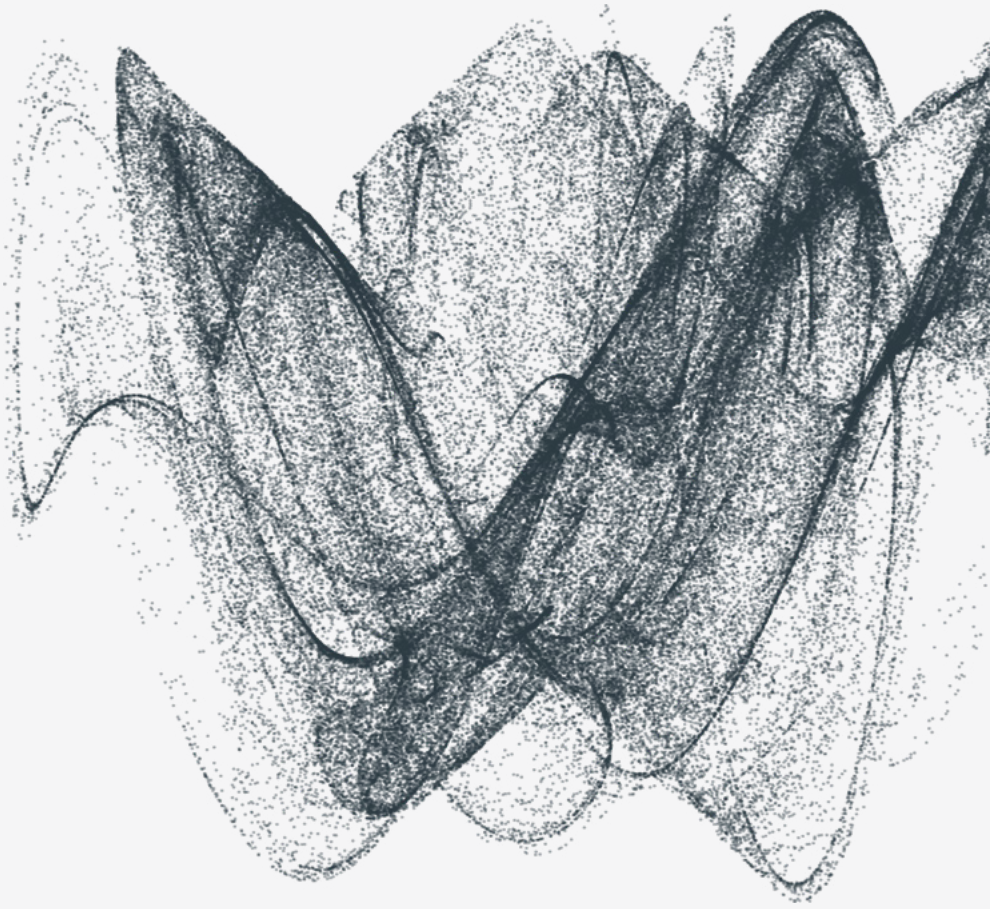
IQVIA™
INSTITUTE
FOR HUMAN DATA SCIENCE



Understanding the Global Landscape of Genomics Initiatives

PROGRESS AND PROMISE

MAY
2020

Introduction

The landscape of initiatives to generate and collect human genomic data is evolving rapidly and is highly diverse, with private and public initiatives across multiple countries. However, to date, there has been no comprehensive directory of genomic and biobank initiatives. To aid understanding of the genomic data landscape at this pivotal point, IQVIA has created a database, using publicly available information, of initiatives that generate and aggregate data on the human genome. With it, stakeholders can gain a better picture of the current genomic landscape of 187 current initiatives and their achievements and assess the value of this explosion in human genomic data to advance Human Data Science and medical research.

This report provides the first examination and segmentation of the global genomic landscape. It reviews the diversity of genomic initiatives, their number, breadth and geographic distribution, as well as key parameters that tie to their utility to improve health outcomes. It further puts forth a view of how those parameters tie to value generation for research or personalized medicine. Elements that provide utility to genomic databases are reviewed and the alignment of current initiatives to these elements is assessed. Finally, the complete set of initiatives is analyzed using a standardized genomic data quality score.

This report utilizes the IQVIA Genomic Initiatives Database — a new database of genomic data initiatives and databases, public and private — collected through a systematic assembly and curation of publicly available information. The publication of this first genomic landscape provides a baseline — a starting point to measure progress on the most important use of human genomic data — to build insight into the origins of human disease and better therapeutics to prevent and treat it.

The study was produced independently by the IQVIA European Thought Leadership on behalf of the IQVIA Institute for Human Data Science as a public service, without industry or government funding. The contributions to this report of Andrea Guaraglia, Ron Miller, Winfred Shaw and dozens of others at IQVIA are gratefully acknowledged.

Find Out More

If you wish to receive future reports from the IQVIA Institute for Human Data Science or join our mailing list, visit iqviainstitute.org

MURRAY AITKEN
Executive Director
IQVIA Institute for Human Data Science

Overview

The evolving genomic landscape

+ Since 1990, the cost of sequencing a whole genome dropped from $2.7 billion to as low as $300, opening new opportunities to build repositories of genomic data.
+ The next decade will supply researchers with vastly increased genomic data resources to gain insights into the molecular mechanisms of human disease and better understand the epidemiological landscape.
+ Genomic initiatives can help healthcare stakeholders identify genetic variants that increase risk for disease, diagnose patients earlier and prevent disease, develop companion diagnostics to personalize treatment with medicines, and accelerate the discovery, repurposing and clinical development of medicines.

Characterizing genomic initiatives

+ By the start of 2020, 38 million genomes had been analyzed using techniques ranging from genotyping to whole genome sequencing, and this number is expected to grow to 52 million by 2025.
+ The IQVIA Genomic Initiatives Database is a useful repository of publicly available information on global initiatives generating and aggregating data on the human genome and can help stakeholders understand how to leverage this data for the advancement of human data science and medical research.
+ There are 187 genomic initiatives globally of which 50% originated in the U.S. and 19% in Europe.
+ Planned national genomic databases are proliferating as countries increasingly appreciate the potential technological and healthcare system benefits of genomic data, with some countries planning to sequence the entire population.
+ Initiatives in the United States are more likely to be privately owned than those in Europe. Only 32% of initiatives are public in the U.S. versus 50% in Europe but the U.S. still has nearly double the number of public initiatives.
+ The medical utility of data varies across initiatives based on the number of genomes collected, the completeness of genomic data, linkage to other healthcare-relevant data, and the disease or populations it covers.
+ Among disease-specific datasets more than half (53%) focus on oncology, 13% on rare diseases and 10% on CNS disorders.
+ North American initiatives dominate by target cohort sizes, with 36 million genomes North America expected to be collected by 2025, but a higher proportion of the data (74%) is genotyped data with more limited applications.
+ Compared with other regions, a higher proportion of the genetic data collected in Europe and Asia will be whole genome (28%) or biological samples (29%) that could be fully sequenced either now or in the future.
+ Europe is expected to fully sequence 1.5 million whole genomes by 2025, more than the 780,000 in North America, but an additional 2.2 million whole exomes will also be sequenced in the latter and will hold high medical value.
+ Only 42% of databases publicly state their genomic data links to patient demographic information or clinical data, among which 28% of initiatives tie to the most valuable EMR/EHR and clinical data.

Overview

+ The value of today's evolving genomic landscape can be assessed by using a "data quality score" that considers the breadth of genomic data, its linkage to other medically relevant data, data consent policies and data access.

+ Analysis of initiatives' data quality versus their target cohort sizes suggests that the next decade will see an increasing number of large genomic databases with strong utility for human data science and medical research.

What initiatives have done with this data

+ Use of data from the UK Biobank for a Genome-Wide Association Study (GWAS) on lung function as well as the China Kadoorie Biobank led to the identification of 43 novel, independent genomic signals for lung function. This study nearly doubled the number of known genomic signals for lung function, improved understanding of patient COPD risk and is helping drug developers find new drug targets or repurpose approved drugs to treat COPD and other respiratory conditions.

+ The Lung-MAP Trial, an 'umbrella trial' launched in June 2014, has significantly increased the amount of genomic data on non-small cell lung cancer. It has furthered scientific understanding of common NSCLC mutations, established a pathway to accelerate drug development to benefit patients and thereby advanced personalized medicine and research.

+ Using genomic data from the Genomics England Initiative, a clinical algorithm was developed to identify patients with early onset neurological conditions who progress rapidly and identify associated genetic variants. By focusing clinical research initially on patients that progress rapidly, this can accelerate drug development and speed targeted therapies to patients who have the greatest need and may benefit the most from treatment.

The future of genomic data

+ Genomic data currently collected by existing initiatives focus on populations in the United States and Europe and do not reflect the genomics of global populations well.

+ Unless there are dramatic increases in the sequencing of populations in Asia, Africa and South America, they will continue to be under-represented in genomic databases. Though Asia's population makes up 60% of the world's population, initiatives currently plan to sequence only six million Asian genomes by 2025 — 12% of the global target.

+ With a continued drop in the costs of genomic sequencing, and the rising interest of governments to build national, whole-population genomic databases, fully-linked, consented and accessible databases that hold the greatest value for medical research are increasingly becoming a reality.

+ Challenges remain in oncology as most tumor samples are collected in a way optimized for histology not whole genome sequencing. Though cancer has been a driver of initial genomic activity, this issue may impact downstream genomic analyses and see genomic data contribute more in other disease areas.

+ The development of interoperability standards and agreements are needed to enable future linkage across databases and accelerate the power of very large, high-quality genomic databases to improve human health in the next decade.

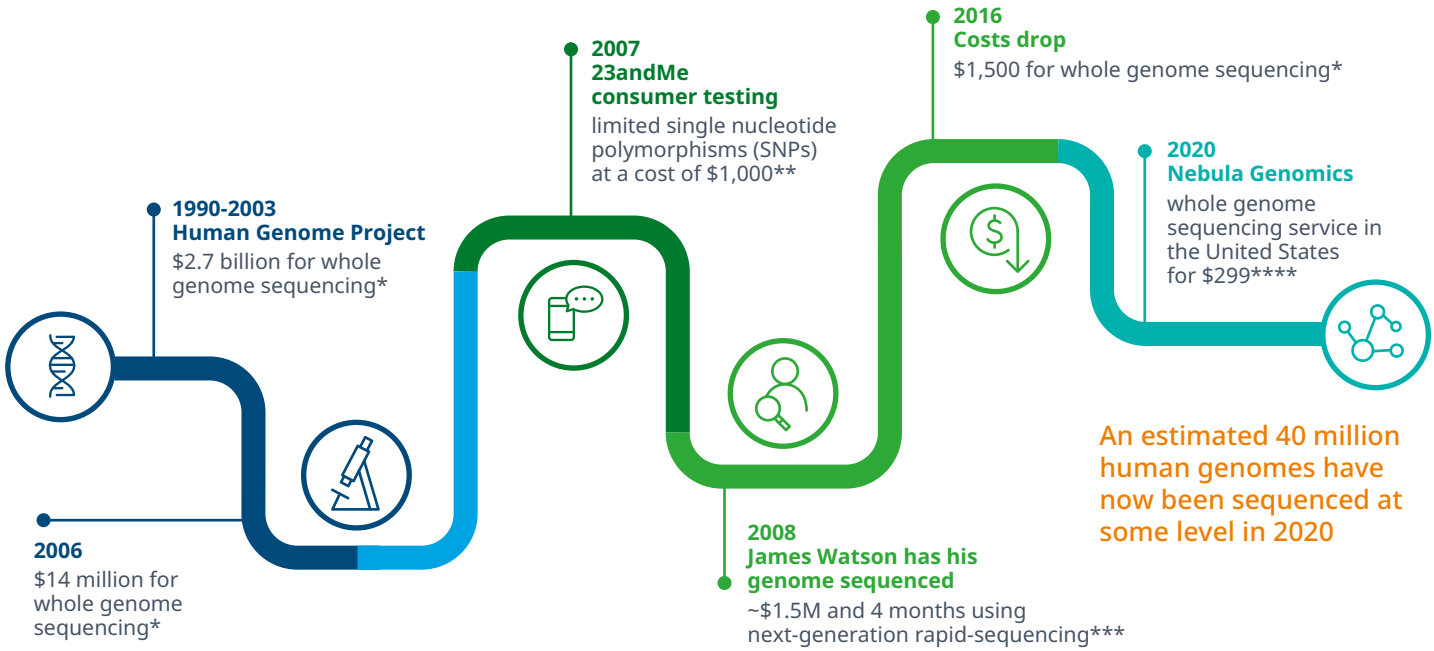
The evolving genomic landscape

In 1990, a thirteen-year long project started: the sequencing of the human genome. The final sequence, delivered in 2003, was a patchwork of multiple individuals' genomes, representative of no single person. The project involved 20 research institutions and cost an estimated $2.7 billion.[1] Since 2003, the cost of genomic sequencing, either in part or in full, has fallen exponentially (see Exhibit 1). In 2006, the cost to generate a full human genome sequence was estimated at roughly $14 million.[1] In the same year, 23andMe was founded and launched a consumer DNA testing service in 2007, initially genotyping only limited single nucleotide polymorphisms (SNPs) at a cost of $1,000. A decade later in 2016, whole genome sequencing cost was estimated to have fallen to $1,500.[1] As an example, by 2020, Veritas Genetics, a U.S. start-up, offered a whole genome sequencing service in the United States for $599,[2] as has Nebula Genomics for $299,[3] and the cost to research organizations sequencing at scale is expected to fall lower still as technologies beyond next generation sequencing emerges. In the future, new technologies will likely bring the cost of whole genome sequencing to the point where it is a trivial cost per individual sequenced.

Dramatic improvements in sequencing technology and equally dramatic reductions in cost have led to an explosion in the number of human genomes sequenced. Almost 40 million human genomes have now been sequenced to some extent, or genotyped.[4] Undoubtedly, the 2020s will see an acceleration in the accumulation of

Dramatic improvements in sequencing technology, and equally dramatic reductions in cost, have led to an explosion in the number of human genomes sequenced.

Exhibit 1: The Evolving Cost of Genomic Sequencing



Source: IQVIA European Thought Leadership, Feb 2020. * NIH. National Human Genome Research Institute. The cost of sequencing a human genome. 2019 Oct 30 **Goetz, T. 23andMe Will Decode Your DNA for $1,000. Welcome to the Age of Genomics. 17 Nov 2007. ***Wadman, Meredith. "James Watson's genome sequenced at high speed." Nature 452 (7189), 788. 2008. **** Nebula Genomics. 30x Whole genome sequencing for $299. Apr 2020.

genomic data. However, it is difficult to make predictions on growth: some organizations that generate this data do not make forecasts at all, while others — often nationally-instituted genomic projects — have set ambitious targets which may or may not be completed within their designated timelines. What is clear, however, is that the generation of human genome data is undergoing a phase of exponential growth. Healthcare stakeholders will exit the 2020s with vastly increased data resources, greatly enhanced opportunities to gain insight into the genetic causes of human disease, and the prospect of leveraging these for the development of treatments and preventive approaches.

THE BENEFITS OF HUMAN GENOMIC DATA

Human genomic databases are built and used for a wide range of purposes. Understanding individual ancestries, researching human origins, forensic enquiry and encouraging life science innovation (often alongside the desire to build a national profile in cutting edge technologies) fall among these, but by far the most important and prevalent applications for genomic initiatives today is gaining insight into the causes of human disease and disability, and developing therapeutics to prevent, treat and cure disease. Over 10,000 human diseases are monogenic, caused by a defect in a single gene, including the vast majority of rare or orphan diseases.[5] Diseases and conditions that cause most of the world's burden of disease mortality and morbidity, including cancers, diabetes, cardiovascular diseases and asthma, involve a complex interplay between genetic and environmental influences in their origins. Infectious diseases such as HIV[6] and malaria[7] also involve a genetic component to susceptibility and treatment response. Genomic databases can help healthcare stakeholders and researchers in a number of ways (see Exhibit 2).

Exhibit 2: Applications of Genomic Databases for Healthcare and Life Sciences Stakeholders

USE	DESCRIPTION
Disease epidemiology	Identify genetic defects and variant-linked conditions, their influence on disease onset and progression, and their relationship to phenotype
Molecular mechanisms of disease	Leverage info on coded protein structures and function, as well as the influence of non-coding regulatory regions and binding factors to clarify disease pathways
Patient risk assessment	Identify defective genes and genetic variants associated with increased individual risk for disease
Informing treatment	Personalize treatment via pharmacogenomics and assessment of patient disease risk
Interpretation of clinical trial results	Understand why a trial may have been more or less successful specific segments of patients to help guide future trials and develop personalized treatments
Clinical trial optimization	Design better clinical trial protocols and recruit patients based on their genetic information
Drug target identification	Prioritize drug targets based on genetic information
Drug repurposing and repositioning	Identify new uses for approved drugs based on a better understanding of disease processes and drug mechanisms of action
Companion diagnostic development	Find genetic markers that impact how a person responds to a drug. Develop a genetic test to measure that marker for use before drug prescription
Personalized medicines market sizing	Measure how many people will benefit from a drug based on their genomic data, use that information to determine the number of potential customers

Source: IQVIA Institute, Mar 2020

Specifically, genomic databases enable researchers and healthcare stakeholders to,

- Understand the epidemiological landscape for a disease, identifying the genetic defects and variants that are linked to a condition, their influence on disease onset and progression, and their relationship to phenotype. Such studies — for example the Pan-Cancer Analysis of Whole Genomes study[8] — contribute both, to a better understanding of how disease originates and affects populations, and potentially to earlier diagnosis and prevention of disease when earlier molecular changes can be identified and detected. They can also help identify new uses for approved drugs based on a better understanding of disease processes.

- Identify defective genes and genetic variants known to be associated with increased risk for disease in specific individuals. For example, tests to establish an individual's genetic propensity for developing breast cancer by analyzing BRCA genes have been available since 1996. Since then, the number of genetic tests has multiplied, with a recent announcement in the United Kingdom of the intent to offer full genome sequencing of U.K. newborns as part of routine screening, with 20,000 babies participating in an initial pilot.[9]

- Allow the personalization of treatment with medicines via pharmacogenomics, where genetic variants that have an impact on patient response to drugs or the pharmacokinetics of medicines are identified and catalogued and may be returned to patients to inform care. A rapidly growing number of recommendations for specific medicine-gene interactions already exist,[10] and the development of genetic companion diagnostic tests to measure those markers before drug prescription for precision medicine is also continuing. Clinically actionable genomic variants may also be used to inform patient treatment directly at institutions such as at the Danish National Genome Center[11] and Dana Farber's Center for Cancer Genome Discovery,[12] among others, some of which directly tie to genomic initiatives.

- Drive the discovery and development of new medicines at all stages by helping to identify druggable targets in disease pathways, guide the development of medicines to edit/replace specific genes (gene therapies), and influence manufacturers' research and development plans — for example, through analyses sizing the potential treatable population or offering better ways to identify patient candidates for clinical trials or evaluating therapeutic outcomes in specific patient populations.

However, not all genomic databases are positioned to fully realize this promise. The degree of research utility that a genomic database has is heavily dependent upon its scale, the type of genomic data collected, population composition, linkage with other phenotypic/clinical information, and its availability for use by researchers.

What is clear is that the generation of human genome data is undergoing a phase of exponential growth. Healthcare stakeholders will exit the 2020s with vastly increased data resources, greatly enhanced opportunities to gain insight into the genetic causes of human disease, and the prospect of leveraging these for the development of treatments and preventive approaches.

Characterizing genomic initiatives

A GLOBAL OVERVIEW OF INITIATIVES

To aid understanding of the genomic data landscape at this pivotal point, IQVIA has created The IQVIA Genomic Initiatives Database, derived from publicly available information on global initiatives generating and aggregating data on the human genome. We believe it is the most comprehensive database of genomic initiatives currently available. The intent of this unique assessment is to depict current and planned achievements, and to enable stakeholders to realize the value of this explosion in human genomic data for the advancement of human data science and medical research.

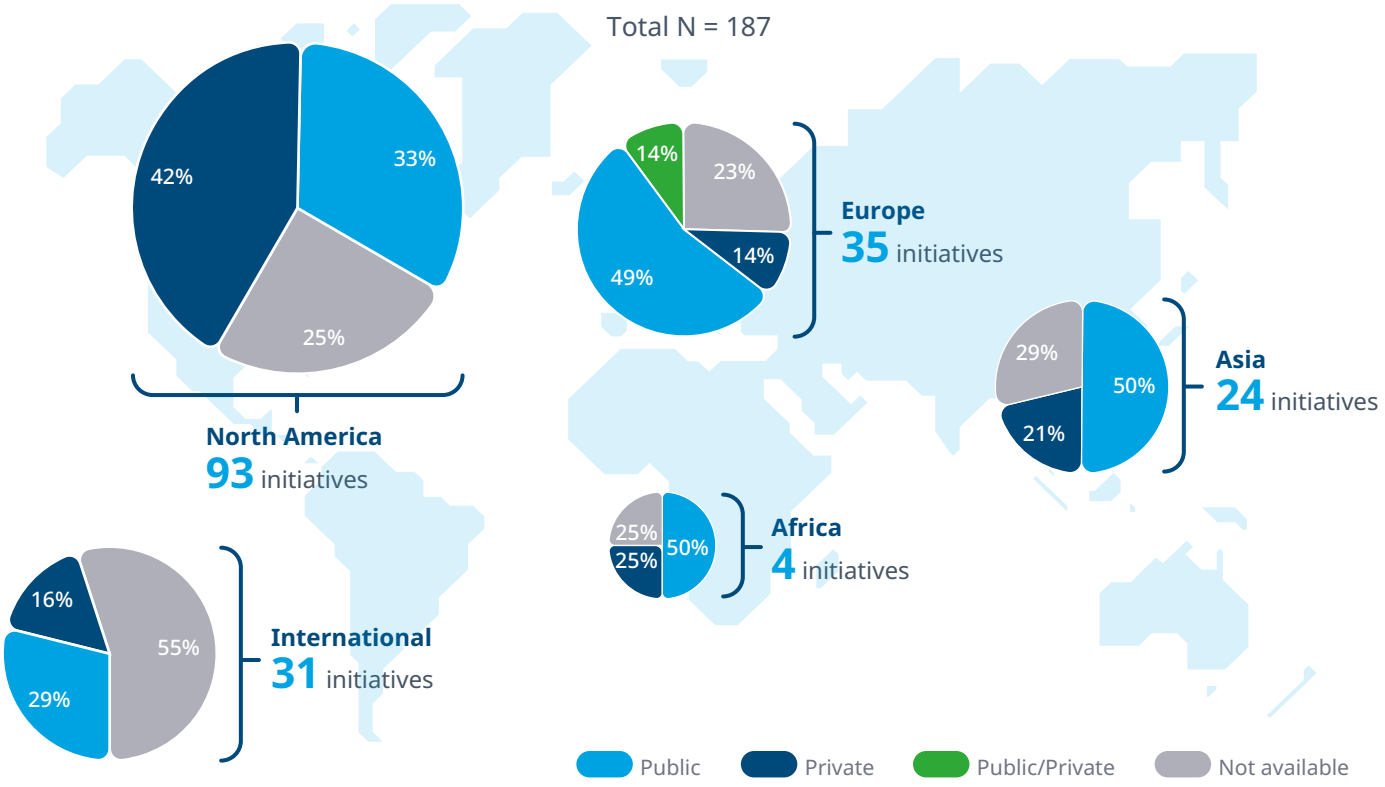
An analysis of the IQVIA Genomic Initiatives Database identifies 187 genomic initiatives for which publicly available information was available at or before the end of 2019. These initiatives are widely distributed globally by geographic origin (not the same as the geographic scope) with 50% of initiatives in the United States and 19% in Europe (see Exhibit 3). Examining ownership status by region, U.S. originated initiatives are more likely to be privately owned (43% of total) than in Europe or the rest of the world, where publicly funded initiatives dominate. Public initiatives make up 32% of all genomic initiatives in the United States versus 50% in Europe. However, governments and healthcare systems are investing in these initiatives in both regions and the United States still has nearly double the number of public initiatives (31 in the U.S. versus 17 in the EU). Privately owned companies with very large genomic databases, such as Ancestry.com and 23andMe, make headlines, but most genomic

Governments and healthcare systems are investing in these initiatives...

THE IQVIA GENOMIC INITIATIVES DATABASE

IQVIA has created a proprietary Genomic Initiatives Database, derived from publicly available information on initiatives that generate and aggregate data on the human genome. The genomic initiatives included in this dataset include private, consumer-oriented companies; non-nationally-based medical research-focused organizations both private and public; nationally based organizations which predominantly have overt medical research focus, and other organizations. To be included in the database the initiative must have wholly or partially generated genomic or genotype information themselves — i.e., are contributing previously un-genotyped/sequenced genomes, or, in the case of biobanks, possess biological material that can be genotyped/sequenced now or in the future. Initiatives that do not sequence or genotype themselves, but instead provide platforms for existing genomic data, have been excluded from the database as they generate no new data. Initiatives that offer a hybrid approach, having their own sequencing capability as well as offering a platform for further analysis of existing genomic data are included. For each initiative, 26 elements of information were captured, including details on the geographic origin of each database, scope, ownership status (public, private and unknown), purpose of the initiative (national, research institute driven, private company, other), type of genomic data collected in terms of population type currently covered or planned to be covered, breadth, linkage, consents and access, and other relevant data. In addition, the database includes publicly available information on the current number of individual genomes sequenced or biobank tissue samples collected and available to be sequenced. We also have collected data on the targets announced for individual genomic initiatives. Publicly available data on all characteristics were not available for all genomic initiatives. As the landscape of initiatives to generate and collect human genomic data evolves rapidly, this database will be updated. For additional information please see the Appendix.

Exhibit 3: Number of Initiatives by Type and Geography



Source: IQVIA Genomic Initiatives Database, Feb 2020

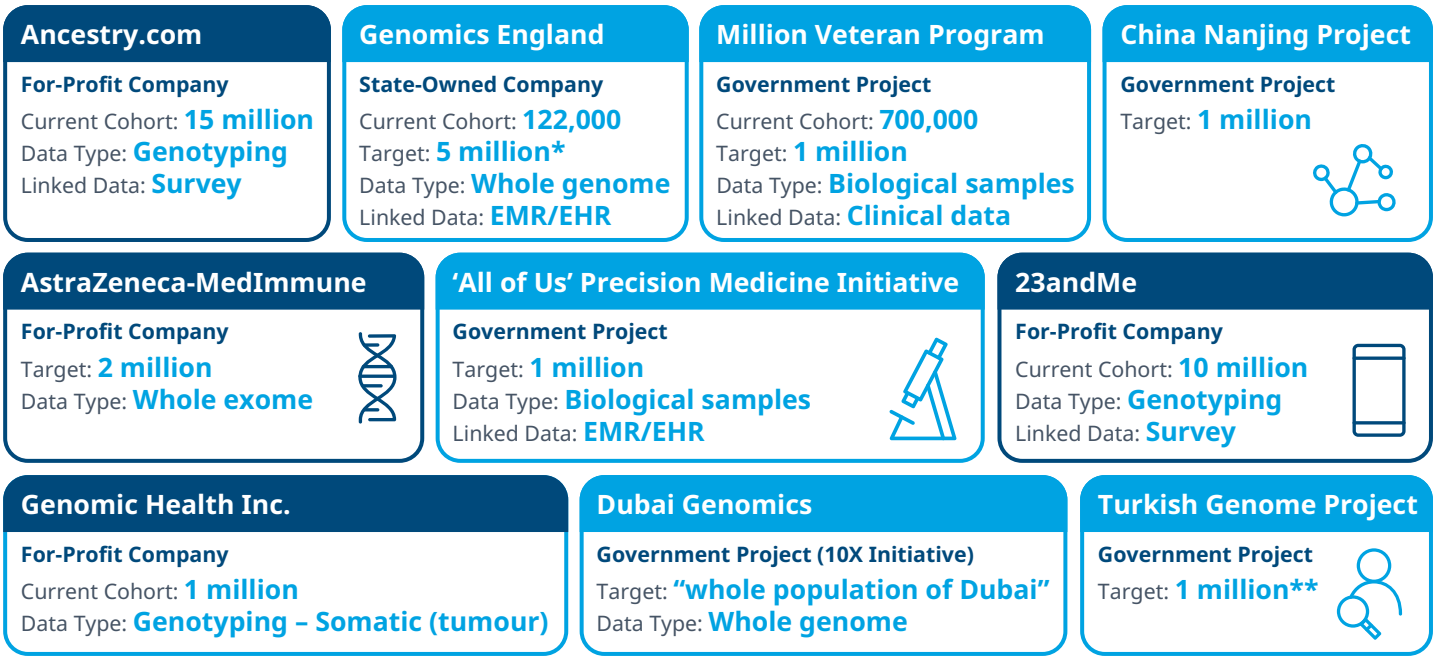
Notes: This map describes the source of initiatives in terms of region of origin and does not necessarily describe their scope. 'International' denotes initiatives which are either (1) international at conception or (2) those outside of N America, Europe, Asia, and Africa. Geography denotes the geographic starting point.

initiatives are publicly funded, such as the U.S. Million Veteran Program, or Genomics England. Governments and healthcare systems investing in these initiatives see genomic data as a route to better healthcare insight and delivery, and a strategic asset for the future.

Among publicly-driven initiatives, planned national genomic databases are proliferating as countries increasingly appreciate the potential technological and healthcare system benefits of building population genomic data at scale. These initiatives are a mixture of wholly public and public/private partnerships. Many have ambitious targets (see Exhibit 4) — for example, the Turkish Genome Project, when announced in 2017 stated that its aim was to sequence one million genomes (to an unspecified level of detail) by 2023, coinciding with the 100th anniversary of the founding of the modern state of Turkey. Dubai Genomics states that it will undertake whole genome sequencing of the entire population of Dubai, citizens and residents alike, which would imply a target of 2.8 million genomes (although that number is not explicitly given). The most ambitious sequencing target of all is from China, where a target of 100 million Chinese Genomes by 2030 was announced by the Chinese Precision Medicines initiative,[13,14] but as accomplishing this at current whole genome sequencing costs would be extremely expensive (i.e., such scope would require the cost of sequencing to fall) and little more specific information has been given, we have not yet recorded it as a target in our analysis. Broadly, publicly-driven initiatives have as their goals population health insights, medical breakthroughs and technology development.

Privately driven initiatives, of which 23andMe and Ancestry.com are flagships, have come from a more consumer-based angle, with their initial focus on

Exhibit 4: Examples of Large Genomics Initiatives Globally



Source: IQVIA Genomic Initiatives Database, Feb 2020

Notes: Public information; where data is not included, it is because it has not been reported publicly (e.g. no current cohort size). *Of which at least 500k will be whole genomes **planned for completion by 2023, the 100th anniversary of the founding of the modern Turkish state. No available information on current progress to this target found.

providing insights into heritage and ethnic origins. However, the focus of these companies is increasingly expanding to encompass more substantive elements of genome-based healthcare insight. 23andMe has sold FDA approved tests for certain conditions since 2017, although, even earlier, the company provided a range of health and disease information but were required to desist from providing health risk-based information in 2014. In late 2019, Ancestry.com announced it, too, would provide tests via a division called AncestryHealth, although in this case the test would be administered by a network of physicians rather than directly via the company itself.

UTILITY FOR EFFECTIVE MEDICAL RESEARCH

The initiatives in the global genomic landscape are very heterogenous in term of population covered, scope and quality of data collected, purpose of data collection and availability for research. To evaluate the genomic landscape and what progress has occurred to develop the medical utility of this data, IQVIA assessed initiatives against a number of criteria including their scope (number of genomes), breadth (the completeness of genomic data collected), linkage (to other healthcare-relevant data) and relevance (disease, population, and other specific criteria of coverage). Overall, databases vary widely by characteristics that matter for research utility.

Population inclusion and initiative location

Population types covered by genomic initiatives include self-selected (e.g., for consumer genetics databases like 23andMe, Ancestry.com, National Geographic), nationally-based (e.g., Dubai Genomics, which aims to sequence the genomes of all Dubai citizens and residents), specific condition (e.g., Undiagnosed Diseases Network, which collects the genomes of individuals with obviously apparent but undiagnosed conditions and aims to collect 8,000 genomes), and hybrids of the above (e.g., Genomics England, which is collecting the genomes of NHS England patients with a focus on those suffering from rare diseases or cancer).

Oncology databases (where both germline and somatic genomes could be collected) make of 53% of all datasets that focus on specific diseases while others focus on rare diseases (13%) and CNS disorders (10%) or hold data on multiple diseases (see Exhibit 5). Many of these are publicly funded databases. Other databases are typically agnostic to a patient's current health status. There's an interplay between the number of genomes collected and the disease area focus — for strongly genetically-determined conditions such as rare diseases or cancers, a relatively small, narrowly-targeted database could be medically very valuable if other data quality criteria are met, but for conditions with a genetic component but an equally important environmental component — for complex diseases and for those with polygenic etiology — much larger databases are required to generate insight, especially if the full scope and impact of the genetic component is not fully understood.

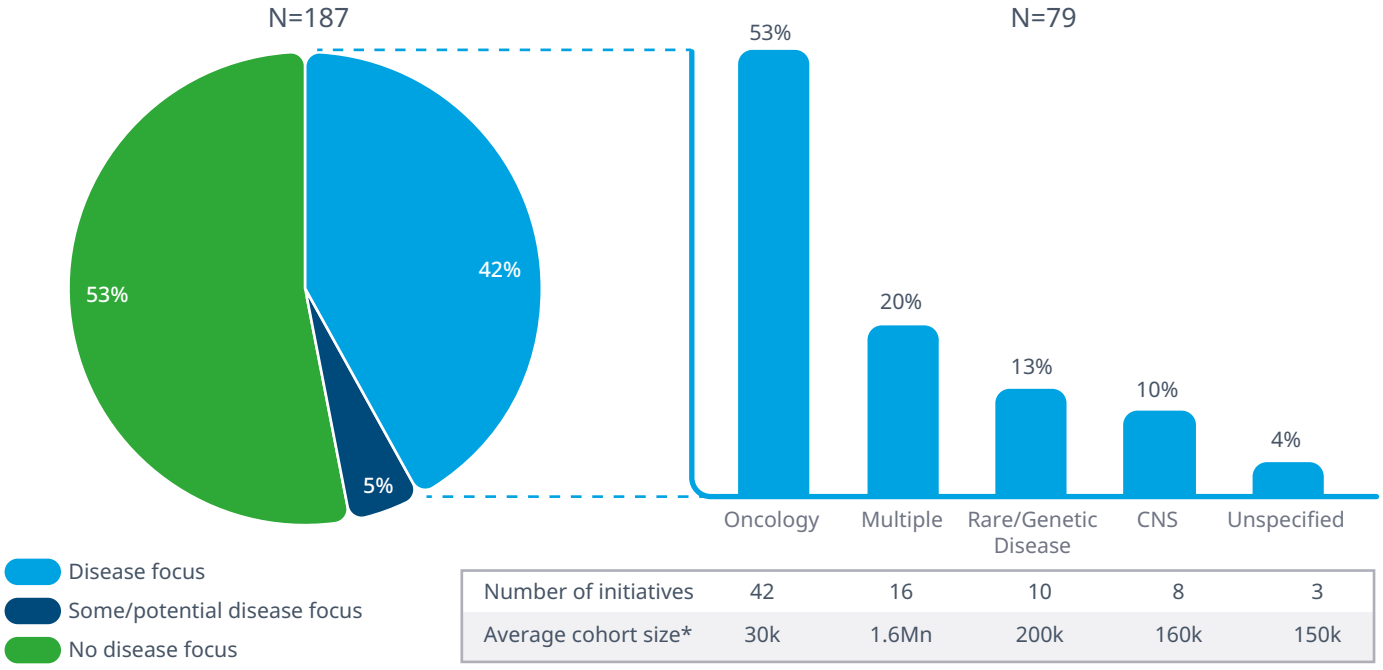
For strongly genetically-determined conditions such as rare diseases or cancers, a relatively small, narrowly-targeted database could be medically very valuable...

Where genomic databases are located can either match the origin of the population for which they hold genomic data or differ from it. A private company such as 23andMe sequences the genomes of individuals from around the world, but it is located in the United States. Publicly funded initiatives such as the U.S. Million Veteran Program (MVP) or Genomics England collect data on a country's population, although that population can, of course, be highly diverse in origin.

Breadth of genomic data

The amount of information collected on an individual human's genome varies hugely between genomic initiatives. A given human's diploid genome is approximately six billion base pairs in size, although the

Exhibit 5: Number of Genomics Initiatives by Disease Area Focus



Number of initiatives	42	16	10	8	3
Average cohort size*	30k	1.6Mn	200k	160k	150k

Source: IQVIA Genomic Initiatives Database, Feb 2020
Notes: Average cohort size based on target size first, current size if target unavailable, rounded to nearest 10,000.

representative sequence required (as chromosomes are present in most human cells in duplicate) for full genome sequencing is half that — approximately three billion base pairs. Only **whole genome sequencing** covers the full genome, or approximately three billion base pairs, but the known protein coding regions of the genome — the exome — make up a fraction of this or about 1.5% of these three billion base pairs (see Exhibit 6). **Whole exome sequencing**, focuses only on these segments of the genome that are directly expressed, substantially reduces the sequencing required, but it may lose the possibility of insight into the "unknown unknowns" — the apparently dark segments of the human genome which more sophisticated or wide-ranging analysis may subsequently find to be relevant to human diseases. It may also lose a lot of the non-coding but regulatory sequences which turn genes on and off in response to environmental or developmental stimuli, losing, in the process, the sensitivity to understand subtle contributions of genetic variants to human disease. Finally, **genotyping** sequences only very small fragments of the human genome, typically significantly less than 1% of an individual's genome. The commonly used **Single Nucleotide Polymorphism (SNP)** genotyping approach focuses on identified points of variation between human genomes, of which there are typically 4–5 million for each individual and over 100 million identified to date in total,[15] although most are in non-coding parts of the genome which have unknown impact on structure or function. SNP chips typically test for between 2–3 million SNPs on an individual's genome and offer a simple yes/no response for presence of the variant.

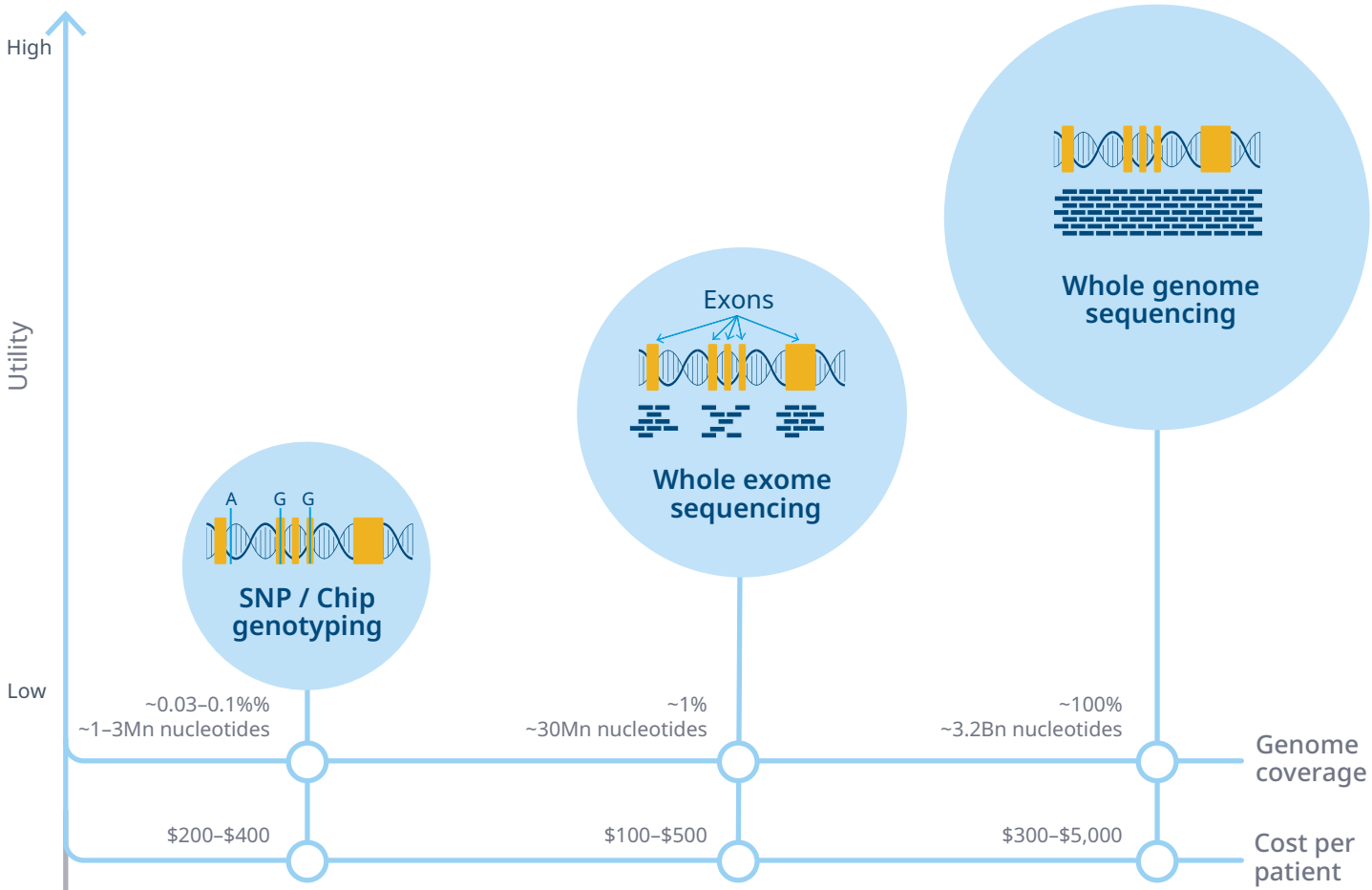
THE IMPACT OF WHOLE GENOME SEQUENCING: IMPROVING POPULATION HEALTH AND PATIENT DIAGNOSIS

In the past, early molecular tests returned only limited information about disease. Now sequencing costs have fallen to the point where wider-scale testing and routine clinical use of whole genome sequencing is feasible. Whole genome sequencing can be used to provide much deeper insights than other sequencing approaches that cover a tiny minority of an individual's genomic data, filling in multiple diagnostic blind spots, including improved information on complex structural variants, fusion genes, telomere length and mitochondrial variants.

This opens new doors to improve the health of populations. For instance, the U.K. now plans to offer whole genome sequencing to all newborns as a routine measure starting in 2020 through an initiative between the Genomics England Initiative and the recently created NHS England Genomic Medicine Service.[16] It is expected to lead to an expansion of the recognized population with various conditions and eligible for treatment, as well as identify patients likely to receive, or not receive, benefit from specific medicines.

Leveraging whole genome sequencing to identify disease-linked variants can lead to expanded recognition of risk factors tied to disease that will benefit the populations and affected individuals screened. As an example, when whole genome sequencing was applied to genetic samples of 33 individuals with putative Inherited Retinal Disease (IRD) who had not obtained diagnosis from targeted next generation panel sequencing who had not been confirmed via targeted next generation panel sequencing[17] (the standard clinical process) they were able to identify 14 clinically-relevant genetic variants tied to the disease, including large deletions and variants in non-protein coding regions of the genome, and confirm a molecular diagnosis of IRD in an additional 11 individuals. This enriched information has implications for drug development. Accounting for population structure and weighting, it was calculated that whole genome sequencing methods could increase diagnosis of IRD by 29%. By increasing the potential diagnosed population, it might increase the number of patients eligible for treatment by pipeline IRD drugs or expand enrollment criteria for clinical trials, and therefrom might spur manufacturers to invest in this area. As whole genome sequencing becomes a routine element of clinical care and clinical research, its power to build insight on and redefine markets will grow significantly.

Exhibit 6: Utility of Data on Genetic Mutations Detected by Chip Genotyping or Sequencing



SNP / Chip genotyping
Looks for known highly-specific short gene sequences to identify mutations/variations of interest

Utility

- Provides partial understanding of patient risk tied only to known affected regions
- Usable for initial drug development work and research

Limitations

- Only confirms if the specific variant(s) tested are present
- Provides limited understanding of molecular pathways of disease and variations on protein structure and function
- Poorer for identifying persons with specific variations within population e.g., trial recruitment
- Poor specificity for treatment decisions in regions associated with disease
- Not particularly useful to identify or study rare diseases tied to novel mutations

Whole exome sequencing
Only sequences every known region that directly makes proteins

Utility

- Researchers can understand variant links to disease causality and risk, molecular mechanism, coded protein structures and function
- Usable for identifying persons with specific variants, such as for trial recruitment, so long as the variant exists within the exome
- Can guide clinical decisions, elements of clinical trials such as design and recruitment, and drug development
- Useful to identify or study rare diseases tied to novel mutations

Limitations

- Not valuable if variants for study or influencing disease exist outside exome
- Non-coding and regulatory regions of the genome are not sequenced that may be subsequently found relevant to disease and are therefore, unable to be studied

Whole genome sequencing
Sequences the entire genome

Utility

- Researchers can understand variant links to disease causality and risk, molecular mechanism, coded protein structures and function, as well as the influence of non-coding regulatory regions, promoter/ repressor/ enhancer actions and binding
- Best for identifying persons with variants, such as for trial recruitment
- Best for guiding clinical decisions, elements of clinical trials such as design and recruitment, and drug development
- Best for understanding rare diseases with novel mutations

Limitations

- Data is very large and is the most expensive

Source: IQVIA Institute, Mar 2020
Notes: Genotyping includes snip chip, snip assay and chip assay.

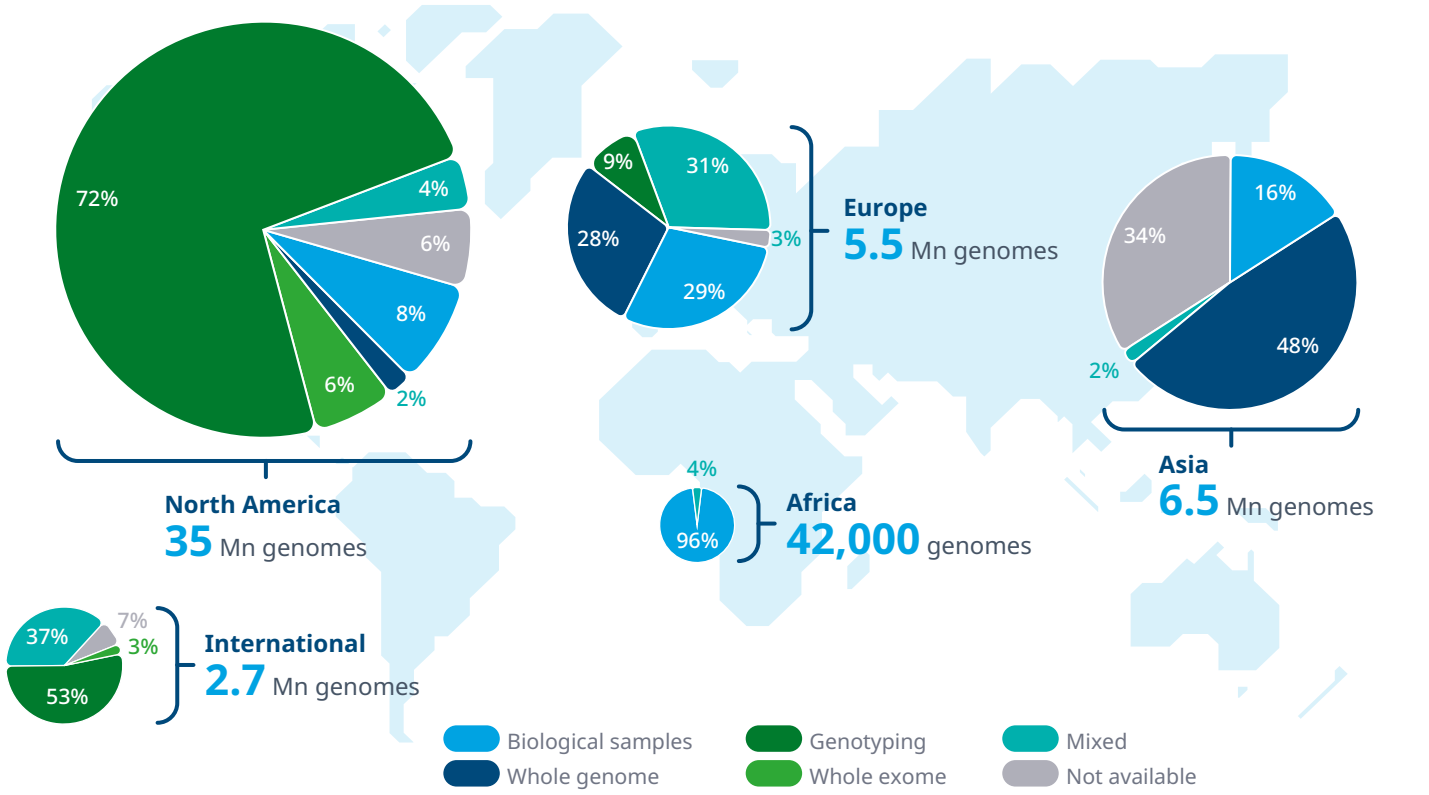
There is a trade-off between the level of information collected on an individual genome and the number of genomes collected. In the period 2000–2016, the cost of whole genome sequencing was inhibitory to widespread use, and therefore, mass market consumer genetic testing focused on relatively inexpensive genotyping, as did research organizations. Since 2016, whole genome sequencing costs moved towards and then proceeded below $1,000 per genome, thus enabling increasingly greater breadth of genomic data to be included in genomic databases.

Databases collecting more detailed genomic data — at exome or whole genome level — currently tend to be smaller in terms of the number of individuals sequenced. The largest are on the order of hundreds of thousands of genomes, but many major initiatives have targets in the millions. U.S. originated initiatives tend to be the largest but at lower breadth, for instance with limited gene variant data (e.g., SNPs), while European initiatives are currently smaller but more likely to collect whole exome or whole genome sequences.

Scope of future targets

The IQVIA Genomic Initiatives Database captures both the current number of genomes collected, genotyped or sequenced by each initiative, as well as the targets announced for future numbers of genomes to be collected. As previously noted, while targets are important to understand to assess the rate of growth of the genomic landscape, some targets are firmer than others, with near-term targets to 2025 more likely to be fulfilled. If they are, the current total of some 38 million genomes genotyped or sequenced to some breadth would grow to 52 million by 2025 (see Exhibit 7).

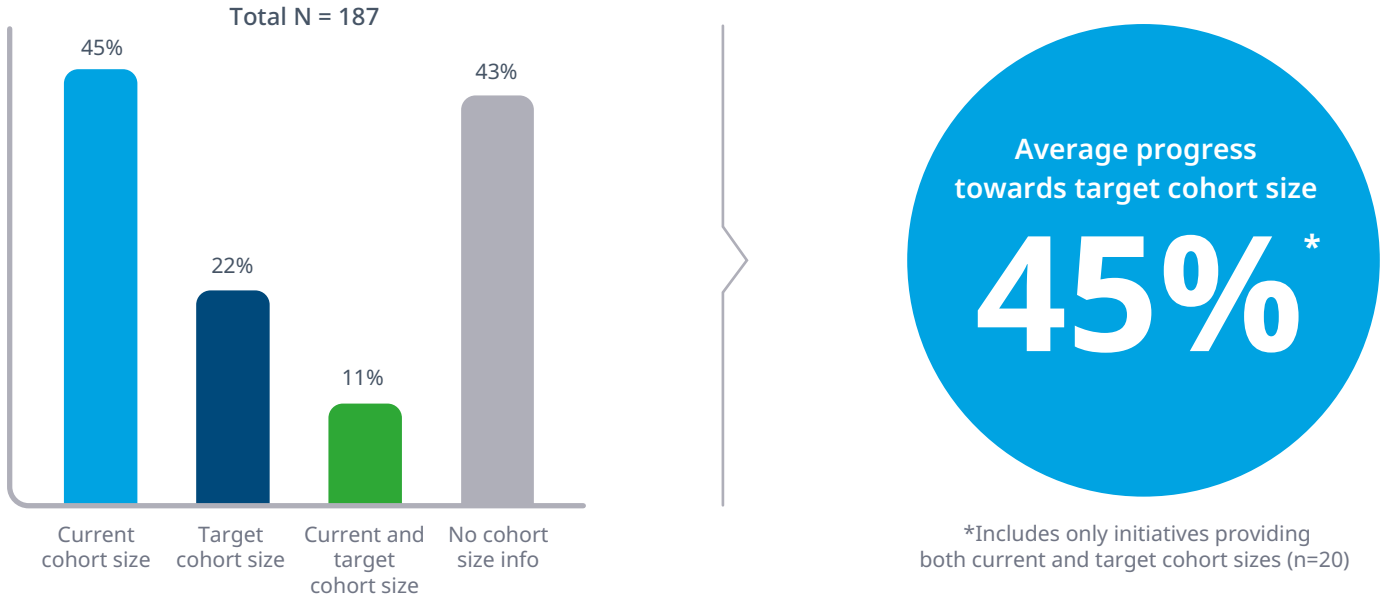
Exhibit 7: Initiatives by their Breadth and 2025 Target Cohort Sizes by Geography



Source: IQVIA Genomic Initiatives Database, Feb 2020

Notes: 'International' denotes initiatives which are international at conception plus one Australian initiative. 'Asia' includes middle east. Where initiatives do not have target cohort sizes, current cohort sizes are used. Only publicly available information is included. This map describes the geographic origin of initiatives sequencing genomes and not where the genomes come from.

Exhibit 8: Percentage of Initiatives Reporting Cohort Size Information and Percentage of targets collected



Source: IQVIA Genomic Initiatives Database, Feb 2020

This would not, however, include growth from the large private consumer genotyping databases, such as 23andMe, where no targets are announced. If these databases continue to grow rapidly, the total number of human genomes genotyped or sequenced at some level and held in a database can expect to be substantially higher than 50 million by 2025. This is still less than 1% of the estimated 8.1 billion world population by 2025.

There is significant geographic imbalance in the quantity and breadth of genomic data forecast to be collected. North American initiatives dominate by target cohort sizes, but the breadth of genomic data is lower (see Exhibit 7). North America, driven by the United States and by the private consumer genetics companies, is by far the largest source of data overall with 35 million genomes by 2025 and of genotyped data (25.1 million genomes), although we cannot break down the actual geographic origin of genotypes generated by 23andMe, which could come from many countries, and were therefore assigned to the U.S. by company location. Europe and Asia will likely have smaller databases by 2025, but a much higher proportion of the genetic data collected will be either whole genome (28%), or biological samples (29%) that could be fully sequenced either now or in the future. This could increase the value of these genomic databases for medical research. If the genomic information initially collected was limited — for example, involved genotyping rather than whole genome sequencing — more extensive genomic data may be subsequently required for future analyses. Biobanks, if they preserve biological material correctly, allow additional sequencing of samples more easily than re-calling individuals for further sample collection.

Europe is expected to fully sequence 1.5 million whole genomes by 2025, more than the 780,000 in North America, but an additional 2.2 million whole exomes will also be sequenced in the latter which hold high medical value. Over half of genomics initiatives provide some level of cohort size information but most of those focus on current numbers, making it challenging to determine how far along these initiatives are versus their 2025 targets. For the 20 initiatives that do provide both current and target numbers, these show average progress of 45% towards those targets (see Exhibit 8).

Linkage to phenotypic and longitudinal information

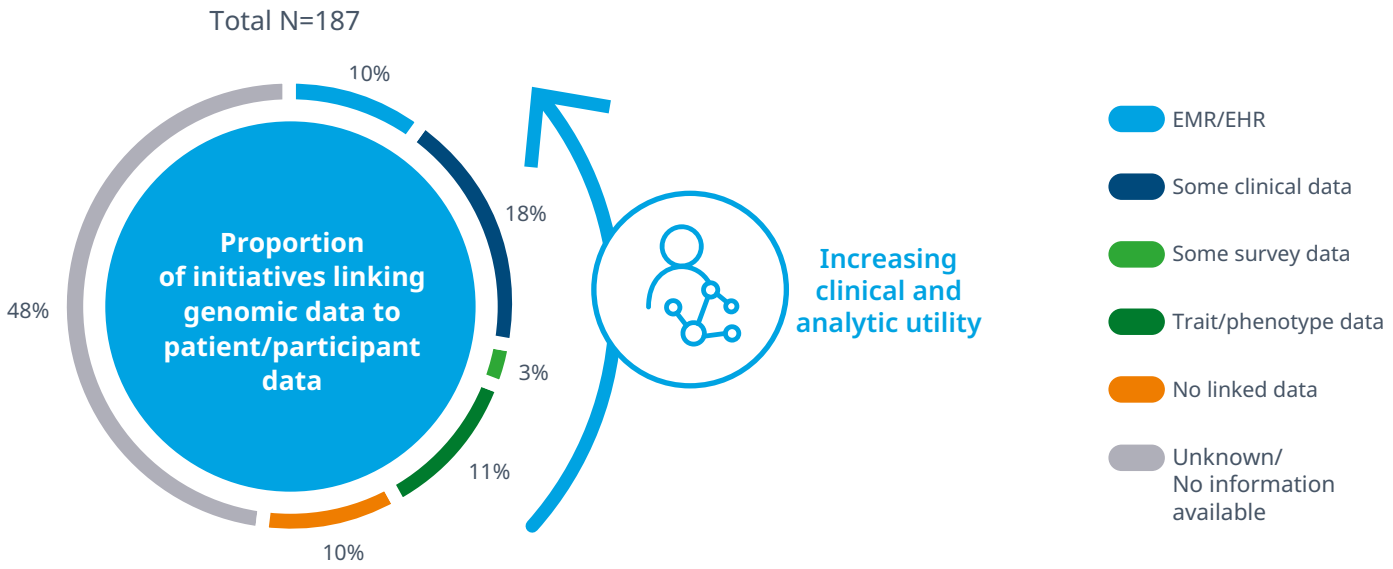
If context is lacking for any of the genomic information collected — for example, an individual's demographics, diagnosed health conditions, treatments, or treatment outcomes — even whole genome sequences may hold limited value for human data science or medical research. This is because a patient's health status or response to care, for instance from cancer treatment, would then fail to be understood as influenced by genotype differences. If such contextual data is captured only at a single point-in-time rather than tracked over time for an individual, the value may also be limited, because phenotypes can evolve over time for, for example in children with rare diseases.

Genomic initiatives vary significantly in the degree of linkage to phenotypic information. Demographic data and questionnaires collected, electively, at a single point in time without the possibility of update, are of least value, while data that is consented and appropriately-protected linkage to full medical records, with the possibility of longitudinal follow-up, are of greatest value. Consumer genetic databases have more limited contextual value than national, or research institution driven initiatives with consented linkage to appropriately protected medical records. Only 42% of databases in the IQVIA Genomic Initiatives Database – a minority – publicly state there is at least some level of linkage to data with clinical or analytic utility on patient demographics or health, and among these only 28% of initiatives tie to the most valuable EMR/EHR and clinical data (see Exhibit 9).

Access and consent

For genomic databases to be appropriate for medical research, they must have measures limiting access to the collected data and must start with the properly-informed consent of individuals donating their genetic information. The access policies of current genomic initiatives are quite variable, some being essentially "closed databases" open only to limited, often private, subscribers. Others allow access at various levels to a wider group of organizations, public and private, and in some cases apply differential approaches to for-profit and non-profit organizations. The question of consent and access becomes more complex still when the pooling of data is considered across different genomic databases.

Exhibit 9: Linkage of Genomic Data to Patient Data Within Initiatives



Source: IQVIA Genomic Initiatives Database, Feb 2020

Most of the databases covered in the IQVIA Genomic Initiatives Database not provide public information on their consent policies. Those that do often seek broad consent, meaning individuals allow their data to be shared at the decision of the data collector, and for some databases, individual donors have the right to delete their data. Access to genomic data also varies. For instance, FinnGen, a Finnish genomic database, is funded by nine pharmaceutical companies which, in return, have access to the data. Publicly funded databases often allow access to academics and other research institutions with an approved research proposal.

Linkage to other databases

Quality databases also are made more valuable by linking them together, which is enabled through common standards and standardized tools. Aggregation or linkage across individual genomic databases to generate even larger genomic datasets could greatly enhance the potential for effective human data science and medical research. In 2018, the European Union announced that 13 European countries would cooperate in linking genomic databases including Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Finland, Italy, Lithuania, Luxembourg, Malta, Portugal, Slovenia, Spain, Sweden, and the U.K.[18] The rewards of such an effort are potentially immense: greater representation across relevant populations, and the possibility of more powerful analyses uncovering insight only possible from larger datasets. There are also significant challenges: creating governance mechanisms to determine the conditions for distributed access to genomic data across borders, developing of common technical specifications for the security and exchange of data (which of course may have different levels dependent on genomic breadth and associated phenotypic data), and facilitating the interoperability of different registries. In the future, common standards for governance, technical specifications and interoperability, deployed across the genomic landscape, would truly accelerate the power of genomic data for human health.

Integration into clinical care and research

As DNA sequencing technologies improve and the cost of genomic sequencing continues to decline, the explosion of data will need to be translated into clinically meaningful and useable insights — for example, insights that can be validated and fed into clinical decision support (CDS) systems for the purpose of improving treatment and outcomes for patients. Accomplishing this will require a significant data infrastructure for storing, analyzing, interpreting, accessing, and sharing the data generated.

Currently, most genomic initiatives will have developed their own custom protocols and systems for handling the data they generate. This lack of standardization can compromise the comparison of sequence findings from different laboratories, leading to difficulties in confirming results or identifying clinically relevant results. Better consistency is required in terms of which reference genome assembly is used to align the sequence data, which statistical thresholds are used to identify variants, which genomic coordinates are used to define a variant's position, as well as which gene and variant naming conventions are used. Additionally, for genomic data to inform clinical decisions, it must be integrated into existing clinical systems such as electronic health records, and workflows, which are not always designed to support this.

Nevertheless, there are examples of efforts towards standardization. One example where large-scale standardization has been implemented is NHS England's Genomic Medicines Services, supported by Genomics England. This effort required the creation of a generalized bioinformatics solution to store and interpret genomic data, which was fully integrated into the health service at a national level. Genomics England also developed a publicly-available knowledgebase in which virtual gene panels related to human disorders can be created, stored and queried. This platform, called PanelApp, includes a crowdsourcing tool through which these panels are reviewed by experts worldwide, and is being used to achieve consensus on which genes have sufficient evidence for disease association.

Another key example of broader international standardization is the work of the Global Alliance for Genomics & Health (GA4GH) whose mission it is "to create frameworks and standards to enable the responsible, voluntary, and secure sharing of genomic and health-related data". In this instance, the focus is on ensuring that data is shared securely and only in the ways it has been consented to by the participants.

THE GENOMIC DATA QUALITY SCORE

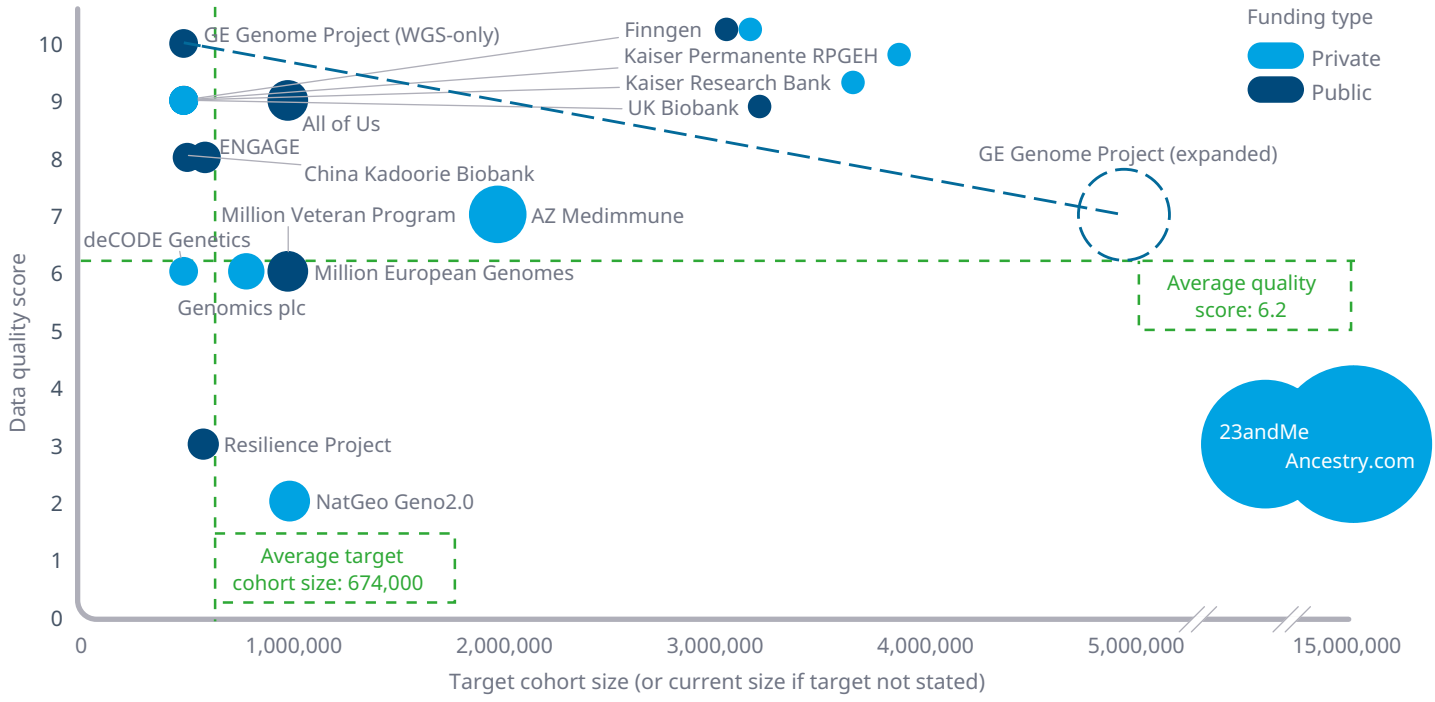
An overall quality score was derived for each initiative in the database for which information on both the type of data collected and the level of linkage to patient data were available. Each initiative was scored 1-5 based on type of data with whole genome sequencing being scored highest (5) and genotyping scoring lowest (1). Intermediate values of 4, 3, and 2 were assigned to biological samples, whole exome sequencing and mixed approaches that include genotyping, respectively. Initiatives were also scored on the level of linkage to patient data - EMR/EHR linkage was scored highest (5) and linkage to trait/phenotype data was scored lowest (1), with linkage to some clinical data scored as a 4 and some survey data scored as a 2. To emphasize the large difference in usefulness between clinical data and survey data, no item was scored as a 3 and initiatives that had no linkage to any patient data were given a score of 0. The combined score is a sum of these two individual scores. As such, the maximum score achievable is 10 (whole genome sequencing linked to EMR/EHR data) and the minimum score is one (genotyping only with no linkage to patient data).

ASSESSMENT OF DATA QUALITY

The value of today's evolving genomic landscape can be assessed using a combined "data quality score" into which the breadth of genomic data, its linkage to other medically relevant data, and consent and access are factored.

Analysis of initiatives data quality score versus the current or known 2025 target cohort size (see Exhibit 10) suggests that the next decade will see significantly increased growth in the number of large genomic databases with strong scores, meaning that their utility for human data science and medical research is high. The reduction in costs to undertake whole genome sequencing makes this possible, but the desire by healthcare systems, researchers, company and other stakeholders to ensure the full potential of genomic research is realized, with the most detailed and linked clinical data is also a major driver. The 2020s will be the age of mass full genome sequencing, and with it, a step up in the insights generated for human health. Genomic data initiatives are proliferating but size, data quality, and data accessibility are hugely variable (see Exhibit 10).

Exhibit 10: Key Genomic Data Initiatives by Target Cohort Size, Data Quality Score, and Funding Type



Source: IQVIA Genomic Initiatives Database, Feb 2020

Notes: Includes genomic data initiatives for which information on cohort size, type of data collected, and linkage to additional data is available and which have a minimum 500,000 target cohort size. Where target cohort size was unavailable, information of current cohort size was charted. GE Genome Project = Genomics England Genome Project.

What initiatives have done with this data

There are a wide range of actual and potential uses for genomic data. Genomic initiatives can help researchers understand disease causality and risk, molecular mechanisms of disease, how diseases originate and affect populations and enable earlier diagnosis and disease prevention, enable precision medicine, and help identify new uses for approved drugs based on a better understanding of disease processes. In the life sciences industry, key applications include drug discovery and repurposing, market redefinition and expansion, clinical trial design and post-launch real world evidence studies.

UNDERSTANDING DISEASE RISK THROUGH GENOMIC ASSOCIATION STUDIES: ACCELERATING DRUG DISCOVERY

Research techniques such as genomic association studies examine genetic links to disease and health. By examining data on linked clinical and functional characteristics alongside genetic data, such studies can provide evidence of associations and drive insight into human disease pathways that underpin drug discovery. They can also enable existing pharmacotherapies to be repurposed for new indications.

Use of genomic initiative data

COPD is the third leading cause of death globally.[19] The UK Biobank, a biobank of over 500,000 individuals, provided genomic data on 48,943 patients as part of a powerful Genome-Wide Association Study (GWAS) called UK BiLEVE[20,21] that examined genetic links to lung function, and COPD susceptibility. This database, as the largest source of both clinical spirometry data and DNA[17], was used to examine 27,624,732 variants and their disease-relevant associations with lung function (forced expiratory volume in one second, FEV_1) as well as COPD exacerbations in a European population. For data on a non-European population, the China Kadoorie Biobank cohort (CKB) was also examined. Through this process, 81 independent genetic variants associated with lung function were identified. The second stage of the analysis involved testing of a further 95,375 more individuals' data taken from the UK Biobank, the SpiroMeta consortium and UK Households Longitudinal Study (UKHLS) in another GWAS to confirm the effect of these variants on COPD susceptibility.

Impact

Notably, this genome-wide association study identified 43 novel independent relevant genetic variants for lung function, almost doubling the number of known genomic signals for lung function to a total of 97. It thereby expanded understanding of molecular pathways that tie to patient risk of COPD and showed that individuals have nearly a fourfold difference in their COPD risk.
The variants, which may have causal effects on lung function, were found to tie to 234 genes playing a role in multiple molecular pathways including elastic fiber pathways, Hedgehog mediated signaling, epigenetic regulation pathways, and the inositol phosphate metabolism pathway.

This data now provides a rich resource for researchers and drug developers to search for gene-encoded proteins tied to lung function and COPD that could become targets for drug development, or identify which drugs could be repurposed to treat COPD or other respiratory conditions among the arsenal of currently approved drugs for other clinical indications. As currently approved drugs and those in development target the protein products of only seven of the 234 genes, the additional targets identified can now be tested to see if they are druggable. This data stems from a collaboration between the UK Biobank and leading life sciences companies led by Regeneron to fund and accelerate exome sequencing of half a million participants.[22] As the routine application of whole genome sequencing in clinical care grows, so too does the potential for genomics to underpin drug discovery.

ADVANCING PERSONALIZED MEDICINE THROUGH INNOVATIVE CLINICAL STUDY DESIGN

Genomic data can be used to guide clinical trial design and in study planning, such as by assessing a study's feasibility based on the number of affected patients. It can also support the development of biomarkers by validating them and building stronger evidence of their utility, as well as be used to identify patients likely to respond to a medicine once it is approved. Studies that include in-depth genomic profiling can help researchers understand what might be driving differences in patient response to a drug or intervention, and once variants have been linked to patient response, genomic data can then be used to identify relevant sub-populations for inclusion in future clinical trials, such as those most likely to benefit from therapies, or those with fast-progressing disease where effects of the drug are likely to be seen clearly and quickly. By first targeting clinical development programs to rapidly progressing patients, and then broadening to other groups of patients once initial safety and efficacy signals are seen, or approvals achieved, this can speed drug development. Retrospective analysis of the genomic ties to trial outcomes, may also support the revival of previously failed clinical trials by providing context and identifying sub-populations of responders. Once a drug is approved, genomic data can be used to develop robust clinical assays to identify patients with disease or those likely to benefit from therapy.

Use of genomic initiative data

As an increasing number of oncogenic mutations were discovered in lung tumors, it became increasing possible to stratify lung cancer into genetically defined subsets. The Lung-MAP Trial, launched in June 2014, is an initiative to accelerate personalized treatment for advanced non-small cell lung cancer, through the first multi-arm or 'umbrella cancer trial governed by a master protocol.[23] It provides an example of how genomic insights can inform clinical trial design for lung cancer and expand access to precision medicine, while building a genomic dataset that can then be used in the future to improve lung cancer care.[24] With Lung-MAP — a public-private partnership of U.S. cancer research organizations, such as the National Cancer Institute, Friends of Cancer Research, and pharmaceutical companies, including Pfizer, Genentech/Roche, AstraZeneca, Amgen, Bristol-Myers Squibb and Foundation Medicine — patients were screened for specific genetic changes in their tumor (i.e. biomarkers) using next-generation sequencing, and those with biomarkers targeted by one of the trial treatments were enrolled in the appropriate sub-study testing the relevant treatment. Patients whose tumor did not match a biomarker-specific trial drug, were able to enroll in a non-match sub-study where immunotherapies or other drugs expected to have activity across multiple molecular subtypes could be used, or treatment combinations could be used for immunotherapy-resistant patients.[23]

Impact

With Lung-MAP, the amount of genomic screening data available on lung cancer patients has significantly increased, with more than 2,800 patients screened through February 2020, thereby advancing both personalized medicine and research.[23] Data from these patients will further scientific understanding of common NSCLC mutations, help discover future treatment targets, serve as a resource for biomarker assay development, and clarify the frequency and geographical distribution of biomarkers. Most critically, the Lung-MAP trial established a pathway to accelerate drug development to benefit patients. The trial, now open to nearly all patients with NSCLC, helped connect hundreds of patients to personalized biomarker-targeted drugs and immunotherapies treatments through testing, with 799 patients paired to a trial therapy, and expanded access to the trial to many clinical sites across the country.[17] This trial structure may have also benefitted patients with rare molecular biomarkers who had difficulties accessing precision trials due to a lack of comprehensive genomic screening or geographic distance from enrolling sites. It likely also helped trial sponsors in the oncology space who have seen trial complexity increase over time[25] and, along with investigators, have found it difficult to start trials and recruit patients.

Finally, this new trial design also demonstrated benefits in accelerating the screening of large patient populations through NCI's National Clinical Trials Network.[23] and new sub-studies were able to be rapidly started under the master protocol to test promising new treatments.[23] This was enabled by the uniform genomic screening within the trial protocol. Over five sub-studies were opened through February 2020 and eight drug-centered sub-studies testing 12 novel therapies were completed since activating in June 2014. Seven others are planned tied to pharmaceutical partners.

ACCELERATING CLINICAL DEVELOPMENT BY UNDERSTANDING UNDERLYING GENETICS OF DISEASE PROGRESSION

Genomic data has the power to define and identify new sub-groups within populations. For instance, when included in an analysis of patients treated with a medicine, genetic data can show that some subsets of patients harbor specific gene variations that might contribute to differential outcomes. By providing context to clinical information, genomic data can add value to real world and randomized studies. Specifically, adding genomic data to real world studies enables researchers to,

- Understand the epidemiology and distribution of genetically-defined patient groups
- Assess the burden of illness and natural history of disease
- Examine the treatment patterns and outcomes of specific genetically-defined patient groups
- Analyze time to treatment for selected populations
- Observe real world outcomes and improving clinical pathway efficiency for patients with known genetic features
- Identify a target group of patients for deeper chart review.

Use of genomic initiative data

As an example of how genomic data can be used in real world studies, de-identified data from Genomics England that links genetic data from sequencing to clinical parameters was used in an IQVIA project. Data scientists developed an algorithm to identify patients with a particular neurological condition whose disease severity progressed rapidly. This was done by examining health changes noted in the de-identified clinical record longitudinally over time. By comparing the genetic characteristics of both slowly and rapidly-progressing patients, specific genetic variants that correlated with the speed of progression among patients were able to be identified.

Impact

This study not only offers a greater understanding drivers of disease progression, but also offers to identify distinct subtypes of a single neurologic disease or segment out distinct patient subpopulations experiencing that neurologic disease. Once genetic differences are identified, they can be assessed and validated to see whether those biomarkers are predictive of which patients will progress rapidly — i.e., as a predictive biomarker — and hence may benefit the most from treatment. The ability to focus on such rapidly progressing patients using genetic data offers to accelerate the future clinical development programs, influence trial design, as well as get targeted therapies to patients who have the greatest need and are likely to obtain the greatest benefit. Subsequently it may also benefit the full breadth of the affected population by speeding drugs to market and building initial data on safety and efficacy in that initial subset.

The future of genomic data

Currently announced future targets for genomic sequencing by 2025 total 50–100 million individuals globally,[a] and the figure is likely to rise rapidly. However, this number hides the fact that the detail, utility and accessibility of that data varies widely. The data is also currently very focused on populations in the United States and Europe — and therefore, does not reflect the genomics of global populations well, although it unsurprisingly does reflect the genomics of the populations of the countries comprising most of innovative prescription medicines spend. Asia is 4.6 billion out of the world's population of 7.7 billion people or 60%, and Asian originated genomic databases currently target only six million genomes to be sequenced, exclusive of the ambitious but long term Chinese target of 100 million genomes by 2030. It's clear that there will be room for growth in genomic databases for the foreseeable future.

Our global landscape of genomic initiatives shows that enormous research promise is being driven by a continued drop in the costs of genomic sequencing, and the rising interest of countries and whole healthcare systems in building genomic databases, meaning the fully linked, consented and accessible databases which will be of greatest value for medical research are becoming increasingly possible at scale.

It should be noted that even with the realization of the promise of more genomic data than ever before challenges will remain. For cancer, challenges in the nature of samples collected from tumors must be overcome – most tumor samples are collected as formalin fixed, paraffin embedded (FFPE) samples, as opposed to fresh frozen tissue. Such FFPE samples are optimized for histology not for whole genome sequencing. DNA extracted from these samples can vary widely in quality due to age, fixation conditions (including length of exposure to formalin), DNA-protein crosslinking, and inhibitors, which may impact downstream genomic analyses. Many adult cancers have a relatively limited range of known or reasonably potential driving/contributing variants, meaning that somatic panel tests may predominate in oncology for the foreseeable future. Therefore, whilst cancer has been a driver for much initial genomic activity, other disease areas may drive it more in the future.

Organizations, public and private, which seek to use genomic data should maintain a global and constantly updated view of genomic initiatives development, encouraging investment in developing this data and also encouraging the development of databases to appropriate standards: in terms of quality and linkage of genomic data within each database, the right ethical governance and transparent, effective consent procedures, and education for individuals and healthcare professionals involved in genomic data collection. They should also push for the development of interoperability standards and agreements to enable future linkage across databases to be designed into them as they are created, so that the power of very large, high quality genomic databases can be accelerated during the 2020s and increase the power of genomic data to improve human health.

There's clearly still a very long way to go — even the most ambitious targets result in a percentage of the global population sequenced at some level by 2030 that is less than 5% and probably less than 2%. Unless there are dramatic increases in the sequencing of the populations of Asia, Africa and South America, genomic databases will continue to under-represent these populations and their ethnicities, while North American and European Caucasian populations will continue to be over-represented. As rare diseases and cancers are strongly driven by genetic make-up, data in these areas matter for global health. The organizations that seek to use genomic data to research human health and develop therapies to treat and prevent human disease should play a role in advocating more equitable coverage of global population genomes.

a. Depending whether Chinese targets of 100 million genomes by 2030 are included, as this ambition is heavily caveated

Appendix

METHDOLOGY

The IQVIA Genomic Initiatives Database

Building a systematic database of initiatives which generate and accumulate human genomic data poses challenges. The number of initiatives is now large, and widely distributed globally. The IQVIA Genomic Initiatives Database was built using publicly available information, and it is entirely possible that there are initiatives for which there is no publicly available data, or for which the publicly available data is not up to date. In such cases, there may be gaps in the database. Additionally, the landscape of initiatives that are generating and collecting human genomic data is evolving rapidly and is highly diverse, with private and public initiatives stretching across multiple countries, thereby posing additional challenges.

The genomic initiatives included in this dataset include private, consumer-oriented companies; non-nationally-based medical research-focused organizations both private and public; nationally based organizations which predominantly have overt medical research focus, and other organizations.

We have chosen to focus on initiatives which publicly available information indicates are generating genomic data, that is, contributing previously sequenced genomes, or, in the case of biobanks, tissue samples which can be sequenced now or in the future. There are also initiatives which do not sequence genomes themselves, but provide platforms existing genomic data, promising enhanced analysis or superior access — Promethease is an example of a company offering this service. Still other initiatives offer a hybrid approach, with their own sequencing capability as well as offering a platform for further analysis of existing genomic data. We've chosen to exclude initiatives which generate no new genomic data from our landscape. We cannot discount the possibility that there is some double counting of genomes in our database if genomic data generated by one initiative is also held by another. Our data was initially collected mid-2019 and then updated in early 2020. Currently, we have some degree of information on 187 initiatives, public and private, which meet our criteria of being initiatives with a database of genomic information which they have wholly or partially generated themselves. For each initiative, we sought to collect up to 26 pieces of information, covering the nature of the initiative, the type of genomic data collected in terms of population, breadth, linkage, consents and access, and other relevant data.

The IQVIA genomic initiatives database additionally categorizes initiatives by the nature of the population currently covered or planned to be covered, including publicly available information on both the current number of individual genomes sequenced at some level, or biobank tissue sample which could be sequenced that have been collected. We also have collected data on the targets announced for individual genomic initiatives. Some types of initiative, principally consumer genetic testing, do not announce targets. Others, often national level initiatives, have stated targets of great ambition, in the millions to hundreds of millions of genomes. Future targets are sometimes vague on the achievement pathway, promised far into the future, or heavily caveated with dependencies on future events (e.g., a dramatic fall in sequencing costs). Therefore, forecasts on the numbers of human genomes to be sequenced should be treated with caution.

Publicly available data on all characteristics was not available for all genomic initiatives. We expect that, as more public initiatives start to get underway, a greater depth of information will be available for more initiatives.

The Genomic Initiatives Database is, of course, a snapshot in time; news about existing initiatives, as well as new initiatives, appear sufficiently frequently that a six-monthly review may be advisable.

References

1. NIH. National Human Genome Research Institute. The cost of sequencing a human genome. 2019 Oct 30. Available from: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost
2. Veritas. Get the most comprehensive genetic testing services there is. 2019. Available from: https://www.veritasgenetics.com/myGenome
3. Nebula Genomics. 30x Whole genome sequencing for $299. Accessed April 29 2020. Available from: https://nebula.org/whole-genome-sequencing/
4. IQVIA estimates as of the beginning of 2020 based on publicly available information
5. WHO. Human genomics in global health. 2020. Available from: https://www.who.int/genomics/public/geneticdiseases/en/index2.html
6. Kaur G, Mehra N. Genetic determinants of HIV-1 infection and progression to AIDS: susceptibility to HIV infection. Tissue Antigens. 2009 Apr. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19317737
7. Marquet S. Overview of human genetic susceptibility to malaria: from parasitemia control to severe disease. Infection, Genetics and Evolution. 2018 Dec. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28579526
8. Gerstung M, Jolly C, Leshchiner I. Dentro SC, Gonzalez S, et al. The evolutionary history of 2,658 cancers. Nature 578, 122–128 (2020). Available from: https://www.nature.com/articles/s41586-019-1907-7
9. Donnelly L. All children to receive whole genome sequencing at birth, under ambitions laid out by Matt Hancock. The Telgraph. 2019 Nov 5. Available from: https://www.telegraph.co.uk/news/2019/11/05/children-receive-whole-genome-sequencing-birth-ambitions-laid/
10. CPIC. Genes-drugs. 2020 Mar 25. Available from: https://cpicpgx.org/genes-drugs/, https://www.pharmgkb.org/
11. Danish National Genome Center. 2019 Available from: https://eng.ngc.dk/
12. Dana Farber Cancer Institute. The Center for Cancer Precision Medicine. Our process. Available from: https://www.dana-farber.org/research/departments-centers-and-labs/integrative-research-centers/center-for-cancer-precision-medicine/our-process/
13. Gross A. China fast becoming top player in booming Asia genomics market. MedTech Intelligence. 2018 Jun 22. Available from: https://www.medtechintelligence.com/column/china-fast-becoming-top-player-in-booming-asia-genomics-market/
14. Dankar FK, Ptitsyn A, Dankar SK. The development of large-scale de-identified biomedical databases in the age of genomics-principles and challenges. Hum Genomics. 2018;12(1):19. Published 2018 Apr 10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5894154/

References

15. NIH. Genetics Home Reference. What are single nucleotide polymorphisms (SNPs)? 2020 Mar 31. Available from: https://ghr.nlm.nih.gov/primer/genomicresearch/snp

16. Meade N, Spink J. Let's grasp this opportunity to examine the potential future of screening. BioNews. 2019 Nov 11.

17. Ellingford JM, Barton S, Bhaskar S, Williams SG, Sergouniotis PI, et al. Whole genome sequencing increases molecular diagnostic yield compared with current diagnostic testing for inherited retinal disease. Ophthalmology. 2016 May;123(5):1143-50. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26872967

18. European Commission. EU countries will cooperate in linking genomic databases across borders. 2018 Apr 24. Available from: https://ec.europa.eu/digital-single-market/en/news/eu-countries-will-cooperate-linking-genomic-databases-across-borders

19. Lozano R, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380:2095–128.

20. Wain LV, Shrine N, Miller S, et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. Lancet Respir Med. 2015;3(10):769–781. doi:10.1016/S2213-2600(15)00283-0

21. Wain LV, Shrine N, Artigas MS, et al. Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. Nat Genet. 2017;49(3):416–425. doi:10.1038/ng.3787

22. UK Biobank. Regeneron announces major collaboration to exome sequence UK Biobank genetic data more quickly. 2018 Jan 8. Available from: https://www.ukbiobank.ac.uk/2018/01/regeneron-announces-major-collaboration-to-exome-sequence-uk-biobank-genetic-data-more-quickly/

23. Herbst RS, Bazhenova L, Neal J, Waqar SN. Lung-MAP: A five-year recap on the first master protocol trial in cancer research. The Cancer Letter. 2020 Feb 21. Available from: https://cancerletter.com/articles/20200221_1/

24. NIH. National Cancer Institute. Lung-MAP: Master protocol for lung cancer. 2019 Jan 31. Available from: https://www.cancer.gov/types/lung/research/lung-map

25. IQVIA Institute for Human Data Science. The changing landscape of research and development - innovation, drivers of change, and evolution of clinical trial productivity. 2019 Apr.

About the authors

SARAH RICKWOOD
VP, EMEA Marketing and Thought Leadership, IQVIA

Sarah Rickwood is Vice President, European Marketing and Thought Leadership in IQVIA, a team she has run for 10 years. She has 28 years' experience as a consultant to the pharmaceutical industry, having worked with most of the world's leading pharmaceutical companies on issues in the US, Europe, Japan, and leading emerging markets. Sarah presents to hundreds of pharmaceutical industry clients every year on a wide range of global pharmaceutical industry issues, and has published white papers on many topics, including uptake and impact of innovative medicines, and challenges for Launch Excellence, the relative strength and prognosis for the developed and the emerging pharmaceutical markets, the global uptake and impact of multichannel marketing, biosimilars market opportunity and the drivers of market establishment, orphan drugs launch challenges, cell and Gene therapies commercialization challenges and opportunity, the evolution of the global Biologics market and Africa pharmaceutical market opportunity and challenges. Sarah holds a degree in biochemistry from Oxford University.

ALEXANDRA SMITH
Consultant, EMEA Thought Leadership, IQVIA

Alexandra Smith is a consultant in the European Thought Leadership team at IQVIA since July 2015, working with IQVIA proprietary healthcare data to provide insights into market trends affecting the pharmaceutical industry as a whole. Alexandra has focused on specific areas such as the diabetes market and the rise of multichannel marketing where she has published several key whitepapers. Prior to joining IQVIA, she obtained a Masters in Biochemistry and a PhD in Developmental Biology from Oxford University.

MURRAY AITKEN
Executive Director, IQVIA Institute
for Human Data Science

Murray Aitken is Executive Director, IQVIA Institute for Human Data Science, which provides policy setters and decisionmakers in the global health sector with objective insights into healthcare dynamics. He led the IMS Institute for Healthcare Informatics, now the IQVIA Institute, since its inception in January 2011. Murray previously was Senior Vice President, Healthcare Insight, leading IMS Health's thought leadership initiatives worldwide. Before that, he served as Senior Vice President, Corporate Strategy, from 2004 to 2007. Murray joined IMS Health in 2001 with responsibility for developing the company's consulting and services businesses. Prior to IMS Health, Murray had a 14-year career with McKinsey & Company, where he was a leader in the Pharmaceutical and Medical Products practice from 1997 to 2001. Murray writes and speaks regularly on the challenges facing the healthcare industry. He is editor of Health IQ, a publication focused on the value of information in advancing evidence-based healthcare, and also serves on the editorial advisory board of Pharmaceutical Executive. Murray holds a Master of Commerce degree from the University of Auckland in New Zealand, and received an M.B.A. degree with distinction from Harvard University.

About the Institute

The IQVIA Institute for Human Data Science contributes to the advancement of human health globally through timely research, insightful analysis and scientific expertise applied to granular non-identified patient-level data.

Fulfilling an essential need within healthcare, the Institute delivers objective, relevant insights and research that accelerate understanding and innovation critical to sound decision making and improved human outcomes. With access to IQVIA's institutional knowledge, advanced analytics, technology and unparalleled data the Institute works in tandem with a broad set of healthcare stakeholders to drive a research agenda focused on Human Data Science including government agencies, academic institutions, the life sciences industry and payers.

Research Agenda

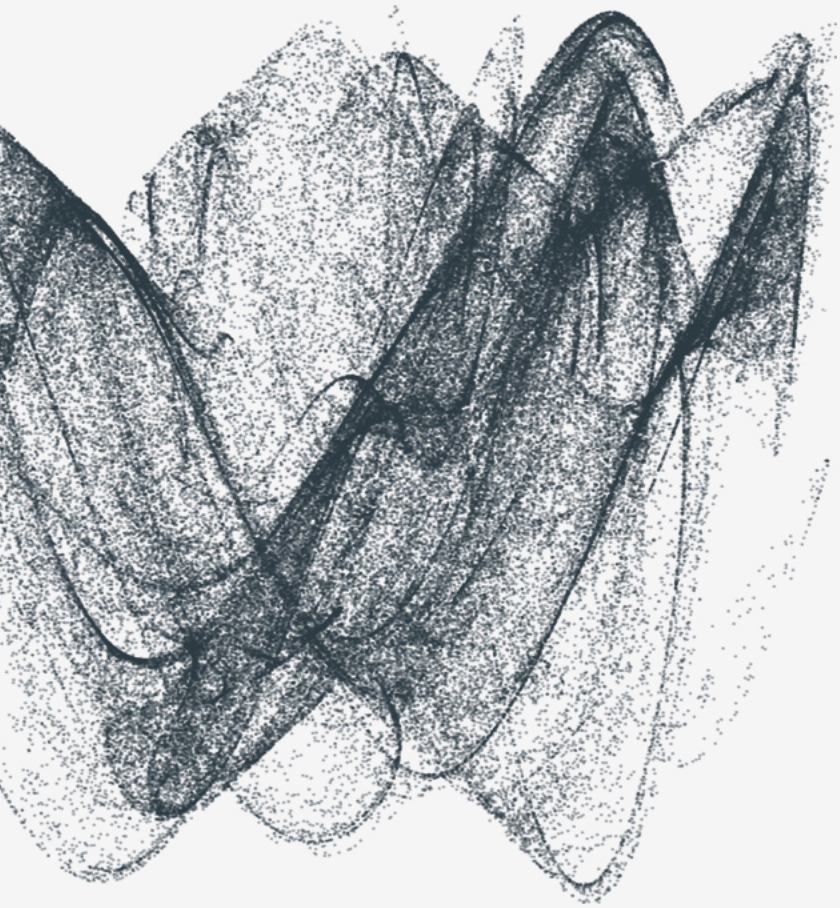
The research agenda for the Institute centers on 5 areas considered vital to contributing to the advancement of human health globally:

- Improving decision-making across health systems through the effective use of advanced analytics and methodologies applied to timely, relevant data.
- Addressing opportunities to improve clinical development productivity focused on innovative treatments that advance healthcare globally.
- Optimizing the performance of health systems by focusing on patient centricity, precision medicine and better understanding disease causes, treatment consequences and measures to improve quality and cost of healthcare delivered to patients.
- Understanding the future role for biopharmaceuticals in human health, market dynamics, and implications for manufacturers, public and private payers, providers, patients, pharmacists and distributors.
- Researching the role of technology in health system products, processes and delivery systems and the business and policy systems that drive innovation.

Guiding Principles

The Institute operates from a set of Guiding Principles:

- Healthcare solutions of the future require fact based scientific evidence, expert analysis of information, technology, ingenuity and a focus on individuals.
- Rigorous analysis must be applied to vast amounts of timely, high quality and relevant data to provide value and move healthcare forward.
- Collaboration across all stakeholders in the public and private sectors is critical to advancing healthcare solutions.
- Insights gained from information and analysis should be made widely available to healthcare stakeholders.
- Protecting individual privacy is essential, so research will be based on the use of non-identified patient information and provider information will be aggregated.
- Information will be used responsibly to advance research, inform discourse, achieve better healthcare and improve the health of all people.



The IQVIA Institute for Human Data Science is committed to using human data science to provide timely, fact-based perspectives on the dynamics of health systems and human health around the world. The cover artwork is a visual representation of this mission. Using algorithms and data from the report itself, the final image presents a new perspective on the complexity, beauty and mathematics of human data science and the insights within the pages.

Artwork on this report is generated using data derived from clinical information taken from Genomics England using Human Phenotype Ontology (HPO). HPO provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. This information in linked to Whole Genome Sequencing (WGS) data allowing genomically-enabled RWE analyses to be conducted.

IQVIA™
INSTITUTE
FOR HUMAN DATA SCIENCE

CONTACT US
100 IMS Drive
Parsippany, NJ 07054
United States
info@iqviainstitute.org
iqviainstitute.org