

Executive summary	2
Introduction	4
Defining genomics	7
Sector scenarios	9
Regulatory issues	16
Next steps	23
Annex A: Key technologies and terms	24
Annex B: Genomic sector economic overview	26
Annex C - Background and context	32

Executive summary

Executive summary

Over the last decade there has been a rapid increase in investment and influence in genomic technologies in a range of sectors. The technology offers well-known benefits in areas like advanced disease treatment and personalised healthcare, and there is a drive to deliver a genomic-focused healthcare system. But the rapid growth of applications in sectors like insurance, education and law enforcement also have the potential to impact us all.

As the UK's data protection regulator, the Information Commissioner's Office (ICO) aims to increase public trust in how organisations process personal information through responsible practice. We want to empower people to safely share their information and use innovative products and services that will drive our economy and our society. In our ICO25 strategy, we committed to set out our views on emerging technologies to reduce burdens on businesses, support innovation and prevent harms.

The genomics chapter of our 2023 Tech Horizons Report set out an initial area of concern around polygenic risk scores; probabilistic assessments of traits and characteristics derived from genomics that could inform how we deliver health care and other services in the future.

This report goes into more depth to consider regulatory and privacy issues raised by the wider development of genomics. We illustrate how these may arise through scenarios that cover potential uses of the technology in and beyond healthcare, including in direct-to-consumer services, insurance, education and law enforcement. The issues we cover include:

- the challenges of understanding when genomic data may be considered personal information. Genomic information relating to the deceased, or once thought unimportant or even unidentifiable in terms of personal information, may have an increasingly large significance as personal data as research rapidly advances;
- the complexities of using and sharing third party information (including both genomic information and inferences derived from it) given how much of a genome and the information derived from it is shared - even on a familial level;
- the associated risks and challenges of anonymising and pseudonymising genomic information in a way that embeds privacy by design without compromising innovative and necessary research; and
- the significant risks of bias and discrimination emerging from the processing of genomic information. Inaccurate models trained on unrepresentative datasets may emerge, or inferences used to support unfair decisions even when using accurate information.

We will address these areas of concern through:

- ongoing engagement with key stakeholders across industry, regulation, academia and civil society. This will include inviting organisations to work with our Regulatory Sandbox to engineer data protection into uses of genomic information;
- monitoring and highlighting developments in this area through our Tech Horizons reports, where we will set out future programmes of work on the issues as they arise. We will pay particular attention to the sharing of third party genomic information and direct-to-consumer genomic services linked to polygenic risk scoring. We have invited views from organisations with an interest in this area; and

- engagement with the public to better understand their knowledge and concerns about the use of genomics and privacy.

Introduction

Genomics is a relatively recent branch of scientific study. It focuses on the interplay of genes with each other and their environments and how these can impact our traits and characteristics. Genomics offers an opportunity to radically advance our understanding, diagnosis, support and treatment of a variety of illnesses and conditions.

The UK government has identified genomic information and uses such as **polygenic risk scoring** ¹ as one of the critical areas that can contribute to the UK becoming a technological superpower. Specifically, the government are focused on a drive to deliver a genomic-focused healthcare system. ² Building on over 70 years of developing research, the increased pace of genomic analysis offers opportunities and risks not just for healthcare, but other diverse sectors such as education and insurance.

Government interest is mirrored by private sector interest. While there is significant overlap between genetics and genomics, we feel that the public should be aware of the importance of genomic data and its potential impact on privacy and data protection, both positive and negative. This will be key to developing public trust in the appropriate use of genomic information, initially for healthcare. It will also make people aware of the potential risks and harms of using direct-to-consumer genomic testing and public sharing of genomic data.

The analysis in this report aims to support our ability to protect people, provide clarity for businesses and enable privacy-positive innovation. It is aimed at:

- organisations and people considering the policy intersections of privacy and genomic data; and
- organisations seeking to deploy new or innovative forms of processing based on genomic data.

This report explores plausible scenarios, developed through stakeholder engagement and research, and use cases for emerging genomic technologies and solutions to increase understanding of possible future uses of genomic information. These cases illustrate potential deployments across sectors including health, education, insurance and law enforcement. While they do not offer examples of ethical use or best practice regarding privacy, the scenarios raise key issues about gathering and using genomic data. We examine these to better understand critical challenges around emerging genomic uses, techniques and privacy.

We intend to address these issues through continuing proactive work with stakeholders and the public, as well as further cross-regulatory work. We are asking for views from interested organisations at the end of this report and in the longer term we are also aiming to create guidance.

Why genomics?

Using genomic information can offer significant benefits and opportunities for people and organisations. They can offer, among other things:

- upstream treatment for medical conditions as varied as cancer and Covid 19; and
- targeted and predictive insights into potential health issues and even behavioural traits such as educational or sporting attainment.

Alongside these opportunities for innovation, we also recognise the scope for future potential harm and the

risks to data rights and privacy.

As well as the drive to develop UK research and treatment capabilities, this rapid development of genomics technologies is driven by a series of opportunities:

- The development of sequencing technologies and genotyping arrays allowing for cheap and rapid measurement of genetic variation across a genome. ³
- Increased accessibility of AI and algorithmic processing of complex data sets that allows for quicker analysis.
- Increased interest in preventive healthcare as a means of delivering rapid and effective treatment.
- A lack of specific regulation in non-health related sectors including direct-to-consumer services.
- A critical mass of genomic information for researchers to investigate.
- A significant increase in both national and global funding.

One of these drivers, the use of AI, indicates an overlap with an area that is already important to us. While we consider novel and heightened risks around this area within this report, you can find specific [AI reports and guidance on our website](#).

While innovations can offer opportunities and challenge the status quo, they can also present new issues and risks that undermine their promised progress. Unexpected sharing of purportedly medical genomic information with insurance companies by organisations such as UK Biobank has raised public concerns about security, transparency and fairness. ⁴ It is important that we consider both the potential benefits and harms of these evolving approaches, so we can respond in a timely and proactive manner.

Processing genomic data poses a significant and specific risk to people's information rights in several ways:

- Its intrinsic nature. You cannot change your DNA or how it links to your **phenotypes** (a person's observable traits). This means that if this genomic data was lost, stolen or inappropriately used, you could not simply replace or vary it. ⁵
- Its links to multiple people. Genomic data will relate to relatives of a person, posing risks and challenges as to how and when this information can be shared and processed.
- Its potential for leading to someone making inaccurate or inappropriate inferences, or both, about people. For example, making an incorrect inference about a person's heritage, characteristics or health. When genomic information is processed in conjunction with AI-based processing and automated decision making, it may be further impacted by an underlying and inappropriate systemic bias.

There are already several significant analyses of emerging uses of genomic technologies and genomic data. These include reports by [the Government Office of Science](#) and the [Ada Lovelace Institute](#). These reports provide excellent overview on scientific context, potential future uses of genomic information and key aspects such as algorithmic processing of this information. What this report provides is a specific focus on the issues and opportunities arising around privacy and data protection.

In addition to briefly examining the legal and regulatory context in **Annex C**, we also consider emerging market indicators about genomic technologies, such as funding and patents. Understanding the broader market is important in assessing which sectors are likely to see markets develop first and what issues may emerge.

At a national level, there is clear evidence that the UK private sector is investing in genomic technology,

with as many as 142 companies focusing on this sector.⁶ On a global scale, investment in genomic analysis and the creation of related patents continues to increase significantly. You can find further details about this in **Annex B**. This growth reflects the potential to develop and deploy genomic-centric approaches in regions where data protection regimes differ significantly from the UK GDPR. In certain cases, use of these approaches may not adhere to the expectations we have for fairness and transparency in the way they use personal information. In turn, this may pose significant challenges if these uses become common in the UK or are used by those with data rights under the UK GDPR. Nonetheless, if companies based overseas do offer their services to people who are based in the UK, through a website for example, they will still fall within the geographical scope of the UK GDPR under Article 3(2)(a).

Following the COVID-19 pandemic, there has been an increased drive to gather and share genomic data via global biobanks.⁷ Ostensibly sharing information for medical research purposes, this is not always the case. The growing size and complexity of the information sets mean that potential commercial uses of information are becoming possible.⁸ A significant section of these commercial interests are represented by direct-to-consumer services, commonly offering ancestry or genealogical analyses or, increasingly, SNP (single-nucleotide polymorphism) based analysis of common traits about lifestyle and wellbeing.⁹

¹ [A glossary of key terminology can be found in Annex A.](#)

² [Genome UK: 2022 to 2025 implementation plan for England](#)

³ [The road to genome-wide association studies](#)

⁴ [Private UK health data donated for medical research shared with insurance companies](#)

⁵ This also links to broader issues regarding freedom of expression and a right to privacy, as an individual cannot reasonably be expected to conceal their face or gait as they might with a password or pin number. See [R \(Bridges\) v CC South Wales](#) for further details.

⁶ See [Annex B for a further economic exploration of this area.](#)

⁷ [Global Biobank Meta-analysis Initiative](#)

⁸ [15 Ways Genomics Influences Our World](#) 

⁹ [Commercial genomics](#)

Defining genomics

What do we mean by genomics and genomic data in this report?

Genomic research uses and shares data from the whole of a person's DNA sequence and structure, rather than an individual gene. Using this definition provides a very clear link to the definition of genetic data as it exists under the UK GDPR, with implications as to how and when this may be considered personal data as well as special category personal data.

Each person's DNA sequence is represented by some 6.4 billion bases (or letters) in our genome. Genomics refers to the study of these individual genes, grouped into either coding or non-coding genes. ¹⁰ Some genes determine traits such as eye colour or blood type. Others, complex traits and characteristics that reflect the interaction of multiple genes (polygenic), variances in a DNA sequence and the impact of environment. Common complex traits are highly polygenic and influenced but there are thousands of common DNA variants.

Other factors that can alter the instructions our genes provide include mutations that generate positive, negative or neutral effects to our characteristics. These may be inherited or experienced in the lifetime of a person. Access to the information encoded in DNA can be shifted or changed through epigenetic factors and quasi-environmental factors. At its largest scale, functional genomics can be considered a study of multiomics, encompassing not only DNA but its possible modulations, RNA transcripts and modifications and protein structures and how these interact with their environment.

Genetics under the UK GDPR

While genomic data is not itself referenced in UK data protection law, genetic data is. Under Article 4(13) of the GDPR, "genetic data" is defined as:



"personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question."

Recital 34 of the GDPR provides further context for understanding the definition set out in Article 4(13), noting that organisations should interpret it as including any type of analysis that enables them to obtain equivalent information. For example, ribonucleic acid (RNA) plays an essential part in the coding, decoding, regulation and expression of genes.

For the purposes of this report, we will broadly consider that genomic data can be substituted for genetic data and therefore always considered [special category data under article 9 of the UK GDPR](#). However, key issues remain around when genomic data is personal information and when inferences drawn from genomic data may fall into other categories such as biometric, health or simply personal (rather than Article 9 special category) data.

For the purposes of this report, **we will define genomic data as:**

“

personal data relating to the inherited or acquired genomic characteristics of a natural person which can give unique information about the physiology, characteristics or the health of that person and which result, in particular, from an analysis of the interplay of the genetic information of a natural person and their environment.”

This definition has been designed to encompass as broad a spread of uses of genomic information as possible, including:

- epigenetic information; and
- polygenic risk scores.

Some information, such as polygenic risk scores, may be considered second order data, posing later challenges as to whether it falls under special category or genomic data. Recital 35 of the UK GDPR suggests that health data should be interpreted as:

“

information derived from the testing or examination of a body part or bodily substance, including from genetic data and biological samples.’

This may cover certain uses of the information being considered. Therefore, this definition will allow us to examine a variety of risks and challenges in the context of privacy and data protection.

This report does not try to set out the enormous depth and variety of genomic technologies. ¹¹ Rather, this report focuses on the uses, rather than the creation, of personal information.

¹⁰ [Completing the human genome sequence](#)

¹¹ [Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances - PMC](#) 

Sector scenarios

There are three main approaches to the use of genomic data that are of particular interest to us:

- Collection and use of genomic information for research and discovery in healthcare and associated research.
- Application and translation of genomic information for non-health related use, such as direct-to-consumer advice for lifestyle and social traits.
- Familial and ancestral analysis.

Our research indicates a significant rise in organisations using genomic information. Research and medical uses of genomic data are already well advanced, but other sectors are likely to increase their uptake of genomic data in the near term. Some sectors, such as the military, and their uses of genomic data are beyond the scope of this report. We have identified the following sectors where we anticipate that uses of genomic data may have a major impact on UK markets in the next two to seven years:

- The medical research sectors will continue to expand upon GWAS and the examination of polygenic diseases and traits.
- The health sector may explore the potential of predictive healthcare (P4 medicine), drawing upon polygenic risk scoring to provide lifestyle and dietary advice as well as preventative treatments.
- The wellbeing, direct-to-consumer and sports sector may utilise genomic data and polygenic risk scoring to build on the rapidly developing market in familial, ancestry and trait tracking as well as dietary advice and prenatal testing.
- The education sector may seek to use polygenic risk scores to identify SEND requirements for students and likely resources for schools.
- The insurance sector may consider using polygenic risk scoring to inform insurance offerings across health, life and driving sectors for example.
- The law enforcement sector may seek to use genomic, rather than genetic, data to identify suspects. Phenotypes inferred from genomic data may also lead to further routes of identification via facial recognition.

Number of years	Sector
2-3 years	<ul style="list-style-type: none">• The medical research sectors will continue to expand upon GWAS and the examination of polygenic diseases and traits.
	<ul style="list-style-type: none">• The health sector may explore the potential of predictive healthcare (P4 medicine), drawing upon polygenic risk scoring to provide lifestyle and dietary advice as well as preventative treatments.
4-5 years	<ul style="list-style-type: none">• The wellbeing, direct-to-consumer and sports sector may utilise genomic data and polygenic risk scoring to build on the rapidly developing market in familial, ancestry and trait tracking as well as dietary advice and prenatal testing. ¹²

5-7 years

- The **education sector** may seek to use polygenic risk scores to identify SEND requirements for students and likely resources for schools.

10 years +

- The **insurance sector** may consider using polygenic risk scoring to inform insurance offerings across health, life and driving sectors for example.
- The **law enforcement sector** may seek to use genomic, rather than genetic, data to identify suspects. Phenotypes inferred from genomic data may also lead to further routes of identification via facial recognition.

While the above diagram provides a very brief overview of our findings, it is helpful to explore actual uses of genomic data within these sectors from a data protection perspective, before examining their potential issues.

Please note that these scenarios are intended to explore possible developments and uses of genomic information. While the scenarios include high-level commentary on aspects of relevant data protection compliance issues, this should not be interpreted as confirmation that the relevant processing is either desirable or legally compliant. This document does **not** provide ICO guidance.

Short-term (2-3 years):

The **health sector** is likely to use genomic data for the greatest impact in the next two to three years. We are likely to see increased collection and use of genomic information to develop **preventative and personalised healthcare**.

The UK government may consider a wider drive to gather citizen genomic data, to provide more effective, efficient and timely healthcare. A nationally funded trusted research environment (TRE) could hold the information. This would allow for controlled, closely monitored access to the high-risk information. Such an approach could provide a world-leading resource for research, allowing scientists to identify the causes of heritable traits relating to complex illnesses and conditions. In turn, this could create a lifelong basis for proactive treatment. Genomic research could use approaches that can pseudonymise information. However, it becomes complex if organisations then use the information for personalised healthcare and direct interventions for patients and the wider public.

Data protection concerns

Pseudonymisation will be a particular challenge, given the desire to link special category information within healthcare records with genomic data and polygenic risk scores via algorithmic analysis. ¹³ If this happens using polygenic risk scoring, organisations could share the probabilities of conditions and diseases with healthcare providers to recommend lifestyle changes and preventive treatments. With the proliferation of wearables and wellbeing tech, people could also provide a greater input of data about their lifestyle and environment. This might allow for more refined results and predictive abilities. It could support future research into links between the polygenic liability of a trait or disease and environmental factors. In turn, this, may lead to organisations focusing more on collecting additional types of information to generate further inferences.

This would change the purpose of processing people’s genomic data from a specific to possibly more generalised, and speculative, purpose. The data controller would need to consider whether this purpose was [compatible with the original purpose](#). They would also need to have a lawful basis for this new processing operation. If they originally collected the information using consent, they would likely need to collect fresh consent unless they informed people beforehand of this potential use. This approach could also raise concerns about fairness, given that people may not have expected the organisation to use their genomic data in this way.

Transparency will be essential for people to understand how organisations are using their information. This scenario may present highly complex data flows between public and private organisations (e.g. from a research environment to a healthcare provider to people receiving the advice and recommendations). Given the potential impacts and high-risk nature of the collected information, we would also expect organisations to have appropriate security measures in place as set out under Article 5 (1)(f).

Organisations would need to mitigate against systemic discrimination emerging through either combined records or through using potentially repurposed AI models previously used for something other than analysing genomic information. We discuss these risks and how data protection law would potentially apply later in this report.

An extension into the use of wearable technologies to gather additional lifestyle information, such as daily activity, may create further challenges. It may not be clear for people and organisations what constitutes health or wellbeing information. There may also be a fundamental challenge to the fairness of access to treatment, with those willing or able to pay for the additional devices accessing more granular and accurate outcomes.

Data controllers will also face a challenge in upholding appropriate data rights. If someone submits a subject access request for their information from a healthcare provider, or another organisation that uses genomic data, then organisations will need to carefully consider what information they can provide. As genomic data can reveal highly sensitive or intimate insights, providers will need to consider whether or how to appropriately limit or redact information sent to third parties such as family members.

Medium-term (3-5 years):

The **education sector** may consider using genomic data to enhance **SEND support** in schools. [14](#)

The government may initiate a private-public partnership to combine publicly-held genomic data with the generation of polygenic risk scoring. This would build on early projects to gather newborn genomic data for long-term preventive healthcare purposes. [15](#) This partnership could generate an assessment of funding for SEND requirements and screening of traits and disorders. There may be a particular focus on traits related to ADHD and dyslexia. In turn, this could be combined with healthcare records to develop further inferences and increase the accuracy of probabilistic scores. Private organisations could act on behalf of the educational trusts to identify likely needs for students, providing additional information to both schools and teachers to support targeted, effective help.

Developing this hypothetical approach, the partner organisation could suggest that parents or schools could pay for additional polygenic risk scores as a direct-to-consumer option. The organisation could provide the information via a third-party app. This would allow parents to research traits around educational attainment [16](#) as well as sporting and musical ability. The organisation would present the results as probabilities.

However, there may be little supporting material to show how they achieved those results and any limitations, such as the impact of the environment or appropriate actions if the user has particularly high or low risk results. As students age, this record could be combined with healthcare records and form part of a permanent, lifelong citizen record.

Data protection concerns

There is a risk of a lack of transparency as information moves between the public and private sector, inhibiting people's ability to enact their data rights and ensure fair treatment. Furthermore, organisations may have to shift their purpose of processing from funding assessment to trait analysis, creating challenges for purpose limitation and lawfulness of processing.

Depending on the level of human involvement, there may be circumstances where decision-making based on polygenic risk scores amounts to automated individual decision-making with legal or similarly significant effect, within the meaning of Article 22. Controllers would need to make sure that they had met the UK GDPR conditions for carrying out such processing, as well as meeting enhanced transparency requirements.
[17](#)

Controllers would also need to put safeguards in place to protect people's rights in such scenarios. This includes allowing people to challenge solely automated decisions and to obtain human intervention in the decision-making.

Accuracy is also an area of significant risk given that educational achievement is a highly polygenic trait, defined by thousands of genetic variants and the environment. Any scores assigned would be entirely probabilistic. [18](#) The links between inherited intelligence (itself a highly contested notion) and environment require significant further research. We do not yet fully understand or can address the highly-complex and person-specific relationship between genomics and environment. A complex example could be a person with a low genetic risk for educational attainment but who lives in a high-risk environment. Rare variants, unusual circumstances and unrecognised needs may lead to unfair treatment of pupils and systemic discrimination, as results are 'ported' rather than targeted. This may, in turn, lead to organisations using ever wider sets of information to try and address the challenge, raising issues of data minimisation and transparency.

Processing children's information will also require particular care given the sensitivity of the information. If organisations use consent for processing, they will face challenges. Ages of consent will shift as pupils grow up and become responsible for their own information. Organisations will also need to think about what it means to fully inform users when they are providing consent.

Furthermore, if power imbalances exist between an organisation and a person, as is likely in the context of a school, then it's unlikely that consent will be appropriate. Organisations may also have issues with data retention and risks of inaccuracy that may stem from limited data sets. Finally, as with all instances of genomic data, third-party data presents a risk of revealing inferences about family members' health and characteristics.

The **insurance sector** may also seek to build on genomic data in this period, if organisations can use polygenic risk scores to create more accurate estimates. [19](#) Possible areas for rapid uptake of polygenic risk scoring include health and life insurance. Drawing directly upon probabilistic risks of inheritable diseases and conditions, from cancer to Parkinson's disease, insurance providers may offer increasingly personalised insurance plans. These could offer highly-targeted services at affordable rates through direct-to-consumer

services.

However, as providers seek to offer increasingly holistic lifestyle analyses, they could extend this approach to cover traits such as risk-taking behaviours, including:

- certain physical activities (sports deemed to be high risk);
- drinking;
- smoking; or
- sexual behaviour. ²⁰

Much of this information is likely to be special category data as defined under Article 9(2) and high risk in its uses and impact.

Some may claim that such an approach offers reasonably accurate and cost-effective products. However, there is also a significant risk that the information becomes increasingly biased. This may be against those with specific perceived genomic traits, leading to systemic discrimination. It may also lead to aggressive pricing against those who are unable or unwilling to provide their information. Ultimately, insurance providers may decide to refuse insurance to those deemed too high a risk, leading to a fundamentally unfair use of personal information.

A significant challenge in this sector would be the need for data minimisation and finding an appropriate purpose. Providers may also struggle to ensure transparency and fairness in the use of information. Insurers might, for example, seek to use explicit consent if they are using special category information. However, this could face challenges when the only other products available to customers are significantly more expensive.

The threshold of accuracy for the use of polygenic risk scoring would be of particular interest in this scenario. Insurance providers might face significant challenges over the information they use to derive risks and, in particular, over the gap between derived polygenic risks and the customers' environments.

²¹ Fundamentally, providers would need to establish a clear sense of an acceptable level of probability for fair and accurate information use which would outweigh potential risks.

Medium to long-term (5-10 years)

The **law enforcement sector** may make more use of genomic data for crime detection and sentencing, using an increasingly broad array of information and inferences to identify potential suspects.

A private sector company could offer genome wide association studies (GWAS) and analysis of crime scene samples to support murder investigations and cold cases by drawing upon sequences derived by WES (whole exome sequencing) and WGS. ²² The company could use the sample's analysis to compare DNA profiles across police databases and predict a suspect's facial features, age (via telomere length) and gender via phenotype information. Combined with AI processing, they could generate suspect e-fits. The company may suggest that, in turn, the police could compare this against social media records to identify a potential suspect. They could also run through facial recognition systems at events and public spaces.

The company may also seek to conduct research on the genomic information they have. They could explore a possible future add-on service that includes providing information on specific disease-related, psychological or character-related traits from the sample. They could do this to create multi-modal

behavioural analyses (such as biometrics or behavioural analysis) of suspects to combine with AI-driven facial recognition tracking.

Data protection concerns

This approach would likely raise significant challenges around data retention and purpose limitation. A specific challenge could be if an organisation was to attempt to use direct-to-consumer databases to identify people of interest, as has already been attempted in the US.²³ It could also raise issues around lawfulness, as any use of personal information for law enforcement purposes must be necessary. This does not mean that the company's use of the information must be essential, but it must be a targeted and proportionate way of achieving the purpose. A lawful basis will not apply if an organisation can reasonably achieve the purpose by some other, less intrusive means. The police might argue that such an approach was necessary in genuinely exceptional cases, which they could not solve by other means. However, they would need to demonstrate that their use of genomic information was proportionate.

Using genomic information in this way is also likely to be fundamentally unfair and beyond the organisation's original purpose. It is also likely to face challenges on the potential for inappropriate bias and discrimination and the significant chilling effect it may have for society. The example involves several processing operations which we would consider to be high risk. The organisation would need to carry out a data protection impact assessment and consider whether and how they could mitigate the data protection risks.

¹² Although this is currently illegal in the UK under HFEA regulations, this report will consider the implications of information generated by such an approach.

¹³ [Chapter 3 anonymisation guidance](#)

¹⁴ This is an area which organisations such as NCOB are intending to examine further in 2024/25 in regards to scientific and ethical practice.



¹⁵ [Newborn Genomes Programme](#)

¹⁶ [Genomics Beyond Health](#)

¹⁷ Article 13(2)(f) and 14(2)(g) require controllers who are carrying out automated processing, including profiling, to provide "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" as part of their transparency requirements.

¹⁸ Essentially, polygenic risk scores are not certain in the information they provide and do not explain all of a trait's heritability, which for educational attainment stands at about 40%. There is scope for the explanatory power of the EA PGI to explain 40% of differences between pupils if all the heritability can be recovered via GWAS. This requires including rare variants (and maybe gene-gene, gene-environment interplay). Stakeholders have noted that research is currently some way from getting to that point but is heading in the direction where this may be achieved.

¹⁹ [Code on Genetic Testing and Insurance](#)

²⁰ Although predictions around this trait are highly contested in the research community. See [Genetic Influences on Adolescent Sexual Behavior: Why Genes Matter for Environmentally-Oriented Researchers](#)  and [The role of sex in the genomics of human complex traits](#)  for some further information.

²¹ Data subjects may also be entitled to request information about how any automated decisions have been taken, and contest those decisions, in line with the safeguards set out under Article 22.

²² [Supercharged crime-scene DNA analysis sparks privacy concerns](#) 

²³ [Cops Used DNA to Predict a Suspect’s Face—and Tried to Run Facial Recognition on It](#) 

Regulatory issues

Issue 1: Regulatory certainty. When is genomic data personal data?

Under the UK GDPR, there is no explicit definition of genomic information as either a specific form of personal information or special category data. This means that, without considering the purpose, organisations should consider personally identifiable genomic information as personal information. Beyond this, organisations need to carefully consider:

- when and why genomic information may be special category data; and
- what risks large-scale classificatory uses of even non-special category personal information may pose.

Some key challenges are listed below.

Identifying personal and special category personal information

Genomic information isn't always counted as personal information. A large part of a genome (roughly 98%) is non-coding-based DNA that may not be easily linked to a known person in isolation, as it is shared with every human being. ²⁴ In other words, it is not expressed as a gene. When genomic information is considered personal information, it would become hard to define. Context will matter hugely as developing research and analytical tools allow a greater understanding of how this genomic information, once dismissed as 'junk DNA', connects to how specific genes and phenotypes are expressed. For more information, [organisations should consider our guidance on identifiers and related factors](#).

While we have made assumptions here when defining genomic information in terms of both personal and special category information. Further clarity is required as to how and when genetic and genomic information overlap for the purposes of data protection regulation. Not all genomic information is genetic, but certain provisions of the UK GDPR provide an indication of how organisations can consider such material under the legislation.

Where genomic information can be clearly defined as personal information, it is almost certainly special category information under Article 9 as it will contain some portion of genetic information which is already defined as special category data under the UK GDPR. Given this, the usual protections for special category data will apply.

There is also a risk that the use of the term 'genomics' to include associated areas of related inferences, such as phenotypes may cause uncertainty as to what information may or may not be special category data. Phenotypes play a critical role in the development and use of genomic research. While medical uses, such as identifying pre-cancerous growths, will be special category data under Article 9(1) of the UK GDPR, other commercial uses may not. While these instances are currently rare, the pace of research is likely to change this. It may become possible to use phenotypes for educational or insurance purposes as explored above. If phenotypes are considered distinct from genetic or genomic definitions, then it is unlikely that they would count as biometric information either, as explored in [our previous report on emerging biometrics](#).

Even if there is no direct correlation between genomic information and special category data, then the uses of genomic information for medical or identification purposes are nevertheless likely to be special category

genomic or health information under Article 9(1). Organisations will therefore need a lawful basis to use this information under Article 6 and an additional condition for processing special category data under Article 9(2). Organisations must identify the most appropriate basis for processing. Consent may be an appropriate lawful basis and an appropriate special category condition, as long as they can meet the necessary requirements for valid consent.

Conditions for processing special category data

If organisations use genomic information, they should consider whether consent is the most appropriate lawful basis. While medical consent remains a distinct and important issue, explicit consent for using personal information is only one of a variety of appropriate special category conditions under the UK GDPR. It is not inherently 'better' than other conditions.

Any wider automatic reliance on consent around the use of genomic information for consumer purposes could also cause confusion and may prove to be inappropriate under the UK GDPR. People may assume they have the right to automatically withdraw consent even when organisations have not used consent as a basis, given the wider dialogue and calls for the use of consent. In fact, organisations may use other appropriate lawful bases. They need to be transparent about which basis they are using and people's rights as a result. This transparency may prove more effective in helping people to truly understand how organisations are using their information and what their rights are.

Organisations should also be aware that the conditions for processing special category data are likely to be quite limited in a purely commercial scenario. Most of them concern public rather than private interests. There is a high threshold for processing special category data in most circumstances, and where there is no real public interest engaged, organisations will likely have to get explicit consent.

Issue 2: Third-party data, historic data, future data and genomic data

Genomic information poses significant challenges around the appropriate handling of third-party data. Familial information is fundamental to genomic information and any disclosure will inherently involve additional people. Organisations can consider pseudonymisation techniques, but these bring their own challenges as discussed below. We provide guidance as to how and when to [appropriately share such third party information](#).

This is a particular issue when considering appropriate disclosures in response to a subject access request. Use of genomic information or inferences derived from this may risk disclosing special category data linked to family members via genomic analysis. This has already been the case with genetic tests for inherited neurodegenerative conditions. ²⁵ In such situations, inappropriate disclosure or withholding of personal information may present high risks of harm to people.

It also opens up the potential for third-party claims to information, which organisations may think relate to one another. However, given the nature of genomic information, it is likely that organisations will need to closely consider when and how both raw information and complex inferences may relate to and identify (or be capable of identifying) associated third parties. They will also need to consider when they should share this information. Organisations can use our guidance on [how and when confidentiality and consent apply to help with this](#).

There is also an increased risk that direct-to-consumer services, like genomic counselling, may provide

high-risk, high-impact information to people without the more traditional supporting structures of the health sector. Organisations using genomic information will need to pay careful attention to the context of their use. They will also have to put sufficient and appropriate security measures in place under Article 5(1)(f) and Article 32. Raw genomic information (i.e. the processed biological sample without further analysis) is currently unlikely to be easily identifiable by members of the public. However, this may rapidly change as access to AI processing increases, along with associated personal information. ²⁶

Under the UK GDPR, information about a living person is classified as personal data. Information about a deceased person is not classified as personal data. Genetic information has already blurred the definition of **historic information**. As defined under Article 4 of the UK GDPR, genetic and now genomic information which wasn't previously linked to a living person can now be linked through modern research techniques. Large-scale collection of genomic information heightens the risk further, given the information's links to multiple living people, even when the person who provided the original sample and information has died.

Organisations retaining, gathering and using what they might assume to be historic information, should take appropriate steps to consider whether they need to consider this information as personal data. This is likely to be highly contextual and will not relate to the biological samples themselves. Rather, processed information and associated inferences such as phenotypes may relate to living people and, given the context of other records, may count as pseudonymised rather than anonymised.

In contrast to historic information is the challenge around potential future discoveries made with genomic information. Article 89 of the UK GDPR makes provisions for research purposes to retain information for both longer periods and a broad purpose. However, organisations using personal information for research purposes must put safeguards in place and, in particular, follow the principle of data minimisation.

Organisations will need to ensure that, if new potential means of using genomic information emerge, they pay close attention to the implications and requirements of a change in purpose of processing. They will also need to consider how this may fall under people's expectation about the fair use of their information.

Issue 3: Inherent identifiability, anonymisation and security of genomic data

Genomic information is highly distinguishable. Linking a genome to a person only requires a few hundred (out of millions) of SNPs. ²⁷ In many instances of business or research, organisations may prefer to either anonymise or pseudonymise personal information, ²⁸ to minimise potential risks and harms to people. While the large-scale nature of genomic information and its associated risks would seem an obvious fit for this, there are challenges to anonymising genomic information. The levels of encryption or data obfuscation through methods like differential privacy may reduce the information's value to GWAS. ²⁹

Approaches that may offer organisations some means of pseudonymisation include:

- **Data transformation** is likely to remain limited in deployment. This is because genomic information is fundamentally open to re-identification. This can happen through multiple pathways and correlations as high dimensionality information, where the number of dataset features are larger than the number of observations made. ³⁰ Of the millions of SNPs within a genome, only a few hundred are required to identify someone. However, ongoing research in k-anonymity may offer future defences against re-identification.
- **Data obfuscation** is achieved by adding noise to genomic information through techniques such as differential privacy. This is likely to have a significant negative impact on effective genomic research.
- **Synthetic data** offers another approach, through the AI-powered generation of large-scale data sets

that do not involve people. However, complexities with this approach emerge as to when and if the analysis and processing undertaken links to an identifiable person.

[For more information on privacy-enhancing technologies and data protection, please see our guide to what these are and how you can use them to meet your UK GDPR obligations and expectations.](#)

Organisations may find it difficult to achieve pseudonymisation as technology rapidly increases in power and accessibility. In the longer term, this would allow smaller organisations and even people to conduct what were once challenging and lengthy analyses. Given this, organisations must consider other forms of appropriate security.

This may include data aggregation achieved through implementing a trusted research environment (TRE) that can limit direct access to genomic information and associated health information. Researchers may submit queries without direct access, however this may pose a challenge to forming open and agile research environments.

Issue 4: Fairness, accuracy and opinion – epigenetics and polygenic risk scoring.

Article 5(1)(d) of the UK GDPR states that personal information must be 'accurate' and rectified promptly where this is not the case. While the UK GDPR does not provide a definition of accuracy, the Data Protection Act 2018 says that 'inaccurate' means 'incorrect or misleading as to any matter of fact'. With genomic information, the extraction of the DNA itself may introduce inaccuracies. Alternatively, the loss of metadata or the introduction of ambiguous or uncertain labelling of aspects of a genome may also cause inaccuracies.³¹

Inferences derived from genomic information can also change due to epigenetic modifications. This happens where the instructions accessed and read from DNA can be altered temporarily or even permanently by a specific environmental factor such as stress or diet.³² Even at the most basic level, we expect organisations to review their genomic information for accuracy, keep it up to date and label it as historic where required.

The probabilistic and contextual nature of second order information, such as polygenic risk scores, means that their accuracy is more likely to be open to debate. Estimating the gap between phenotypic information and genetic information that exists in many areas of study will remain highly challenging. Personal information may be produced that is probabilistic rather than absolute, based upon combined genomic data and potentially limited phenotype information.

The use of increasingly sophisticated AI models to estimate physical and behavioural aspects of a person brings its own risks around accuracy and transparency.³³ However, being based on statistical analysis rather than subjective opinion may limit the risk around personal information and accuracy under the UK GDPR. Nonetheless, AI systems must still be sufficiently statistically accurate for their purposes. This is to comply with the fairness principle and where organisations are using the models for automated decision making.³⁴

In the scenarios above, covering sectors ranging from healthcare to SEND provision to criminal charges, high-impact decisions may be taken overlaying human opinion on second order information produced by this type of processing. This in turn produces another set of personal (third order) information. There is a

risk that the line between factual information (open to challenge as inaccurate) and opinion (which can be noted as challenged by a person but which remains a fundamentally accurate note as to an opinion held) becomes blurred, which impacts on our ability to assess accuracy.

[Our guidance on accuracy](#) sets out measures to ensure that information is accurate or can be updated promptly and appropriately. It also provides details on the use of opinion to inform decisions and how to record challenges. Organisations will need to ensure the adequacy and accuracy of underlying information and to communicate how they reached their decisions in a transparent and robust manner.

It is also likely that it will be increasingly difficult for people to understand when organisations hold inaccurate personal information or when organisations have made inaccurate inferences. This is because genomic information itself is highly complex and even inferences are likely to require significant technical knowledge to interpret. Combined with the challenges of 'black box' style algorithmic processing, the challenges to transparency may be significant. Organisations should follow our guidance on the lawful fairness and transparency principle to ensure they meet our expectations. Furthermore, if the algorithmic processing amounts to automated decision-making within the meaning of Article 22, there are enhanced requirements relating to transparency.

Issue 5: Genomic determinism and discrimination

As increasingly large data sets are derived and analysed, new forms of discrimination may emerge. Without robust and independent verification of these models, there is a risk that these approaches will be rooted in systemic bias, providing inaccurate and discriminatory information about people and communities. In many instances, this information may then feed into automated systems, raising further questions over Article 22 processing and transparency.

There are also concerns about inappropriate discrimination arising from the current reliance on genomic data sets based on ancestrally European sources. ³⁵ This focus is likely to generate inaccurate information and inferences about other communities and genetic ancestries. As a result, combining broader healthcare records with genomic information and phenotypes to develop outcomes for predictive treatments may reflect and enhance embedded bias and existing discrimination. ³⁶

Active, rather than systemic, discrimination may also emerge. This may see specific traits, characteristics and information becoming seen as undesirable by organisations or groups, without being considered a protected characteristic. Alternatively, this may feed upon the perceived 'accuracy' of polygenic risk scoring, in which probabilistic tendencies become viewed as a guaranteed outcome. People may experience unfair treatment in the workplace or services they are offered based on previously unrecognised characteristics or existing physical or mental conditions. The UK GDPR already sets out requirements that may mitigate these issues. These include (but are not limited to):

- the fairness principle;
- protections for special category data;
- requirements for data protection by design and default; and
- protections for automated decision-making and profiling.

In the face of the above risks, organisations should consider [our guidance on addressing fairness, bias and discrimination](#).

In non-medical contexts, genomic information may not be classified as special category data, reducing the legal safeguards and restrictions around its processing. This may result in organisations failing to implement best practice around technical security in order to ensure that genomic information and its associated inferences remain safe from loss or theft.

Issue 6: Data minimisation, purpose limitation and genomic information

Genomic information may pose particular challenges around data minimisation and purpose limitation, particularly in direct-to-consumer services where providers hold significant sections of genomes and, in the future, potentially whole genomes with the intention of future findings becoming accessible for consumers. Organisations will need to consider carefully whether processing entire genomes is necessary for their purposes, both in the initial analysis of raw information and in the longer term.

Multimic analysis combines large omics information sets (that define the entire biological processes of a biological system) to generate new insights and inferences. This approach highlights the complex interplay in the shift of purpose, and organisations will again need to consider what information they need for a specific purpose. Fundamentally, this reinforces the significant challenges in gathering and using personal information in a rapidly moving area.

Issue 7: AI and genomics

A constant theme throughout the emerging uses of genomic information has been the algorithmic processing of large-scale, complex information. There is significant discussion on the implementation of AI in genomics, for both current and future research which carries significant data protection implications.

³⁷ Key elements to the increased demand of the use of algorithmic processing of genomic information include:

- the size and complexity of genomic datasets;
- the need for rapid specialist insights and inferences derived from the complex data in an intelligible format and;
- correlation between health records and genomic information to interpret phenotypes.

Varying dataset formats, a wide means of gathering information without a universal agreed practice and standards outside of the medical and research sectors and an ancestrally Eurocentric focus on genomic data risks the embedding of fundamental discrimination as noted above. This may only be enhanced through the use of multi-purpose models not initially trained for the purpose of analysing genomic data.

The use of polygenic risk scores to make decisions about individuals may, depending on the level of human involvement, as well as the extent to which these scores are determinative of outcomes for individuals, amount to automated individual decision-making within the meaning of Article 22. ³⁸ [Our guidance about automated decision-making and profiling](#) sets out that a decision that has a 'similarly significant' effect is something that has an equivalent impact on a person's circumstances, behaviour or choices. The scenarios set out above, for both preventative healthcare and SEND provision highlight potential areas in which rapid and potentially automated decisions may be made with significant consequences based upon probabilistic predictions.

Where automated processing does amount to solely automated individual decision-making or profiling

within the meaning of Article 22, this may present a significant challenge to an organisation's activities. Organisations may only carry out such processing when they meet one of the conditions in Article 22(2) (where this is necessary for a contract, is required or authorised by domestic law, or where someone gives their explicit consent). They can also only carry out automated individual decision-making based on special category personal information in very limited circumstances. People who are subject to such decisions have rights under the UK GDPR to obtain meaningful human intervention in the decision-making. They must have the opportunity to express their point of view and challenge decisions. Organisations must consider what appropriate intervention may look like for each situation.

There are also increased transparency requirements for organisations undertaking individual automated decision-making. They must provide people with meaningful information about the logic involved, as well as the significance and envisaged consequences (Articles 13(2)(f) and 14(2)(g)).

²⁴ [The human genome is, at long last, complete](#)

²⁵ [Huntington's disease: Woman with gene fails in bid to sue NHS](#)

²⁶ [The GDPR and genomic data](#)

²⁷ [Classifying single nucleotide polymorphisms in humans](#)

²⁸ Noting that "pseudonymised personal information" is still personal data within the meaning of UK GDPR.

²⁹ [Sociotechnical safeguards for genomic data privacy](#)

³⁰ Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010)

³¹ Cudai, Claudia, Antonella Galizia, Filippo Geraci, Loredana Le Pera, Veronica Morea, Emanuele Salerno, Allegra Via, and Teresa Colombo. 2021. "[AI Applications in Functional Genomics](#)." *Computational and Structural Biotechnology Journal* 19: 5762–90.

³² [Environmental exposures influence multigenerational epigenetic transmission](#)

³³ Dias, Raquel, and Ali Torkamani. 2019. "[Artificial Intelligence in Clinical and Genomic Diagnostics](#)." *Genome Medicine* 11 (1): 70.

³⁴ See UK GDPR Recital 71 for further details.

³⁵ Kessler, M., Yerges-Armstrong, L., Taub, M. et al. [Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry](#). *Nat Commun* 7, 12521 (2016).

³⁶ Chen, Irene Y., Peter Szolovits, and Marzyeh Ghassemi. 2019. "[Can AI Help Reduce Disparities in General Medical and Mental Health Care?](#)" *AMA Journal of Ethics* 21 (2): E167-179.

³⁷ [DNA.I](#) - Ada Lovelace Institute

³⁸ Article 22 regulates the circumstances in which solely automated decisions with legal effects or similarly significant effects on individuals may be taken.

Next steps

We understand the need for further work in this area from a regulatory perspective, due to the range of potential uses of genomic information on the near horizon. As part of this process, we will continue to scrutinise the market and identify key stakeholders who are seeking to develop or deploy uses of genomic information. This will help us to continue building our knowledge and understanding of the issues raised, particularly in direct-to-consumer areas.

In particular, we understand that organisations and researchers desire greater clarity in understanding when genomic information may be personal information. We don't consider that all genomic information is personal information, given the definition of personal information set out under the UK GDPR and the amount of genomic information that may not relate to an identifiable person. We will continue to work with stakeholders and experts to understand how and when these areas do overlap and how different uses and emerging techniques may shift the boundaries. In turn, this will better enable us as a regulator to offer guidance and advice to organisations working with genomic information to ensure that they can work and innovate with confidence in privacy by design practices.

Across our ICO tech future reports on biometrics, neurotechnologies and genomics it has become clear that there is also a need for a better understanding what exactly is health information under the UK GDPR. Our guidance on health data and its use in the workplace provides some of our most recent thinking in this area. We are keen to build upon this and will be reviewing our guidance relating to health data to ensure it provides useful examples that reflect the rapidly shifting and complex context created by broader emerging technologies.

We will also continue to work with stakeholders and others to explain the importance of privacy by design and compliant use of personal information. We may also consider forums for key stakeholders to discuss emerging techniques on appropriate data sharing or anonymisation techniques for genomic information.

We also want to hear from organisations with an interest in direct-to-consumer genomic services and services linked to polygenic risk scoring. We are particularly interested in views on the potential creation and development of standards in this area as well as those who may be interested in working with our Regulatory Sandbox to embed privacy forward practices in uses of genomic information.

In the longer term, we will continue to monitor and call out developments in this area through our Tech Horizons report and set out future programs of work in that document. We also want to continue to work with critical stakeholders. We want to hear from organisations who are working in this sector, whether it is in the development of novel or advanced uses of genomic information, their deployment or through thinking about their implications in a policy based or regulatory context. We would very much like to hear from you as we continue to develop our knowledge and thinking in this area. You can contact us here:

[\[email protected\]](#).

Annex A: Key technologies and terms

Annex A: Key technologies and terms

Epigenetics – Layer of chemical information that sits on top of DNA that regulates how DNA is used (but not the DNA itself), when genetic information is used and what proteins are produced. Epigenetic profiles aren't always stable and can be impacted by environmental factors. A key focus of epigenetic research is the underlying hereditary causes of cancer and cognitive decline.

Functional genomics – research aimed at analysing and understanding the core aspects of genomics (including DNA, RNA, proteins and metabolites, along with their modifications) that link the observable characteristics of a person (the **phenotype**) to the correlation of underlying genetic characteristics (the **genotype**) and the contextual environmental conditions.

Genetics – The study of a section of DNA which in turn, carries out a specific function. Genes are arranged along chromosomes and consist of a sequence of DNA that is transcribed to produce a protein or an RNA product. These products carry out biological functions inside or outside the cell or regulate the activity of other genes.

Genomics, genome – the study of an entire unique DNA sequence belonging to (in this instance) a person that provides the fundamental instructions as to how we develop, heal and behave.

Genome Wide Association Study (GWAS) – An analytical method in which broad areas of a genome are compared to traits and characteristics to find areas of possible correlation. Common DNA variations (typically single-nucleotide polymorphisms, or SNPs, but it can be other types of variation) are compared to a trait or behaviour in a population to see if any variant correlates with the trait. GWASs do not directly show which gene or DNA variant is causally responsible for variation in the trait. They indicate locations on the genome from where the signal is originating. These regions are then followed up in further experiments designed to assess the likely pathways they act on, and their impact on gene function and regulation.

Genomic testing – a wide array of genomic testing is possible and this can include:

- **Predictive testing**
- **Diagnostic testing**
- **Forensic testing**

Phenotypes – the observable traits or aspects of a person as determined by their genomes and environment. For example, physical traits such as height or weight, quantifiable behaviours or symptoms.

Multomic analysis – large scale biological analysis drawing upon multiple data sources including (but not limited to) genomic, epigenomic, microbiome and proteomics to achieve higher quality predictive analysis. The approach is usually combined with AI-based processing given the large quantities of data generated.

Non-invasive prenatal tests (NIPT) – a test undertaken during pregnancy that seeks to identify genetic and genomic markers that may identify particular conditions or chromosomal markers such as Down's Syndrome.

Polygenic risk scoring (PRS) – a cumulative measure of genetic liability for a person based on the

summed effects of many thousands of risk variants distributed throughout the genome.

Whole genome sequencing (WGS) – the process of revealing a complete DNA sequence of an entire genome. This can be achieved through:

- **Short read sequencing** in which fragments of DNA (usually some 100-500 base pairs) are analysed separately before bioinformatic techniques are used to create a full sequence.
- **Long read sequencing** that uses larger fragments (usually 10,000 – 100,000 base pairs) without the need for stitching results together from smaller samples.

Annex B: Genomic sector economic overview

Key points

Genomics activities can be broadly split into **five key areas**: Sampling; Sequencing; Analysis; Interpretation and Application. In 2021/22:

- There were **149 UK sites where genomics has been identified as a primary activity**. The **largest group** of these **relate to application activities**, where genomic information is used to provide diagnostics or inform drug development (67 sites, 45% of sites).
- The total **turnover from UK genomics sites was £3.6 billion**. The share of turnover attributable to sequencing sites has grown significantly, rising from 50% of total turnover in 2008/09 to 88% in 2021/22.

In terms of investment:

- The UK's genomics sector has **received over £300 million in public grant and research funding**.
- Between 2011 and 2021, the sector also **raised £3.3 billion in equity funding** – roughly 10% of the overall private funding in the UK life sciences sector. There is **limited comparable data** on the scale of activity and levels of investment in the genomics sector internationally.

Context – Overview of Genomics Activity

Genomics involves the study of genomes (the complete set of DNA within an individual) using a wide range of rapidly evolving technologies and techniques. ³⁹ It forms part of the UK's life sciences sector – a sector which was estimated to contribute nearly £13 billion to UK GDP in 2020. ⁴⁰

Genomics activities can be categorised into five broad areas. These are highlighted below with associated revenue volumes.

Activity Definition UK turnover in 2021/22

Activity	Definition	UK turnover in 2021/22
Sampling*	The collection of a human DNA sample (i.e. blood or saliva) for laboratory analysis	-
Sequencing	Decodes the genome. Large scale sequencing is heavily reliant on high tech equipment. At £3.2bn sequencing accounts for	£3.20 billion (88% of total)

	the majority of industry turnover in the genomics sector.	industry turnover)
Analysis	After sequencing, data is generated in various forms; data which needs to be analysed and standardised via software and other methods. Comparison with phenotypic data (an individual's observable traits, such as height and eye colour etc) can also help draw out information at the interpretation stage	£0.02 billion (<1% of industry turnover)
Interpretation	Turning the data and information received into previous stages into information for clinicians and pharmaceutical companies	£0.04 billion (1% of industry turnover)
Application	This is the final stage in the processing chain where genomic information is used to provide diagnostic treatment, targeted therapies or inform drug development. Other applications include direct to consumer genetic testing, such as ancestry research, pre disposition to disease or identifying skills such as recognising musical pitch. ⁴¹	£0.40 billion (11% of industry turnover)
Total		£3.6 billion**

Source: Turnover data from UK Office for Life Sciences. ⁴² Figures may not sum due to rounding. *No turnover data was available for sampling activities. ** This figure relates to turnover for the UK's genomics sector and is not directly comparable to the earlier figures referenced for the broader life-sciences sector.

The UK genomics sector is complex, and relies on the interaction of different public and private organisations to function effectively. Some examples of these in a health context include:

- **Private genomics companies** which provide diagnostic tools and therapies to the NHS, and receive data generated from patient interactions in return.
- **Academic spin outs** which have led to the creation of a number of genomics firms, which are often involved in research and development.
- **Pharmaceutical firms** regularly partner with genomics firms to speed up drug development.
- **Government agencies** and **private investors** are also involved and provide funding and support, to finance projects and provide firms with opportunities to scale. ⁴³

There is also a wide range of applications beyond the health sector including: ⁴⁴

- Agriculture - using plant genomes for increased crop yields, or other desirable traits (developing crops that are drought-resistant or require fewer pesticides);
- Environmental conservation - understanding the genetic makeup of endangered species can help inform conservation efforts;
- Climate change mitigation - genomic technologies are being applied to understand and engineer organisms that can capture carbon dioxide or produce biofuels; and
- Bio-based materials – producing sustainable alternatives to plastics and other materials.

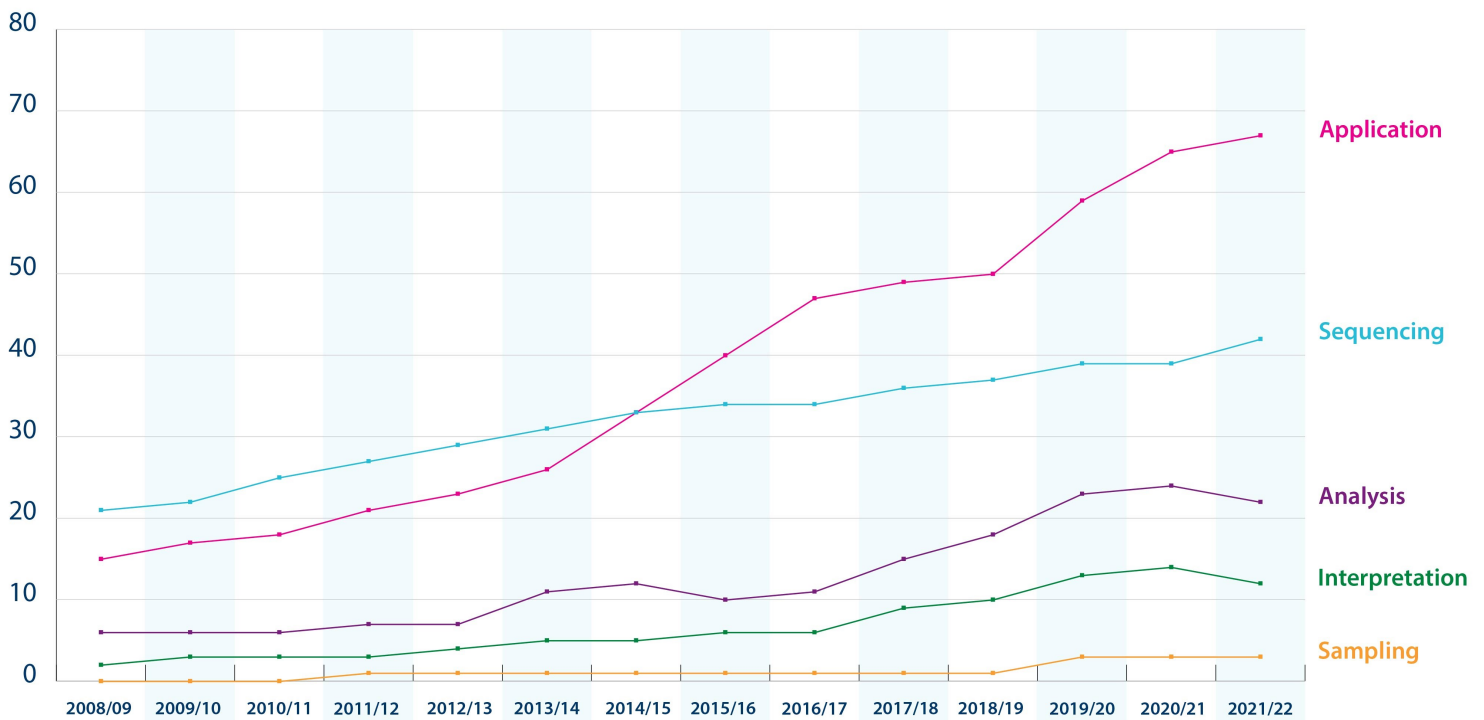
Genomics activity in the UK

Business base

There are a range of estimates around the scale of genomics firm level activity in the UK. Genomics Nation estimate that in 2022 there were 142 genomics companies in the UK, 41 of which were university spin outs. ^{4.5} Similarly, the UK Office for Life Sciences found that in 2021/2022 there were 149 sites where genomics has been identified as a primary activity. ^{4.6}

The largest group of these sites relate to application activities, where genomic information is used to provide diagnostics or inform drug development (67 sites, roughly 45% of sites). Sequencing activities account for the second highest number of sites (42 sites, roughly 28% of sites), followed by analysis (22 sites, roughly 15% of sites) and interpretation (12 sites, roughly 8% of sites). Sampling and unclassified activities make up the remainder (3 sites, roughly 2% respectively). This is highlighted in Figure 2 below showing a moderately growing trend across the genomic business base over time, with application activities showing the highest rate of growth.

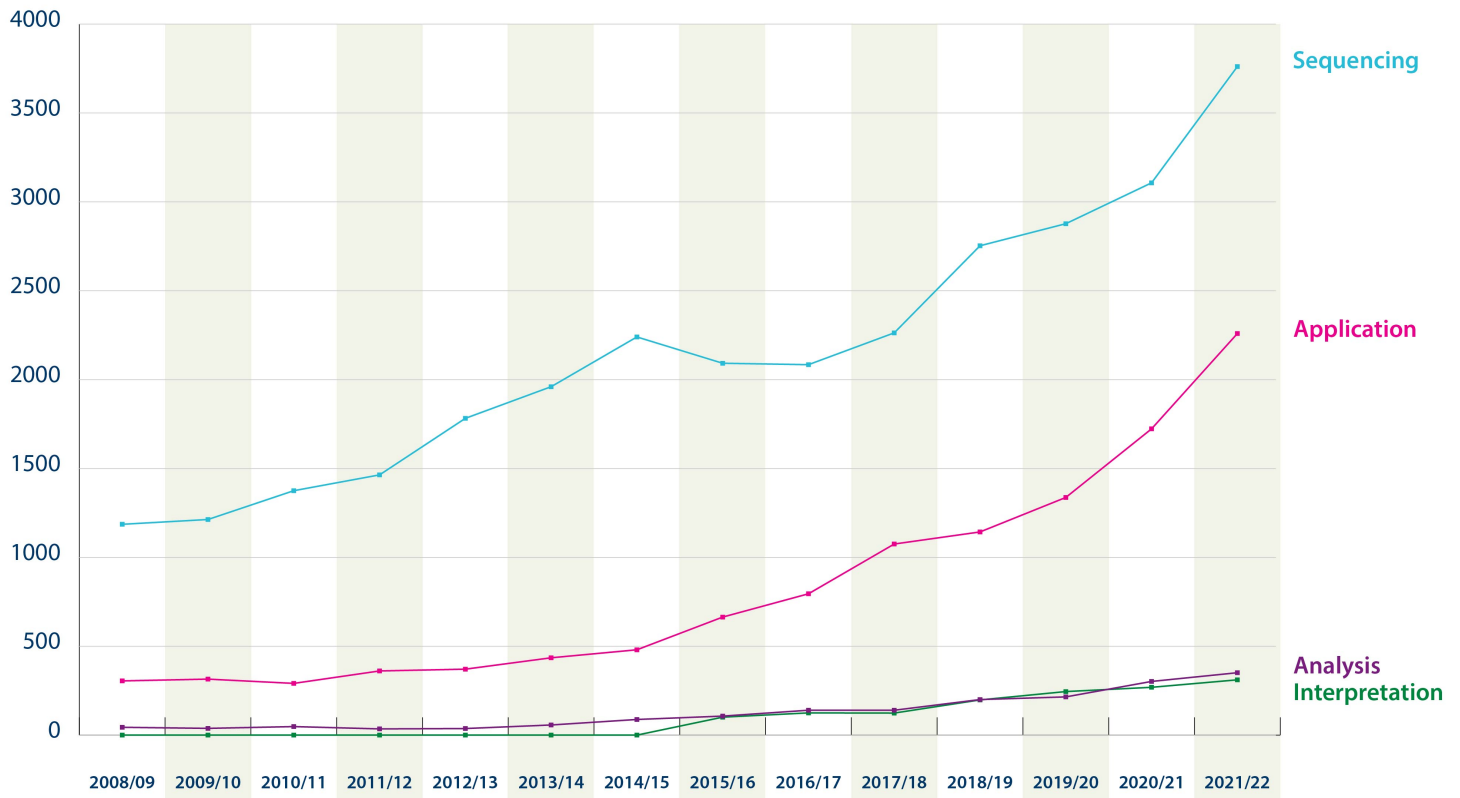
Figure 2 - Number of UK genomic sites (2008/9-2021/22)



Employment

Employment at these sites has more than doubled from 3,200 in 2016/17 to 6,800 in 2021/2022. ^{4.7} The majority of these employees work in sequencing sites (3,761 employed at these sites, 56% of total employees). Application sites account for the second highest number of employees (2,259 employed at these sites, 33% of total employees) followed by Analysis and Interpretation (351 and 311, roughly 5% each). No employment data was available for sampling sites.

Figure 3 - Employment at UK genomics sites (2008/9-2021/22)

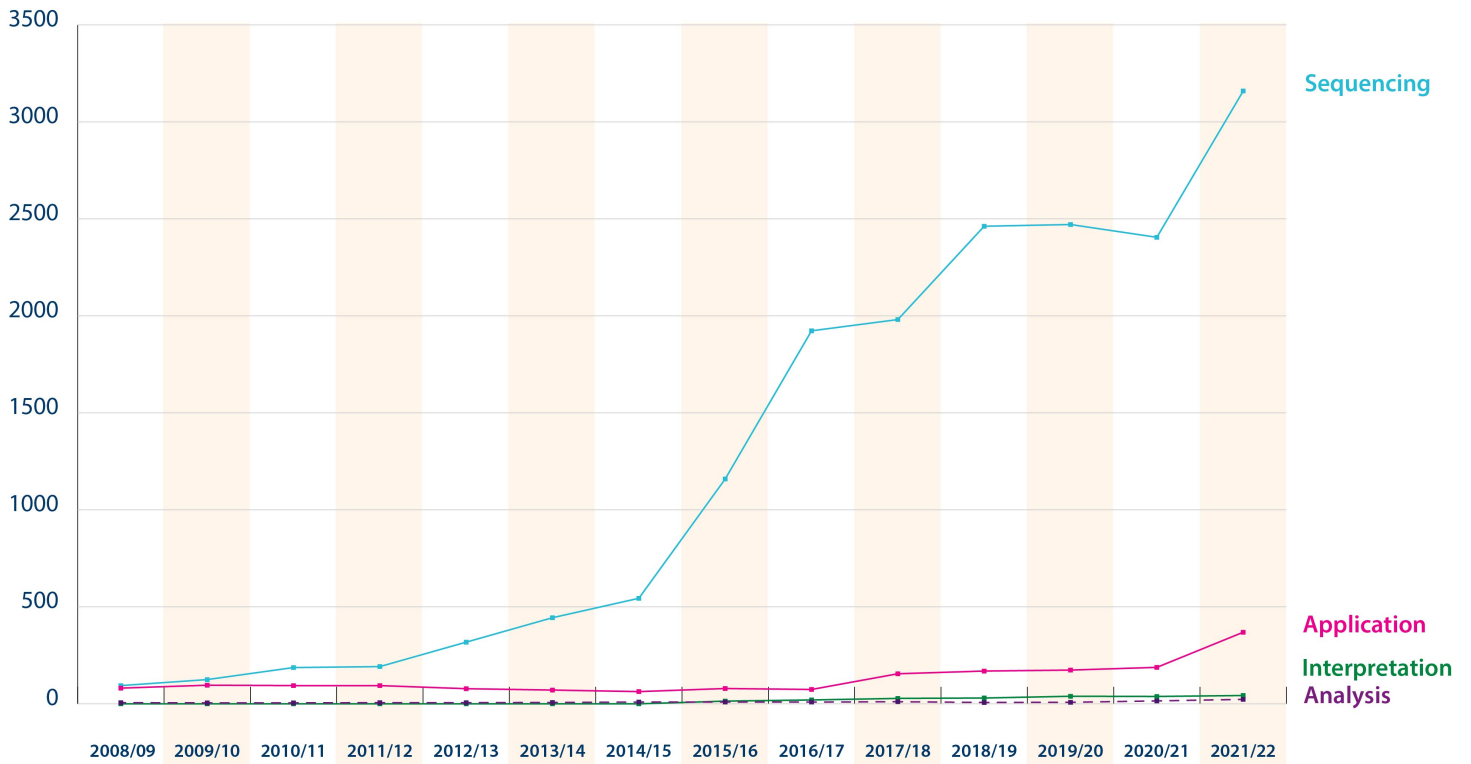


Source: UK Office for Life Sciences

Turnover

In 2021/22, the total turnover from UK genomics sites was £3.6billion – nearly double the level of industry turnover generated in 2011/12, as shown in Figure 4. Sequencing activity has consistently accounted for the largest part of the industry turnover. The share of turnover attributable to sequencing sites has grown significantly, rising from 50% of total turnover in 2008/09 to 88% in 2021/22. [48](#)

Figure 4 – Turnover (£m) from UK genomics sites (2008/09-2021/22)



Source: UK Office for Life Sciences

Public and private investment

Public funding for genomics mainly takes the form of government grants and research funding. Between 2011 and 2021, the sector received £151 million in grant funding – equivalent to around a quarter of the total grant funding awarded to the UK’s life-sciences sector. ⁴⁹ The UK Government has also allocated £175 million for genomics research. This includes:

- £105 million for a research programme, led by Genomics England in partnership with the NHS, to speed up diagnosis and treatment of rare genetic diseases in newborns.
- £22 million for Genomics England to tackle health inequalities in genomic medicine
- £26 million for a cancer research programme, led by Genomics England to improve the accuracy and speed of diagnosis for cancer patients and up to £25 million of Medical Research Council-led funding for a 4-year functional genomics initiative.

There is limited public data available on levels of private investment in the genomics sector.

Between 2011 and 2021, the genomics sector raised £3.3 billion in equity funding – roughly 10% of the overall private funding in the UK life sciences sector. Just over half (£1.7 billion) was raised from private equity and venture capital firms with the remainder coming from sources such as corporate investors. ⁵⁰

Data Availability

While it is important that the figures in this paper are considered in a broader context, there is limited comparable robust data available on levels of activity and investment in genomics internationally to allow

benchmarking.

- ³⁹ [Information Commissioner's Office \(2022\) Tech Horizon Report](#). (Accessed 19 April 2024).
- ⁴⁰ [Oxford Economics \(2024\) UK life-sciences are set for growth](#). (Accessed 26 April 2024).
- ⁴¹ [Information Commissioner's office \(2023\) Tech horizons report](#). (Accessed 5 March 2024).
- ⁴² [Office for Life Sciences \(2023\) Bioscience and health technology sector statistics](#). (Accessed 11 March 2024).
- ⁴³ [UK Bioindustry Association \(2023\) Genomics Nation](#). (Accessed 28 February 2024).
- ⁴⁴ [UK Bioindustry Association \(2023\) Genomics Nation](#). (Accessed 28 February 2024).
- ⁴⁵ [UK Bioindustry Association \(2022\) Genomics Nation](#). (Accessed 28 February 2024)
- ⁴⁶ Sites is an encompassing term used by the UK Office for Life Sciences. It includes genomics firms, other life sciences firms, university spin outs etc. The term sites refers to locations which are classified as level 1 or 2 genomic activity. [Office for Life Sciences \(2023\) Bioscience and health technology sector statistics](#).
- ⁴⁷ [Office for Life Sciences \(2023\) Bioscience and health technology sector statistics](#). (Accessed 11 March 2024).
- ⁴⁸ [Office for Life Sciences \(2023\) Bioscience and health technology sector statistics](#). (Accessed 11 March 2024).
- ⁴⁹ [UK Bioindustry Association \(2021\) Genomics Nation](#). (Accessed 28 February 2024).
- ⁵⁰ [UK Bioindustry Association \(2021\) Genomics Nation](#). (Accessed 28 February 2024)

Annex C - Background and context

Over a hundred years passed between the initial discovery of DNA and the identification by Franklin, Crick and Watson of the helical structure of DNA in 1953. It then took nearly 50 years of research before the entire human genome was sequenced in 2003. In the subsequent 20 years, genomic technologies had developed rapidly. A genome sequencing that once took 13 years to deliver can now be completed in a matter of days. ⁵¹

Yet it is not only the speed of analysis that has changed; the scale of data collected is also likely to expand. ⁵² A short read sequence of a genome might create up to 160 gigabytes (GB) of data, with a long-read sequence creating up to 500GB. ⁵³ As these numbers are scaled up by hundreds of thousands and potentially into the millions, stakeholders have noted storage issues, its energy and environmental impact and security challenges for both organisations and people who might wish to hold their genomic information.

The information and inferences that you can draw from this information has also significantly increased. It is about genetic variation, which is the differences in DNA sequence between people. These differences can be common (variations which arose a long time ago and spread through the population) or rare (arose more recently and have not spread, perhaps because they are damaging to health). Genomics seeks to identify common and rare variants that correlate with disease and understand how they work. Sometimes the variants are not in genes (in non-coding region of the genome) and have subtle effects that are hard to map to genes or biological processes and pathways. Yet discoveries will continue to be key to critical healthcare and medical therapies to treat illnesses and conditions such as macular degeneration, type 2 diabetes, prostate cancer and treatments for COVID-19. ⁵⁴

The UK's own recent pursuit of genomic research begins with the [2011 UK Life Sciences Strategy](#), followed by the subsequent launch of [100,000 Genomes Project](#). The project delivered the DNA sequences of 100,000 NHS patients with either cancer or rare conditions as part of a development of emerging treatments. ⁵⁵ In 2013, Genomics England was created to oversee the project and in 2016, the NHS Genomic Medicine Service (GMS) was created to build upon this work and to integrate genomic medicine into UK healthcare. ⁵⁶ The project was completed in 2018. As well as these bodies, research organisations such as UK Biobank, Our Future ⁵⁷ and the National Institute for Health Research (NIHR) are also supporting further work on the use of genomics in the UK for health and non-health related purposes. ⁵⁸

In terms of actually analysing genomic information, there is increasingly a move to use genome wide association studies (GWAS) as a means of focusing on the '**missing heritability**' problem for common traits. ⁵⁹ However, while GWAS is the paradigm of choice, it typically only assesses common DNA variants using SNP arrays. This focuses analyses and inferences upon commonly recognised traits and characteristics. A whole genome sequencing (WGS) is needed for an assessment of all types of variants, currently posing some barriers in terms of cost and time and well as analysing larger quantities of personal data.

⁵¹ [Super-speedy sequencing puts genomic diagnosis in the fast lane](#)

52 [Big Data: Astronomical or Genomical?](#)

53 [Storage and Computation Requirements](#)

54 [Genomics Beyond Health - full report](#)

55 [100,000 Genomes Project](#)

56 [NHS Genomic Medicine Service](#)

57 [NHS Genomic Medicine Service](#)

58 [Genomics Beyond Health - full report](#)

59 'Missing' refers to the gap between what all common variants in a GWAS capture ('SNP heritability'), and twin-based h^2 estimates. This missing gap indicates other types of variants not assessed in the GWAS might be important, such as structural variants and rare(r) variants.