© Modeling and Artificial Intelligence in Ophthalmology 2021; 3:101-140
Review article https://doi.org/10.35119/maio.v3i1.115

Cloud-based genomics pipelines for ophthalmology: reviewed from research to clinical practice

David C.S. Wong[1], Maximiliano Olivera[2], Jing Yu[3], Anita Szabo[4], Ismail Moghul[5], Konstantinos Balaskas[2,4], Robert Luben[2,4], Anthony P. Khawaja[2], Nikolas Pontikos[2,4]*, Pearse A. Keane[2,4]*

[1]School of Clinical Medicine, University of Cambridge, Cambridge, UK; [2]Moorfields Eye Hospital NHS Foundation Trust, London, UK; [3]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK; [4]Institute of Ophthalmology, University College London, London UK; [5]UCL Cancer Institute, University College London, London, UK

**The authors contributed equally.*

Abstract

Aim: To familiarize clinicians with clinical genomics, and to describe the potential of cloud computing for enabling the future routine use of genomics in eye hospital settings.
Design: Review article exploring the potential for cloud-based genomic pipelines in eye hospitals.
Methods: Narrative review of the literature relevant to clinical genomics and cloud computing, using PubMed and Google Scholar. A broad overview of these fields is provided, followed by key examples of their integration.
Results: Cloud computing could benefit clinical genomics due to scalability of resources, potentially lower costs, and ease of data sharing between multiple institutions. Challenges include complex pricing of services, costs from mistakes or experimentation, data security, and privacy concerns. *Conclusions and future perspectives:* Clinical genomics is likely to become more routinely used in clinical practice. Currently this is delivered in highly specialist centers. In the future, cloud computing could enable delivery of clinical genomics

Correspondence: Pearse A. Keane, Moorfields Eye Hospital NHS Foundation Trust, London, UK.
E-mail: p.keane@ucl.ac.uk

services in non-specialist hospital settings, in a fast, cost-effective way, whilst enhancing collaboration between clinical and research teams.

Keywords: clinical genomics, cloud computing, hospital, ophthalmology, sequencing

1. Introduction

For over three decades, researchers have been anticipating the potential for genomics to revolutionize clinical practice.[1–5] The completion of the Human Genome Project[6,7] has led to groundbreaking discoveries that are being applied for the prevention, diagnosis, and management of disease.[8] However, real-world uptake in medicine is still limited.

The majority of research has involved genome-based discovery of links between genetic variants and diseases, with less than 2% of the literature examining how to apply these into clinical practice.[5,8] The American College of Medical Genetics and Genomics promotes standardized reporting of clinically actionable genes. Their most recent recommendations included a minimum set of 59 genes where variants should be reported in genomics studies. The aim was to identify and manage highly penetrant genetic disorders by detecting potentially pathogenic variants in genomic data.[9,10] This list is by no means exhaustive and, for example, the *RPE65* gene, which is associated with retinal dystrophy and is now treatable, is currently not on this list.[11] Further to this, because our knowledge of healthy genetic variation is limited, especially in individuals of African descent who are under-represented in genomic studies,[12] it is often challenging to determine whether a genetic variant in one of these genes is truly pathogenic and hence should be acted on. Therefore, there is a great need for creating large comprehensive and ethnically diverse databases of genetic variation through genomic initiatives to facilitate the interpretation of genetic variants.

In the UK, there are a growing number of large genomic initiatives that aim to discover the genetic cause of cancers and rare diseases in UK National Health Service (NHS) patients from participating hospitals.[13] These genomics studies generate very large amounts of data which currently require research-specialist organizations such as Genomics England, the Broad Institute, or resources such as university High Performance Computing (HPC) for analysis and storage. Due to their size and complexity, these data, although generated from patients, are rarely integrated back into hospital systems, which limits their utility for clinical care beyond research. This also presents challenges around data security and privacy when genomics information is analyzed and transferred between organizations.

Cloud computing may overcome some of these technical challenges and allow health care organizations like the NHS to integrate genomic analysis back into the health care setting for patient benefit.[14-17] The National Institute of Standards and Technology (NIST) defines cloud computing as “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources”.[18] These resources include data managed storage, computing power, and networks for sharing, often provided as a pay-as-you-go service. This avoids the need to invest in additional IT staff and hardware locally, providing extremely flexible and scalable resources instead. The potential long-term cost savings and easily scalable up-to-date infrastructure therefore make cloud computing attractive to hospitals or clinics with limited or ageing IT resources and few dedicated members of staff for support. In the UK, cloud adoption is growing. The government has had a “cloud first” policy since 2017, encouraging digital services to be run on cloud platforms from inception.[19] NHS Digital has since expanded this scope, supporting the development of networked services for health and social care.[20]

In health care, cloud computing has already been applied in six broad domains:[14] telemedicine, medical imaging, public health, hospital management, therapy, and secondary use of data. However, as of yet, few hospitals have leveraged cloud technology for integrating genomics data. One notable example is the partnership between Google and Mayo Clinic.[21] The Mayo Clinic undertakes both research and clinical sequencing projects, and using Google Cloud Platform (GCP) has paved the way towards scaling these to involve hundreds of thousands of patients.[22] In addition, the Broad Institute has collaborated with GCP to build tools such as the Genome Analysis Toolkit (GATK), which will greatly contribute to the integration of clinical genomics and research.

In this review, we first describe current genomics practices before outlining our vision of how cloud computing may facilitate the integration of genomics into routine clinical practice in the future and specifically in ophthalmology. Our focus is specifically on eye hospitals for two main reasons. Firstly, ophthalmic clinical practice is heavily image-based, and therefore certain eye hospitals may likely already have cloud-based systems in place for imaging and telemedicine. Secondly, genomics is already an important part of eye health care with the development of gene therapies to treat specific inherited conditions but also, in the future, for routine management of many conditions.[23]

2. Genomics in clinical practice

2.1. Introduction to genomics

Genomics refers to the sequencing of the entire DNA sequence from an individual, which is composed of 3.9 billion base pairs. DNA is split into chromosomes, of which there are 23 pairs in humans, and these are contained within the nucleus of most cells in our body. Broadly, the DNA sequence is made of coding regions, which define genes that code for proteins (the units of functionality in our bodies), as well as noncoding regions, which may have a range of functions, including a regulatory function on the expression of genes and hence the creation of proteins. The variation in DNA, genomic variation, is what makes us unique, but also what makes some of us more vulnerable to certain diseases. This genomic variation is therefore the subject of much research and may be useful clinically. The process of identifying clinically relevant genomic variation is known as "genetic testing" and is a routine part of certain ophthalmology subspecialties such as inherited eye diseases. The methods for performing genetic testing have matured considerably over the last four decades so that we now have fast, reliable, cheap, and high-throughput technology for sequencing DNA, known as next-generation sequencing (NGS).[24,25]

Inherited eye diseases are a major cause of irreversible blindness in many countries, among both pediatric and working-age populations.[26-28] Inherited retinopathies alone affect around 1 in 2,000 people worldwide.[29] These debilitating disorders have traditionally been thought to be incurable, but many therapeutic approaches are being developed, often targeting genetic defects.[30] Hence, genetic testing is a necessary first step to enable these gene-targeted treatments. For example, patients with Leber's congenital amaurosis (a severe inherited cause of vision loss) harboring mutations in the *RPE65* gene may be treated with gene replacement therapy, which has been shown to be efficacious and safe.[11] Additionally, genetic diagnosis has a direct impact on family planning, especially for X-linked disorders such as retinitis pigmentosa.[31] As a result, genetic testing is becoming an increasingly popular investigation in ophthalmology.[32]

The examples described so far often rely on targeted genetic testing, which usually only tests specific regions of DNA for a specific type of variation. The data produced is thus of limited use beyond the specific condition tested and does not enable future discovery of other types of variation beyond the examined region. However, thanks to the drop in the cost of NGS, genomics approaches previously limited to large research projects such as the UK Biobank[33] are now becoming part of health care. Genomic data, if sufficiently comprehensive, may be a lifelong source of information for patients. In the UK, the NHS is rolling out genomic testing into clinical practice, spearheaded by the NHS Genomic Medicine Service.[34] This is matched by a worldwide investment of > 4 billion USD in the integration of genomics into health care systems.[35] The stage is therefore set for genomics to become integrated into routine clinical practice. This is becoming a reality in oncology[36] and we believe that ophthalmology will follow suit imminently.

We will first outline current clinical pathways that use genomics in ophthalmology and provide a summary of genomic data analysis and technical considerations. We will then show how these considerations can be met in a health care setting with cloud computing.

2.2. The genomics pathway in ophthalmology

Clinical genomics is currently predominantly targeted at specific known or suspected hereditary conditions[37] such as retinitis pigmentosa[38-43] and its related syndromic variants,[44-47] macular dystrophies,[48-50] and some metabolic diseases.[51-53] These inherited retinal dystrophies account for not only severe visual impairment in young people, but also a huge socioeconomic impact on society.[54] The current workflow in clinical genomics is summarized in Figure 1; currently, results are obtained in 2–6 months or longer depending on individual circumstances.[32] The clinical utility of genetic screening for inherited retinal diseases includes potential gene-therapy treatment, genetic counselling, and eligibility for clinical trials.[55]

2.2.1. The decision to proceed with genomics

Doctors may order genome sequencing to help fine-tune diagnosis or treatment plans. This necessitates a referral to a specialist, tertiary center for clinical genomics. In England, this is provided by the NHS Genomic Medicine Service (similar services operate in Scotland, Wales, and Northern Ireland), in 13 Genomic Medicine Centers. Patients must fulfil specific criteria for referral, specified in the National Genomic Test Directory,[56] which includes clinical indications, genes, and test methods that are currently NHS-approved. These are primarily rare conditions with many associated genes such as Bardet-Biedl syndrome (Table S1, S2).

A multidisciplinary team (MDT) typically discusses the case and determines what sort of genetic study is most appropriate. The MDT consists of nursing staff, an attending physician, and several subspecialists (*e.g.*, pediatricians, clinical geneticists, fertility specialists, endocrinologists, etc.). The decision to go ahead with a genetic test involves genetic counselling with the patient and their family. It is important to define the patient's expectations and motivations of the proposed genetic testing, including the desire to enroll in research studies. Genetic counsellors often recruit patients to clinical trials, and must therefore be transparent about their roles in the study team as well as the effect this may have on patient decision making.[57] Some important points of discussion during genetic counselling are described in depth elsewhere.[58]

In addition to genetic counselling, those ordering and providing the genetic test should consider whether any issues apply relating to the following ethico-legal principles: consent, disclosure of information, confidentiality, and data protection.[59] When sequencing the whole genome, thousands of genetic variants will be found, some of which may be pathological, but the majority of which will have no effect on health. The ethico-legal implications of this are profound and currently unsolved: some view it as an ethical obligation to report and act on incidental pathological findings, whilst others hold higher value to patient autonomy and the "right not to know", especially when it comes to children who may not be Gillick competent to consent.[60] It is therefore critical that carefully

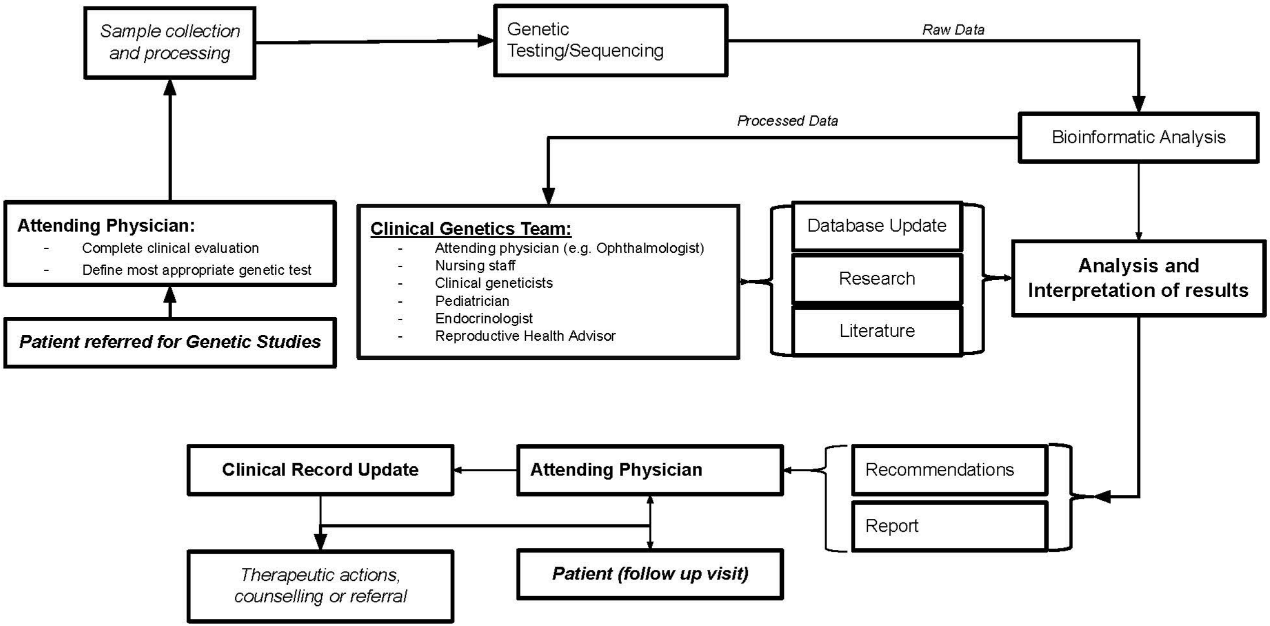

Sample collection and processing
Genetic Testing/Sequencing
Raw Data
Processed Data
Bioinformatic Analysis
Attending Physician:
- Complete clinical evaluation
- Define most appropriate genetic test
Patient referred for Genetic Studies
Clinical Genetics Team:
- Attending physician (e.g. Ophthalmologist)
- Nursing staff
- Clinical geneticists
- Pediatrician
- Endocrinologist
- Reproductive Health Advisor
Database Update
Research
Literature
Analysis and Interpretation of results
Clinical Record Update
Attending Physician
Recommendations
Report
Therapeutic actions, counselling or referral
Patient (follow up visit)

Fig. 1. Current clinical genomics pipeline in a specialist center. The Clinical Genetics Team is the central player in the current workflow; they are responsible for genetic counselling of the patient, sample collection, interpretation of sequencing results, and providing a report with recommendations to the patient's clinical team. This process may take weeks to months to complete and involves referral to a highly specialist center. In England, this is provided by NHS Genomic Medicine Centers, with sequencing carried out in Genomic Laboratory Hubs.

informed consent is obtained before taking samples, including a plan of action in the event that incidental pathological variants are discovered.

In most cases, a peripheral blood sample will be enough to perform any genomic study. In England, samples are sent to be sequenced at one of the seven NHS Genomic Laboratory Hubs. To supplement this, the attending physician should consider the need for complementary tests such as ocular coherence tomography (OCT), fundus autofluorescence (FAF), electrophysiology, blood tests, and neuroimaging. A detailed family history is obtained, and this is combined with the complete and detailed clinical picture. Here, artificial intelligence approaches that predict gene-phenotype correlations from retinal scans[61] as well as from standardized descriptive phenotypes using the Human Phenotype Ontology (HPO)[62,63] can be used for selecting the type of genetic test. The HPO project provides a standardized and controlled vocabulary linking phenotypes with information about genes (*e.g.*, Bardet-Biedl syndrome: Table S1, S2).[64,65]

2.2.2. Types of genomics technologies used in clinical practice

Several types of genomics studies are available to clinicians. These include array-based studies that test variations in segments of DNA through probe hybridization. These have traditionally been the first-line genomic studies of choice in clinical practice.[66,67] However, these methods are template-based and hence cannot detect novel variation. More versatile and diagnostically useful NGS techniques (whole genome and whole exome sequencing) are now increasingly available to clinicians.[68]

Whole genome sequencing (WGS) includes both coding and noncoding DNA regions, giving a complete view of the genome. Currently, this is achieved by short-read sequencing, where short sequences of approximately 100–200 base pairs each are aligned to the human reference genome, eventually covering the entire genome. However, whole exome sequencing (WES) is still more commonly used in clinical genomics.[56] WES is similar to WGS, but only involves sequencing of the coding regions of the genome (roughly 1% of the whole genome) and is therefore currently cheaper.[69] WES also allows for better accuracy of sequencing (higher sequencing depth) due to the lower coverage of the genome. Gene panel testing, as offered by the diagnostic labs, are in fact performing WES on a subset of genes.

2.2.3. Reporting results of genetic studies

After sequencing, clinical genetic reports are validated by a senior clinical scientist and forwarded to the referring clinician. Reports include a summary of sequencing findings (*i.e.*, any pathogenic, likely-pathogenic variants, or variants of unknown significance), clinical implications, and recommended further testing for the patient or their family. The referring clinician is then responsible for explaining test results, aided by a genetic counsellor, and offering appropriate management or treatment;[70] this discussion should cover several key areas. Often, patients are informed about ongoing or future clinical trials[71] that may be of benefit. Due to the hereditary nature of many genetic conditions, family screening is often encouraged in the form of clinical and genetic testing. Another key consideration is family planning and reproductive choices. This is usually considered on a case-by-case basis, taking into account the cultural and religious beliefs of the patient and their family.[32] Finally, genomic studies may yield inconclusive results. In these cases, results from several family members could provide valuable information about a variant's likelihood of pathogenicity and its inheritance, particularly if the analysis includes both affected and not affected individuals.

3. Introduction to cloud computing

Cloud computing is the use of storage and computational services accessed via the Internet (*i.e.*, the "cloud"), instead of directly owning and maintaining the hardware (Fig. 2). Applications using this infrastructure are scalable, location-independent, and have significantly lower overhead costs. In particular, the maintenance and security of high-performance hardware is carried out by the expert cloud provider, who also enables the users to rapidly increase or decrease the amount of computing

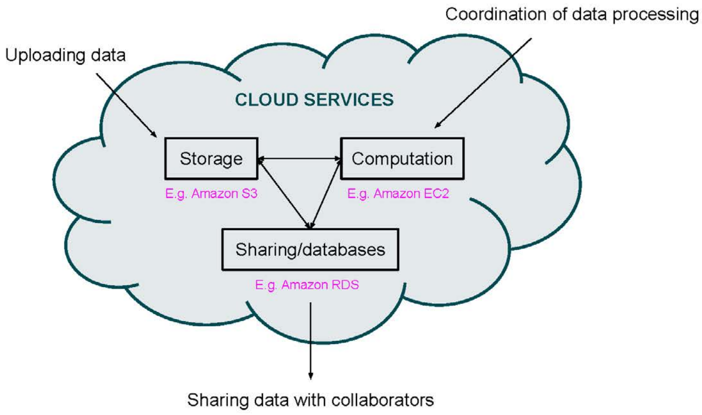


Fig. 2. General concepts in cloud computing. Cloud services generally consist of products to assist with the storage, computation, and sharing of data. Examples offered by Amazon Web Services are shown in magenta; further examples are provided in Table 1. This data may be uploaded to a remote data center via a regular internet connection. An individual in a different geographic location may be able to configure the computational resources in the cloud (*e.g.*, CPUs, GPUs) to analyze the data. The raw data or the results of analysis may be accessed using databases and easily shared with collaborators from any geographical location with internet access.

Table 1. Examples of cloud tools offered by the current main cloud service providers. Information is available through each company's website, and subject to change.

Provider	Storage	Computation	Sharing/Databases
GCP	Cloud Storage	Compute Engine	BigQuery
AWS	Amazon S3, Glacier	Amazon EC2	Amazon RDS
Azure	Azure Data Lake Storage	Azure Virtual Machines	Azure databases
IBM	IBM Cloud Object Storage	IBM Cloud Bare Metal/ Virtual Servers	IBM Cloudant
Alibaba Cloud	Storage Capacity Unit	Elastic Compute Service	ApsaraDB for PolarDB

resources being used at any moment. This makes them ideally suited for processing large amounts of scientific data and for providing a reliable service.

3.1. Services

Cloud providers can deliver different types of services depending on the needs of the user. The most common service models are infrastructure, platform, or software as a service. For example, an Infrastructure as a Service (Iaas) cloud might provide access to a server, a Platform as a Service (PaaS) cloud might provide an operating system, and a Software as a Service (SaaS) cloud might provide data analysis software. Further discussion about the nuances of each service model is beyond the scope of this review and described in detail elsewhere.[72]

3.2. Availability and accessibility

Cloud providers deliver their service through many data centers across the world, usually split into "regions". A cloud can be deployed as a public, private, or hybrid service according to who is running the data centers. For example, popular public cloud providers include Google Cloud Platform (GCP), Microsoft Azure, IBM Cloud, Alibaba Cloud, and Amazon Web Services (AWS). These companies maintain their own data centers and lease their resources to users, usually on a pay-as-you-go basis. This is simpler and may be cheaper than running a private cloud, where the user maintains their exclusively owned data center. However, the advantage of a private cloud is that it ensures the organization's direct control over the security and privacy of their software and data. Virtual private clouds (VPCs) are a way to benefit from the security of a private cloud with the simplicity of a public cloud service. For example, Amazon VPC allows the creation of an isolated section of their public cloud for specific organizations so that computing resources are not shared with other users. VPCs are essential to health care organizations to ensure security and confidentiality of patient data, and a growing number of hospitals now have access to these.

3.3. Compute and storage

Cloud service providers typically offer tools for data storage, computation, and sharing. Some examples are shown in Table 1. In addition to these primary tools, cheap alternatives are often provided. For example, “Preemptible virtual machine instances” from GCP and “Spot instances” from AWS are products that can provide computing power at a much lower price. However, instances may be stopped at any time by the cloud service provider when the computing resources are required for maintenance or by another user with a longer-term plan. These temporary instances are therefore most suited to batch analysis jobs that can be paused.

4. Genomics analysis in the cloud

In this section, we provide more technical details about genomics, including the analysis techniques typically employed when processing sequencing data, the data produced (Table 2), and the hardware and infrastructure requirements for this. This sequence of bioinformatic analysis steps is commonly referred to as a “genomics pipeline”, which is illustrated in Figure 3. A vast array of tools exists to complete each of these steps and a detailed analysis is beyond the scope of this review. Instead, we provide a broad overview of techniques, using selected examples to illustrate how genomics pipelines work and how these can be enhanced by cloud computing.

4.1. Storage and data access

Files containing genomic data can be very large (Table 2). Depending on its use, data may be stored as long-term archival storage (*i.e.*, in the range of several years), short-term storage ((in the range of weeks to months), or storage only for the duration of data analysis. Different hardware is required for each use-case. For very long-term archiving, magnetic tape storage is often used since it is offline, energy-efficient, and extremely reliable. The downside is that it is time-consuming and inefficient to subsequently access these archives. Solid state drives (SSDs) provide the fastest reading and writing speeds, but this is the most expensive form of storage, while hard disk drives (HDDs) provide large amounts of space at a much lower cost, which is useful for longer-term storage. During typical analysis, most files are stored in standard networked hard drives, and specific files are moved onto SSD storage when required for analysis, then deleted shortly afterwards. Most cloud providers offer these different types of storage.

4.2. Next-generation sequencing

NGS techniques such as WGS and WES typically generate millions of short-read sequences that need to be processed before clinically relevant findings can be made. If we imagine that each individual genome is a book, then the process

Table 2. File types and typical sizes per file type, from least to most processed

File type	WES	WGS
FASTQ	8 GB	40 GB
BAM	16 GB	120 GB
VCF	100 MB	400 MB
CSV	200 MB	2 GB

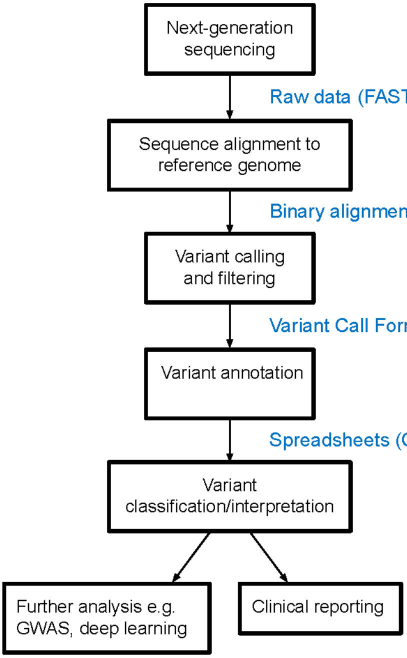

Next-generation sequencing
Raw data (FASTQ) - 40GB
Sequence alignment to reference genome
Binary alignment mapping file (BAM) - 120GB
Variant calling and filtering
Variant Call Format file (VCF) - 400MB
Variant annotation
Spreadsheets (CSV/TSV) - 2GB
Variant classification/interpretation
Further analysis e.g. GWAS, deep learning
Clinical reporting

Fig. 3. A genomics pipeline. Next-generation sequencing is the commonest method of DNA sequencing for genomics currently. This generates raw short-read data which is sent to high-performance computing clusters or uploaded to the cloud if cloud-based platforms are used. This raw data consists of millions of short sequences that must be pieced together by alignment to a reference genome. The aligned genome is then analyzed, and any differences to the reference genome are recorded as variants. There are classified as significant, insignificant, or unknown significance, with respect to biological function. This information can be used for guiding clinical decisions and is also used for further research studies. Typical file sizes for a genome sequenced at a depth of 30x to 50x are indicated in blue.

is akin to finding interesting variations of a hand-copied book against the original copy (the reference genome). Since the copied book (the genome to be sequenced) cannot be read as a whole, it must be shredded into small pieces before being read, usually hundreds of letters (bases) long. Firstly, each shredded piece is positioned using the original as a reference (sequence alignment), only after which variations to the original copy can be identified (variant calling). Each variation will be labelled based on its location, context, and frequency in the human population, and categorized before it can be evaluated for its pathogenic potential (annotation), which can then be used to prioritize variations of interest (variant prioritization). Quality control can be applied at each step to ensure reliability.

Companies such as Illumina, Macrogen, Novogene, and BGI provide a sequencing service. The raw data files produced from short-reads produced from a sequencing run need to be downloaded from the sequencing services. These data files are represented in text format using the FASTA format or its extension, the FASTQ format, which also contains read-quality information. File sizes for an individual genome are typically 8 GB and 40 GB for WES and WGS, respectively, for 100x coverage. These files are usually processed once to produce Sequence Alignment Mapping (SAM) files (see below) and can usually be archived or even deleted afterwards. FASTA and FASTQ files can therefore be stored in long-term storage such as Glacier from AWS (Table 1).

4.3. Alignment

The FASTA or FASTQ short-read sequence data is next mapped to a version of the human reference genome to produce SAM files. The SAM file assigns each short-read to a location on the human genome. This file can be used to look for large genetic variants (insertions or deletions), to ‘phase’ variants (identify whether two or more variants come from the same parent), or to validate a genetic variant. These files are accessed frequently for manual inspection to confirm whether called variants are supported by the aligned reads. SAM files are typically compressed into Binary Alignment Mapping (BAM) files to save space. Typical sizes of BAM files are 16 GB for WES and 120 GB for WGS depending on the depth of coverage. Since BAM files contain extra information about where the reads map on the human genome, their quality, their orientation and their pairing, the file sizes are larger than the FASTQ format.

Sequence alignment is the most time-consuming step of the genomics pipeline, which may typically take up to several hours or even days to complete. Fortunately, tools now exist that take advantage of distributed computing to spread the computational workload between many computers (*e.g.*, HPC nodes), thus speeding up the process.

Cloud-based workflows also enable optimization with systems such as DRAGEN (Dynamic Read Analysis for GENomics), which is a specialized platform provided by Illumina, consisting of hardware and software dedicated to genomics analysis. In the DRAGEN-GATK collaboration, developers from Illumina and The Broad Institute closely collaborate, taking advantage of hardware acceleration from DRAGEN and analysis software from GATK.[73]

4.4. Variant calling

Once aligned, identification of small-scale mismatches (*i.e.*, “variants”) against the reference genome, also known as variant calling, can be obtained using software such as the GATK[74] or Google DeepVariant.[75] Both these approaches are popular, as they offer machine learning-based methods of calling and filtering variants.

The data format produced by variant calling is the Variant Call Format (VCF). It is a standard format used to store the location and associated information of genomic variants. The data are stored in a human-readable manner. Typical uncompressed file sizes for VCF are 100 MB for WES and 400 MB for WGS. It can be efficiently indexed for fast search over the Internet. This means it can be stored on cloud storage such as S3 from AWS. Large VCF files can now be stored using dedicated cloud-based distributed databases, such as Hail.is, that allow storage in large data tables for fast column-wise and row-wise access and scalable analysis.[76]

4.5. Variant annotation

Following variant calling, annotation of variants using software such as the Variant Effect Predictor[77] retrieves information about variants for external databases. This includes information about the frequency of the variant in the general population, the affected gene, and the predicted effect on the protein or on the gene expression. Annotation is crucial for developing clinical insights, and several large databases such as Clinvar/Clingen,[78] dbSNP,[79] and the Genome Aggregation Database (gnomAD)[80] contain relevant annotations for thousands of previously discovered variants. Some of these datasets are available on Amazon S3, which means they can be efficiently shared.

Following variant annotation, files are in tabular format such as comma-separated values (CSV) or tab-separated values (TSV). This is the best format for human viewing using Excel or for further analysis using the R programming language. These files can also be stored in databases for querying. Typical uncompressed file sizes for CSV files containing WES data are 200 MB and 2 GB for WGS. These tend to be larger than VCF files as they can contain extra information in the format of variant annotation (*e.g.*, gene name, transcript consequence, allele frequency, pathogenicity prediction, etc.).

These can be loaded and queried using distributed cloud databases such as BigQuery from GCP[81] or Athena from AWS, which allow for fast searching and filtering of variants using standard search queries at a much larger scale than can be achieved using local hardware.

4.6. Variant interpretation and clinical reporting

Following variant annotation, as part of the reporting of results (section 2.2.3), the potential clinical relevance of variants is derived from its annotation as well as other sources of information such as the patient's phenotype or family history discussed during the MDT meeting. The American College of Medical Genetics and Genomics and the Association for Molecular Pathology created a framework for variant classification to establish consistent standards and guidelines that can be applied to all variants in relation to Mendelian disorders.[82] According to this framework, a variant can be classified into one of the following five categories: "Pathogenic", "Likely Pathogenic", "Uncertain Significance", "Likely Benign", and "Benign". To classify a variant into one of these classes, 28 evidence criteria are defined, each of which supports either pathogenic or benign classification at various levels. The combination of the evidence criteria defines the variant's final classification (Fig. 4).

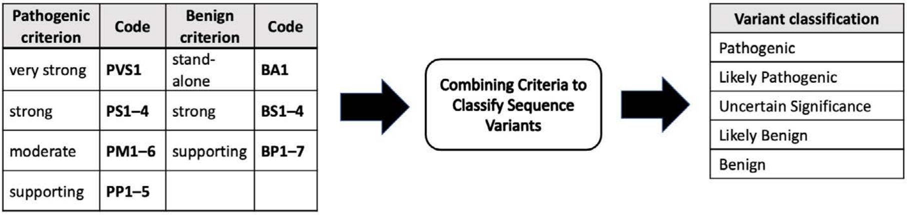
These 28 criteria span across different evidence types (*e.g.*, reported evidence, population, and computational data), so that assessment of the potential pathogenicity of a variant takes into account its frequency in the general population as well as the consensus of several computational tools. For example, if the allele frequency of a variant in gnomAD is high, it is likely too common to cause a rare disease. As a result, this observation of a high variant allele frequency will support the benign classification of the variant. If, on the other hand, the variant is absent or only present at extremely low frequency in a population database, then this finding will support the pathogenic classification of the variant.

Variant classification is a crucial step during clinical genomics investigations because it reveals the variants which are likely to be responsible for the patient's clinical presentation, and molecular diagnoses are crucial for targeted treatments.[32] However, this step may also reveal pathogenic variants not directly related to the current clinical presentation. Therefore, careful genetic counselling is required when reporting these incidental findings. Conversely, such incidental variants may be useful in the future as they may allow early screening and treatment of disease.

Since variant classification is heavily reliant on aggregating information from various sources, it would greatly benefit from application programming interface (API) connectivity. An API is a web-accessible link that developers can use in their code to connect to various databases, obtaining up-to-date information and allowing automatic collaboration between researchers across the world.[83,84]

4.7. Further analysis: statistics, machine learning, and artificial intelligence

Along with the interpretation of individual variants for clinical reporting, large amounts of genomics data allow for statistical and machine learning approaches to variant interpretation. For example, genome-wide association studies (GWAS) compare genomics data from thousands of individuals, grouped as either cases or controls, in order to identify statistically significant variants that are more commonly present in cases than in controls.[85] GWAS approaches work well for common variants



Pathogenic criterion	Code	Benign criterion	Code
very strong	**PVS1**	stand-alone	**BA1**
strong	**PS1–4**	strong	**BS1–4**
moderate	**PM1–6**	supporting	**BP1–7**
supporting	**PP1–5**		

Variant classification
Pathogenic
Likely Pathogenic
Uncertain Significance
Likely Benign
Benign

Fig. 4. The process of variant classification. The American College of Medical Genetics (ACMG) and the Association of Molecular Pathologists (AMP) defined 28 evidence criteria to assess a variant located on a gene that has a definitive role in a Mendelian disease. Each of these criteria is assigned a criterion code that summarizes the type of impact and the level of strength attributed to the evidence criterion. The first letter in the code is either “P” which stands for pathogenic or “B” for benign, after which the abbreviation of the strength level is shown, where “P” stands for supporting, “M” for moderate, “S” for strong, “VS” for very strong, and “A” for standalone. The numbers after these letters refer to different criteria, they do not indicate any difference in strength within the same type and level of evidence group. After assessing a variant to each of these criteria, the criterion codes that are met by the variant are combined following a set of rules provided in the ACMG-AMP guidelines, resulting in a five-tier classification system of the following terms: “Pathogenic”, “Likely Pathogenic”, “Uncertain Significance”, “Likely Benign”, and “Benign”.

such as single nucleotide polymorphisms (SNPs) with a population frequency of at least 5%. Another related approach are gene-based burden tests, which follow the same principle but where the aggregate burden of variants per gene is compared between clinical cases and controls.[86] Gene burden tests are more appropriate when working with rare variants with a population frequency of less than 5%. Public databases such as gnomAD can facilitate gene burden testing by providing an easily accessible set of controls.[80,86]

When applying statistical tests, nongenetic factors such as ancestry can skew the results due to genetic relatedness. This means that some genetic variants will appear more frequently because of shared ancestry rather than disease phenotype. Linear mixed effect models (LMMs) are statistical models that are commonly used to correct for these complicated hidden structures in GWAS studies by modelling the relatedness between individuals based on the genetic variants detected.[87-90] Gene burden testing and LMMs have significant computational requirements that can be addressed with cloud infrastructure, as recently evidenced by the Azure FaST-LMM service.

Large GWAS have been applied to ophthalmology for glaucoma, diabetic retinopathy, keratoconus, and other conditions.[91] Once a GWAS is complete, a number of SNPs are identified as being significantly associated with a disease. These can then be taken forward to build a polygenic risk score (PRS), which is a tailored measure of genetic risk for an individual to develop a condition such as glaucoma. The PRS is not currently used in clinical practice but may be a useful triage tool for prioritizing monitoring of high-risk patients in the future. GWAS are published in databases such as the NHGRI-EBI GWAS catalog,[92] where the data are available for future reference.

Deep learning is also increasingly used to process large volumes of genomic data and generate new insights; this is thoroughly reviewed elsewhere.[93,94] Genomic analytical tools using deep learning are often utilized to improve or complement the variant calling or subsequent analytical steps to correctly identify the variants' clinical significance in relation to the observed disorder. An example is DeepVariant, which is used for the variant calling process. DeepVariant greatly decreases the systematic errors and biases that are frequent when using standard variant-calling tools.[75]

Deep learning is also used to decipher the connection between the observed presentation of a disorder in a patient and its genetic cause (phenotype-to-genotype mapping) by assessing the predicted results of the identified candidate pathogenic variants.[95] Besides identifying the genetic cause of a disorder, another equally important aim in clinical genomics is being able to predict the risks of developing disorders later in the patient's life (genotype-to-phenotype prediction). In most cases, alongside the inherited genetic features, there are several nongenetic risk factors such as environmental exposures and lifestyle choices that together determine the risks of developing diseases.[96.97] In the future, genotype-to-phenotype predictions will most likely take into account genetic and nongenetic health data such as blood tests and imaging to reflect the complex interaction between risk factors. Machine learning methods are already being developed to integrate data from multi-omics studies, for example including genetics, proteomics, and metabolomics to discover new biomarkers for disease.[98]

Deep learning applications have traditionally been limited by hardware requirements. Modern deep learning is typically heavily dependent on graphics processing units (GPUs), which enable thousands of calculations to progress simultaneously at great speed. Specialized hardware such as Google's tensor processing units (TPUs) are continually being developed to cater for the increasing technical demands of modern deep learning.[99]

5. The benefits and challenges of cloud for genomics

5.1. Benefits of cloud integration in genomics

There are clear benefits to how cloud integration accelerates genomics, as evidenced by success stories in the USA.

5.1.1. Success stories

The Broad Institute was launched in 2004 aiming to improve human health using insights from the Human Genome Project.[100] This collaboration between

MIT, Harvard, and affiliated hospitals involves several disciplines, ranging from computer scientists to scientists and health care professionals. Broad Institute scientists have been advancing much further than the Human Genome Project, sequencing many genomes in-house to understand biological and pathological processes. Progress has increased dramatically over the last decade, with the rate of data generation doubling each year. The GCP partnership with the Broad Institute has allowed for significant improvements in their genomics pipelines. This has resulted in a 4-fold increase in the speed of processing and analyzing sequence data compared to when using in-house infrastructure. In addition, the cost of running a genome across the whole pipeline is estimated to be around 5 USD,[101] demonstrating that costs can be minimized with appropriate optimizations in the pipeline. The Broad Institute has made much of their data and analysis tools available for use worldwide through gnomAD[80,102] and Hail,[76] highlighting the collaborative power of cloud deployment.

In the hospital setting, a cloud computing service provided by AWS has been used to develop highly predictive models on electronic health record (EHR) data in a secure manner.[103] This is a strong proof-of-concept that this technology may enable both storage and analysis of EHR data, be integrated into the EHR itself, and deploy machine learning algorithms as decision support tools. Furthermore, in the biomedical setting, data backup performance is faster and more consistent when using cloud storage compared with the use of noncloud systems.[104] This suggests that large amounts of sensitive data like genomic sequencing may be best handled using a cloud-based approach. To our knowledge, the Mayo Clinic is the only group of hospitals that is currently using a cloud-based genomics pipeline to contribute to patient care. Little information is publicly available, but it is clear that collaboration with cloud providers is being used to deliver precision medicine in the form of recommendations based on genomic sequencing data.[21,22]

5.1.2. Scalability and extensibility

Cloud provides a scalable and highly secure computing and storage system. Depending on the requirements, multiple analysis programs running in Docker containers on the Google Compute Engine or Amazon Elastic Compute Cloud (EC2) can be instantly launched to accelerate analysis (Table 1). Software tools such as DISSECT have been developed to take advantage of these computing clusters to accelerate genomics and epidemiology studies.[105]

Cloud-native platforms such as Terra, Illumina Basespace, and Lifebit allow biomedical researchers to conveniently build and run pipelines for processing genomic data without the need to set up and configure cloud infrastructure (Table 3). These platforms usually optimize resource management, thus reducing costs by, for example, making use of unused computing power (*e.g.*, “spot-instances” on AWS). Cloud-native platforms make it easy for researchers without coding knowledge to run existing genomic pipelines and assemble their own through

Table 3. Examples of cloud-native platforms. These allow researchers to quickly build genomics pipelines without manual configuration of cloud infrastructure.

Platforms	Cloud provider
Terra	GCP
Illumina Basespace	AWS
Lifebit	AWS, GCP, or other private infrastructure
Galaxy	In-house clusters, German Network for Bioinformatics Infrastructure (de.NBI)
SevenBridges	AWS, GCP, or other private infrastructure

a web interface. However, if more bespoke analysis is needed, pipelines can be assembled using specifically designed programming languages for building pipelines, such as SnakeMake,[106] Nextflow,[107] or CWL.108

5.1.3. Data sharing

Clinical genomics requires the storage, analysis, and sharing of large amounts of data. Productivity and collaboration between research institutions is greatly enhanced by tools such as gnomAD, which is a collection of > 125,000 exomes and 15,708 genomes from human sequencing studies, publicly available and hosted by the Broad Institute in collaboration with GCP.[80,102,109] GnomAD is invaluable for genomics research; this was illustrated elegantly when researchers used a genomics analysis pipeline running in GCP to analyze data within gnomAD, and discovered > 400,000 structural variants (rearrangements of large sections of DNA), many of which may be clinically relevant.[110] Further to this, specialized websites and APIs can designed to allow researchers to query and visualize data collaboratively.[111,112]

5.1.4. Integration with medical imaging in ophthalmology

Ophthalmology relies heavily on visual pattern recognition from imaging data, and so it has been a prime target for the development of deep learning algorithms that readily automate and scale this process.[61,113-116] The deployment of large-scale deep learning systems enables the transfer and storage of large amounts of information. As we have seen, cloud computing fits these requirements very well.[117,118] Furthermore, imaging devices may be integrated into cloud-based systems to enable more efficient data upload and analysis, for example via GCP's Cloud Healthcare API.[119] In the future, these image analysis workflows could be easily integrated with genomics pipelines in the cloud, which is likely to improve clinical diagnostics and personalized care.[120]

5.2. Challenges to implementing cloud computing in health care

As the significant benefits of cloud computing are increasingly apparent, this will stimulate further adoption. However, barriers to widespread adoption remain that need to be addressed.

5.2.1. Information governance

The biggest challenge to the widespread adoption of cloud computing in hospital settings is the balance of data availability with security and privacy.[121] Data security refers to the protection of data from unauthorized access or manipulation; this is achieved through technical tools like encryption and physical security of server hardware. For example, cloud providers such as GCP and AWS combine physical security of data centers and hardware redundancy to ensure security.[122-124]

Regulatory compliance is an important and expensive undertaking. In particular, ISO 127001, which is a security standard for computing infrastructure, is prohibitively expensive to achieve with one's own data centers. Unlike most hospital systems, cloud providers such as AWS and GCP are already compliant with such regulations; therefore, using cloud services may help adopt new information technology at a fraction of the cost.

Data privacy refers to the limitation of data collection, storage, and usage to protect individuals. In the context of genomics, this refers to the prevention of misuse of genetic information to perpetuate social stigma or target marketing campaigns. One of the most notable legal instruments that exist to protect data privacy is the European Union's General Data Protection Regulation (GDPR), which came into force in 2018.[125] The GDPR aims to facilitate the flow of personal information and protect the fundamental rights of individuals to privacy. It stipulates that personal data should not be processed unless there is at least one legal basis to do so: if the data subject (*e.g.*, the person who has their genome sequenced) has given consent, to fulfil a contractual obligation with the data subject, to comply with other legal obligations, to protect the vital interests of the data subject or other person, for the public interest, or for the legitimate interests of a third party. The principles are therefore vague and subject to some degree of freedom of interpretation.

Furthermore, the "right of erasure" is a key right that data subjects have. For example, a few years after having their genome sequenced, a person may wish to have their data deleted permanently; this must be fulfilled by the "data controller" (the person or company holding the information). This can cause significant technical problems due to the complexity of backups and may pose serious challenges when data has already been used (*e.g.*, as part of a research paper).[126] Cloud computing systems must fulfil these requirements, which could be challenging because the information is often stored in multiple physical locations. Hospitals that wish to implement cloud-based systems should therefore ensure compliance with local and national laws in the design phase, so that legal requirements, such as erasure, may be easily fulfilled. Partnerships with cloud providers should also specify the geographic locations of servers to ensure transparency of data handling. Ultimately, clear communication and informed consent from patients are likely to be the most important legal instrument in enabling implementation of cloud-based clinical genomics; this could be delivered during genetic counselling appointments, for example.[72]

Cloud platforms are designed so that users can access and control computing resources remotely, for example, via secure shell (SSH) or a browser window. Therefore, to protect data security and privacy, information must be protected from unauthorized access on the cloud server, the user's device, and when in transit between the two. Security systems must be in place to reduce the risk of data breaches; the stakes are also very high because patients may be identified using their genetic information.[72] Cloud providers typically implement systems such as encryption of data both at rest on hard drives and when in transit. Techniques such as federated learning enable analysis of data to occur on remote devices, therefore eliminating the need for data transfer in the first place.[127] Data is usually decrypted to carry out mathematical operations during analysis. However, recent advances have shown that techniques such as homomorphic encryption may allow data to be analyzed in an encrypted state, thus protecting the information at all times.[128] Similar techniques are used by GCP in the form of Confidential Computing and Differential Privacy.[129,130] During the implementation of such systems in hospitals, additional arrangements should be made to limit access to data by cloud service providers. In addition, one must consider measures to return or securely destroy information in the event of contracts ending or regulation breaches. The NHS has published guidance for the use of cloud computing services in healthcare, outlining these considerations and relevant local regulations.[131]

5.2.2. Cost

Another barrier to cloud adoption is cost; prices for complex combinations of cloud services are not always transparent or clearly explained. Cloud providers such as AWS and GCP usually have regions (*e.g.*, eu-west1) and whilst data transfer within a region is free or cheap, data transfer between regions is expensive. Furthermore, certain types of subservices like AWS Marketplace which allow providers to license their products to AWS have a different billing system: hourly rates which can confuse end-users and lead to increased cost. Transparency and user satisfaction could be improved by offering live billing as a default rather than end-of-the-month bills, or the ability to set a maximum amount that a user is willing to spend per month for peace of mind. One existing solution that cloud providers offer to help better manage costs are reserved instances, which allow users to pay upfront for some infrastructure such as servers at a discounted rate. Another solution are better resource management tools that allow users to make use of unused resources, such as “spot-instances”, on the condition that their jobs can be preempted at any time. Cheaper alternatives exist. A new storage service known as Wasabi offers long-term storage at a much-reduced rate.[132] Mythic Beasts is a UK cloud provider offering cheaper access to computing power. Nonetheless, the daunting prospect of nontransparent costs remains one of the main barriers to widespread adoption. It is also often difficult to judge the likelihood of failure and subsequent data loss when using smaller cloud service providers.

Furthermore, during the research process pipelines may be run multiple times due to mistakes or experimentation. This affects students and experienced researchers alike. When running pipelines on local servers, only time is lost, but this may be expensive when run on the cloud. One possible solution may be to run experiments on local HPCs, but implement completed workflows on the cloud to leverage speed and ease of sharing data or analysis results.

6. The clinical need and the way forward

Cloud platforms enable health care organizations to integrate genomics into routine care whilst also facilitating research.

6.1. The need for efficient querying and linkage to clinical data

In the UK, research hospitals such as Moorfields Eye Hospital NHS Foundation Trust have contributed largely to genome initiatives such as Genomics England, which now contains the genomes of over 100,000 individuals from the UK. However, this data is currently only accessible through the Genomics England embassy systems, a remote desktop application, and is not therefore integrated with the hospital’s clinical data. This means research staff and clinicians at the hospital cannot query the genomic data effectively. The lack of integration makes it challenging to make the most of this data, such as establishing gene-phenotype correlations, verifying clinical results in light of new data, and to conduct meaningful research into new genetic causes of disease. For patients, this can result in a lower diagnostic yield. For example, if a new gene is found to be associated with a retinal disease, then the clinical researchers may want to query all existing patients with a similar condition to see whether they have any variants in that gene.

6.2. Efficient reanalysis of data and development of tools for new insights

It is important to understand that our knowledge of the human genome and what constitutes genomic variation is continuously expanding. Therefore, there is always a need to revisit previous raw genomic data and reanalyze it in different ways. For example, there are several possible reference maps for the human genome and several types of analyses that can be done to discover new types of variation. To note one example, the previously described FASTQ and BAM files can be reanalyzed at a later point to discover new types of genomic variation (*e.g.*, structural rearrangements). In light of new information and new tools, it is therefore important for hospitals to be able to conduct their own research, which is pertinent to their needs, in the same way that is being achieved with image analysis. For example, if a substantial amount of the genomic analysis has been using an older version of the reference genome, clinical researchers may want to realign their data to the newest build. Another example may be a new tool to interpret noncoding variants or to more effectively identify structural variants that might explain the cause of the disease in a subset of individuals. Both these examples would be easily accomplished using cloud systems due to the flexibility of storage options and scalable computing power that may be organized into pipelines.

6.3. Triage and surveillance

A future application of genomics in the health care setting is triaging and surveillance of common treatable conditions via genetic risk profiling. PRS, which are derived from GWAS, can estimate an individual's life-long risk for age-related macular degeneration, glaucoma, or diabetic retinopathy. Although they are currently used only in research settings, they may be adopted into clinical practice in the future. As PRS will also likely need updating in light of new discoveries, reanalysis from raw data will likely be necessary. Hospitals running cloud-based EHRs would be perfectly primed for this, since detailed clinical data would be easily accessible.

Genetic testing can also be useful to study the genetic basis of drug response, known as "pharmacogenomics",[133-135] which is important to avoid adverse drug reactions and maximize efficacy when planning management. For example, anti-vascular endothelial growth factor (anti-VEGF) injections are the mainstay treatment for exudative age-related macular degeneration; some genotypes show increased response to this treatment, whereas some show a reduced efficacy.[136] Genetic variation in the VEGF signaling pathway may also explain variations in response to treatment of proliferative diabetic retinopathy.[137] In the future, this may be useful for tailoring pharmacological treatments based on genomic data.

6.4. Ethics and regulation

With these new exciting applications of genomics in healthcare, careful ethical and regulatory oversight are needed, since large amounts of clinical data are being used for research.[138] The increased accessibility of clinical data for research poses important questions for consent. The UK Biobank recruited a large cohort of approximately 500,000 people between 2006 and 2010 for a prospective study examining the lifestyle, genomic, and environmental determinants of serious illnesses. Data was released in 2012; since then, it has become a major open-access resource for researchers.[139] The UK Biobank has also adapted research consent ethics for the genomic era;[140] for example, the participants consented for their data to be used for third-party research projects generally, but consent was not obtained for specific uses of their data. [139] This allowed maximum flexibility for third parties (industry or academia) to use this data for research purposes. This precedent could be applied in future to health care data shared on the cloud, thus facilitating the integration of clinical care with research.

When using clinical data, incidental findings could be made, particularly when analyzing a whole genome. The question of what to do with these findings has been a controversial topic in genomics for a long time.[141,142] The American College of Medical Genetics and Genomics recommended in 2013 that clinical genomics tests should not only test for genes of interest, but also conduct a search for a set of variants deemed to be of medical value, with no option for the patient to decline unwanted information; this was widely seen as unethical due to the disregard for patient autonomy.[60,143,144] However, due to the huge positive clinical impact genomics could have, others have advocated for a balanced approach by regulators, arguing that the use of genomic data for research should be permitted without explicit consent, provided that mechanisms protecting data security and privacy are put in place.[145]

These issues will be compounded upon the introduction of cloud computing for genomics pipelines because it will greatly enhance the accessibility of genetic data. Moving forwards, it will be crucial that the delicate balance between patient privacy and data accessibility for research is negotiated with care, involving discussions with patients, clinicians, researchers, and cloud service providers.

6.5. A vision of a future cloud-based clinical genomics pipeline

Initiatives for making genomic data available to the research community have been driven by projects such as the Personal Genome Project. For example, the Personal Genome Project UK provides genomics data (microarray, WGS, and WES) but also transcriptomics and methylation data.[146.147] All this data is under a Creative Commons license that places them in the public domain, allowing them to be downloaded without any registration, and as such has been integrated into cloud providers and genomic web platforms (Table 3). These open datasets can be a first step towards prototyping cloud-based clinical genomics pipelines.

For the UK, large amounts of genomics data currently from NHS patients resides in the Genomics England embassy and other genomic studies. Bringing this data back into a health care setting via cloud-enabled hospitals would allow for linkage to detailed clinical data and leveraging genomics for clinical use.

To illustrate how a clinical genomics pipeline might operate in a hospital setting, we now consider a hypothetical patient who presents to an eye hospital in 2030 with a common eye problem (Fig. 5). Mrs. XX is referred to her local hospital due to impaired vision. Her ophthalmologist takes a full history and carries out a thorough examination, and specialist ophthalmic nurses perform automated perimetry, OCT, and retinal fundus imaging. From this clinical evaluation, Mrs. XX is diagnosed with a rare form of primary open-angle glaucoma, which was a

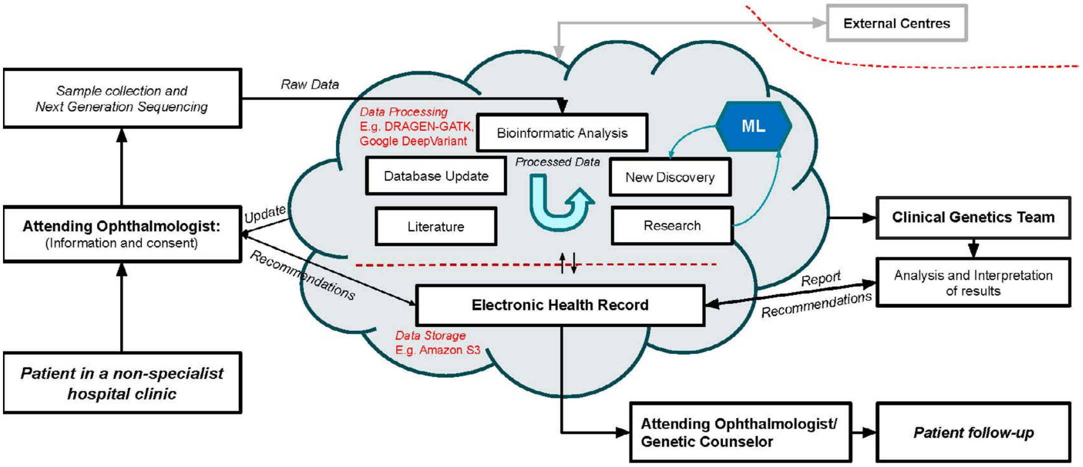


Fig. 5. A proposed cloud-integrated clinical genomics pipeline. Bioinformatic analysis is automatically performed based on the most current database of known significant variants. Variants of uncertain significance could be stored for further analysis or research. Cloud-based platforms such as DRAGEN-GATK may be used for variant calling, annotation, and classification. Machine learning (ML) algorithms may be used to discover new patterns and recommend particular treatments if applicable, as well as facilitating data processing steps such as sequence alignment and variant calling. If a discovery or recommendation is made, it will be integrated into the cloud system and be immediately available to clinicians and researchers with prior authorization. The attending physician and the whole clinical genetics team constantly get feedback and alerts from the cloud system. The EHR has a key role, as all data obtained from clinical or laboratory testing is constantly updated, being part of the cloud. The EHR has multiple data safety and privacy protections, with strictly controlled interactions with the rest of the cloud. Cloud-based systems allow access to authorized personnel in external centers such as specialist hospitals or research groups. This could decrease the need for referral to specialized centers for clinical evaluation and treatment.

cause of irreversible vision loss back in 2021. However, effective treatments are now available for some genetic forms of open-angle glaucoma. The ophthalmologist carefully explains the value of genetic testing to Mrs. XX whilst checking that she understands fully and provides written information. The discussion covers therapeutic options, including gene therapy and drug combinations, which may be tailored to suit the genetic variants present in Mrs. XX's genome. Mrs. XX is also reassured that many technical and regulatory measures are in place to protect the security of her genetic data as well as her privacy and that of her family members. Family planning is also discussed, since the results may affect her decision to have children, and Mrs. XX understands that her data may be used for research purposes. Mrs. XX then decides to go ahead with a blood test for genetic testing, although she had the option for further appointments with a genetic counsellor to go into more details of the implications of genetic testing. Specifically, Mrs. XX consents to the use of her genetic information to guide her management plan, and to contribute to research in a secure manner that ensures that she is not identifiable. She also specifies that her data should be permanently erased after 5 years if her disease progression has halted.

The blood samples are collected in the clinic and sent to a dedicated laboratory where the genetic material is extracted and undergoes NGS. The raw data is uploaded onto a cloud platform integrated with the hospital's EHR, where it automatically undergoes bioinformatic analysis designed by the hospital clinical genetics team, and a preliminary report is generated. The process of analyzing the genomic data in the cloud platform costs the health system approximately 5 USD, similar to many other blood tests. The hospital team receives this report immediately via the EHR, and a multidisciplinary meeting comes to an agreement that Mrs. XX has genetic variants in lipid metabolism genes that may be targeted to treat her glaucoma,[148] so funding is put into place for gene therapy. Mrs. XX is booked in for an appointment a week after her first visit, and these findings are discussed with her. After extensive genetic counselling, she then consents to gene therapy, thus halting her disease progression and saving her sight from further deterioration. Authorized research groups are able to access the genetic data via the cloud and use homomorphic encryption to analyze the data securely, whilst discovering new insights about the pathogenesis of glaucoma which guide future research into potential therapies.

7. Conclusions and future directions

Genomics is already revolutionizing diagnosis and targeted management in clinical practice whilst also becoming exponentially cheaper. The Human Genome Project,[6,149,150] an international venture with a total cost of 3 billion USD, reached the goal of sequencing the euchromatic regions of the genome (92.2% of the total genome with a 99% accuracy) from a small group of human donors. Current costs of genome sequencing fell well below Moore's Law,[151,152] such that the price of a complete WES is currently below the 1000 USD barrier.[152] Indeed, in a staggering feat of optimization, GCP and The Broad Institute have reduced the cost of sequencing and running the GATK Best Practices pipeline to roughly 5 USD per genome.[101]

Cloud-based systems have many features that may facilitate the clinical application of genomics studies. The on-demand, scalable nature of cloud services is especially useful when managing the large amounts of data involved in genomic experiments. This may also be a more economical option than continually maintaining local services that may not be in use at all times, particularly if data is generated or analyzed in a batch manner. The cloud could enable hospitals with strained budgets to minimize their need for maintaining physical infrastructure, security, or recovery of hardware. Cloud technology also offers wide access to stored data, and this could greatly facilitate analysis and interpretation, which is a highly multidisciplinary clinical pathway. In addition, the benefits extend to research, since more data could be made available to research groups that may use "data-hungry" analytic techniques, and the insights generated may give feedback to improve clinical practice.

Cloud computing is perfectly primed to facilitate the widespread adoption of genomic analysis in clinical practice. Under this model, data storage and computational power are much more scalable and cost-effective for hospitals than local computing solutions, which require significant maintenance and down-time. In addition, collaboration between institutions such as hospitals, sequencing companies, and research groups is easily achieved on cloud platforms, eliminating the need for complex data transfer arrangements. Such systems may be readily incorporated into EHRs, thus greatly improving the accuracy and speed of patient care. This collaborative environment may also allow all health care professionals to have access to the most up-to-date information about any pathological genetic findings, irrespective of their location. Hence, all patients may have the same amount and quality of information in their genomic reports, regardless of whether they are in a specialist hospital or not, or whether they are in a developed country or not. With the combination of technological advancement and falling costs, cloud-based clinical genomics will soon become routinely used beyond suspected genetic or hereditary conditions.

Declarations

Ethics and consent to participate

Not required.

Competing interests

Dr. Keane has acted as a consultant for DeepMind, Roche, Novartis, and Apellis, and is an equity owner in Big Picture Medical. He has received speaker fees from Heidelberg Engineering, Topcon, Allergan, and Bayer. Dr. Olivera has received travel grants from Allergan and Alcon. Dr. Khawaja has acted as a consultant for Aerie, Allergan, Google Health, Novartis, Reichert, Santen, and Thea. Dr. Pontikos and Dr. Moghul are co-founders and directors of Phenopolis Ltd, a software company active in health care research.

Funding

Dr. Keane is supported by a Moorfields Eye Charity Career Development Award (R190028A) and a UK Research and Innovation Future Leaders Fellowship (MR/T019050/1). Dr. Olivera is supported by a Spanish Vitreoretinal Society long-term scholarship. Dr. Pontikos is supported by a Moorfields Eye Charity Career Development Award (R190031A). Ismail Moghul is supported by the Biotechnology and Biological Sciences Research Council (Grant No. BB/M009513/1). Dr. Khawaja is supported by a UK Research and Innovation Future Leaders Fellowship, a Moorfields Eye Charity Springboard Award, and an Alcon Research Institute Young Investigator Award.

Acknowledgements

Dr. Pontikos would like to thank the Archer family in memory of Marion Woods.

References

1. Dulbecco R. A turning point in cancer research: sequencing the human genome. Science [Internet]. 1986 Mar 7;231(4742):1055–6. https://doi.org/10.1126/science.3945817
2. Collins FS. Shattuck lecture--medical and societal consequences of the Human Genome Project. N Engl J Med [Internet]. 1999 Jul 1;341(1):28–37. https://doi.org/10.1056/NEJM199907013410106
3. Guttmacher AE, Collins FS. Genomic medicine--a primer. N Engl J Med [Internet]. 2002 Nov 7;347(19):1512–20. https://doi.org/10.1056/NEJMra012240
4. Manolio TA, Chisholm RL, Ozenberger B, et al. Implementing genomic medicine in the clinic: the future is here. Genet Med [Internet]. 2013; Apr;15(4):258–67. https://doi.org/10.1038/gim.2012.157
5. Roberts MC, Kennedy AE, Chambers DA, Khoury MJ. The current state of implementation science in genomic medicine: opportunities for improvement. Genet Med [Internet]. 2017 Aug;19(8):858–63. https://doi.org/10.1038/gim.2016.210
6. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science [Internet]. 2001 Feb 16;291(5507):1304–51. https://doi.org/10.1126/science.1058040
7. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature [Internet]. 2001 Feb 15;409(6822):860–921. https://doi.org/10.1038/35057062

8. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? Genet Med [Internet]. 2007 Oct;9(10):665–74. https://doi.org/10.1097/GIM.0b013e31815699d0
9. Kalia SS, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med [Internet]. 2017 Feb;19(2):249–55. https://doi.org/10.1038/gim.2016.190
10. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med [Internet]. 2013 Jul;15(7):565–74. https://doi.org/10.1038/gim.2013.73
11. Russell S, Bennett J, Wellman JA, Chung DC, Yu Z-F, Tillman A, et al. Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. Lancet [Internet]. 2017 Aug 26;390(10097):849–60. https://doi.org/10.1016/S0140-6736(17)31868-8
12. Dorschner MO, Amendola LM, Turner EH, et al. Actionable, pathogenic incidental findings in 1,000 participants' exomes. Am J Hum Genet [Internet]. 2013 Oct 3;93(4):631–40. https://doi.org/10.1016/j.ajhg.2013.08.006
13. Turro E, Astle WJ, Megy K, et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature [Internet]. 2020 Jul;583(7814):96–102. https://doi.org/10.1038/s41586-020-2434-2
14. Griebel L, Prokosch H-U, Köpcke F, et al. A scoping review of cloud computing in healthcare. BMC Med Inform Decis Mak [Internet]. 2015 Mar 19;15:17. https://doi.org/10.1186/s12911-015-0145-7
15. Ali O, Shrestha A, Soar J, Wamba SF. Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review. Int J Inf Manage [Internet]. 2018 Dec 1;43:146–58. Available from: http://www.sciencedirect.com/science/article/pii/S0268401218303736
16. Gao F, Sunyaev A. Context matters: A review of the determinant factors in the decision to adopt cloud computing in healthcare. Int J Inf Manage [Internet]. 2019 Oct 1;48:120–38. Available from: http://www.sciencedirect.com/science/article/pii/S0268401218307266
17. Zandesh Z, Ghazisaeedi M, Devarakonda MV, Haghighi MS. Legal framework for health cloud: A systematic review. Int J Med Inform [Internet]. 2019 Dec;132:103953. https://doi.org/10.1016/j.ijmedinf.2019.103953
18. Mell P, Grance T. The NIST definition of cloud computing [Internet]. National Institute of Standards and Technology; 2011. Report No.: 800-145. Available from: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf
19. UK Government Cloud First policy [Internet]. Government Digital Service; 2017 Feb [cited 2021 Mar 3]. Available from: https://www.gov.uk/guidance/government-cloud-first-policy
20. Internet First Policy - NHS Digital [Internet]. NHS Digital; 2020 Jan [cited 2021 Mar 3]. Available from: https://digital.nhs.uk/services/internet-first/policy
21. Kurian T. How Google and Mayo Clinic will transform the future of healthcare [Internet]. 2019 [cited 2020 Dec 22]. Available from: https://cloud.google.com/blog/topics/customers/how-google-and-mayo-clinic-will-transform-the-future-of-healthcare

22. Sheffi J, Vaisipour S. Accelerating Mayo Clinic's data platform with BigQuery and Variant Transforms [Internet]. 2020 [cited 2020 Dec 22]. Available from: https://cloud.google.com/blog/products/data-analytics/genome-data-analytics-with-google-cloud
23. Singh M, Tyagi SC. Genes and genetics in eye diseases: a genomic medicine approach for investigating hereditary and inflammatory ocular disorders. Int J Ophthalmol [Internet]. 2018 Jan 18;11(1):117–34. https://doi.org/10.18240/ijo.2018.01.20
24. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature [Internet]. 2017 Oct 19;550(7676):345–53. https://doi.org/10.1038/nature24286
25. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet [Internet]. 2016 May 17;17(6):333–51. https://doi.org/10.1038/nrg.2016.49
26. Liew G, Michaelides M, Bunce C. A comparison of the causes of blindness certifications in England and Wales in working age adults (16-64 years), 1999-2000 with 2009-2010. BMJ Open [Internet]. 2014 Feb 12;4(2):e004015. https://doi.org/10.1136/bmjopen-2013-004015
27. Solebo AL, Rahi J. Epidemiology, aetiology and management of visual impairment in children. Arch Dis Child [Internet]. 2014 Apr;99(4):375–9. https://doi.org/10.1136/archdischild-2012-303002
28. Solebo AL, Teoh L, Rahi J. Epidemiology of blindness in children. Arch Dis Child [Internet]. 2017 Sep;102(9):853–7. https://doi.org/10.1136/archdischild-2016-310532
29. Sohocki MM, Daiger SP, Bowne SJ, et al. Prevalence of mutations causing retinitis pigmentosa and other inherited retinopathies. Hum Mutat [Internet]. 2001;17(1):42–51. Available from: https://doi.org/10.1002/1098-1004(2001)17:1<42::AID-HUMU5>3.0.CO;2-K
30. Vázquez-Domínguez I, Garanto A, Collin RWJ. Molecular Therapies for Inherited Retinal Diseases-Current Standing, Opportunities and Challenges. Genes [Internet]. 2019 Aug 28;10(9). https://doi.org/10.3390/genes10090654
31. Neveling K, Collin RWJ, Gilissen C, et al. Next-generation genetic testing for retinitis pigmentosa. Hum Mutat [Internet]. 2012 Jun;33(6):963–72. https://doi.org/10.1002/humu.22045
32. Méjécase C, Malka S, Guan Z, Slater A, Arno G, Moosajee M. Practical guide to genetic screening for inherited eye diseases. Ther Adv Ophthalmol [Internet]. 2020 Jan;12:2515841420954592. https://doi.org/10.1177/2515841420954592
33. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature [Internet]. 2018 Oct;562(7726):203–9. https://doi.org/10.1038/s41586-018-0579-z
34. Turnbull C, Scott RH, Thomas E, et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ [Internet]. 2018 Apr 24;361:k1687. https://doi.org/10.1136/bmj.k1687
35. Stark Z, Dolman L, Manolio TA, et al. Integrating Genomics into Healthcare: A Global Responsibility. Am J Hum Genet [Internet]. 2019 Jan 3;104(1):13–20. https://doi.org/10.1016/j.ajhg.2018.11.014
36. Turnbull C. Introducing whole-genome sequencing into routine cancer care: the Genomics England 100 000 Genomes Project. Ann Oncol [Internet]. 2018 Apr 1;29(4):784–7. https://doi.org/10.1093/annonc/mdy054
37. Pletcher BA, Toriello HV, Noblin SJ, et al. Indications for genetic referral: a guide for healthcare providers. Genet Med [Internet]. 2007 Jun;9(6):385–9. https://doi.org/10.1097/gim.0b013e318064e70c

38. Ali MU, Rahman MSU, Cao J, Yuan PX. Genetic characterization and disease mechanism of retinitis pigmentosa; current scenario. 3 Biotech [Internet]. 2017 Aug;7(4):251. https://doi.org/10.1007/s13205-017-0878-3

39. Parmeggiani F, Barbaro V, De Nadai K, et al. Identification of novel X-linked gain-of-function RPGR-ORF15 mutation in Italian family with retinitis pigmentosa and pathologic myopia. Sci Rep [Internet]. 2016 Dec 20;6:39179. https://doi.org/10.1038/srep39179

40. Kabir F, Ullah I, Ali S, et al. Loss of function mutations in RP1 are responsible for retinitis pigmentosa in consanguineous familial cases. Mol Vis [Internet]. 2016 Jun 10;22:610–25. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27307693

41. Ullah I, Kabir F, Iqbal M, et al. Pathogenic mutations in TULP1 responsible for retinitis pigmentosa identified in consanguineous familial cases. Mol Vis [Internet]. 2016 Jul 16;22:797–815. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27440997

42. Athanasiou D, Aguila M, Bellingham J, et al. The molecular and cellular basis of rhodopsin retinitis pigmentosa reveals potential strategies for therapy. Prog Retin Eye Res [Internet]. 2018 Jan 1;62:1–23. Available from: http://www.sciencedirect.com/science/article/pii/S1350946217300769

43. Sun Y, Li W, Li J-K, et al. Genetic and clinical findings of panel-based targeted exome sequencing in a northeast Chinese cohort with retinitis pigmentosa. Molecular genetics & genomic medicine [Internet]. 2020;8(4):e1184. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/mgg3.1184

44. Iannaccone A, Breuer DK, Wang XF, et al. Clinical and immunohistochemical evidence for an X linked retinitis pigmentosa syndrome with recurrent infections and hearing loss in association with an RPGR mutation. J Med Genet. 2003 Nov;40(11):e118. https://doi.org/10.1136/jmg.40.11.e118.

45. Whatley M, Francis A, Ng ZY, et al. Usher Syndrome: Genetics and Molecular Links of Hearing Loss and Directions for Therapy. Front Genet [Internet]. 2020 Oct 22;11:565216. https://doi.org/10.3389/fgene.2020.565216

46. Carelli V, Sabatelli M, Carrozzo R, et al. “Behr syndrome” with OPA1 compound heterozygote mutations. Brain [Internet]. 2015 Jan;138(Pt 1):e321. https://doi.org/10.1093/brain/awu234

47. Katsanis N, Ansley SJ, Badano JL, et al. Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. Science [Internet]. 2001 Sep 21;293(5538):2256–9. https://doi.org/10.1126/science.1063525

48. Kersten E, Geerlings MJ, Pauper M, et al. Genetic screening for macular dystrophies in patients clinically diagnosed with dry age-related macular degeneration. Clin Genet [Internet]. 2018;94(6):569–74. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/cge.13447

49. Altschwager P, Ambrosio L, Swanson EA, Moskowitz A, Fulton AB. Juvenile Macular Degenerations. Semin Pediatr Neurol [Internet]. 2017 May;24(2):104–9. https://doi.org/10.1016/j.spen.2017.05.005

50. Rahman N, Georgiou M, Khan KN, Michaelides M. Macular dystrophies: clinical and imaging features, molecular genetics and therapeutic options. Br J Ophthalmol [Internet]. 2020 Apr;104(4):451–60. https://doi.org/10.1136/bjophthalmol-2019-315086

51. Chan B, Adam DN. A Review of Fabry Disease. Skin Therapy Lett [Internet]. 2018 Mar;23(2):4–6. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29562089

52. Nozu K, Nakanishi K, Abe Y, et al. A review of clinical characteristics and genetic backgrounds in Alport syndrome. Clin Exp Nephrol [Internet]. 2019 Feb;23(2):158–68. https://doi.org/10.1007/s10157-018-1629-4

53. Savige J, Ariani F, Mari F, et al. Expert consensus guidelines for the genetic diagnosis of Alport syndrome. Pediatr Nephrol [Internet]. 2019 Jul;34(7):1175–89. https://doi.org/10.1007/s00467-018-3985-4

54. Retina International. The socioeconomic impact of inherited retinal dystrophies [Internet]. 2019 [cited 2021 Jul 13]. Available from: https://www2.deloitte.com/au/en/pages/economics/articles/socioeconomic-impact-inherited-retinal-dystrophies.html

55. Lenassi E, Clayton-Smith J, Douzgou S, et al. Clinical utility of genetic testing in 201 preschool children with inherited eye disorders. Genet Med [Internet]. 2020 Apr;22(4):745–51. https://doi.org/10.1038/s41436-019-0722-8

56. NHS National Genomic Test Directory [Internet]. NHS England. 2020 [cited 2021 Mar 2]. Available from: https://www.england.nhs.uk/publication/national-genomic-test-directories/

57. Berrios C, James CA, Raraigh Ket al. Enrolling Genomics Research Participants through a Clinical Setting: the Impact of Existing Clinical Relationships on Informed Consent and Expectations for Return of Research Results. J Genet Couns [Internet]. 2018 Feb;27(1):263–73. https://doi.org/10.1007/s10897-017-0143-2

58. Patch C, Middleton A. Genetic counselling in the era of genomic medicine. Br Med Bull [Internet]. 2018 Jun 1;126(1):27–36. https://doi.org/10.1093/bmb/ldy008

59. Report of the Joint Committee on Genomics in Medicine. Consent and confidentiality in genomic medicine: Guidance on the use of genetic and genomic information in the clinic [Internet]. Third Edition. RCP, RCPath and BSGM; 2019. Available from: https://www.rcplondon.ac.uk/projects/outputs/consent-and-confidentiality-genomic-medicine

60. Wolf SM, Annas GJ, Elias S. Point-counterpoint. Patient autonomy and incidental findings in clinical genomics. Science [Internet]. 2013 May 31;340(6136):1049–50. https://doi.org/10.1126/science.1239119

61. Fujinami-Yokokawa Y, Pontikos N, Yang L, et al. Prediction of Causative Genes in Inherited Retinal Disorders from Spectral-Domain Optical Coherence Tomography Utilizing Deep Learning Techniques. J Ophthalmol [Internet]. 2019 Apr 9;2019:1691064. https://doi.org/10.1155/2019/1691064

62. Cipriani V, Pontikos N, Arno G, et al. An Improved Phenotype-Driven Tool for Rare Mendelian Variant Prioritization: Benchmarking Exomiser on Real Patient Whole-Exome Data. Genes [Internet]. 2020 Apr 23;11(4). https://doi.org/10.3390/genes11040460

63. Pontikos N, Murphy C, Moghul I, et al. Phenogenon: Gene to phenotype associations for rare genetic diseases. PLoS One [Internet]. 2020 Apr 9;15(4):e0230587. https://doi.org/10.1371/journal.pone.0230587

64. Köhler S, Carmody L, Vasilevsky N, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res [Internet]. 2019 Jan 8;47(D1):D1018–27. https://doi.org/10.1093/nar/gky1105

65. Sergouniotis PI, Maxime E, Leroux D, et al. An ontological foundation for ocular phenotypes and rare eye diseases. Orphanet J Rare Dis [Internet]. 2019 Jan 9;14(1):8. https://doi.org/10.1186/s13023-018-0980-6

66. Miller DT, Adam MP, Aradhya S, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet [Internet]. 2010 May 14;86(5):749–64. https://doi.org/10.1016/j.ajhg.2010.04.006
67. South ST, Lee C, Lamb AN, Higgins AW, Kearney HM, Working Group for the American College of Medical Genetics and Genomics Laboratory Quality Assurance Committee. ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013. Genet Med [Internet]. 2013 Nov;15(11):901–9. https://doi.org/10.1038/gim.2013.129
68. Clark MM, Stark Z, Farnaes L, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ Genom Med [Internet]. 2018 Jul 9;3:16. https://doi.org/10.1038/s41525-018-0053-8
69. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet [Internet]. 2014 Jan;59(1):5–15. https://doi.org/10.1038/jhg.2013.114
70. Ziccardi L, Cordeddu V, Gaddini L, et al. Gene Therapy in Retinal Dystrophies. Int J Mol Sci [Internet]. 2019 Nov 14;20(22). https://doi.org/10.3390/ijms20225722
71. US National Library of Medicine, National Institue of Health, Clinical Trials [Internet]. [cited 2021 Jan 22]. Available from: https://www.clinicaltrials.gov
72. Carter AB. Considerations for Genomic Data Privacy and Security when Working in the Cloud. J Mol Diagn [Internet]. 2019 Jul;21(4):542–52. https://doi.org/10.1016/j.jmoldx.2018.07.009
73. Van der Auwera G. DRAGEN-GATK Update: Let's get more specific [Internet]. 2021 [cited 2021 Mar 2]. Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360039984151-DRAGEN-GATK-Update-Let-s-get-more-specific
74. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics [Internet]. 2013;43:11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43
75. Poplin R, Chang P-C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol [Internet]. 2018 Nov;36(10):983–7. https://doi.org/10.1038/nbt.4235
76. Hail Team. Hail Overview [Internet]. 2020 [cited 2020 Dec 22]. Available from: https://hail.is/docs/0.2/overview/index.html
77. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. Genome Biol [Internet]. 2016 Jun 6;17(1):122. https://doi.org/10.1186/s13059-016-0974-4
78. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen--the Clinical Genome Resource. N Engl J Med [Internet]. 2015 Jun 4;372(23):2235–42. https://doi.org/10.1056/NEJMsr1406261
79. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res [Internet]. 2001 Jan 1;29(1):308–11. https://doi.org/10.1093/nar/29.1.308
80. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature [Internet]. 2020 May;581(7809):434–43. https://doi.org/10.1038/s41586-020-2308-7
81. Pan C, McInnes G, Deflaux N, et al. Cloud-based interactive analytics for terabytes of genomic variants data. Bioinformatics [Internet]. 2017 Dec 1;33(23):3709–15. https://doi.org/10.1093/bioinformatics/btx468

82. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med [Internet]. 2015;17(5):405–23. Available from: https://www.nature.com/articles/gim201530?ux=07df2189-4e01-4c08-8ef3-5619cff0ca61&ux-2=3739b439-66b5-4bf5-921e-0916eef236a7&ux3=&uxconf=Y

83. Sobreira NLM, Arachchi H, Buske OJ, et al. Matchmaker Exchange. Curr Protoc Hum Genet [Internet]. 2017 Oct 18;95:9.31.1–9.31.15. https://doi.org/10.1002/cphg.50

84. Buske OJ, Schiettecatte F, Hutton B, et al. The Matchmaker Exchange API: automating patient matching through the exchange of structured phenotypic and genotypic profiles. Hum Mutat [Internet]. 2015 Oct;36(10):922–7. https://doi.org/10.1002/humu.22850

85. Lee JJ, Wedow R, Okbay A, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet [Internet]. 2018 Jul 23;50(8):1112–21. https://doi.org/10.1038/s41588-018-0147-3

86. Guo MH, Plummer L, Chan Y-M, Hirschhorn JN, Lippincott MF. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. Am J Hum Genet [Internet]. 2018 Oct 4;103(4):522–34. https://doi.org/10.1016/j.ajhg.2018.08.016

87. Wang L, Jia P, Wolfinger RD, et al. An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. Bioinformatics [Internet]. 2011 Mar 1;27(5):686–92. https://doi.org/10.1093/bioinformatics/btq728

88. Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet [Internet]. 2012 Sep;44(9):1066–71. https://doi.org/10.1038/ng.2376

89. Runcie DE, Crawford L. Fast and flexible linear mixed models for genome-wide genetics. PLoS Genet [Internet]. 2019 Feb;15(2):e1007978. https://doi.org/10.1371/journal.pgen.1007978

90. Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet [Internet]. 2010 Apr;42(4):348–54. https://doi.org/10.1038/ng.548

91. Hardcastle AJ, Liskova P, Bykhovskaya Y, et al. A multi-ethnic genome-wide association study implicates collagen matrix integrity and cell differentiation pathways in keratoconus. Communications Biology [Internet]. 2021 Mar 1;4(1):266. https://doi.org/10.1038/s42003-021-01784-0

92. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res [Internet]. 2019 Jan 8;47(D1):D1005–12. https://doi.org/10.1093/nar/gky1120

93. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med [Internet]. 2019 Nov 19;11(1):70. https://doi.org/10.1186/s13073-019-0689-8

94. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature [Internet]. 2015 May 28;521(7553):436–44. Available from: https://doi.org/10.1038/nature14539

95. Hsieh T-C, Mensah MA, Pantel JT, et al. PEDIA: prioritization of exome data by image analysis. Genet Med [Internet]. 2019 Dec;21(12):2807–14. https://doi.org/10.1038/s41436-019-0566-2

96. Lee A, Mavaddat N, Wilcox AN, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors [Internet]. Vol. 21, Genetics in Medicine. 2019. p. 1708–18. https://doi.org/10.1038/s41436-018-0406-9

97. Inouye M, Abraham G, Nelson CP, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. J Am Coll Cardiol [Internet]. 2018 Oct 16;72(16):1883–93. https://doi.org/10.1016/j.jacc.2018.07.079

98. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnol Adv [Internet]. 2021 Jul;49:107739. https://doi.org/10.1016/j.biotechadv.2021.107739

99. LeCun Y. 1.1 Deep Learning Hardware: Past, Present, and Future. In: 2019 IEEE International Solid- State Circuits Conference - (ISSCC) [Internet]. 2019. p. 12–9. https://doi.org/10.1109/ISSCC.2019.8662396

100. Broad Institute [Internet]. [cited 2021 Mar 2]. Available from: https://www.broadinstitute.org/

101. Sheffi J. In our genes: How Google Cloud helps the Broad Institute slash the cost of research [Internet]. Google. 2018 [cited 2020 Dec 22]. Available from: https://www.blog.google/products/google-cloud/our-genes-how-google-cloud-helps-broad-institute-slash-cost-research/

102. gnomAD Production Team. gnomAD v3.1 [Internet]. 2020 [cited 2020 Dec 22]. Available from: https://gnomad.broadinstitute.org/blog/2020-10-gnomad-v3-1/

103. Ehwerhemuepha L, Gasperino G, Bischoff N, Taraman S, Chang A, Feaster W. HealtheDataLab - a cloud computing solution for data science and advanced analytics in healthcare with application to predicting multi-center pediatric readmissions. BMC Med Inform Decis Mak [Internet]. 2020 Jun 19;20(1):115. https://doi.org/10.1186/s12911-020-01153-7

104. Chang V, Wills G. A model to compare cloud and non-cloud storage of Big Data. Future Gener Comput Syst [Internet]. 2016 Apr 1;57:56–76. Available from: https://www.sciencedirect.com/science/article/pii/S0167739X15003167

105. Canela-Xandri O, Law A, Gray A, Woolliams JA, Tenesa A. A new tool called DISSECT for analysing large genomic data sets using a Big Data approach. Nat Commun [Internet]. 2015 Dec 11;6:10162. https://doi.org/10.1038/ncomms10162

106. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics [Internet]. 2012 Oct 1;28(19):2520–2. https://doi.org/10.1093/bioinformatics/bts480

107. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol [Internet]. 2017 Apr 11;35(4):316–9. https://doi.org/10.1038/nbt.3820

108. Amstutz P, Crusoe MR, Tijanić N, Chapman B, Chilton J, Heuer M, et al. Common Workflow Language [Internet]. Common Workflow Language working group; 2016. Available from: https://w3id.org/cwl/v1.0/

109. Google Cloud and the Broad Institute are providing free access to the Genome Aggregation Database [Internet]. [cited 2021 Feb 26]. Available from: https://cloud.google.com/blog/topics/healthcare-life-sciences/google-cloud-providing-free-access-to-genome-aggregation-database

110. Collins RL, Brand H, Karczewski KJ, et al. A structural variation reference for medical and population genetics. Nature [Internet]. 2020 May;581(7809):444–51. https://doi.org/10.1038/s41586-020-2287-8

111. Pontikos N, Yu J, Moghul I, et al. Phenopolis: an open platform for harmonization and analysis of genetic and phenotypic data. Bioinformatics [Internet]. 2017 Aug 1;33(15):2421–3. https://doi.org/10.1093/bioinformatics/btx147

112. Dolman L, Page A, Babb L, et al. ClinGen advancing genomic data-sharing standards as a GA4GH driver project. Hum Mutat [Internet]. 2018 Nov;39(11):1686–9. https://doi.org/10.1002/humu.23625

113. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol [Internet]. 2019 Feb;103(2):167–75. https://doi.org/10.1136/bjophthalmol-2018-313173

114. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med [Internet]. 2019 Jan;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7

115. Hogarty DT, Mackey DA, Hewitt AW. Current state and future prospects of artificial intelligence in ophthalmology: a review. Clin Experiment Ophthalmol [Internet]. 2019 Jan;47(1):128–39. https://doi.org/10.1111/ceo.13381

116. Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. NPJ Digit Med [Internet]. 2021 Jan 8;4(1):5. https://doi.org/10.1038/s41746-020-00376-2

117. Wang SY, Pershing S, Lee AY, AAO Taskforce on AI and AAO Medical Information Technology Committee. Big data requirements for artificial intelligence. Curr Opin Ophthalmol [Internet]. 2020 Sep;31(5):318–23. https://doi.org/10.1097/ICU.0000000000000676

118. Kagadis GC, Kloukinas C, Moore K, et al. Cloud computing in medical imaging. Med Phys [Internet]. 2013 Jul;40(7):070901. https://doi.org/10.1118/1.4811272

119. Cloud Healthcare API documentation [Internet]. Google Cloud Platform. [cited 2021 Mar 2]. Available from: https://cloud.google.com/healthcare/docs

120. Yan Q, Weeks DE, Xin H, et al. Deep-learning-based Prediction of Late Age-Related Macular Degeneration Progression. Nat Mach Intell [Internet]. 2020 Feb;2(2):141–50. https://doi.org/10.1038/s42256-020-0154-9

121. Chen D, Zhao H. Data Security and Privacy Protection Issues in Cloud Computing. In: 2012 International Conference on Computer Science and Electronics Engineering [Internet]. 2012. p. 647–51. https://doi.org/10.1109/ICCSEE.2012.193

122. Google. Handling genomic data in the cloud [Internet]. Google Cloud Platform; 2019. Available from: https://cloud.google.com/files/genomics-data-wp.pdf

123. Google. Google Cloud security foundations guide [Internet]. Google Cloud Platform; 2020. Available from: https://services.google.com/fh/files/misc/google-cloud-security-foundations-guide.pdf

124. Cloud Security - Amazon Web Services [Internet]. Amazon Web Services. [cited 05 Mar, 2021]. Available from: https://aws.amazon.com/security/

125. Crutzen R, Ygram Peters G-J, Mondschein C. Why and how we should care about the General Data Protection Regulation. Psychol Health [Internet]. 2019 Nov;34(11):1347–57. https://doi.org/10.1080/08870446.2019.1606222

126. Politou E, Michota A, Alepis E, Pocs M, Patsakis C. Backups and the right to be forgotten in the GDPR: An uneasy relationship. Computer Law & Security Review [Internet]. 2018 Dec 1;34(6):1247–57. Available from: https://www.sciencedirect.com/science/article/pii/S0267364918301389

127. Li T, Sahu AK, Talwalkar A, Smith V. Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Process Mag [Internet]. 2020 May;37(3):50–60. https://doi.org/10.1109/MSP.2020.2975749

128. Yi X, Paulet R, Bertino E. Homomorphic Encryption. In: Yi X, Paulet R, Bertino E, editors. Homomorphic Encryption and Applications [Internet]. Cham: Springer International Publishing; 2014. p. 27–46. https://doi.org/10.1007/978-3-319-12229-8_2

129. Guevara M. How we're helping developers with differential privacy [Internet]. Google Developers. 2021 [cited 2021 Mar 2]. Available from: https://developers.googleblog.com/2021/01/how-were-helping-developers-with-differential-privacy.html
130. Porter N, Golan G, Lugani S. Introducing Google Cloud Confidential Computing with Confidential VMs [Internet]. Google Cloud Blog. 2020 [cited 2021 Mar 2]. Available from: https://cloud.google.com/blog/products/identity-security/introducing-google-cloud-confidential-computing-with-confidential-vms
131. NHS and social care data: off-shoring and the use of public cloud services [Internet]. NHS Digital; 2018 Apr [cited 2021 Mar 5]. Available from: https://digital.nhs.uk/data-and-information/looking-after-information/data-security-and-information-governance/nhs-and-social-care-data-off-shoring-and-the-use-of-public-cloud-services
132. Hot Cloud Storage [Internet]. Wasabi. [cited 2021 Mar 7]. Available from: https://wasabi.com/hot-cloud-storage/
133. Wang L, McLeod HL, Weinshilboum RM. Genomics and drug response. N Engl J Med [Internet]. 2011 Mar 24;364(12):1144–53. https://doi.org/10.1056/NEJMra1010600
134. Weinshilboum RM, Wang L. Pharmacogenetics and pharmacogenomics: development, science, and translation. Annu Rev Genomics Hum Genet [Internet]. 2006;7:223–45. https://doi.org/10.1146/annurev.genom.6.080604.162315
135. Shastry BS. Pharmacogenomics in ophthalmology. Discov Med [Internet]. 2011 Aug;12(63):159–67. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21878193
136. Dedania VS, Grob S, Zhang K, Bakri SJ. Pharmacogenomics of response to anti-VEGF therapy in exudative age-related macular degeneration. Retina [Internet]. 2015 Mar;35(3):381–91. https://doi.org/10.1097/IAE.0000000000000466
137. Agarwal A, Soliman MK, Sepah YJ, Do DV, Nguyen QD. Diabetic retinopathy: variations in patient therapeutic outcomes and pharmacogenomics. Pharmgenomics Pers Med [Internet]. 2014 Dec 12;7:399–409. https://doi.org/10.2147/PGPM.S52821
138. Jefferson ER, Trucco E. Chapter 20 - The challenges of assembling, maintaining and making available large data sets of clinical data for research. In: Trucco E, MacGillivray T, Xu Y, editors. Computational Retinal Image Analysis [Internet]. Academic Press; 2019. p. 429–44. Available from: https://www.sciencedirect.com/science/article/pii/B9780081028162000216
139. Conroy M, Sellors J, Effingham M, et al. The advantages of UK Biobank's open-access strategy for health research. J Intern Med [Internet]. 2019 Oct;286(4):389–97. https://doi.org/10.1111/joim.12955
140. Wolf SM, Crock BN, Van Ness B, et al. Managing incidental findings and research results in genomic research involving biobanks and archived data sets. Genet Med [Internet]. 2012 Apr;14(4):361–84. https://doi.org/10.1038/gim.2012.23
141. McGuire AL, Caulfield T, Cho MK. Research ethics and the challenge of whole-genome sequencing. Nat Rev Genet [Internet]. 2008 Feb;9(2):152–6. https://doi.org/10.1038/nrg2302
142. Kohane IS, Masys DR, Altman RB. The incidentalome: a threat to genomic medicine. JAMA [Internet]. 2006 Jul 12;296(2):212–5. https://doi.org/10.1001/jama.296.2.212
143. McGuire AL, Joffe S, Koenig BA, et al. Point-counterpoint. Ethics and genomic incidental findings. Science [Internet]. 2013 May 31;340(6136):1047–8. https://doi.org/10.1126/science.1240156

144. Allyse M, Michie M. Not-so-incidental findings: the ACMG recommendations on the reporting of incidental findings in clinical whole genome and whole exome sequencing. Trends Biotechnol [Internet]. 2013 Aug;31(8):439–41. https://doi.org/10.1016/j.tibtech.2013.04.006

145. Johnson SB, Slade I, Giubilini A, Graham M. Rethinking the ethical principles of genomic medicine services. Eur J Hum Genet [Internet]. 2020 Feb;28(2):147–54. https://doi.org/10.1038/s41431-019-0507-1

146. PGP-UK Consortium. Personal Genome Project UK (PGP-UK): a research and citizen science hybrid project in support of personalized medicine. BMC Med Genomics [Internet]. 2018 Nov 27;11(1):108. https://doi.org/10.1186/s12920-018-0423-1

147. Chervova O, Conde L, Guerra-Assunção JA, et al. The Personal Genome Project-UK, an open access resource of human multi-omics data. Sci Data [Internet]. 2019 Oct 31;6(1):257. https://doi.org/10.1038/s41597-019-0205-4

148. Choquet H, Wiggs JL, Khawaja AP. Clinical implications of recent advances in primary open-angle glaucoma genetics. Eye [Internet]. 2020 Jan;34(1):29–39. https://doi.org/10.1038/s41433-019-0632-7

149. The Human Genome Project [Internet]. [cited 2021 Jan 22]. Available from: https://www.genome.gov/human-genome-project

150. Green ED, Watson JD, Collins FS. Human Genome Project: Twenty-five years of big biology. Nature [Internet]. 2015 Oct 1;526(7571):29–31. Available from: https://doi.org/10.1038/526029a

151. Humphries C. A Moore's Law for Genetics [Internet]. [cited 2021 Jan 22]. Available from: https://www.technologyreview.com/2010/02/23/205915/a-moores-law-for-genetics/

152. Wetterstrand KA. DNA Sequencing Costs: Data [Internet]. National Human Genome Research Institute. [cited 2021 Mar 5]. Available from: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Appendix

Table S1. Human Phenotype Ontology (HPO) associations for Bardet-Biedl syndrome (https://hpo.jax.org/app/browse/disease/ORPHA:110). Over 13,000 clinical phenotypes (*e.g.*, hypertension, hearing impairment) are described in the database, each with a unique identifier grouped by categories representing body systems. When searching by diseases, a brief description of the disease, its identifier, and genetic associations with the corresponding identifier are also shown.

HPO_TERM_ID	HPO_TERM_NAME	CATEGORY
HP:0000822	Hypertension	Cardiovascular
HP:0001395	Hepatic fibrosis	Digestive System
HP:0000365	Hearing impairment	Ear
HP:0000368	Low-set, posteriorly rotated ears	Ear
HP:0000135	Hypogonadism	Endocrine
HP:0000639	Nystagmus	Eye
HP:0000512	Abnormal electroretinogram	Eye
HP:0000580	Pigmentary retinopathy	Eye
HP:0008736	Hypoplasia of penis	Genitourinary system
HP:0008724	Hypoplasia of the ovary	Genitourinary system
HP:0000028	Cryptorchidism	Genitourinary system
HP:0000003	Multicystic kidney dysplasia	Genitourinary system
HP:0000100	Nephrotic syndrome	Genitourinary system
HP:0004322	Short stature	Growth
HP:0001513	Obesity	Growth
HP:0000494	Downslanted palpebral fissures	Head and neck
HP:0000470	Short neck	Head and neck
HP:0000426	Prominent nasal bridge	Head and neck
HP:0001162	Postaxial hand polydactyly	Limbs
HP:0006101	Finger syndactyly	Limbs

HPO_TERM_ID	HPO_TERM_NAME	CATEGORY
HP:0003202	Skeletal muscle atrophy	Musculature
HP:0001249	Intellectual disability	Nervous System
HP:0002167	Neurological speech impairment	Nervous System
HP:0002230	Generalized hirsutism	Skin, Hair, and Nails
HP:0010747	Medial flaring of the eyebrow	Skin, Hair, and Nails

Table S2. Human Phenotype Ontology gene associations for Bardet-Biedl Syndrome. Genes associated with the syndrome are listed with identification numbers and symbols.

GENE_ENTREZ_ID	GENE_SYMBOL
79738	BBS10
55212	BBS7
123016	TTC8
157657	C8orf37
129880	BBS5
583	BBS2
22954	TRIM32
80184	CEP290
11020	IFT27
585	BBS4
54585	LZTFL1
51057	WDPCP
4867	NPHP1
166379	BBS12
8195	MKKS
84100	ARL6
92482	BBIP1
27241	BBS9
54903	MKS1
26160	IFT172
10806	SDCCAG8
582	BBS1