progress made by various national genome sequencing initiatives.

Sequencing Technology

Over the past decade, technological improvements in high-throughput sequencing technologies have resulted in a growing worldwide capacity to easily generate nucleotide sequences [2]. The most common current technology, implemented in instruments made by Illumina, uses the sequencing-by-synthesis method and is able to sequence as many as 18,000 whole human genomes a year. Large-scale initiatives such as the UK 100,000 Genome Project or the Human Longevity Inc. initiative to build a facility scaling up to 100,000 genomes a year [10] primarily use this robust technology.

Two other approaches generate longer read lengths than the sequencing-by-synthesis method, making them well-suited for unsolved problems in genome, transcriptome, and epigenetics research. The first is nanopore sequencing, in which a single DNA molecule is guided through a barrier with pores that allow only a single nucleotide to pass through at a time, the electrical charge of which is then measured and recorded. The MinION handheld device released by Oxford Nanopore Technologies in mid-2014 is the first and only nanopore sequencer on the market and has been since deployed widely. Its applications range from the metagenomic detection of Ebola viral pathogens from clinical samples in the field during the 2014-2015 epidemic, with an unprecedented less-than six hours sample-to-answer turnaround time [11]; to the rapid detection of antimicrobial resistance in outbreak situations where strain identification can be obtained within 30 minutes of sequencing; and, using about 500 reads, initial drug-resistance profiles within two hours and complete resistance profiles within 10 hours [12]. A second long read length method is the Single-Molecule Real-Time sequencing (SMRT) technology used in Pacific Biosciences machines [13]. These can generate reads greater than 10,000 bases with over 99.999% accuracy, enabling the production of finished bacterial genomes, the resolution of structural variations, and a better resolution of single nucleotide variants.

Reference Data

The generation of the reference datasets enables comparative analyses with parallel data from disease-centric studies to identify variants and processes that are linked to disease. These result in a better understanding of the underlying mechanisms by which various diseases occur, an improved ability to predict which treatments will work best for specific patients, and improved approaches such as genome-based strategies for the early detection, diagnosis, and treatment of disease.

Since the availability of the first complete draft of the human genome, many large experimental studies and associated computational resources have been published in efforts to further basic biological knowledge. These include the ENCODE project to understand the function of genes and the elements that regulate genes throughout the genome [14], the 1000 Genomes Project to generate comprehensive genetic variation maps of individuals from multiple populations [15], advances in mass spectroscopy and electron microscopy which generate 3-dimensional structures of proteins and enable the identification of functions, protein capture experiments which explore how DNA and proteins interact with one another and with the environ-

ment to create complex living systems, and large-scale epigenome maps from healthy and diseased human cells [16].

The Genome Reference Consortium that coordinates, builds, and manages reference genomes for several model organisms, develops novel methods such as graph-based algorithms to represent complex allelic diversity. Significant effort is required to maintain up to date reference genome sequences, represent alternative loci, and effectively resolve difficult regions of the genome [17].

Functional and Clinical Interpretation

When identifying causative gene mutations, the first step is for rare diseases to catalogue all the nucleotide differences or variations in a patient's genome compared to a reference genome, and for cancers to catalogue the differences between the healthy and tumor genomes. The more complex, and naturally the more valuable, next step is to understand the clinical significance of the variants, their inheritance patterns, and the strength of their association to the disease or phenotype [18].

Understanding the clinical and functional significance of each variant requires complex bioinformatics analyses and the integration of numerous other data types:

- Data such as gene structure information is necessary to determine whether the variant lies in the coding or non-coding portion of the genome;
- Coding variant, protein structure, and functional data are needed to determine the impact of the mutation on protein function;
- Transcriptomics and proteomics data are required to determine cell and tissue expression profiles;
- Mutation experimental data from human cell or model organisms and disease variation information are needed to understand linked phenotypes;
- Protein interaction network and biological pathway knowledge are required to learn more about the function and relationship with other proteins;
- Data from clinical trials and pharmaceutical agents are also needed to know if there have been medicines developed that target this protein or biological pathway;
- If available, longitudinal phenotypic information at the individual and population levels is also needed.

All of the above require the ready availability of curated, structured reference information.

Much of the data required to determine clinical significance is deposited and curated in bespoke biological repositories such as those hosted by the European Bioinformatics Institute (EMBL-EBI) [19] or the NIH National Center for Biotechnology Information (NCBI). However, data integration for genome-wide bioinformatics analyses and conversion of data to knowledge rely on the continuous development of analytical pipelines and systems. Tools such as the Ensembl Variant Effect Predictor is a powerful toolset for the analysis, annotation, and prioritization of genomic variants in coding and non-coding regions, providing access to an extensive collection of genomic annotations [20].

Clinical Data Environment

The integration of EHRs in patient care is also needed to link molecular and clinical data. One important component required for data integration is the careful curation and mapping of data to controlled vocabularies or ontologies. For genomic data integration with clinical information, data from primary care, hospitals, outcomes, registries, and social care records should be first recorded using controlled clinical terminologies, such as SNOMED Clinical Terms and the Human Phenotype Ontology [21]. Ontologies as such are not ever complete and end-users such as clinicians will need to work with ontology developers to continuously improve the precision and accuracy of terminologies [22].

Using standard terminologies for recording clinical data is however just the primary step. Clinical data is usually generated and held across a wide variety of point of care settings such as acute hospitals, general practitioners, community hospitals, mental health, and social care. Integration of health data from different sources can facilitate the efficient and timely use of clinical information gathered throughout a patient's journey, and hence should minimise duplication and enable greater continuity of care. The implementation of EHRs such as EPIC within hospitals has driven greater standardisation and efficiency. However there were significant challenges to overcome in the deployment and operationalisation of such a system [23]. For research, organisations like CDISC work to create standards in order to support the acquisition, sharing, submission, and archiving of clinical research data.

The end of patent battles on genetic tests by commercial companies, such as Myriad Genetics and others, over BRCA1 and BRCA2 mutation detection[1], coupled with cheaper high-throughput sequencing and high profile celebrity activities (e.g. the Angelina Jolie effect [24]), have resulted in an increased demand for new approaches to diagnostic tests. These include developments in cost-effective accurate mutation detection strategies and a standardized, systematic approach, to the reporting of test results, especially in cancer [25]. Rahman and colleagues have developed a rapid, robust, large-scale, cost-effective testing pathway that can be utilized within the hospital framework and easily adapted across other healthcare systems [26].

As the volume of genomic data grows with associated clinical data, it is also useful to note that aggregation and reanalysis of such data will result in a new and improved understanding of clinical value over time [27]. For example, a novel variant discovered in a patient today may have little or no information associated with it. However as genomic data grows and this variant is analyzed in conjunction with other similar variants, more statistically significant results can result in greater confidence of this variant being associated, or not, with a disease.

Data Storage and Sharing

New models for data storage and sharing have also emerged so that experts across the continuum from molecular biology to medi-

[1] http://www.sciencemag.org/news/2015/01/end-road-myriad-gene-patent-fight

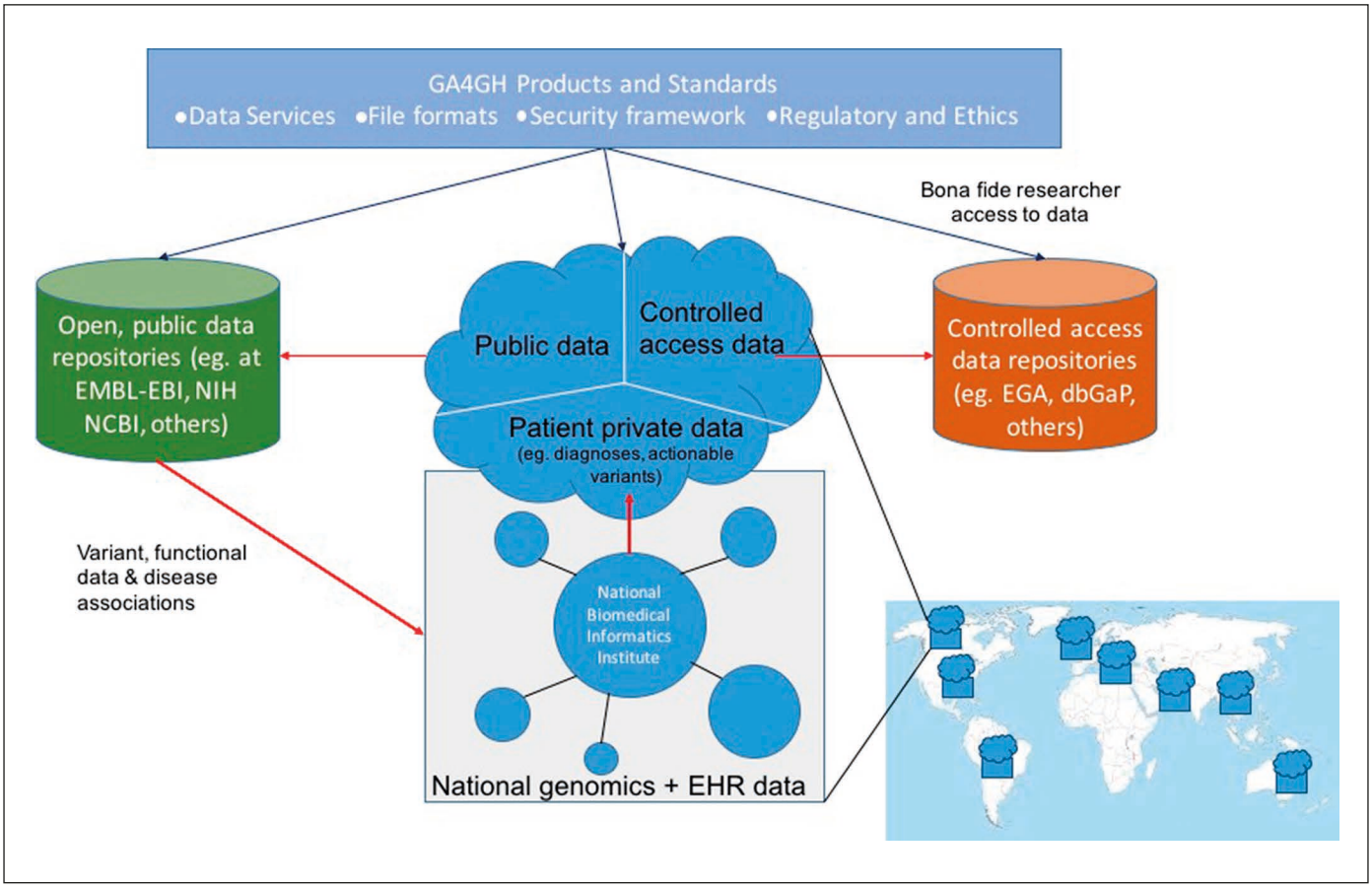

Fig. 1 National and International Data Sharing Mechanisms. Data from national genome sequencing initiatives and related clinical data will naturally need to be stored within centralized national storage systems. Anonymized data from the analyses of these projects can be made available for research use, either via controlled-access data repositories or open, public data repositories, depending on consent and data category. Products and Standards developed by GA4GH can be applied by the three different resource types to ensure interoperability of medical genomics data.

cine can access, use, and deposit data. A pilot project in rare disease, the Deciphering Developmental Disorders (DDD) study, aimed to determine the feasibility of translating new high-throughput genomic technologies into clinical practice, and of elucidating the underlying genetic architecture of developmental disorders. This study, which utilized whole exome sequencing to diagnose 27% of 1,133 previously investigated yet undiagnosed children with developmental disorders [28], also established a unique database model in the DECIPHER[2] database.

Contributing to the DECIPHER database is an international community of academic departments of clinical genetics and rare disease genomics that now numbers more than 250 centers who have uploaded more than 18,000 cases. Each center maintains control of its own patient data (which are password protected within the center's own DECIPHER project) until consent is given to share the data with chosen parties in a collaborative group or to allow anonymous genomic and phenotypic data to become freely viewable within genome browsers. Once data is shared, consortium members are able to gain access to the patient report and to contact other members to discuss patients of mutual interest. After data analysis, pertinent genomic variants are returned to individual research participants via their local clinical genetics team. The DDD study also demonstrated that the systematic recording of relevant clinical data, curation of a gene–phenotype knowledge base, and development of clinical decision support software were crucial for scalable prioritization and review of possible diagnostic variants [29]. Most of the diagnostic variants identified in known genes were novel and not present in current databases of known disease variations.

The Global Alliance for Global Health (GA4GH) established in 2013 is developing a common framework of approaches for adoption in order to accelerate progress in human health, drive efficiencies, and lower costs. The goal of the Alliance is to create data standards and strategies for storage and analysis of medically relevant genomic data, and to catalyze the creation of data sharing standards and methods to ensure worldwide interoperability of medical genomics data [30]. GA4GH includes institutions like EMBL-EBI that play a key role in facilitating the transfer of knowledge and expertise in data management and analysis of big data

[2] https://decipher.sanger.ac.uk/index

projects. Currently 452 institutions and companies across 42 countries participate to the Alliance, which emphasizes the global interest, demand for standardization, and need for joint working solutions[3]. The strategic goal of GA4GH is to build a system of servers, to create standard markup languages, and to develop resources and applications similar to the implementation of the World Wide Web for users to access genomics information. A schematic view of how this may look in practice is shown in Figure 1.

National Sequencing Initiatives: Approach and Progress

The use of genomics as a healthcare diagnostic tool is becoming increasingly more common due to the desperate need to understand the underlying causes of diseases and to provide more efficacious medicines to patients. The disease areas that are likely to significantly benefit from the use of genomic diagnostics are those where the identification of causative gene mutations is more straightforward, such as in rare diseases, cancers, and infectious diseases, as well as the detection of chromosomal abnormalities in non-invasive prenatal testing. Over the last few years, several national initiatives have been launched with the strong support of the leaders of respective countries. A complete list can be found in Table 1. The scope, approach, and goals of each of these are quite different, some of which are described below.

The UK 100,000 Genomes Project was formally announced in December 2012 by the Prime Minister as part of the UK Government's Life Sciences Strategy[4]. The delivery of the project was charged to a newly established limited company owned by the UK Department of Health, Genomics, England, with rare diseases, cancers, and infections chosen as the disease areas to focus on. Since then, the program has developed and executed an aggressive delivery plan including the setup of several genomic medicine centers across the country, a data infrastructure, a panel of annotation suppliers, a clinical interpretation partnership with researchers, relationships with the pharmaceutical industry and biotechnology companies, as well as a skills and education program for national health service employees. With over 18,000 genomes sequenced to date, several cases of rare disease diagnoses of previously undiagnosable genetic conditions are coming to light[5] such as a rare mutation in the SLC2A1 gene which caused a patient's Glut1 deficiency syndrome. This mutation was narrowed down from the 6,414,934 variants initially observed when compared to the reference sequence. Besides clinical interpretation, the project is also addressing technical and logistical challenges, such as obtaining tumor DNA of sufficient quality and quantity to meet healthcare pathology test standards, and the mapping of clinical data to standard ontologies using robust and user-enabled approaches.

In January 2015, U.S. President Obama announced the Precision Medicine Initiative (PMI), a national, large-scale, research enterprise with one million or more volunteers from diverse social, racial/ethnic, ancestral, geographic, and economic backgrounds, from all age groups and health statuses as well as a dedicated cohort of oncology patients. June 2016 saw the kick-off of this initiative with the establishment of six recruiting centers that aimed to enroll 10,000 participants in the first year, starting in November, then 35,000 a year through 2020 to reach a total of 150,000[6]. The goals of the program included developing criteria and standards for the incorporation of rapidly evolving technology and mobile health technologies into cohort design, both for baseline and ongoing data collection, as well as interoperable and standardized EHRs plus various genomics and imaging approaches[7].

Table 1 Key aspects of national genomics healthcare initiatives

Program Name	Start Date	Number of genomes	Disease areas	Investment
Estonian Genome Project	2001	1 million individuals over 5 years	Random selection	1.5 billion Estonian kroons
Genome Denmark	March 2011		Cancer, pathogen, reference genome	80 million DKK
Iceland (deCode Genetics)	-	10,000	All disease areas	-
UK 100,000 Genomes Project	December 2012	100,000	Rare disease, cancer, infection	£300 million
US Precision Medicine Cohort	January 2015	>1 million volunteers	All disease areas	$215 million
Scottish Genomes Partnership	January 2015	> 3000 citizens	Cancer, childhood illnesses, rare genetic diseases, disorders of the central nervous system and population studies	£21 million
Genomic Medicine France 2025	April 2015	235,000 genomes annually by 2020	Rare disease, cancer, diabetes, other common diseases, reference genome	€1 billion
Finland	July 2015	-	All disease areas	€50 million
China Precision Medicine Initiative	March 2016	100 million genomes over 15 years	All disease areas	US$9.2 billion
Germany	Planned			€360 million

[3] http://genomicsandhealth.org

[4] www.genomicsengland.co.uk

[5] https://www.genomicsengland.co.uk/first-children-recieve-diagnoses-through-100000-genomes-project/

[6] http://www.sciencemag.org/news/2016/07/president-obama-s-1-million-person-health-study-kicks-five-recruitment-centers

[7] https://www.nih.gov/sites/default/files/research-training/initiatives/pmi/pmi-working-group-report-20150917-2.pdf

The substantial learnings from previously established programs, such as the Electronic Medical Record and Genomics (eMERGE) network on approaches to integrate genomic variant information within electronic medical records, will be incorporated into PMI protocols [31, 32]. Researchers are also working on integrative models based on current knowledge of genomics and epigenomics and the relevant biochemistry and cellular-tissue physiology to predict how to obtain data from these very large cohorts. Such predictions could specify which clinical parameters to measure and at what intervals [33]. eMERGE is also studying the ethical, legal, and social issues involved in the use of EHRs for genomics research, such as privacy, confidentiality, and communications to the public, as well as the return of actionable genomic test results to EHRs for use in clinical care.

The French national sequencing initiative, France Médecine Génomique 2025, has an investment of €670 million over the first five years, accompanied by commercial contributions worth another €230 million, to sequence 235,000 genomes annually by 2020. Launched in April 2015, the French project will begin by sequencing genomes not only from patients with rare diseases and cancers, but also from some forms of diabetes, which is noted as an urgent priority for research to help develop better targeted treatments.

Applications of the Secondary Use of Patient Data

The launch of several national sequencing initiatives will create a 'longitudinal life course of electronic health' database of all participants, based upon a flow of electronic health data from primary care, hospitals, outcomes, registries, and social care records.

These extensive records will provide the opportunity to evaluate genomics in the context of rich and extended phenotypes including biochemical parameters, health outcomes, mortality data, and pharmacogenomics. Analysis of these data beyond the purpose of primary diagnosis will allow researchers to move past the primary phenotype of the disease that led to the patient's enrolment to evaluate genomics in the context of other continuous traits, diseases, and response to therapy.

The use of genomic data in secondary research has advanced the development of new tools and algorithms such as those used to model the genetic diversity and evolutionary patterns of individual cancers [34]. Studies that aim to use biobanks and integrate different data types have successfully stratified patients and identified potential biomarkers of drug response. In a recent study by Folkersen et al. in Rheumatoid Arthritis (RA), a biobank was used to test the claim that the current state-of-the-art precision medicine will benefit RA patients. High-throughput RNA sequencing, DNA genotyping, extensive proteomics, and flow cytometry measurements, as well as comprehensive clinical phenotyping, led to the identification of a small set of biomarkers available in peripheral blood that predict clinical response to tumor necrosis factor (TNF) blockade [35].

Genomic data has also enabled very large-scale projects, such as the Pan-Cancer Analysis of Whole Genomes (PCAWG) study, which is an international collaboration to identify common patterns of mutation in more than 2,800 cancer whole genomes from the International Cancer Genome Consortium. PCAWG aims to generate genomic, transcriptomic, and epigenomic changes in 50 different tumor types and/or subtypes. Such analyses have shown the value of integrating genomics data from the same patient, will lead to novel targets and disease mechanisms, and should in turn drive enhanced diagnostic and therapeutic yields for individual patient benefit [36-38].

As well as research applications, the development of new or repurposed drugs stands to substantially benefit from more disease-associated genomic data. One of the main reasons for the high rate of attrition in late-stage clinical trials is thought to be the lack of drug efficacy [39]. Often the incorrect gene or protein is selected as the drug target in early drug development, where the premise is that perturbation of this protein by a compound will significantly change the course of disease [40]. Recent publications have shown that genetic data that link a target to a phenotype or disease have higher success rates in the clinic [41].

A study by Bagley et al. combined data from electronic medical systems with disease-associated genetic variants data to study the relationship between disease co-occurrence and commonly shared genetic architectures of disease. The study looked at 35 disorders, medical records for over 1.2 million patients, and variants from over 17,000 publications, and found specific shared genes between disease classes that were not previously thought to be related, such as autoimmune and neuropsychiatric disorders [42].

It is clear that current definitions and categorization of diseases and phenotypes do not necessarily reflect true molecular relationships and underlying biological relationships. Public-private initiatives such as Open Targets, a collaboration between Biogen, the EMBL-European Bioinformatics Institute, GlaxoSmithKline, and the Wellcome Trust Sanger Institute (http://www.opentargets.org), aim to provide comprehensive and up-to-date relevant genetics and high-throughput genomics data for drug target selection and validation [43, 44]. Chen and Butte have provided a comprehensive review on the availability of public data and the analytical tools across various data types for target selection and drug discovery [45].

Processed data from genome sequencing projects and the knowledge extracted from these can also be integrated into existing reference data in a manner that enables further analyses while protecting the privacy of patients within each country or system.

Data that can be captured include:

- Molecular profiles that characterize differences between diseased and normal states, or that provide sub-classification of a disease;
- Annotation of variants of clinical importance in different diseases;
- Annotation of variants and genes, and their association with drugs;
- Biomarkers (e.g. protein, RNA, metabolites, and complete metabolomes) for diagnosis and disease monitoring;
- Reference images that can link molecular data to disease phenotypes;
- Human pathogen data and their virulence components.

Such data will provide broaden references dataset for further clinical research as well as develop resources that are fundamentally about understanding biology.

Ongoing Challenges: Technological, Societal, and Economic

Bioinformatics analysis leading to clinical interpretation is an expensive part of the pipeline. The importance of data sharing and common standards is emphasized by Muir et al. [46] who highlight that storage and computation costs have not decreased as quickly as sequencing costs. They concluded that "if the sequence data generated by individual labs is not processed uniformly and sequence databases are not made easily accessible and searchable, then the analysis of aggregated datasets will be challenging". Reducing costs will require simultaneous improvements in data sharing and use of common standards. We describe below some of the technological, economic, and societal challenges that need resolution to fulfil this vision.

Standardization

If we look back at the historic peaks and troughs of molecular biology, major technological innovations have preceded moments of great discovery. For example, the invention of high-throughput Sanger sequencing technology enabled the completion of the human genome [47]. Innovation in data generation is always followed by a surge of data analysis methods and tools. At the moment, the bottlenecks in genomic medicine lie at the data analysis and interpretation end of the pipeline. Interpretation in this scenario is a critical step since a patient's diagnosis status and potential treatment options are crucially dependent on interpretation, and not on the raw or processed sequence data.

Efforts to identify gold standard methods and evaluate the performance of data analytical methods are currently emerging. These include work by Tokheim and colleagues who compared eight different algorithms that attempted to identify, from variation data, which gene variants drove cancer driver genes and which were simply passenger mutations [48]. The analysis found that most approaches had a high rate of false positives and more work was needed to develop a gold standard method.

Electronic Medical Records (EMR) represent a convenient source of coded medical data, but the lack of standards and the variation among the different systems can introduce inaccuracies and biases when this data is used for analyses such as calculating disease prevalence, incidence, age of onset, or disease comorbidity [49]. There is a further need for standardisation around clinical data capture and communication, which addresses the quality, completeness, and adoption of standards. Analysis would also improve if there was standardization in the way data are collected from participants across hospitals and clinics. Far more can be interpreted from a genome sequence when an accurate patient record is available. The challenges lie in being able to collect this information from busy clinicians, and the data also needs to be integrated across the various points of patient care. Clinical decision support systems that have been approved by regulatory bodies and in which significance and confidence levels around genetic findings are systematically inferred and reported [50] are also needed.

Data Storage and Sharing

Currently, much of the human genomic data generated so far is deposited into public databases for broad research reuse. Human genomic and phenotypic data from clinical or research studies which would require a researcher to have a bespoke signed agreement with the originating body (via a Data Access Committee or other mechanism) are largely stored in controlled-access repositories such as the European Genome-Phenome Archive (EGA), the NIH database of Genotypes and Phenotypes (dbGaP), or held by the originating body. However, given the differing ethical and legal systems of each country, these systems are not scalable nor are they appropriate for the growing volume of genomic data from national health studies.

Managed storage systems which follow national legislation and which allow access to data for research purposes are essential. Researcher access to genomic databases is necessary to create a research community that will be connected and may contribute directly to national health services and patient care systems. Analyses that utilize the aggregated data from hundreds of thousands or millions of patients from multiple healthcare systems will add much more to our knowledge of the genetic basis of disease than multiple individual studies using small sample cohorts from individual healthcare systems.

Large-scale cohorts are particularly important for very rare diseases for which patient numbers in any one country may be too small to provide adequate data to identify causative genes. For rare diseases, the sample size needed to infer whether observed variants that are associated with a disease are causative and statistically significant will often require combining data from patients in multiple countries: cross-border sharing of data is essential so that virtual or physical data aggregation and sufficiently powered analyses can be performed.

Researcher Access to Data

New data sharing mechanisms are also needed to minimize the movement of large volumes of data and allow instead for analyses to take place at the point where data are stored. Cloud computing frameworks allow remote storage, with analysis scripts uploaded to the cloud and analysis performed remotely on virtual machines physically located at the remote site "next to" the data. This greatly reduces data transfer requirements since only the scripts and analytical results are transferred to and from the analysts' institution or desktop: data resides permanently in the cloud [51]. The cost structure for computational and analysis resource in genomics points towards the efficiency of developing a single, large center for data analysis and processing. Such a resource would, through virtual access, be combined with a more widely distributed network of expertise [22]. Initiatives such

as the European Open Science Cloud will further the creation of infrastructures to enable data sharing and service provision across borders and disciplines[8].

With health data for large numbers of people, it will be critical to find ways to protect individuals' privacy and the confidentiality of their health information. However, if data is to be shared by various communities, the correct legal and ethical frameworks must be in place. It is critical to find ways to protect participants' privacy and the confidentiality of their health information while simultaneously enabling research to take place. Current practices for researcher access to data that include paper-based agreements between users, institutions, and data access committees must be replaced by electronic mechanisms. These processes, at the interface between basic research and clinical research, should be strengthened and explicitly funded.

Individuals will also need to understand the risks and benefits of participating in genomics diagnostic and research. Understanding what data is collected and generated is also important. Much like the legislation needed to protect consumer data after the advent of web-based purchases and mobile technology, patients should be made aware of the use and implications of generating personal genomics information. Requirements to collect consent for research should be more harmonized and regulatory tools developed as described in the recently published code of practice [52]. The new EU data protection framework, in the form of the European General Data Protection Regulation (GDPR), will place a number of direct obligations on data controllers, which will drive better forms of consent collection and withdrawal.

Biomedical Informatics Coordination

We believe the overall ideal endpoint at the national level is the development of a 'Biomedical Informatics Institute' to act as a driver and coordinating center for health and biomedical informatics research in each country. This center should seek to act in conjunction with existing medical research and informatics organizations to form a seamless and integrated network with hospitals, research organizations, and local and international health initiatives to maximize the utility of genomics and electronic health data. In bigger nations, this institute would itself likely be a network, but with a center of gravity, or hub, at or within one institute.

Such centers would be the natural partners for research bioinformatics organizations such as EMBL-EBI or NCBI. In European countries, the development of biomedical informatics institutes or networks may be coordinated through an ELIXIR[9] Node: ELIXIR is the European life-science infrastructure for biological data. Research bioinformatics institutes can then be responsible for handling and providing both open (public) and controlled access data and bioinformatics services that are shared between researchers (including clinical researchers), whereas the national biomedical informatics institute can be responsible for data and services that need to stay within the national framework (see Figure 1).

Much research and development is needed in areas such as: the development of analytical methods, tools, and standards to link and extract value from increasingly complex, disparate, diverse, and numerous data sets; the development of secure interoperable research environments and data flows to provide a technology framework to federate existing platforms that will connect diverse health and biomedical data assets; the development of partnerships with owners and controllers of data, regional and national health and social care partners, academia and industry; and the development of skills and capacity in the discipline of medical informatics, training researchers with interdisciplinary skills in core data science and medical research. Alongside these, the economic cost and impact of delivering precision medicine in an effective and affordable ways also need to be considered.

Future Landscape

It is difficult to imagine exactly what the biomedical industry will look like in a few years' time, but it is certain that the surge in biological data flows will continue. This increase in data will be from large innovative research projects, such as the International Human Cell Atlas Initiative[10] that aims to create comprehensive reference maps of all human cells, and also from devices, apps, wearables, and implantable technologies. Translational BioInformatics methods, tools, and resources will need to evolve to include algorithms for streaming data capture, real-time data aggregation, machine learning, predictive analytics, and visualization solutions in order to integrate health monitoring data with EMRs and genomics data [53].

If genomics medicine approaches are to become part of routine healthcare, doctors and other healthcare providers will require better grounding in molecular genetics and biochemistry. They will increasingly find themselves needing to interpret the results of genetic tests, understand how that information is relevant to treatment or prevention approaches, and convey this knowledge to patients. Education and skills in the data sciences is much needed [54]. Programs to ensure the long-term generation of proficient investigators who understand the multi-disciplinary nature of genomics in clinical practice and research, should be established, and will perhaps even form a new medical discipline.

Open data that allows data reuse and data integration has made possible great advances in molecular biology over the last few decades. These advances range from recombinant DNA drugs, animal cloning, gene therapy, and forensic science to stem cell therapy. Although the primary objective of genomic data generated for healthcare purposes is for disease diagnoses, treatment, and prevention, the availability of these data for use in secondary research can result in a better understanding of disease mechanisms and will lead to improvements in treatment strategies. Moreover, the crossover of bioinformatics into healthcare will further enable fundamental discoveries about the big questions of biology.

[8] http://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf#view=fit&pagemode=none

[9] http://www.elixir-europe.org/

[10] https://www.humancellatlas.org/

It is worth saying that we are only at the very start of this health revolution brought about by genome sequencing. If we compare the time the human genome was sequenced in 2001 [47] following the first bacterial genome sequence in 1995 [55], to today, it is not a stretch of the imagination to envision human genome sequencing as a part of standard care pathways and real-time biomedical and health care analytics in the clinical setting. The systems and processes we put in place today must support the future and not just represent our present reality.

Acknowledgements
The authors wish to thank Charles E. Cook for critical comments on the manuscript.

Conflicts of Interest
Ewan Birney is a paid consultant to both Oxford Nanopore Technologies and GlaxoSmithKline. Work on this review was not supported by Oxford Nanopore Technologies or GlaxoSmithKline and the opinions expressed in it are the author's.

References

1. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet 2014;30(9):418-26.
2. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;17(6):333-51.
3. Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med 2001;40(4):346-58.
4. Altman RB. Towards Clinical Bioinformatics: Redux 2015. Yearb Med Inform 2016(Suppl1):S6-S7.
5. Denny JC. Surveying Recent Themes in Translational Bioinformatics: Big Data in EHRs, Omics for Drugs, and Personal Genomics. Yearb Med Inform 2014:199-205.
6. Tenenbaum JD. Translational Bioinformatics: Past, Present, and Future. Genomics Proteomics Bioinformatics 2016;14(1):31-41.
7. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 1989;245(4922):1066-73.
8. Brodlie M, Haq IJ, Roberts K, Elborn JS. Targeted therapies to improve CFTR function in cystic fibrosis. Genome Med 2015;7(1):101.
9. Wainwright CE, Elborn JS, Ramsey BW, Marigowda G, Huang X, Cipolli M, et al. Lumacaftor-Ivacaftor in Patients with Cystic Fibrosis Homozygous for Phe508del CFTR. New Engl J Med 2015;373(3):220-31.
10. Telenti A, Pierce LCT, Biggs WH, di Iulio J, Wong EHM, Fabani MM, et al. Deep sequencing of 10,000 human genomes. Proc Natl Acad Sci 2016;113(42):11901-6.
11. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med 2015 Sep 29;7:99.
12. Cao MD, Ganesamoorthy D, Elliott AG, Zhang H, Cooper MA, Coin LJM. Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinIONTM sequencing. Gigascience 2016;5(1):32.
13. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13(5):278-89.
14. genome AieoDeith. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489(7414):57-74.
15. The Genomes Project C. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
16. Stunnenberg HG, Abrignani S, Adams D, de Almeida M, Altucci L, Amin V, et al. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell 2016;167(5):1145-9.
17. Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin C-S, et al. Extending reference assembly models. Genome Biol 2015;16(1):13.
18. Wright CF, Middleton A, Burton H, Cunningham F, Humphries SE, Hurst J, et al. Policy challenges of clinical genome sequencing. BMJ 2013;347.
19. Cook CE, Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, et al. The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic Acids Res 2016;44(D1):D20-6.
20. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17(1):122.
21. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res 2017;45(D1):D865-D76.
22. Bowdin S, Gilbert A, Bedoukian E, Carew C, Adam MP, Belmont J, et al. Recommendations for the integration of genomics into clinical practice. Genet Med 2016;18(11):1075-84.
23. Nguyen L, Bellucci E, Nguyen LT. Electronic health records implementation: An evaluation of information system impact and contingency factors. Int J Med Inform 2014;83(11):779-96.
24. Evans DGR, Barwell J, Eccles DM, Collins A, Izatt L, Jacobs C, et al. The Angelina Jolie effect: how high celebrity profile can have a major impact on provision of cancer related services. Breast Cancer Res 2014;16(5):442.
25. Wallace AJ. New challenges for BRCA testing: a view from the diagnostic laboratory. Eur J Hum Genet 2016;24(S1):S10-S8.
26. George A, Riddell D, Seal S, Talukdar S, Mahamdallie S, Ruark E, et al. Implementing rapid, robust, cost-effective, patient-centred, routine genetic testing in ovarian cancer patients. Sci Rep 2016;6:29506.
27. Health TGAfGa. A federated ecosystem for sharing genomic, clinical data. Science 2016;352(6291):1278-80.
28. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet 2015;385(9975):1305-14.
29. Study DDD. Large-scale discovery of novel genetic causes of developmental disorders. Nature 2015;519(7542):223-8.
30. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. Nature 2015;526(7573):336-42.
31. Brown SA, Jouni H, Marroush TS, Kullo IJ. Disclosing Genetic Risk for Coronary Heart Disease: Attitudes Toward Personal Information in Health Records. Am J Prev Med 2017 Apr;52(4):499-506.
32. Cutting E, Banchero M, Beitelshees AL, Cimino JJ, Fiol GD, Gurses AP, et al. User-centered design of multi-gene sequencing panel reports for clinicians. J Biomed Inform 2016;63:1-10.
33. Iyengar R, Altman RB, Troyanskya O, FitzGerald GA. Personalization in practice. Science 2015;350(6258):282-3.
34. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science 2014;346(6206):251-6.
35. Folkersen L, Brynedal B, Diaz-Gallo LM, Ramskold D, Shchetynsky K, Westerlind H, et al. Integration of known DNA, RNA and protein biomarkers provides prediction of anti-TNF response in rheumatoid arthritis: results from the COMBINE study. Mol Med 2016;22.
36. Tirode F, Surdez D, Ma X, Parker M, Le Deley MC, Bahrami A, et al. Genomic landscape of Ewing sarcoma defines an aggressive subtype with co-association of STAG2 and TP53 mutations. Cancer Discov 2014;4(11):1342-53.
37. International Cancer Genome Consortium PedBrain Tumor P. Recurrent MET fusion genes represent a drug target in pediatric glioblastoma. Nat Med 2016;22(11):1314-20.
38. Kirby MK, Ramaker RC, Gertz J, Davis NS, Johnston BE, Oliver PG, et al. RNA sequencing of pancreatic adenocarcinoma tumors yields novel expression patterns associated with long-term survival and reveals a role for ANGPTL4. Mol Oncol 2016;10(8):1169-82.
39. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotech 2014;32(1):40-51.
40. Waring MJ, Arrowsmith J, Leach AR, Leeson PD, Mandrell S, Owen RM, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat Rev Drug Discov 2015;14(7):475-86.
41. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti

P, Shen Y, et al. The support of human genetic evidence for approved drug indications. Nat Genet 2015;47(8):856-60.
42. Bagley SC, Sirota M, Chen R, Butte AJ, Altman RB. Constraints on Biological Mechanism from Disease Comorbidity Using Electronic Medical Records and Database of Genetic Variants. PLoS Comput Biol 2016;12(4):e1004885.
43. Barrett JC, Dunham I, Birney E. Using human genetics to make new medicines. Nat Rev Genet 2015;16(10):561-2.
44. Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res 2017 Jan 4;45(D1):D985-D994.
45. Chen B, Butte AJ. Leveraging big data to transform target selection and drug discovery. Clin Pharmacol Ther 2016;99(3):285-97.
46. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol 2016;17(1):53.
47. Consortium IHGS. Initial sequencing and analysis of the human genome. Nature 2001;409(6822):860-921.
48. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci 2016;113(50):14330-5.
49. Bagley SC, Altman RB. Computing disease incidence, prevalence and comorbidity from electronic medical records. J Biomed Inform 2016;63:108-11.
50. Overby CL, Kohane I, Kannry JL, Williams MS, Starren J, Bottinger E, et al. Opportunities for genomic clinical decision support interventions. Genet Med 2013;15(10):817-23.
51. Health GAfGa. A federated ecosystem for sharing genomic, clinical data. Science 2016;352(6291):1278-80.
52. Bahr A, Schlünder I. Code of practice on secondary use of medical data in European scientific research projects1. International Data Privacy Law 2015;5(4):279-91.
53. Shameer K, Badgeley MA, Miotto R, Glicksberg BS, Morgan JW, Dudley JT. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. Brief Bioinform 2016 Jan;17(1):43-50.
54. Greene AC, Giffin KA, Greene CS, Moore JH. Adapting bioinformatics curricula for big data. Brief Bioinform 2016;17(1):43-50.
55. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269(5223):496-512.

Correspondence to:
Dr. Ewan Birney
European Bioinformatics Institute (EMBL-EBI)
European Molecular Biology Laboratory
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD
United Kingdom
E-mail: birney@ebi.ac.uk
Website: http://www.ebi.ac.uk