FUTURE OF HEALTH

Using the African Genome

AN INDUSTRY REPORT BY

ONE BIO

AFRICA'S BIOTECH INCUBATOR

MAY 2021

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The future of healthcare is precision medicine, where the ability to create specific therapeutics based on the genetic makeup of an individual will allow us to provide tailored care. A strong understanding of the genetic markers of disease is critical to achieving this ultimate goal, however there is something that hinders our progress: the current dearth of African genome sequences.

The sequencing of genomes was traditionally concentrated heavily in Western Europe, leading to a large number of Caucasian-based sequences, leaving the African genome wholly underrepresented in comparison. This situation not only leads to poor clinical outcomes for non-Caucasian populations, but is also fairly counterproductive for genomic research as a whole, since the African genome has been shown to be the most diverse in the world. Within it are large numbers of genetic variants not found in other populations of people, which may hold the key to our understanding of human genetic diseases.

Increasing the number of sequenced African genomes will be critical, not only for improving diagnosis and clinical outcomes in individuals of African descent, but also for bettering our understanding of genetic disease and thus improving drugs and therapeutics for all. While there is a newfound understanding of the importance of sequencing African genomes, there is still not an enormous presence industry-wide. The only company that has dedicated it’s focus to the African genome is 54gene, a start-up founded in 2019 and located in Nigeria that has received US$9.5 million in funding to date.

The lack of competition, however, represents a multitude of opportunities in the space. The total global genomics market is predicted to grow rapidly to as much as US$82.6 bn by 2027 (the highest predicted CAGR of 19.5%). The emergence of COVID-19 and subsequent rush to sequence the virus and create an effective vaccine are expected to positively impact market growth, as many governments are pouring resources into genetic research.

African genomic data is primed to disrupt a number of areas in the industry, including genomics research, detection & prevention of disease, and AI drug discovery and development, among others. There are many exciting areas for future development that will positively impact multiple populations of people and aid in our overall understanding of complex disease. Ultimately, the African genome will play a critical role in our steps towards precision medicine and represents an enormous opportunity to improve healthcare for all.

INTRODUCTION

As we become more technologically and medically advanced, our ability to research and understand complex genetic diseases will improve, as well as our ability to accurately diagnose and treat such conditions. Not only will the discovery and development of novel drugs and therapeutics improve treatment options – there is a distinct opportunity to create personalized medications and treatments for the individual person depending on their unique genetic, environmental, and lifestyle factors. In other words, we can use a patient's genome to accurately recommend the best possible treatment for their ailments, known as the emerging field of precision medicine.

While our understanding of genomics has increased vastly over the years, there is still much we don't know when it comes to how certain complex groups of genes encode for specific disease phenotypes. Our tools to measure individual genomic information have greatly improved, in both speed, accuracy, and cost. The Human Genome Project in 2001 marked the first total sequence of a human genome, and took 13 years to complete. Now, individual genomes can be sequenced in a number of hours, and our knowledge of specific SNPs or gene clusters gives us a more informed understanding of potential phenotypes.

There are key pieces of research that are missing, however, that hinder our progress within genomics. It is undisputed that the African genome is extraordinarily underrepresented as compared to the genomes of Caucasian (and increasingly Asian) populations. It is estimated that only 1-3% of all genomes sequenced are from individuals of African descent, whereas there have been hundreds of thousands of Caucasian genomes sequenced through various efforts (e.g. the UK Biobank contains 500,000 genomes of European individuals).

Modern humans, *homo erectus,* evolved in Africa around 200 thousand years ago (kya) meaning Africa represents the root of human genetics (Figure 1). It is estimated that 99% of human evolution occurred in Africa, before various pockets of the population migrated over 50,000 years ago. Those smaller populations went on to form various races and ethnic groups we know today. Importantly, those smaller populations that left thousands of years ago did not take with them the breadth of genetic diversity as the rest of the African continent, and instead represent a much more narrow set of genes. On the other hand, the African genome contains virtually all of those specific variants[1] and many more, making it the most rich and diverse genome in the world.

[1] A genetic "variant" refers to a specific region of the genome (DNA sequence) that differs between two genomes.

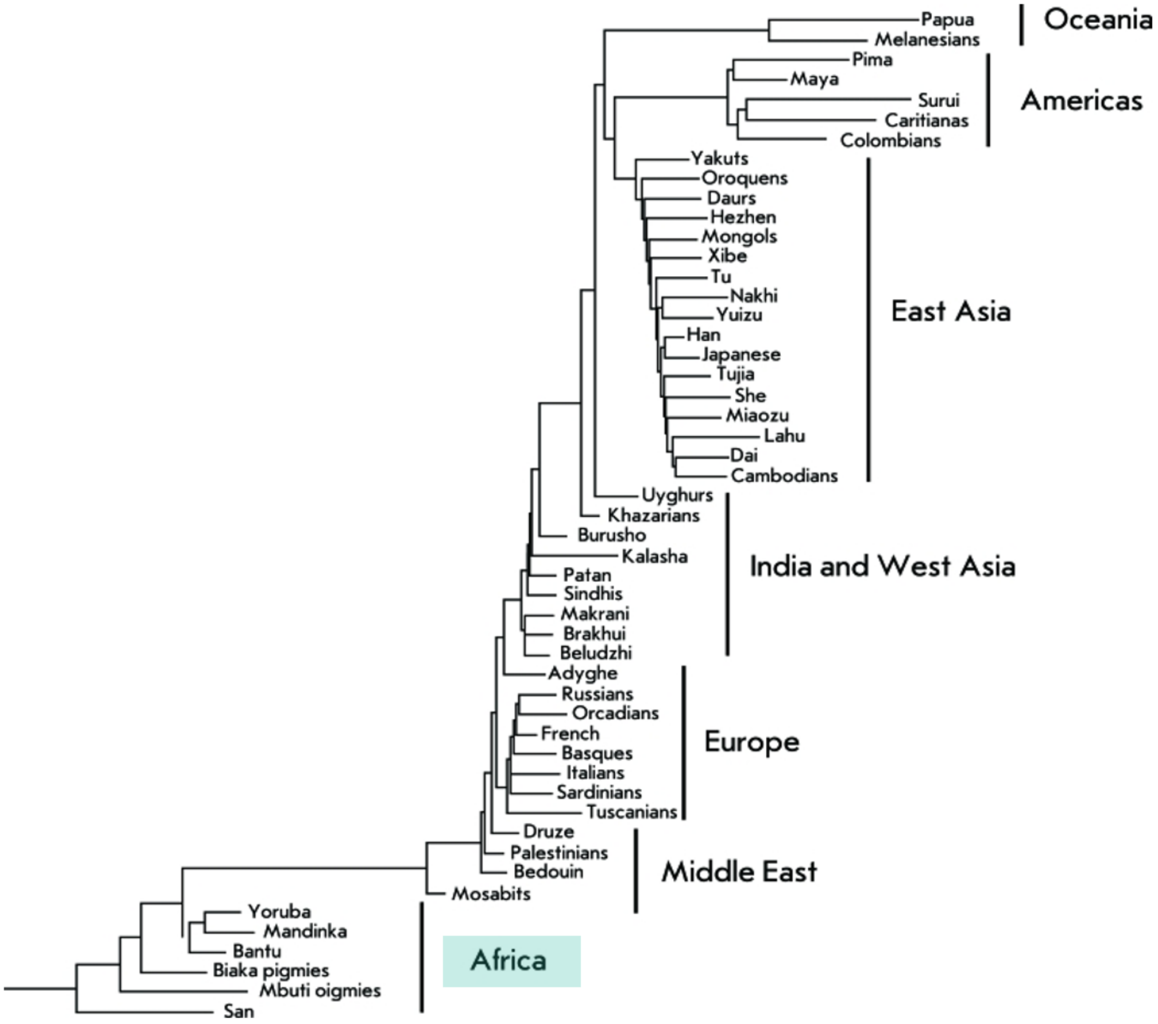


Figure 1. Geographic structure of human genetic diversity (Stepanov, 2010).

It can therefore be concluded that genomes from other populations don't include as many gene variants that may code for specific disease outcomes, which impedes our overall understanding of highly complex genetic diseases. Despite this fact, the vast number of sequenced genomes and reference genomes[2] are Caucasian, which means these disease-variants are missed at a much higher rate in African populations. Unfortunately, this lack of genetic information can lead to misdiagnosis, inefficacies in treatment, or a lack of therapeutic options for caucasians and non-Caucasian populations, especially African populations (Figure 2).

[2] Reference genomes are used by researchers who need to align and assemble experimental or patient genome sequence data.

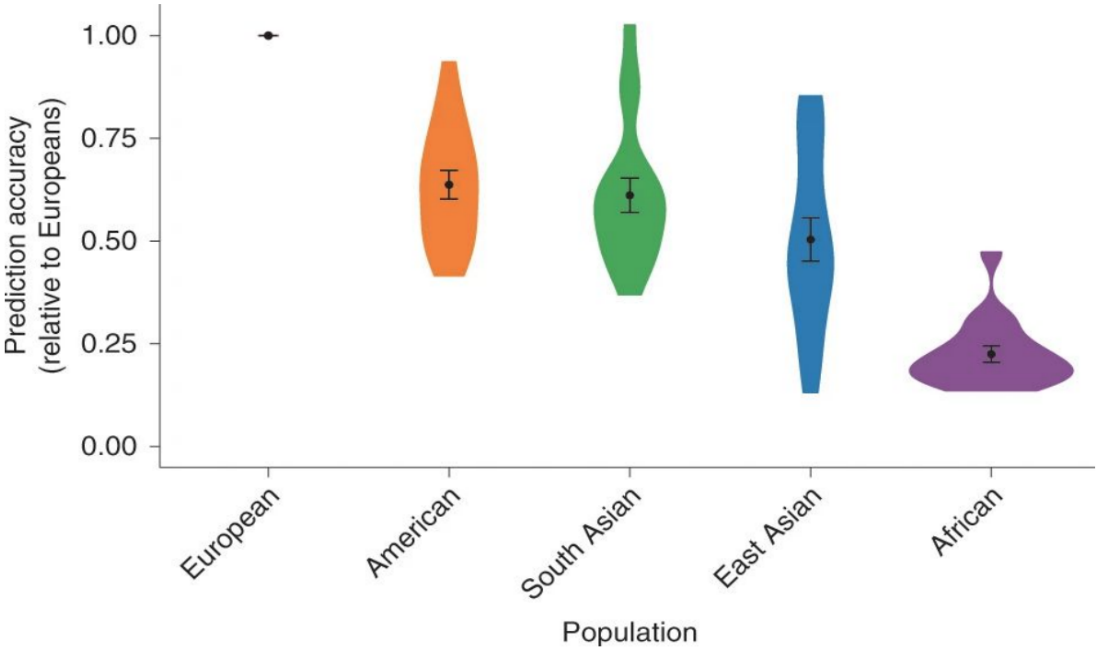


Figure 2. Predictive accuracy of reference genomes across various populations of peoples.

Despite technological advancements and some institutional and funding support, there is still an enormous lack of African genomics data, and diverse genetic data in general. By 2018, the majority of genome-wide association studies[3] (which aim to identify genetic variants associated with complex traits) were conducted in European (52%) or Asian (21%) populations. When further viewing the studies' distribution by ethnicity, the vast majority surveyed were of European (78%) or Asian (10%) descent. Only 2% were African, and other other ethnicities represented ≤1% of GWAS (Figure 3). This finding is striking, and showcases that there is much left to be done in terms of genomic data collection in underrepresented countries and ethnicities, particularly that of African genomics.

[3] Genome-wide association studies (GWAS) is an approach used in genetics research to associate specific genetic variations with particular diseases.

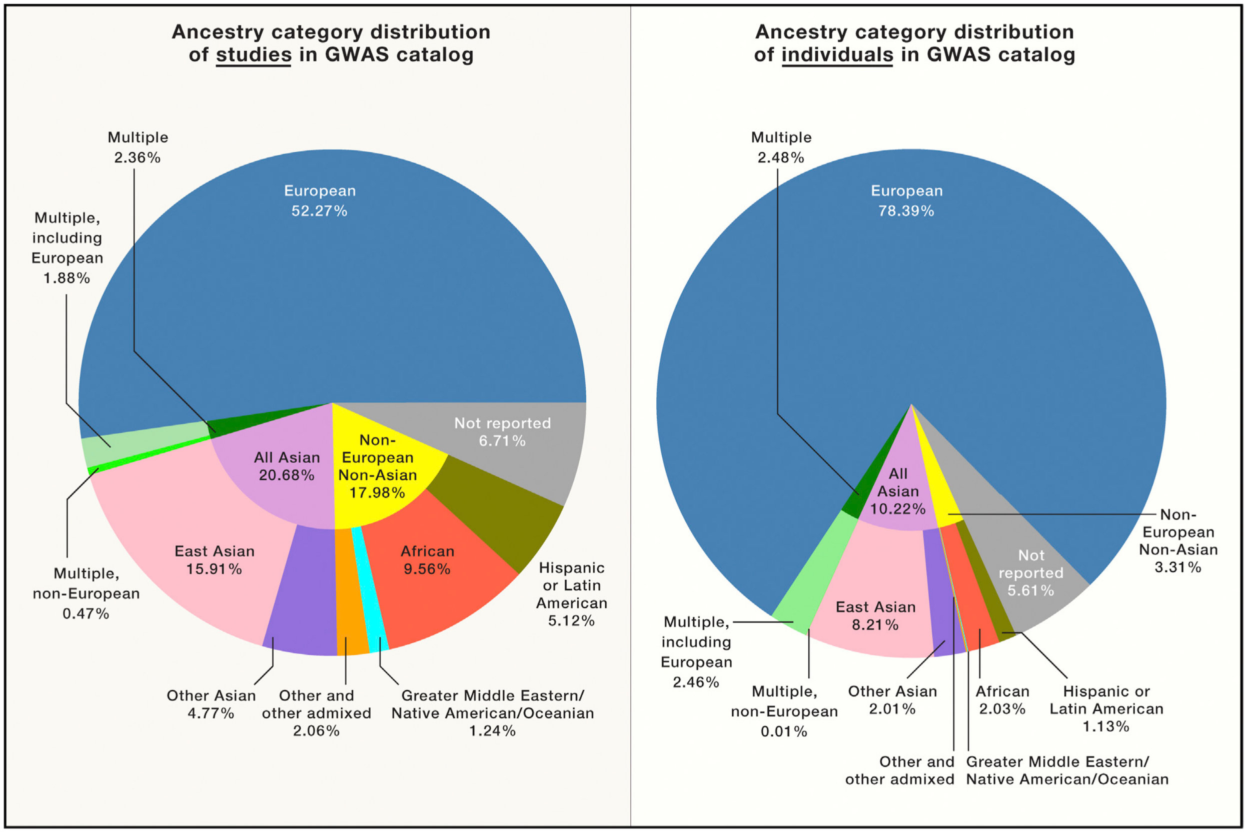


Figure 3. Summary of GWAS Studies by Ancestry for Studies in the GWAS Catalog through January 2019 (Sirugo et al., 2019).

The staunch lack of genomic representation for African populations is a hindrance to all research and development pertaining to genomics, whether it be discovering the underlying causes of complex diseases or pursuing clinical improvements such as precision medicine. According to a 2019 study, “analysis of diverse populations is critical for delineating the genetic architecture of complex diseases, turning the promise of precision medicine into reality for all individuals” (Tucci & Accay, 2019). In other words, increasing the representation of the African genome will provide a diversity of genetic data that will aid in our understanding of the basis of many genetically complex diseases, regardless of ethnic group. This knowledge will allow scientists to conceive of better drug discovery and health care strategy, which is ultimately beneficial to everyone.

The diversity of African genomes cannot be understated. A recent study in *Nature* (Oct 2020) conducted through H3Africa performed whole-genome sequencing analysis on 426 individuals from 50 ethnolinguistic groups in Africa (including previously un-sampled populations) representing the most extensive study of African genomes reported to date. The research uncovered over 3 million new genetic variants. Though the study itself was focused more on population genetics and historical migration patterns, one of the lead researchers, Zané Lombard stated: “From a health information point of view, we also showed that there were 62 new loci that we found to be under positive selection, and that gives you an idea of how our genome interacts with environments and how environmental forces like viral infections, et cetera, can drive genomic variation.” These implications are significant and confirm the importance of genetic diversity in helping to better inform our understanding of health- and disease-related information.

As the genomics industry (and within it the market for precision medicine) continues to grow rapidly, having access to African genome sequences will undoubtedly be advantageous for companies looking to utilize such genomic information for research, drug development, clinician reports, etc. Researchers, academics, start-ups and drug companies have become attune to the many advantages the African genome confers, but few have begun to focus on the space, leaving ample room within the market for more companies and partnerships to emerge.

INDUSTRY OVERVIEW

The African genome is clearly a potent source of crucial genomic information that can be helpful across a number of areas. Important to recognize is that the sequencing of the African genome provides data points that can be further utilized for other purposes down the line, e.g. genomics research, clinical diagnosis tools, drug R&D. The collection/storage, processing, and usage of this genomic data thus represent the areas where the African genome will be invaluable and where various opportunities may lie.

Genomic Data Collection & Storage

The first step in accessing the power of African genetic data is by sequencing and documenting African genomes. For context, a whole genomic sequence contains all the genetic information from an individual and determines the specific order of the base pairs (bp) of the four chemical building blocks (A–T and C–G) that make up a DNA molecule. The human genome contains around 3.1 billion bp, showcasing how sequencing genomes produces billions of data points. All that is required to sequence a genome is a biological sample, such as blood, saliva, epithelial cells, bone marrow, etc. – even a small amount of such samples can provide enough genetic material for a whole genome sequence.

The process of sequencing of genomes, while once time consuming and expensive, has become increasingly accessible, meaning African genomes can be sequenced quickly and economically. The first human genome took 13 years and US$1 bn in funding to sequence. Now, a full human genome can now be sequenced in about a day (a notable decrease in processing time) and costs have decreased sharply to approximately US$800 to US$1200 for a whole genome sequence (Figure 4).

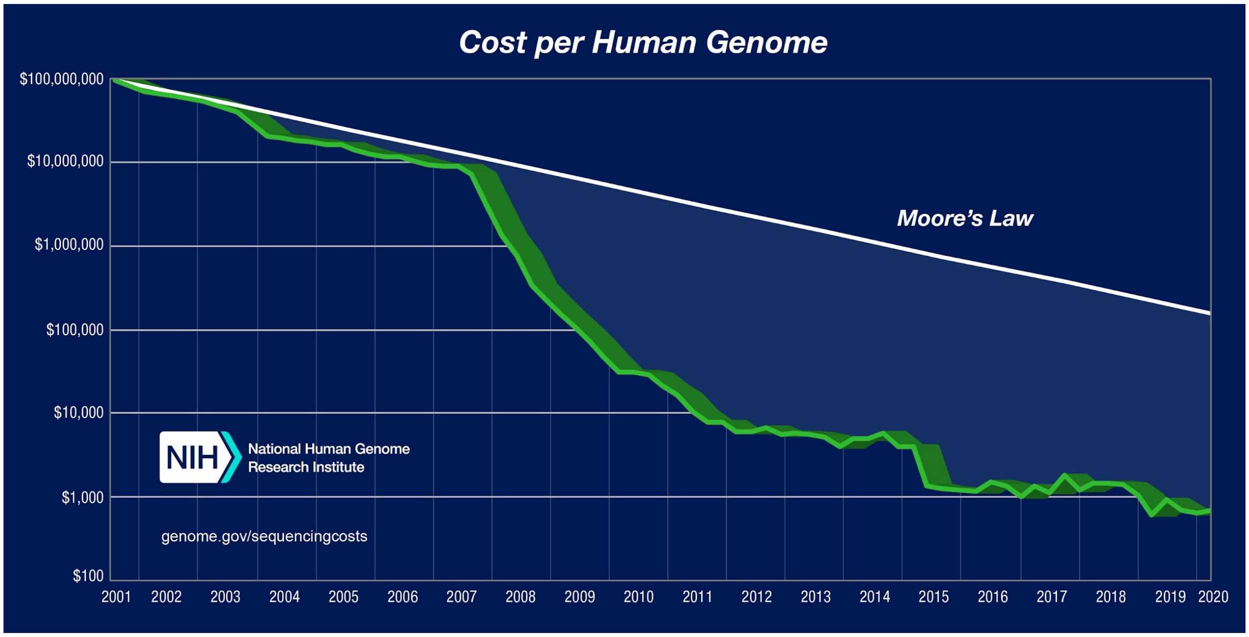


Figure 4. Cost per individual genome sequence over time, NIH.

Technological tools such as gene chips[4] (Figure B1), while not as thorough as a full sequence, can bring down the cost even further to less than US$100 per sample. This greater access to genomic sequencing means enormous amounts of genetic data can be generated fairly quickly and cheaply. A large number of samples and sequences are important for gaining a more holistic understanding of the human genome and future applications (e.g. diagnostic tools, drug development). Thus becomes the question of sample access and more importantly, storage.

Currently, biological samples are stored by biobanks, a type of biorepository where specimens such as blood, urine, skin cells, organ tissue, and other bodily materials are collected from participants, sequenced, and stored (Figure 5). There are over 120 biobanks worldwide, some with very small numbers of patient populations and samples, and others with hundreds of thousands of participants. Biobanks can be public or private, and many are funded by governments. Some examples of the largest biobanks include the UK Biobank (500,000 participants), the China Kadoorie Biobank (510,000+ participants), and the Biobank Japan Project (200,000 patients). Currently, the immediate focus of such banks is on disease understanding and drug development. Future industry routes will turn more to personalized and automated healthcare approaches, by uncovering key insights into the genetic components of complex human diseases.

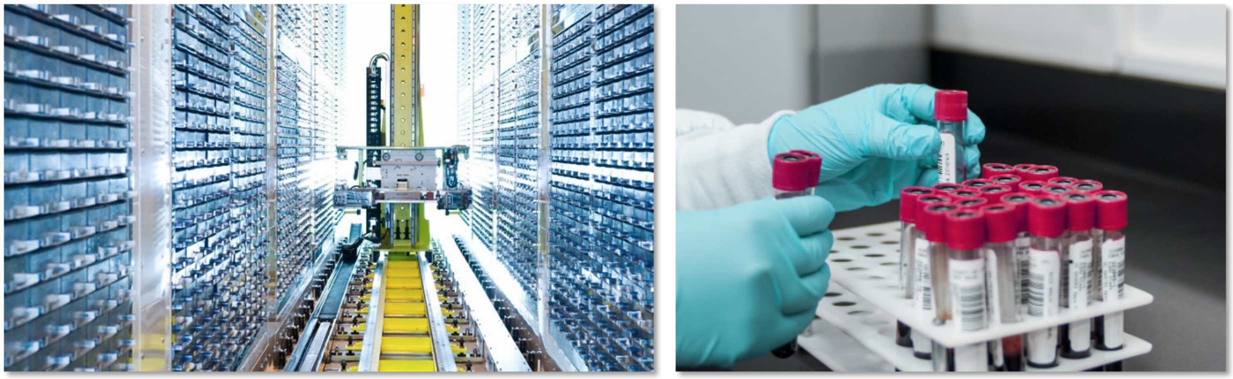


Figure 5. Photo of UK Biobank sample storage (left) and individual blood samples (right).

In terms of African genome storage, consortiums such as the Human Hereditary and Health in Africa (H3Africa), the International HapMap, and 1000 Genomes project have shown a growing engagement of African genomes in their databases, but are critically limited by government funding. H3Africa, for example, was initially supported with $176m from the United States National Institute of Health (NIH) and Wellcome Trust, but is expected to run out of funding within the next two years. Government funding is critical for success – over a decade ago China committed 2% of its GDP towards genetic study (e.g. cutting taxes and providing houses for scientists), which resulted in it surpassing Europe in biotech investment in 2018. By contrast, in 2016 Nigeria promised to dedicate 1% of its GDP towards science and technology advances (worth US$3.8 bn), but the budget has remained at only around US$750 mn.

[4] Gene chips, or DNA microarrays, are a collection of microscopic DNA samples on a small surface. They are used to measure the expression level of a large number of genes simultaneously or genotype multiple regions of a genome. They can be used in lieu of full genome sequences. *See Appendix B.*

Genomic Data Processing & Applications

How genomic data is processed and applied is just as important as how it is sequenced and stored. The main applications of genomic data come in the form of research into complex genetic diseases and subsequent diagnostic tools and drugs/therapeutics R&D for such diseases. Within R&D specifically, AI-powered drug discovery is an emerging area.

It is within various types of applications that the full power of the African genome is realized. Overall, African genomic data has many distinct advantages when it comes to data application and usage, due to its diversity, complexity, and sheer number of novel data points (i.e. variants) and associations waiting to be uncovered. Such properties make it invaluable in diagnostics, and drug & therapeutics research, discovery, and development. As such, the African genome will allow us to take valuable steps towards precision medicine.

Diagnostics

There are a number of different diagnostic areas in which genomics plays a key role: point-of-care (POC) diagnostics, *in vitro* diagnostics (IVD), molecular diagnostics, and genetic testing. Though these represent distinct markets, there is much overlap between these categories within technologies and aims. Importantly, genetic data plays a critical role in each. As the African genome has many novel variants, African genomic data thus represents a great potential for improving diagnostic outcomes in all individuals.

Point-of-Care (POC) diagnostics represent diagnostic testing that occurs by the patients' side, rather than in a lab. This has clear benefits, as POC testing is quick, leading to faster treatment and decreased health service costs (Figure 6). Previously the field was mostly associated with biochemical testing, e.g. the measurement of glucose (~40% of the industry), hemoglobin or blood cell counts, but has now expanded to include infectious disease, e.g. COVID-19 diagnostic testing (Figure 7). Genomic data and technologies are useful in this particular context as they can be used in combination with epidemiological research to help identify disease sources and transmission, which can ultimately inform infection control strategies.

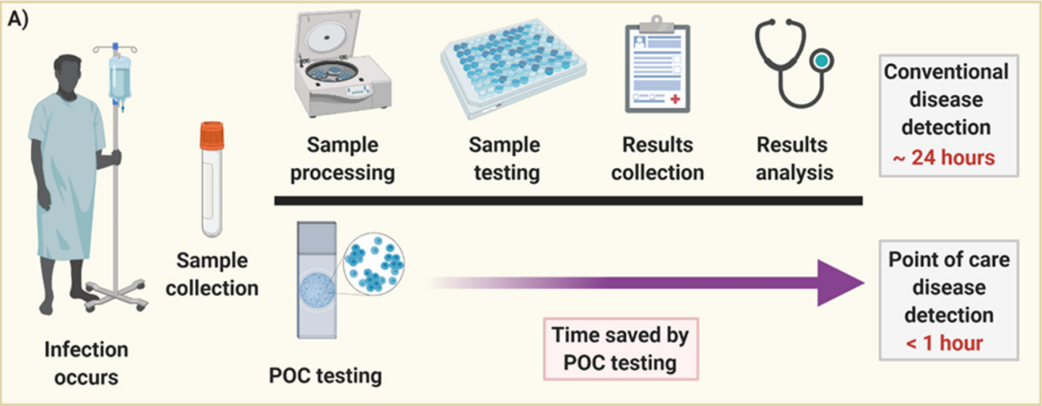


Figure 6. Schematic of disease detection using conventional methods of detection (e.g. IVD) compared to POC testing approaches (taken from Rezai et al.)

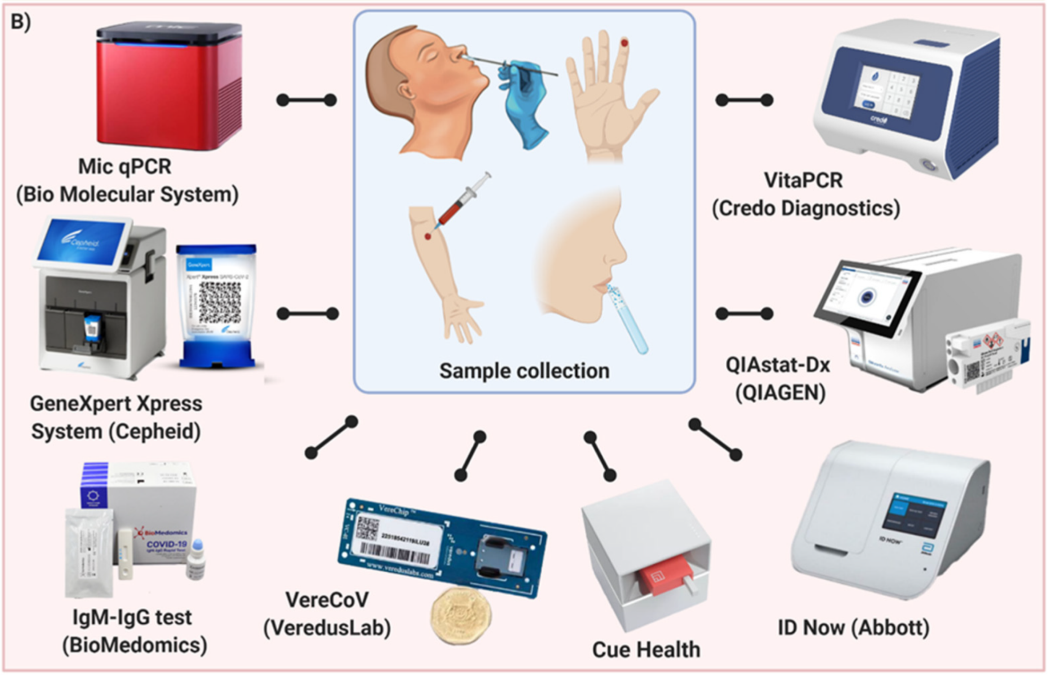


Figure 7. Various types of rapid POC devices currently commercially-available and FDA approved for COVID-19 detection (taken from Rezai et al.)

IVD represents diagnostic testing, done in a laboratory, of samples (e.g. blood or tissue) that have been taken from the human body. The aim is to detect diseases or other conditions, and the technique can be used to monitor a person's overall health to point to cure, treat, or prevent disease. IVD is an important technique of precision medicine because it can help identify patients who will likely benefit from specific treatments or therapies. Genomics plays an important role in these diagnostic technologies, such as next-generation sequencing (NGS) which attempts to aid in the diagnostics of germline diseases with the identification and curation of novel genetic variants.

Molecular diagnostics represents the broadest category of diagnostics and is referred to as the detection of genomic variants used to determine the susceptibility of an individual to certain disease and existing disease states. Tools used for this aim are those such as DNA sequence analysis, gene expression profiling, and detection of biomarkers. Tools such as NGS or genome-wide association studies (GWAS) provide valuable insights to the mechanisms of disease and genomic biomarkers.

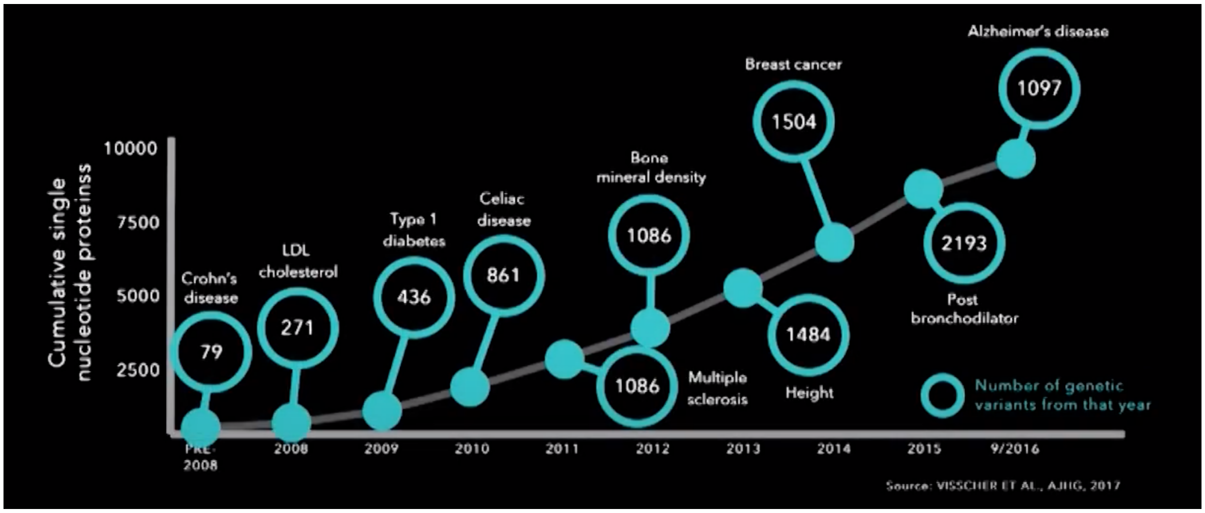
Finally, genetic testing, also called DNA testing, is simply a type of medical testing that detects changes in chromosomes, genes, or proteins. Results from such tests can confirm a suspected genetic condition and help to control a person's chance of developing or passing on a genetic condition. Specifically, such tests come in the form of kits and panels and testing is performed by taking blood samples from patients. There are multiple types of tests, including predictive & presymptomatic testing, carrier testing, prenatal & newborn testing, diagnostic testing, pharmacogenomic testing, and others. Such tests can also be important in diagnosing various cancers, cardiovascular disease, and other genetic diseases.

Drug and Therapeutics R&D

Many drugs currently available are referred to as “one size fits all” – in other words, for certain ailments there is only one drug available or readily prescribed to all individuals. There is a clear issue with this, as not all people can be expected to react similarly to the same drug. It can be hard to predict how a patient might react to a prescribed formula, whether it is beneficial, non-responsive (no reaction), or ultimately adverse (negative side effect). In fact, adverse reactions are a significant cause of hospitalizations in the U.S. Thus, an important area of precision medicine is *pharmacogenomics*, which is the study of how genes affect a patient’s response to certain drugs. This is a relatively new field within precision medicine that combines pharmacology (the study of drugs) with genetics in order to develop safe, effective medications and doses that are tailored to variations in a person’s genes.

By utilizing knowledge gained from various genomics projects (e.g. the Human Genome Project), scientists can measure and predict how inherited differences in genes affect the body’s response to certain drugs. Such genetic differences can be used to not only predict whether a certain medication will be effective or result in an adverse reaction (a more clinical approach), but it can also be used in the future to help develop new drugs tailored to treat a wide variety of health problems such as cardiovascular disease, Alzheimer's disease, cancer, HIV/AIDS, and asthma, among others.

This process, however, is not as simple as it sounds. While our tools for investigating genetic data are improving, it is still quite difficult to understand the basis for many diseases because the underlying cause cannot be accurately identified. This can be seen when looking for causal variants of disease: while there are some cases in which individual variants can be discovered as causal for a certain disease (e.g. BRCA genes for breast cancer), most diseases are not so simple (Figure 8). Furthermore, because variants can be inherited in groups across a population, they can be correlated to a disease without being relevant to that disease at all, leading to the misidentification of many “causative” mutations. This problem stems directly from lack of diversity in sequenced genomes, as many of the same populations are studied over and over, with no significant or pointed findings.



Source: VISSCHER ET AL., AJHG, 2017

Figure 8. Number of discovered genetic associations/variants correlated with certain phenotypes, e.g. diseases (taken from Daphne Koller WiDS 2020 presentation).

The diversity of African data specifically would allow causative variants and potential drug targets to be interrogated from many different angles, as genomes with different combinations of variants are common in Africa. In fact, when speaking on the aforementioned *Nature* (2020) study, researcher Zané Lombard commented on how “variants that were previously shown to be likely pathogenic were actually observed quite commonly in some of the population groups that we looked at. What this tells us is that those variants, because of how frequently they occur in these populations, are probably not having the kind of pathogenic impact that they previously were predicted to have.” Such findings significantly alter our perception of potential disease-causing variants and clearly illustrate how the genomic diversity of African populations can aid in our overall understanding of disease.

AI-driven Drug Discovery

The current route for drug discovery and development is inefficient and expensive. The average cost to research and develop a successful drug is estimated to be around US$2.6 bn, according to the Pharmaceutical Research and Manufacturers of America (PhRMA). The enormously high price tag is a result of failures – oftentimes, thousands or even millions of compounds are screened and assessed early in R&D, with only a few being approved to move forward in the process. The likelihood of a drug being further approved after undergoing clinical trials is less than 12%. Furthermore, the average time horizon for introducing a new drug, from experimental stages to market, is 12 years.

AI offers a unique opportunity to significantly bring down the cost of development. Machine learning, for example, is a subset of AI that works to identify patterns in data in the form of a model that can be used to make predictions about new data. This not only significantly cuts down on time, but further invites insights that would be lost with normal research methods. As CEO of Insitro, Daphne Koller, explains: “Machine learning is now doing amazing things if you give it enough data. We finally have the opportunity to create biological data at scale.” Deep learning algorithms can be used to process complex and large genomic datasets (Figure 9).

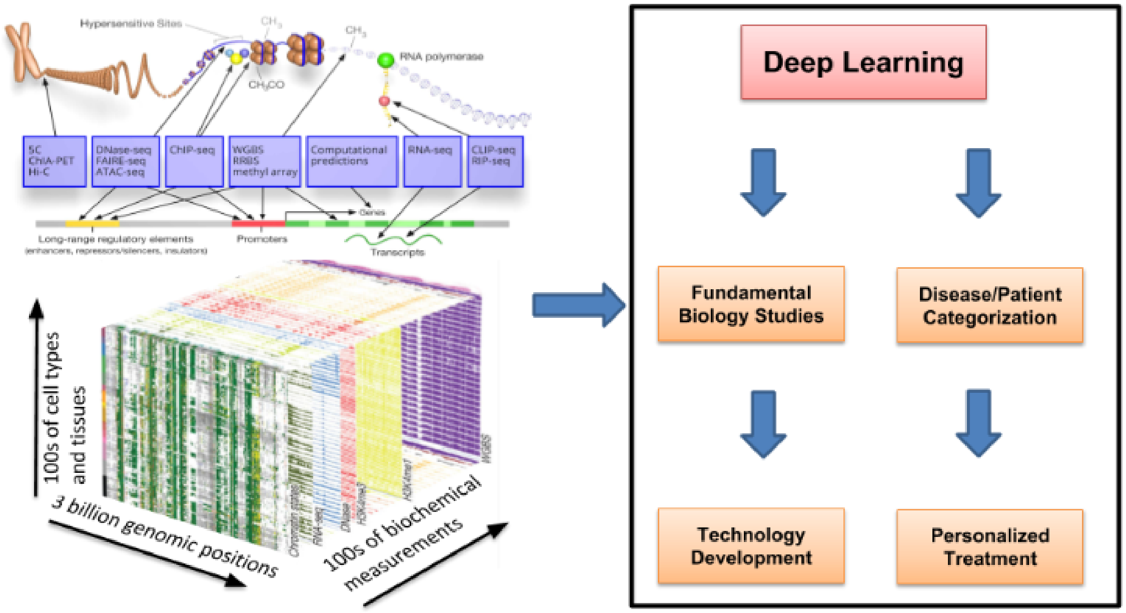

Hypersensitive Sites
CH3
CH3
CH3CO
RNA polymerase
5C
ChIA-PET
Hi-C
DNase-seq
FAIRE-seq
ATAC-seq
ChIP-seq
WGBS
RRBS
methyl array
Computational predictions
RNA-seq
CLIP-seq
RIP-seq
Genes
Long-range regulatory elements
(enhancers, repressors/silencers, insulators)
Promoters
Transcripts
100s of cell types and tissues
3 billion genomic positions
100s of biochemical measurements
Deep Learning
Fundamental Biology Studies
Disease/Patient Categorization
Technology Development
Personalized Treatment

Figure 9. Schematic of deep learning converging with genomics.

The advantage ultimately comes down to being able to rapidly sift through and glean insights from enormous amounts of data. In that way, biology and big data science seem to be the perfect pair: biological research, such as genomic sequencing, can generate datasets with millions of distinct points. This amount of data would take a human researcher an innumerably long time to analyze; ML and AI, on the other hand, are able to quickly process such datasets and uncover associations and patterns humans would have missed. It's an incredibly efficient method of research, and will soon help to shorten the process of drug discovery.

There is a distinct advantage the African genome offers within AI drug discovery. Such discoveries are predicated on access to enormous amounts of data – ML models, for example, will generally contain a few different datasets, each used to fulfill various roles in the system. The more data you provide to the ML system, the faster that model can learn and improve. Big data is therefore advantageous to improving ML model outcomes.

Diversity of data points is also ideal when setting up ML models. MIT's Stefanie Jegelka explains: "We want to pick sets that are diverse. [For example,] if you have a large data set and you want to explore — say, a large collection of images or health records — and you want a brief synopsis of your data, you want something that is diverse, that captures all the directions of variation of the data."

The global pandemic has further highlighted the need for an AI-focused approach to genomics and drug discovery. The biotech company Healx, for example, is using its AI platform to develop drug combinations from approved drugs to find treatments for COVID-19. This requires analysis of the eight million possible pairs and 10.5 billion drug triples stemming from the 4,000 approved drugs already on the market. Healx's AI platform, Healnet, overcomes this challenge by analysing data to predict combination therapies most likely to succeed in the clinic.

GLOBAL GENOMICS MARKET

The global genomics market best encapsulates where the African genome can fit into and disrupt, and includes a number of important segments that overlap with other global markets, the most relevant being (i) diagnostics, (ii) drug discovery & development, (iii) AI and computational tools, and lastly (iv) precision medicine. The following subsections dive deeper into these specific segments as their own independent markets.

An interesting segment sometimes covered independently is genealogy and ancestry, however, however, we have included this segment in diagnostics under “genetic testing.” Other segments that are irrelevant to the usage of the African genome include agricultural and animal research, biofuels, marine research, and forensics.

In 2019, the global genomics market was estimated to be between US$13.4 bn and US$18.85 bn. Some reports predicted a CAGR between 7.7% and 11%, leading to an estimated total market size around US$31.1 bn in 2027 (Figure 10). A newer report from Business Fortune Insights estimates a market size of US$82.60 bn by 2027, with a much higher CAGR of 19.5% between 2020 and 2027, which they credit to heightened research surrounding the emergence of the Sars-Cov-2m, commonly known as COVID-19 or coronavirus. Governments, academic institutions, pharmaceutical and biotech companies have accelerated genomic research to ascertain a better understanding of the virus and quickly create an effective vaccine – this collaboration between the public and private sectors for the foreseeable future will grow the market at a faster rate than previously imagined.

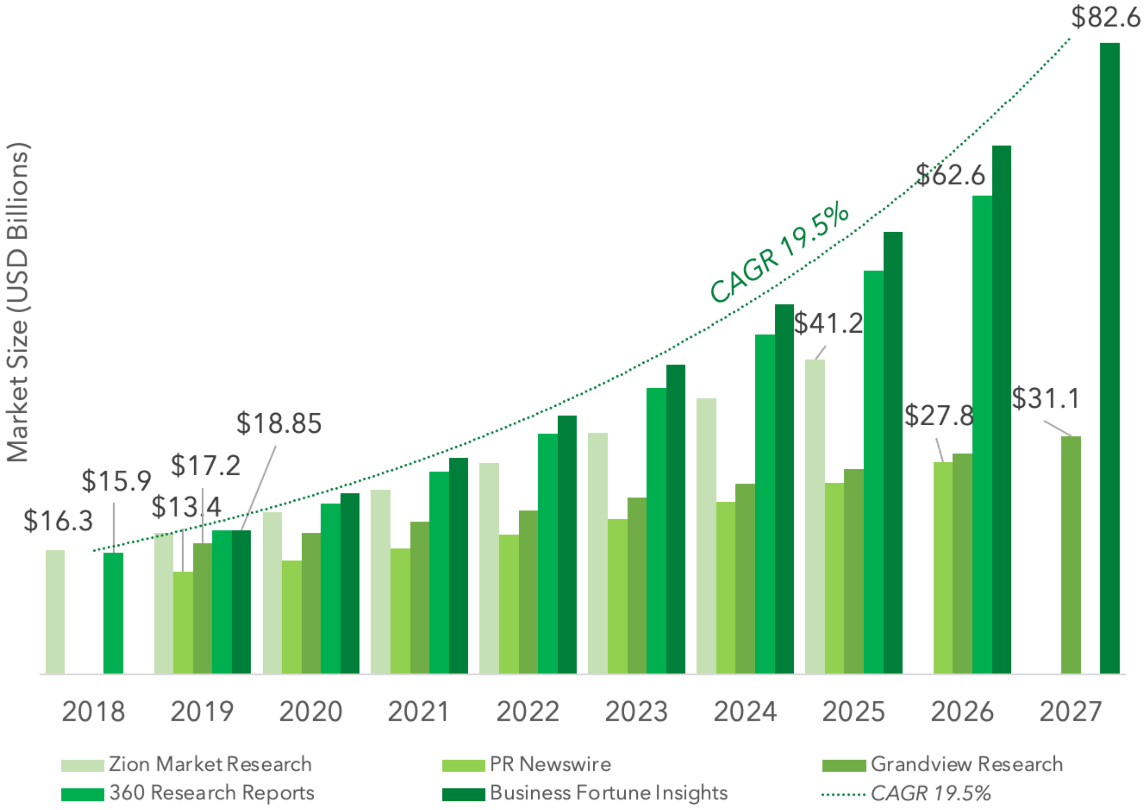


Figure 10. Forecast of the Global Genomics Market from 2018 through 2027.

North America represents the largest market segment, hovering at almost half the market, which is likely due to rising patient awareness, substantial investments in research by government organizations, and advanced healthcare infrastructure. However, Asia-Pacific represents the fastest growing region with an estimated CAGR of 9.1%, owing to increasing adoption and awareness for the latest genomics technologies in the emerging countries of this region. China is playing a pivotal role in regional market growth through initiatives such as the Precision Medicine Initiative (PMI) for the use of genomics in healthcare in 2017.

Key end-users are pharmaceutical and biotechnology companies, hospitals and clinics, academic and government institutes, and clinical and research laboratories. Pharma and biotech companies in particular are expected to dominate the global market throughout the forecast period owing to an increasing number of genetic research studies.

I. Diagnostics

There are a number of different markets that comprise the overall market for diagnostics, in which genomics plays a key role: point-of-care (POC) diagnostics, *in vitro* diagnostics (IVD), molecular diagnostics, and genetic testing. Though each market has slightly different products and aims, genomics plays a role in each, with varying degrees of prominence.

Point of Care (POC)

The overall market size was estimated between US$27-33 bn in 2020, expected to grow at a CAGR around 11.4%, to eventually reach US$50.6 bn or greater (Figure A1). This is up significantly from a World Health Organization (WHO) report that estimated the market at US$12-13 bn in 2016, highlighting significant growth potential. Specifically, the COVID-19 pandemic resulted in a positive demand shock for POC diagnostics – it is estimated the market grew at an enormous CAGR of 73.5% during the year of 2020, majorly due to the dissemination of COVID-19 testing kits and will subsequently return to a CAGR between 9 and 12% in the following years.

Other major growth drivers include technological advancements in POC devices, the rising incidence of infectious disease, and increase in investments by key market players. Demand for POC diagnostics kits increases accordingly with the rapid increase of acute and chronic diseases worldwide. Restrictions on growth include stringent regulatory policies in certain countries (e.g. FDA restriction in the US) and inadequate adoption of the technique, as most POC tests are administered by clinical personnel with limited laboratory knowledge.

In vitro Diagnostics (IVD)

The overall global IVD market was estimated between US$70-83.4 bn in 2020, with a slightly lower estimated CAGR between 2 and 6%, to reach a maximum predicted value of US$113.5 bn by 2027 (Figure A2). Growth of the industry has similarly been positively impacted by COVID-19, with a key overlap with the POC market for POC IVD devices, for testing during the pandemic; unsurprisingly, infectious disease was the dominant market segment in 2020 with around 41% of overall revenues.

While the pandemic fueled enormous short-term growth with the majority of tests being approved for emergency use, the increase in other conditions such as cancer, autoimmune diseases, and inflammatory diseases are expected to drive long-term growth. Some further key drivers of the market overall are the development of automated diagnostic systems for labs and hospitals, an increasingly large geriatric population, and an increasing number of IVD products being brought to market by key players. Some challenges for the industry are inadequate reimbursement for tests (a major growth constraint) as well as a changing regulatory landscape that is expected to become more stringent.

The IVD market produces reagents, analytical instruments, and accessory products that are used to perform diagnostic laboratory tests. Reagents[5] were found to be the dominant segment in 2020 at approximately 65% of revenues, which is attributed to the rise of rapid, accurate, and sensitive devices. Other analytical instruments (i.e. equipment and machines), such as commercial kits and robots in PCR laboratories are expected to drive growth – *in vitro* diagnostic software is furthermore used in many devices, such as point-of-care analyzers, laboratory-based analyzers, handheld personal in vitro diagnostics, and others.

Molecular Diagnostics

The global molecular diagnostics market was valued between US$10-18 bn in 2020 and is predicted to reach up to US$32 bn by 2026, registering a CAGR between 9 and 13%. One lone report, however, predicts the market already exceeds US$36 bn and will eventually reach US$46 bn by 2028 (Figure A3).

The increasing global prevalence of infectious disease and various cancers drives the market and creates demand for new diagnostic procedures and products. Molecular diagnostics represents a class of techniques that allows for the examination of biological markers in an organism’s genome and determines how their cells express their genes as proteins. Such techniques include polymerase chain reaction (PCR), DNA sequencing, and next generation sequencing (NGS); instrumentation, reagents, consumables and software are considered within the market as well. Demand for these techniques and products comes from various medical segments, including oncology, pharmacogenomics, infectious diseases, genetic testing, neurological disease, cardiovascular disease, microbiology, and others, and the end-user market includes hospitals, laboratories, blood banks, home health agencies, etc.

The overall molecular diagnostics market was dominated in 2019 by PCR, the most widely used technology. Advancements in PCR technologies such as multiplex PCR[6] are some of the highest impact drivers in this market segment, and usage of the technique has increased. Sequencing, on the other hand, is projected to see the highest growth rate in coming years due to multiple factors, such as the possible decrease in cost and rise in the portability of DNA sequencers. The development of NGS techniques may further contribute to its development, as that will provide high throughput analyst and sequencing of genetic data.

[5] Reagents are solutions of highly-specific chemical or biological substances that are able to react with target substances in the samples)(which are solutions of highly-specific chemical or biological substances that are able to react with target substances in the samples.

[6] Use of polymerase chain reaction (PCR) to amplify several different DNA sequences simultaneously, allowing for the detection of multiple targets in a single reaction well, with a different pair of primers for each target. The technique requires two or more probes that can be distinguished from each other and detected simultaneously.

Genetic Testing

The global genetic testing market surpassed US$13 bn in 2019 and is expected to grow at a CAGR between 10 and 12% through 2027, reaching up to US$31.9 bn (Figure A4). Genetic testing involves testing the genome to identify alterations in chromosomes, genes or proteins – this information can be used for diagnostic purposes, with the genetic disease testing segment comprising 25% of the market. Other uses of genetic testing is genealogy/ancestry investigation (e.g. 23andMe, Ancestry.com), where consumers send in small biological samples (e.g. saliva) that are used to sequence their DNA in order to ultimately ascertain a comprehensive familial genetic history. However, testing is most often used for diagnostics, prenatal screening, and pharmacogenomics as a predictive tool. Advances in sequencing techniques have reduced the sequencing time, and cost of genetic testing. For example, microarrays substantially reduce sequencing time by utilizing microchips and innovations such as exome sequencing and NGS have reduced the cost of genetic testing. These advances in genetic testing services, as well as a growing incidence of chronic diseases and increased demand for personalized treatments will propel market growth.

II. Drug Discovery and Development

The global drug discovery market was estimated between US$50-70 bn in 2020 and is expected to grow at a rapid CAGR between 8 and 10%. By more conservative estimates, the overall market will eventually reach over US$80 bn in 2026-7, however it has the potential to surpass US$100 bn by 2025 (Figure 11).

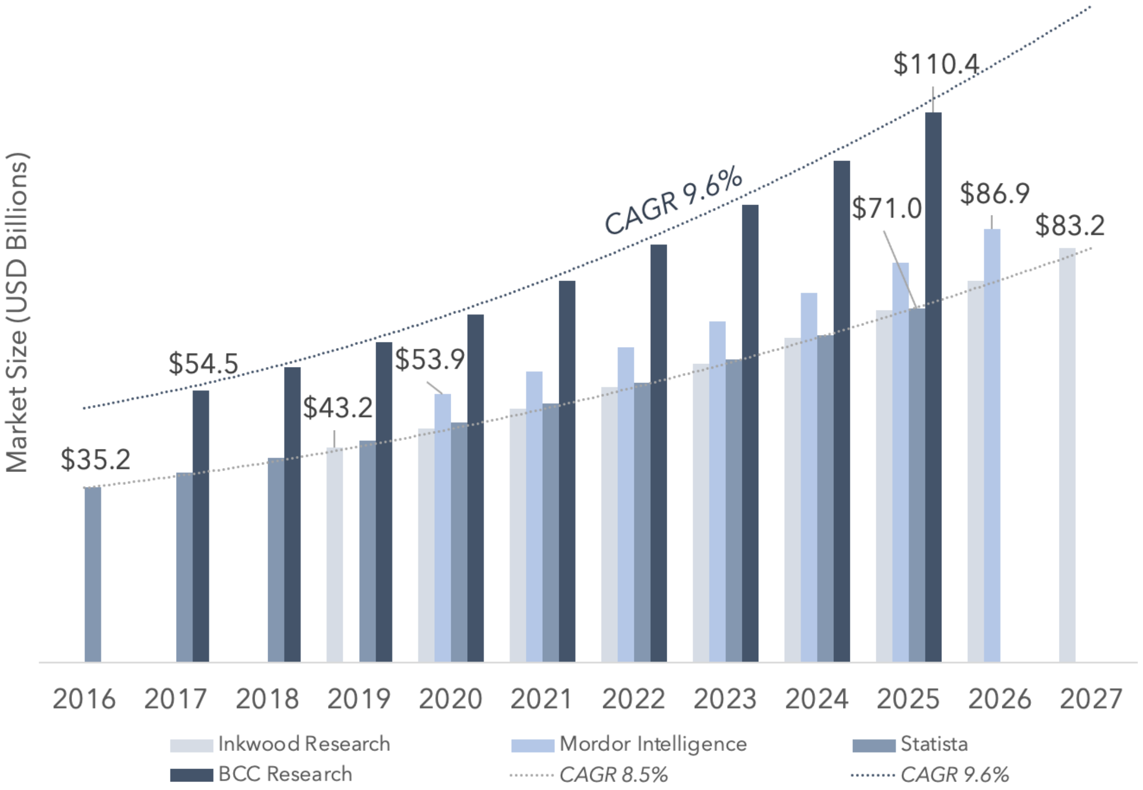


Figure 11. Forecast of the Global Drug Discovery Market from 2016 through 2027.

Due to the COVID-19 pandemic, health systems rapidly invested into drug R&D to create vaccines to combat the virus. Research groups around the world collaborated to identify drugs for treatment of COVID-19, potential compounds were screened, the FDA fast-tracked molecules under clinical trials and drug approval; accordingly, it is expected that the drug discovery market will be positively impacted. Other major drivers include the increasing instance of diseases, such as CNS disorders and cardiovascular disease, rising healthcare costs, and upcoming patent expirations of blockbuster drugs. These factors, combined with an upsurge in the use of advanced technologies (e.g. high throughput, bioinformatics, combinatorial chemistry), will drive growth in the overall market.

Biologics is the drug type expected to grow the fastest in coming years. Such drugs offer multiple benefits, such as fewer side effects, more potent and effective action, and the potential to cure diseases rather than merely alleviate symptoms. Their efficacy and safety has driven wide adoption and rapid growth in this segment.

For market share, North America is expected to dominate, as it currently leads the world in spending on drug R&D. It is also the largest market for bulk drugs and finished dosage formulas, and the U.S. is expected to become a major competitor for the biosimilars market. The overall global market is neither fragmented nor consolidated, with a few major players including Pfizer, GlaxoSmithKline, Merck, Agilent Technologies, and Eli Lilly.

Drug discovery and development represents the largest segment of the Global Genomics Market, due to the direct application of genomics data in drug R&D. It is furthermore an area of great opportunity for the application of African genetic data, making it a critical market to pay attention to in the coming years.

III. AI and Computational Tools

According to a 2019 BCG report, by 2022 spending on general AI-related tools in healthcare is expected to top US$8 bn annually across seven key areas: remote prevention and care (US$2.1 bn), diagnostics support (US$1.2 bn), treatment pathways and support (US$2.8 bn), drug discovery & development (US$1.3 bn), operations (US$500 mn), marketing and sales, and support functions. The areas in which genomics are most relevant are diagnostics, treatment, and (most importantly) drug development, which together comprise US$5.3 bn. It is likely these values will inflate as a result of the pandemic.

When considering genomics specifically, the Global Artificial Intelligence in Genomics market size was valued at US$142.5 mn in 2019 and is expected to grow at an enormous CAGR of around 39.0% from 2020 to 2028, estimated to eventually be valued at US$2.76 bn. AI and computational tools are significantly evolving the genomics industry, making genomics research cheaper, faster, and more accurate. Growing collaboration among hospitals, research institutes, and biopharmaceutical companies will boost the growth and adoption of AI within the genomics market globally. The huge size of the genome, regulatory factors, high cost, technology limitations, and prediction norms are some of the factors substantially affecting market growth.

Genomic medicine is making an impact in fields such as oncology, pharmacology, infectious diseases, and more healthcare verticals – machine learning and AI offer an important role in determining personalized treatment plans, clinical care, and research activities, leading to a better understanding of disease cause and treatment. Major companies such as Sanofi, Pfizer and Genentech, have already shown their interest and support for these technologies.

IV. Precision Medicine

The precision medicine market is unique because while it is an important segment of the genomics market, it is also considered a separate market. On its own, it is estimated to be valued between US$53.3 bn and US$70 bn for 2020, which is much higher than the genomics market. This is because it includes other segments. Many reports project high levels of industry growth, with a CAGR of up to 12.5%. The overall market is expected to reach a size of US$278.61 bn by 2030 (Figure 12).

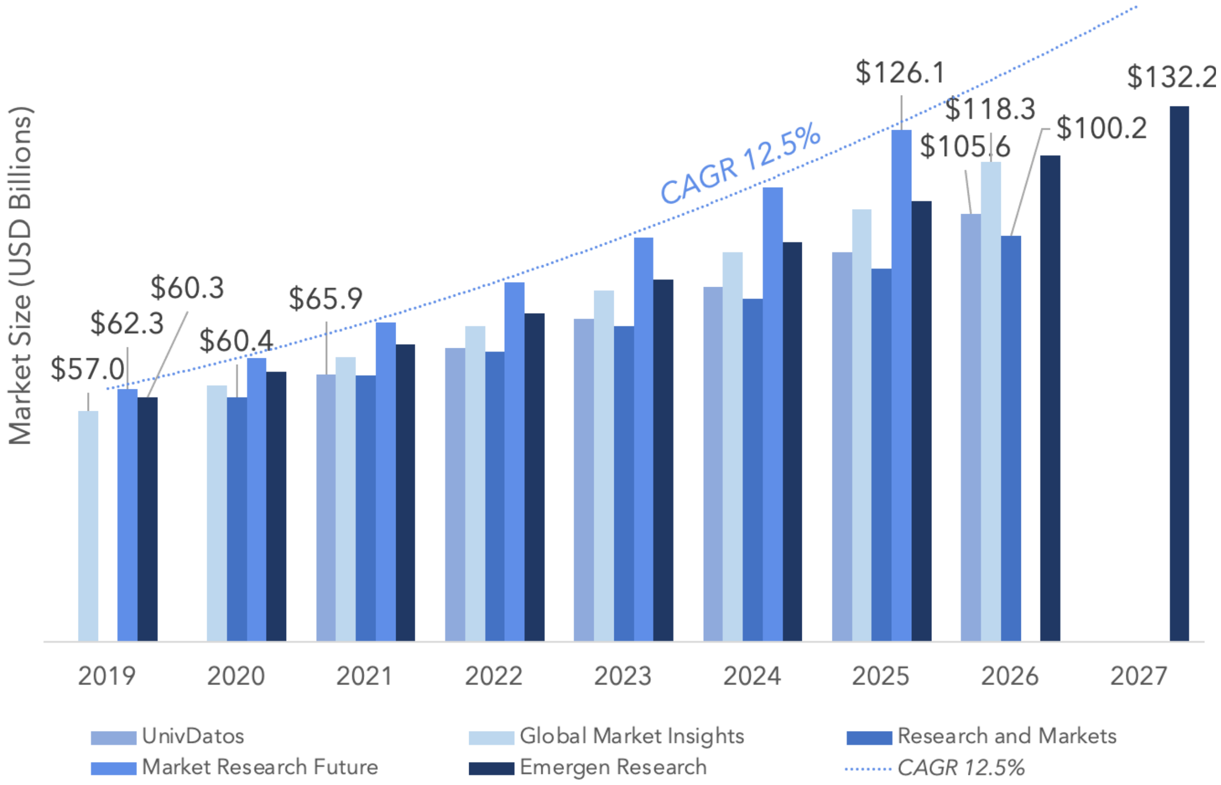


Figure 12. Forecast of Global Precision Medicine from 2019 to 2027.

Growth in the precision medicine market is propelled by an increasing demand for personalized treatment, technological innovation and advancement (including Biomarker-based tests/kits, next gene sequencing, and precise imaging), and government support and regulations. The introduction of cost-effective genomic and molecular biology testing methods, increasing prevalence of cancer and rare diseases, and rising use of big data in precision medicine are also key factors for boosting market growth. In the coming years, technologies such as gene sequencing, biomarkers, big data analytics, and companion diagnostics are anticipated drivers. Currently, the majority of the market share is held by North American companies, located predominantly in the United States, however Asia-Pacific is the region expected to see the quickest growth.

INDUSTRY PLAYERS

Genomics Companies

African Genomics Companies

While many large and established companies exist within the genomics and precision medicine markets, the number of companies that are currently sequencing and researching the African genome is quite limited.

One of the most notable companies in this space is 54gene, self-described as “an African genomics research, services and development company,” founded by Abasi Ene-Obong[7] in 2019. The “54” in its name represents the 54 countries on the African continent, however, the company itself is based in Nigeria. The company partakes in both sample storage and research initiatives: it created Africa’s first private biobank, the 54gene Biobank™, last year and currently focuses its research on genomic studies in non-communicable diseases (e.g. cancer, neuro-degenerative diseases, Sickle Cell, etc.) and infectious diseases (i.e. disorders caused by bacteria, viruses, parasites, or fungi). 54gene plans to create a genetic database of at least 100,000 Nigerians (via sequencing and using existing medical data). They are in a prime position to work along the drug discovery pipeline, from contract research for larger pharma companies to clinical trials to potentially drug R&D themselves. Overall, the goal of the company is to equalize precision medicine (Figure 13).

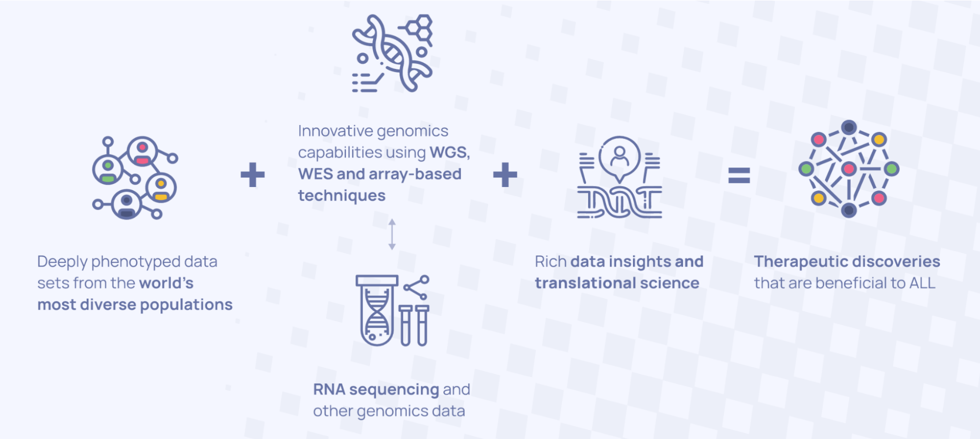

Deeply phenotyped data sets from the world's most diverse populations
+
Innovative genomics capabilities using WGS, WES and array-based techniques
RNA sequencing and other genomics data
+
Rich data insights and translational science
=
Therapeutic discoveries that are beneficial to ALL

Figure 13. 54gene workflow schematic (taken from website).

The company recently raised US$15 mn in Series A funding this past April, bringing their total VC funding to US$19.5 mn. Notably, 54gene’s early success is also a result of collaborative efforts with other research institutions and companies. It is in a consortium with H3Africa which has created strong research initiatives and collaboration. The H3Africa consortium is self-described as a group that “facilitates fundamental research into diseases on the African continent while also developing infrastructure, resources, training, and ethical guidelines to

[7] Abasi Ene-Obong is a 35-year-old Nigerian native with an MBA (bioscience management), an MSc in human molecular genetics, and a PhD in cancer biology. He has experience working in health care organizations (e.g. Gilead Sciences, IMS Health, PwC) in Nigeria, the U.K. and the U.S.

support a sustainable African research enterprise – led by African scientists, for the African people." The consortium has 51 projects in its pipeline currently and most recently published a study in *Nature* showcasing 3.4 million unique gene variants that had never been previously identified. Other notable collaborators and partnerships include Illumina, the Bill & Melinda Gates Foundation, the World Economic Forum, and Paradigm4, among others. 54gene further has strategic investment partners such as Adjuvant Capital, KdT Ventures, Better Ventures, etc.

Another important African genomics company is Artisan Biomed, based in Cape Town. Founded in 2015 with US$20.5mn in funding, Artisan Biomed is a non-profit medical testing & clinical laboratories company that aims to develop Precision Medicine in South Africa, utilizing four distinct solutions:

1. Make molecular applications available to patients in SA to improve disease and health management in a cost-effective way – leverage 'omics' and pathology testing capabilities.
2. Offer 'omics' services to enhance clinical and translational research by creating an integrated/high-quality environment for sample testing, data utilization and product dev
3. Develop new medical diagnostic applications for "People of African Descent" by using advanced computational methodologies for the analysis of data generated from routine testing, clinical research and focused R&D initiatives
4. Offer training programs designed to empower clinical researchers, medical experts and other stakeholders in the use of Precision Medicine applications

The company is a subsidiary of the Centre for Proteomic & Genomic Research (CPGR), one of Africa's first fully integrated 'omics' service providers and was based on an initiative by the Department of Science and Technology (DST). It is supported with funding by the Technology Innovation Agency (TIA) and has entered into a partnership with Lancet Laboratories.

A further company of interest is Next BioSciences, a B2C biotech company established in 2005 that is involved in stem cells, genetics, biologics, and pathology. Located in Johannesburg, the company offers a number of products and services, from stem cell banking to genetic screenings and COVID testing, and brings in around US$7.56 mn in revenue each year according to Dun & Bradstreet. Notably, Next Biosciences owns Netcells, the largest private stem cell bank in Africa, specifically focused on umbilical cord stem cells for use in regenerative medicine. In 2016, the company merged with a local laboratory Genesis Genetics, making it the only company in Africa to offer genetic and metabolic screening.

Tech-driven Genomics Companies

Other companies within the genomics space specifically utilize cutting edge technological tools, such as AI, machine learning, and precision biology. Many recently-founded companies utilizing these combined technologies have received enormous amounts of funding.

A recent example is Insitro, a San Francisco based AI drug discovery platform founded by Daphne Koller, a highly accomplished computer scientist and Stanford professor with a track-record in entrepreneurship. Insitro uses state-of-the-art bioengineering technologies that enable it to generate high-throughput, functional genomic datasets and align them with patient data via novel machine learning methods, thereby building predictive models that can accelerate target selection and the design and development of effective therapeutics. The company’s founder and its methods are so convincing that it has received US$400 mn in Series C funding from investors such as CPP Investments, T. Rowe Price, Andreessen Horowitz, BlackRock, Foresite Capital, and others, bringing its total investment to US$743 mn overall.

Many other companies, such as Deep Genomics and Insilico Medicine occupy this space and focus on improved drug discovery and diagnostics using machine learning and AI-platforms (Table A1). According to Deep Pharma Intelligence, the Pharmaceutical AI sector is “heating up” as Big Pharma and contract research orgs are increasingly competing for AI partnerships. Unsurprisingly, the COVID-19 pandemic has accelerated the adoption of AI and is acting as a positive catalyst for the industry – in fact, prominent COVID vaccine producers such as Pfizer, Astrazeneca, and Johnson & Johnson all invested heavily in multiple AI-focused companies in 2020, such as Insilico, Benevolent, and Exscientia respectively, among others (Figure 14). There are many other Pharma corporations focused on investing in or partnering with AI-focused biology companies, so this industry sector will be important to watch closely in the coming years.

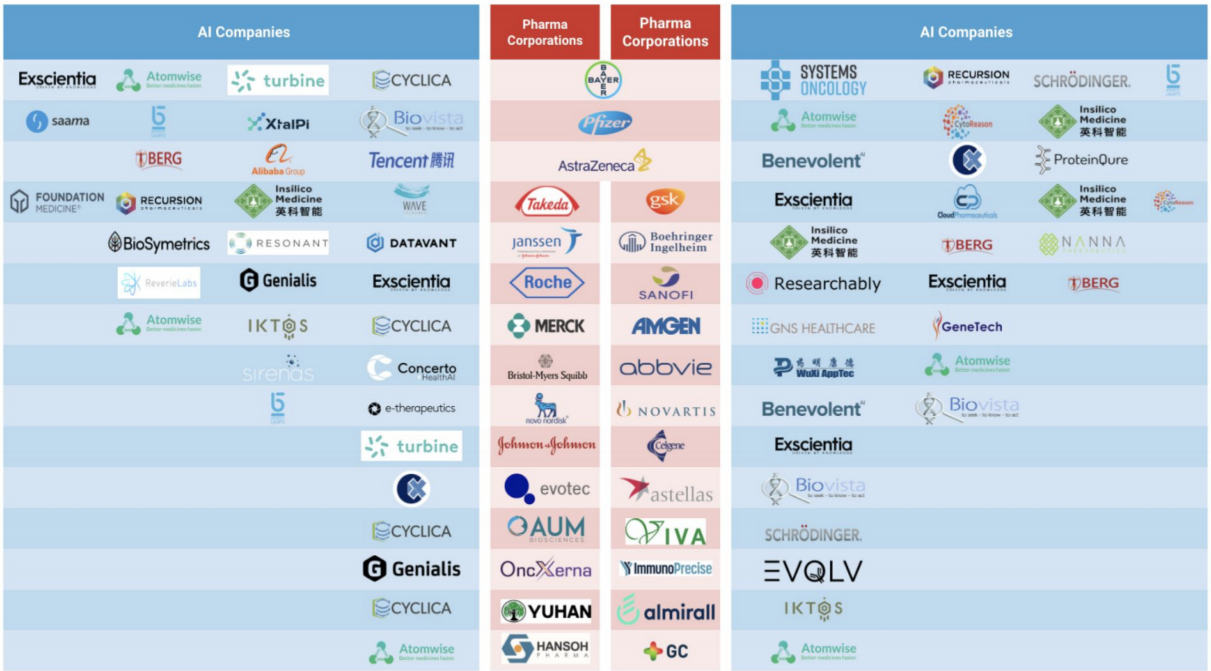


Figure 14. Pharma AI Deals Structure 2020 (Deep Pharma Intelligence).

Potential Partnerships

While the number of companies and research facilities sequencing and storing African genomics data is small, the number of groups interested in utilizing that information is vast. Because of the utility of the genetic data, pharmaceutical companies of varying size, drug discovery platforms (AI and traditional), biotech startups, genomics companies, and academic research facilities are all interested in gaining access to African genomics data, and some will be willing (and able) to pay a hefty price. To illustrate, Ene-Obong has already commented on the wide-ranging international interest 54gene has received. This opens a number of interesting opportunities and also important ethical questions.

While figuring out the ethical ramifications and methods of sharing African DNA data with international companies will need to be carefully considered, these more established companies may be invaluable for providing the resources necessary to thoroughly research African genomics and work toward medical advancement for all. Large pharmaceutical companies have the infrastructure, funding, and connections that are needed for effective drug and therapeutics R&D. As mentioned in the previous section, there are a number of established pharma companies in the space that are actively partnering with AI and genomics-focused companies (Figure A5, Figure A6, Table A2).

OPPORTUNITIES

Power of the African Genome

The diversity of African data allows causative variants and potential drug targets to be interrogated from many different angles during the first and most affordable phase of a long term drug discovery project.

When investigating the causes of disease or looking for new drug targets, there are sometimes individual mutations that have an obvious causal link to a disease. The immense success of research into the BRCA genes, for example, and the cancer treatments that followed are a testament to this. However, most investigations yield less obvious results. Variants are inherited in groups across a population, which means a variant can be strongly correlated with a disease without being causative or relevant at all. Often hundreds of "potentially causative" mutations are identified because of this. To investigate which of these variants would make a good diagnostic or drug target would take decades and cost billions.

An ideal scenario would be to eliminate potential variants by finding them either in healthy individuals or not finding them in individuals who do have the disease. The lack of diversity outside of Africa means that the studies to find genomes that satisfy these criteria require many participants and still yield several potential variants for expensive downstream analysis. The highly diverse populations of Africa allow variants to be analysed from several different angles. Genomes with different combinations of variants are common in Africa so study sizes are much smaller and more economical. The investigative power of these studies are much higher too, with a higher confidence in fewer potential drug targets. This reduces the downstream costs and potential failure of a drug discovery project dramatically.

Detection and Prevention

As we step into the future of healthcare with precision medicine, there is ample opportunity to improve disease understanding and drug targeting in non-white populations using African genetic data. This opportunity within pharmacogenomics will have lasting positive health impacts on many millions of people and, furthermore, such streamlined care will save millions in unnecessary healthcare costs.

Timely diagnoses and treatments require identification of biomarkers that would be useful for early detection, prevention, and treatment of diseases that are associated with specific medical conditions (e.g. cancer, cardiovascular disease, neurological disorders, etc.). In the diagnosis process, markers can determine staging, grading, and selection of the initial therapy. During treatments, they can be used to monitor therapy, select additional therapies, or monitor recurrent diseases. The ability to determine such biomarkers for diagnosis and prognosis of disease will undoubtedly have a meaningful impact, and is an important step forward into precision medicine.

A clear opportunity within the IVD market, then, is the development of condition-specific markers and tests, specifically using the African Genome. Technological advancements in genomics, proteomics, and molecular pathology have helped introduce new biomarkers with potential clinical value. By sequencing and researching the African genome, more biomarkers for disease are bound to be uncovered or potential variants thought to be pathogenic will be disproven (as was seen in the Choudhury et al. Nature study). The integration of biomarkers and the availability of biomolecular tools are expected to help in the development of a new range of condition-specific tests.

POC genetic testing is another area of interest – with increasing incidents of infectious disease, genomics will play an important role in pinpointing the cause of disease, diagnostics, strain tracking, and mitigating viral spread.

AI and Drug Development Platforms

As previously laid out, current methods of drug discovery and dev are inefficient and expensive. The average cost of R&D of a successful drug is US$2.6 bn and takes 12 yrs to get to the market. CEO of Lantern Pharmaceuticals, Panna Sharma, states: “We live in an era where drugs have gotten obnoxiously more expensive and everything else in multiple other industries has gotten cheaper. Computers, computing power, clothes. The only thing that’s gotten more expensive is real estate and drugs - pharma and healthcare in the United States.” These sentiments are echoed by Daphne Koller, who founded her company Insitro with the hope of discovering new therapeutics by using machine learning and modeling to find the most efficient routes to drug discovery, significantly lowering the cost of R&D.

AI offers a unique opportunity to significantly bring down the cost of development while rapidly discovering new potential drugs and therapeutics. Combining the power of the African Genome with the power of AI computing would undoubtedly yield promising results in the future of disease understanding and treatment for all. Investment into this space is high, which means there is ample opportunity to receive funding. Thus, investing in and implementing a high-performance computational platform or partnering with existing companies and providing crucial genetic data may be a potential area of opportunity in the future.

RISKS & ETHICAL CONSIDERATIONS

______While there are clearly many benefits to sequencing and utilizing African genomic data, there are a number of key risks and ethical concerns must be considered. Researchers must keep in mind a series of serious ethical, legal and social issues, such as informed consent, benefit sharing, confidentiality, ownership, commercialization and public participation, in order to ensure that the African genome is researched respectfully and that the resulting benefits positively impact all.

Risks

Storage, usage, and dissemination of genetic information represent key risks when sequencing and using DNA. Biobanks are currently responsible for keeping a patient's genetic record confidential and distributing that information to researchers and institutions in a proper, legal manner. There are no current common-practices employed by all biobanks and biorepositories, so there is an inherent risk to each individual. In fact, the World Health Organization has called for the global governance of biobanks, due to the differences in values, practices, and other requirements across borders.

Secure data, especially genetic, is of incredible importance towards building safe and sustainable biobanks and repositories. U.S. citizens, for example, are wary of sharing personal DNA samples after it was discovered that 23andMe shared genetic data with third party institutions, including academic institutions, pharmaceutical companies, and other large corporations. Many gene-testing companies have policies that allow them to share anonymised data with third-party entities that many consumers are likely unaware of, which is disingenuous and leads to a lack of trust from the public.

Ethical Considerations

The ethical considerations of using the African genome come down mainly to consent, compensation, and exploitation. The fear of exploitation (e.g. unfair distribution of risks and benefits) makes many low- to middle-income countries hesitant about foreign researchers accessing and using their human biological samples and associated data. Namely, the fraught history of exploitation of resources and labor on the African continent by foreign nations cannot be minimized – the African genome is undoubtedly a valuable resource, and how it is measured and utilized must be carefully considered. It is critical that any future medical and scientific advancements brought about by African DNA are accessible and beneficial to the African continent, and that future progress is not wrongfully withheld from those that contributed to such advancements.

Another important consideration is that of consent and compensation. The issue of consent centers around participants understanding what they are agreeing to. Many African peoples do not know or understand what DNA is and language barriers may prevent effective explanation – given this, can they fully consent to providing their DNA? The answer is doubtle, especially when the biological sample may be taken in exchange for money. Furthermore, it is important to keep in mind what is considered "fair" compensation – oftentimes, individuals are provided with a small initial payment for a biological sample, but never see any compensation from the enormous amounts of profits that may be generated down the line. It will therefore be necessary to consider various types of payment amounts and/or structures – this may not look the same across the board, but will be important to keep in mind.

Historical examples of ethical misconduct show why these considerations are so grave. One of the most infamous and historical examples of misconduct is that of Henrietta Lacks, an African-American woman with “immortal cells” that had the unique ability to grow indefinitely. Scientists harvested these cells without her knowledge before she died, and her immortal cell-line has been used in almost every major medical advance in the past half-century. A distinct lack of consent is problematic, but even further, neither she nor her descendents have been compensated, despite the overwhelming number of advancements (and subsequent profits) brought about by her cell line, leading to many debates over payment.

Even recently, Sanger Institute was accused of misusing African genetics on the basis of lack of consent. In 2018, Sanger was accused by whistleblowers of commercializing a new gene chip without the consent of hundreds of African people whose donated DNA was used to develop the chip. It further did not gain a proper legal agreement with its partner institutions in Africa, and in March 2019 Stellenbosch University demanded that Sanger return the samples. Bioethicist Jantina de Vries from the University of Cape Town in South Africa said, “What happened at Sanger was unethical. Full stop.” This kind of misconduct erodes trust between African scientists and other international entities that may be able to do good work and assist with resources, tools, and infrastructure that are not as accessible in parts of Africa.

REFERENCES

54gene <https://www.54gene.com>

Aboshiha, A.; Gallagher, R.; Gargan, L. Chasing Value as AI Transforms Health Care, Mar 2019. BCG.

Advances in Genomics and Proteomics in the $8.62 Billion Molecular Diagnostics Market, Apr 2020. The Business Research Company, found via PR Newswire.

Africa's people must be able to write their own genetic agenda, 2020. Nature Editorials.

Algorithmia. The importance of machine learning data, 2020. Algorithmia Blog.

Armstrong, R. Can this AI platform change the future of cancer treatments?, Dec 2019. European Pharmaceutical Manufacturer.

Artisan Biomed <https://www.artisanbiomed.com>

Azodi, C.; Tang, J.; Shiu, S-H. Opening the Black Box: Interpretable Machine Learning for Genetics. *Cell Press Reviews* (2020). https://doi.org/10.1016/j.tig.2020.03.005

Bajpai, P. Artificial Intelligence and Drug Discovery: The Companies Leading the Way, Apr 2020. Nasdaq.

Bright, J. African genomics startup 54gene raises $15M led by Adjuvant Capital, 2020. TechCrunch.

Chen, H. and Pang, T. A call for the global governance of biobanks, 2014. World Health Organization.

Choudhury, A.; Aron, S.; Botigué, L.R. *et al.* High-depth African genomes inform human migration and health. *Nature* **586**, 741–748 (2020). https://doi.org/10.1038/s41586-020-2859-7

D'Onfro, J. A Star Professor – And Her Radical, AI-Powered Plan to Discover New Drugs, 2019. Forbes.

Faggella, D. 7 Applications of Machine Learning in Pharma and Medicine, Mar 2020. Emerj Artificial Intelligence Research.

Genomics Market 2019 Size, Share, Growth | Global Industry Research Report, 2026, Jul 2019. 360 Research Reports, via FortuneBusiness Insights.

Genomics Market Size is Projected to Reach USD 115.6 Billion by 2026 - Valuates Reports, Aug 2020. PharmiWeb.com, found via PRNewswire.

Goyal, S. Point of care genomic tests in infectious disease, June 2019. University of Cambridge.

Global Artificial Intelligence in Genomics Market Assessment, 2020. Insight Ace Analytic.

Global Genomics Market Size, Status, and Forecast, 2020-2026, Aug 2020. Valuates Reports.

H3Africa. <https://h3africa.org>

Hardesty, L. Data diversity: Preserving variety in subsets of unmanageably large data sets should aid machine learning, 2016. MIT News.

Hildreth, C. Next Biosciences and Genesis Genetics Combine to Consolidate Stem Cell Banking and Genetic Testing Within Africa, 2016. BioInformant.

insitro Raises $400 Million in Series C Financing, Mar 2021. Business Wire.

In vitro Diagnostics, Oct 2019. FDA.

Koller, D. Machine Learning: A New Approach to Drug Discovery | Daphne Koller | WiDS 2020, Mar 2020. ICMEStudio, YouTube <https://www.youtube.com/watch?v=V6bSlPNwrKo&t=1444s>

Koslov, M. New Genome Sequences Reveal Undescribed African Migration, Oct 2020. The Scientist.

Luh, F. & Yen, Y. FDA guidance for next generation sequencing-based testing: balancing regulation and innovation in precision medicine. *npj Genomic Med* 3, 28 (2018).

Macdonald, Fiona. Point-of-care testing in genomics, Jan 2017. Genomics Education Programme.

Maxmen, A. The next chapter for African genomics, Feb 2020. Nature News Feature.

Molecular Diagnostics. ScienceDirect.

Munshi, N. How unlocking the secrets of African DNA could change the world, 2020. FT Magazine, Genomics.

Orchard-Webb, D. 10 Largest Biobanks in the World, May 2018. Biobanking.com

Pandya, J. Biobanking Is Changing The World, Aug 2019. Forbes.

Point of care (POC) Diagnostics Market Size, Share & COVID-19 Impact Analysis, By Product..., By End User…, and Regional Forecast, 2021-2028, May 2021. Business Fortune Insights.

Point Of Care Diagnostics & Testing Market Size, Share & Trends Analysis Report By Product..., By End-use..., By Region, And Segment Forecasts, 2021 - 2028, Jan 2021. Grandview Research.

Point of Care/Rapid Diagnostics Market by Product..., Platform..., Mode of Purchase..., Enduser... - Global Forecast to 2025, Feb 2021. Markets and Markets.

Precision Medicine Market By Technology…, By End-Use…, By Application, By Region Forecasts to 2027, Aug 2020. Emergen Research.

Precision Medicine Market - Forecasts from 2021 to 2026, Jan 2021. Research and Markets.

Rahman, L. Why we need to take AI in drug discovery seriously, May 2020. European Pharmaceutical Manufacturer.

Rezaei, M.; Bazaz, S.R.; Zhand, S.; Sayyadi, N.; Jin, D.; Steward, M.P.'; Warkiani, M.E. Point of Care Diagnostics in the Age of COVID-19. *MDPI Diagnostics* 11, 1 (2021). https://doi.org/10.3390/diagnostics11010009

Sennaar, K. Machine Learning in Genomics – Current Efforts and Future Applications, Nov 2019. Emerj Artificial Intelligence Research.

Stokstad, E. Major U.K. genetics lab accused of misusing African DNA, Oct 2019. Science Magazine.

Tucci, S. & Akey, J.M. The long walk to African genomics. *Genome Biol* 20, 130 (2019). https://doi.org/10.1186/s13059-019-1740-1

Ugalmugle, S.; Swain, R. Genetic Testing Market Size By Test Type..., By Application..., Industry Analysis Report, Regional Outlook, Application Potential, Competitive Market Share & Forecast, 2020 – 2026, Feb 2020. Global Market Insights.

Ugalmugle, S.; Swain, R. Precision Medicine Market Size By Technology…, By Application…, By End-Use…, Industry Analysis Report, Regional Outlook, Application Potential, Competitive Market Share & Forecast, 2020 – 2026, Feb 2020. Global Market Insights.

Williams, J. Exploring the World's Largest Biobanks, 2018. Global Engage.

World Health Organization. Chapter 3: Overview of the diagnostics market // Ensuring innovation in diagnostics for bacterial infection: Implications for policy [Internet], 2016. NCBI.

APPENDIX A. Supplementary Figures.

Figure A1. Point-of-care (POC) global market size estimates, 2020 - 2028.

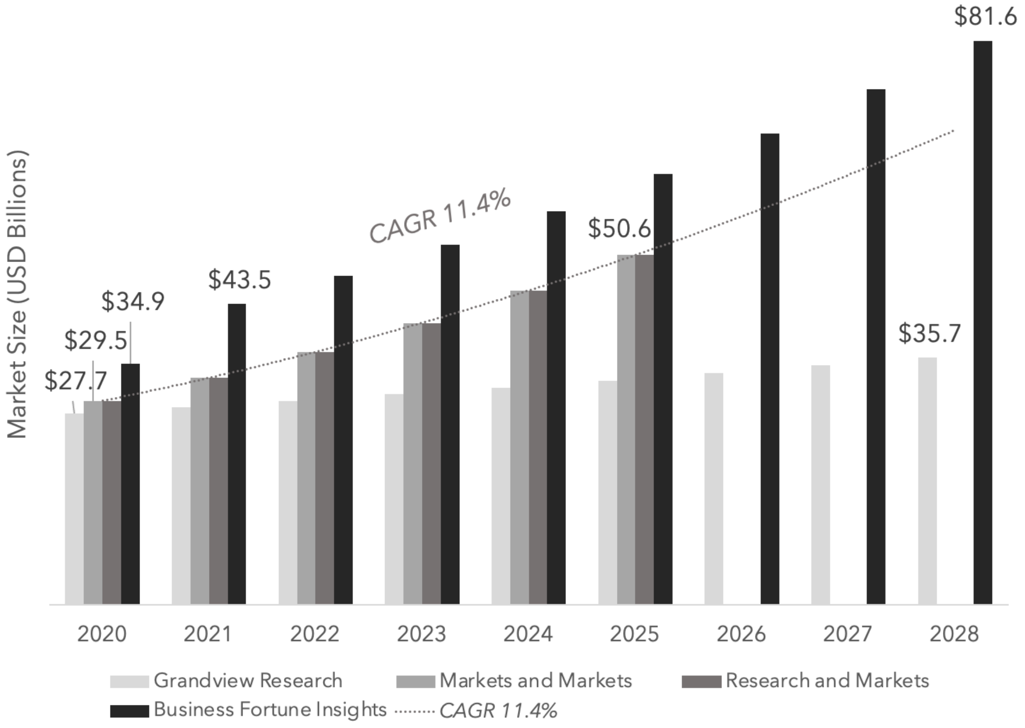


Figure A2. *In vitro* diagnostics (IVD) global market size estimates, 2019 - 2027.

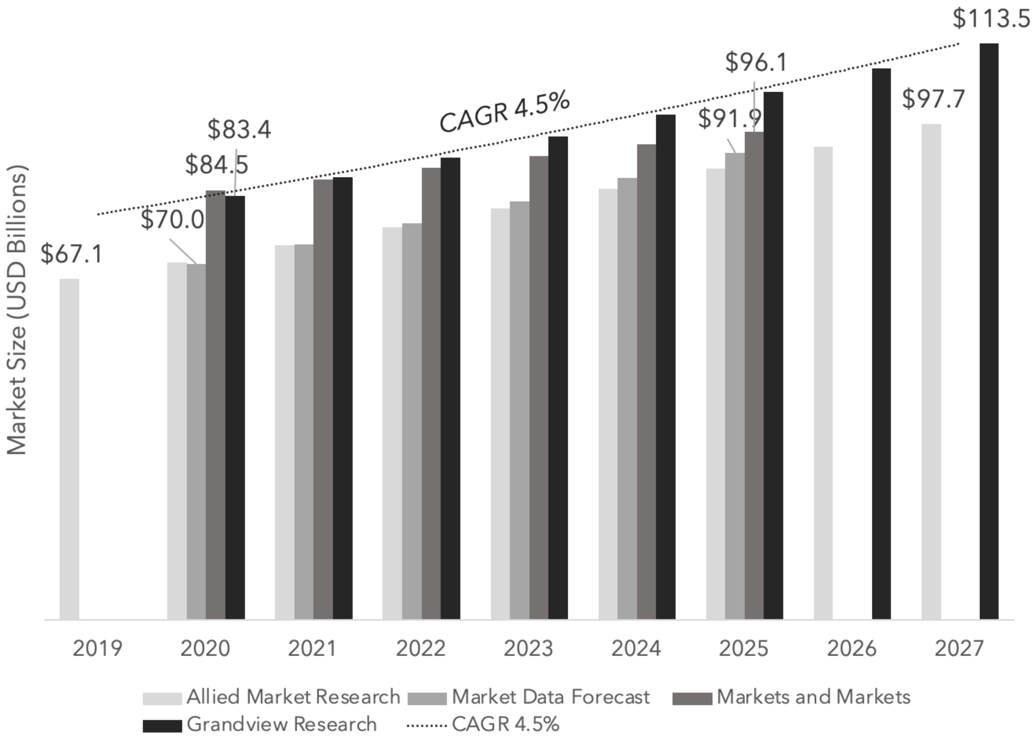


Figure A3. Molecular diagnostics global market size estimates, 2020 - 2028.

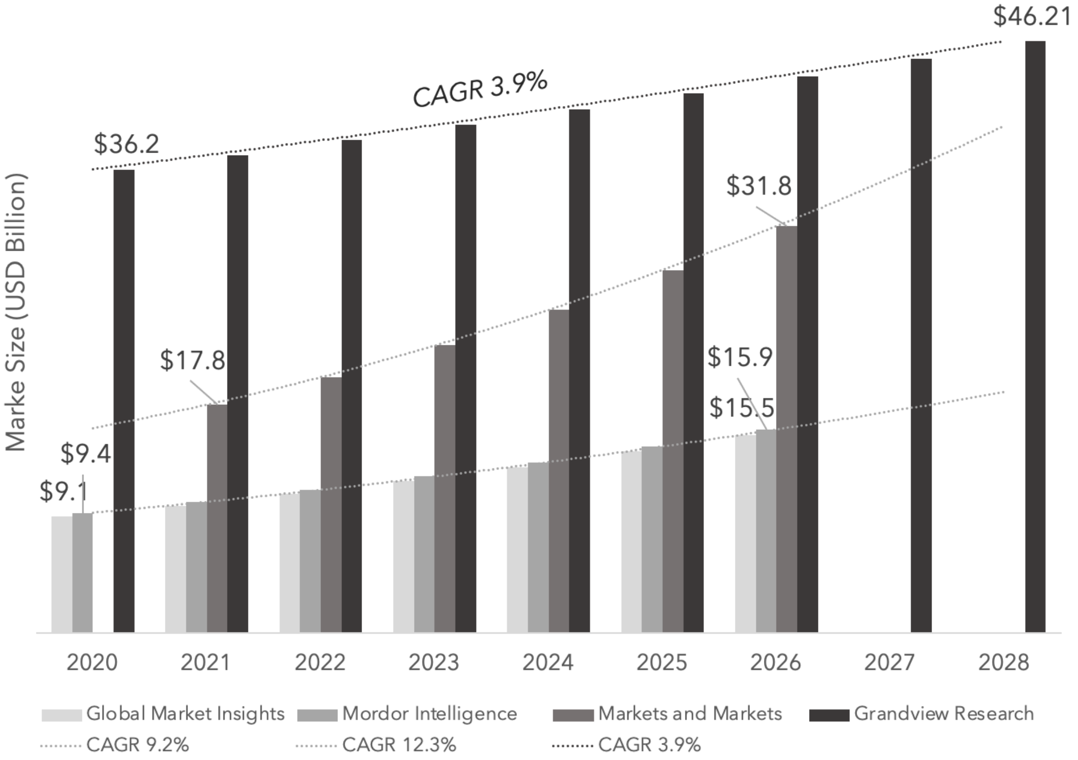


Figure A4. Genetic testing global market size estimates, 2019 - 2027.

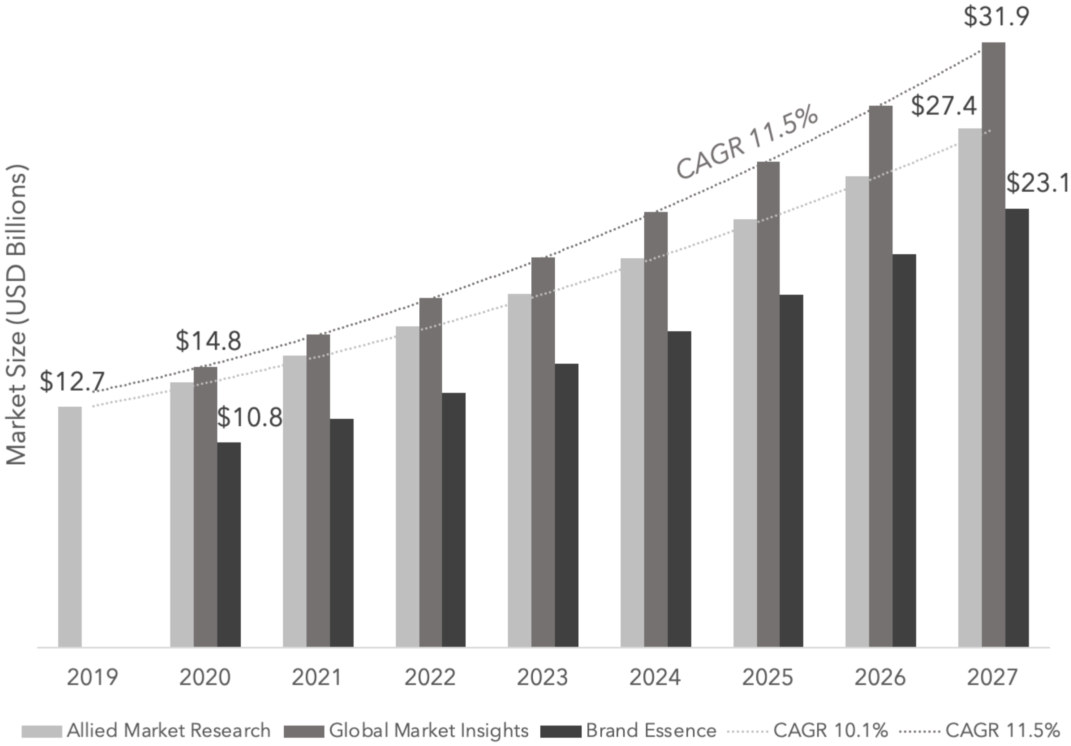


Figure A5. Big Pharma Private Market Investments, 2015-18 (CBInsights). *Interactive figure.*

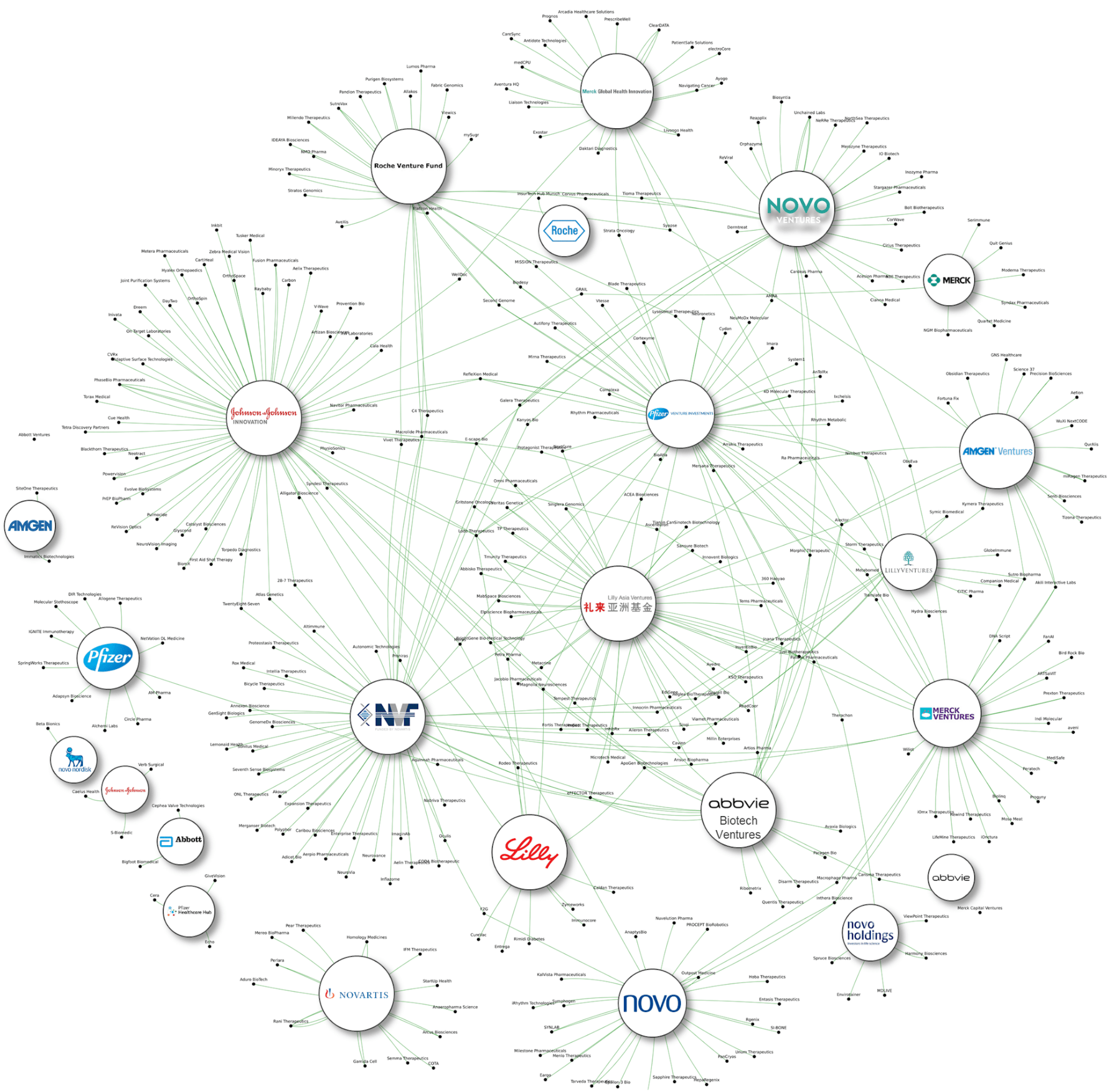


Figure A6. Investment Digest: The 2020 Overview of Pharmaceutical Artificial Intelligence Sector (Deep Pharma Intelligence). *Interactive Figure*.

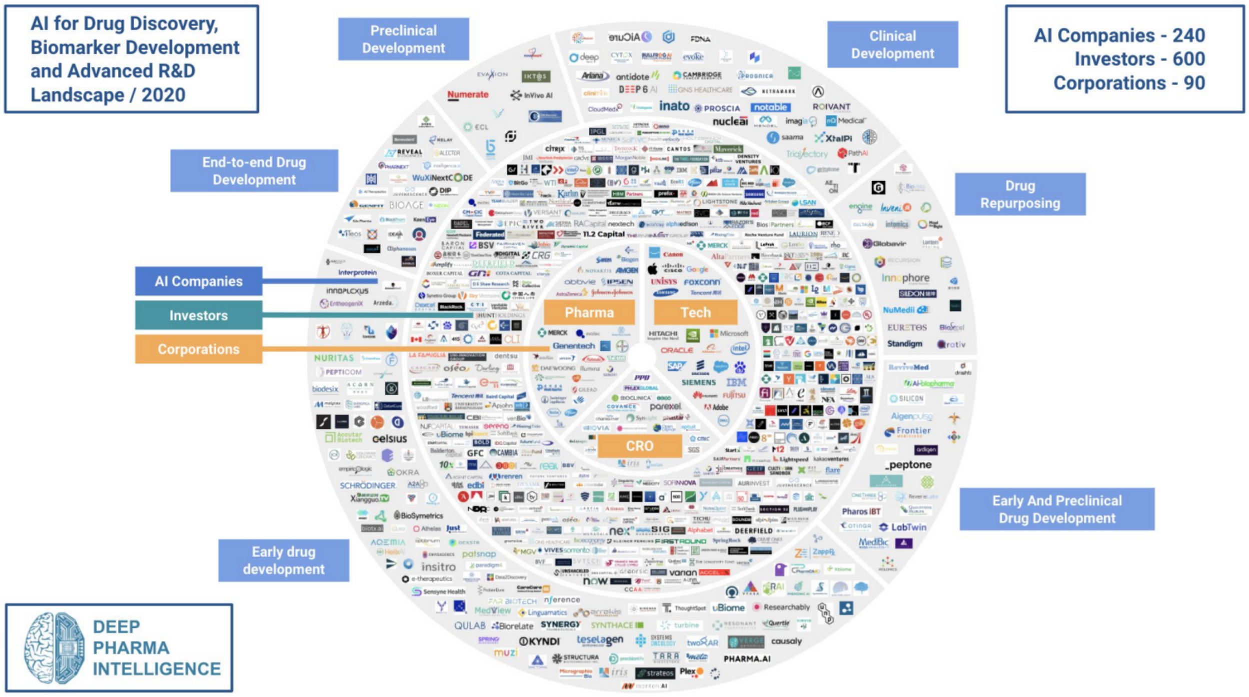


Table A1. Tech-driven Companies in the Genomics Market.

Company	Founded	Funding (USD)	Information
Benevolent AI	2013 London	$292 mn	Apply AI, machine learning and other advanced technologies to reinvent the ways drugs are discovered and developed. Mission is to re-engineer drug discovery and deliver life-changing medicines for patients.
Berg	2006 Framington, MA	NA	Clinical-stage, AI-powered biotechnology company using a proprietary intelligence platform: Interrogative Biology® AI coupled with patient biology to accelerate clinical identification and pursue promising therapeutic targets -- faster discovery & development of treatments, more effective precision treatments for individuals, and a reduction in costs to our healthcare systems.
Deep Genomics	2015	$56.7 mn	AI-powered Discovery Platform

	Toronto		The AI platform utilizes biology, from disease biomarkers all the way down to the level of DNA, RNA and the molecular machinery of the cell, to rapidly detect the best drug candidates
Exscientia	2012 Scotland	$103.7 mn	Pharmatech company using Artificial Intelligence (AI) to drive drug discovery; uses the power of AI with the experience of seasoned drug hunters.
Healx	2014 Cambridge	$67.9 mn	AI-powered, patient-inspired biotech specialising in rare diseases -- leverages public and proprietary biomedical data and features the world’s leading knowledge graph for rare diseases, combined with patient insights & drug discovery expertise
Insilico Medicine	2014 Hong Kong	$51.3 mn	Biotech company that combines genomics, big data analysis, and deep learning for *in silico* drug discovery
Insitro	2018 San Francisco, CA	$243 mn	Data-driven drug discovery and development company. Uses machine learning (ML) and high-throughput biology to transform the way that drugs are discovered and delivered to patients
Lantern Pharmaceuticals	2013 Dallas, TX	$8.7 mn	Advancing & accelerating the development of precision oncology therapeutics using A.I., genomics and machine learning to analyze biomarker signatures to develop personalized drug therapies Response Algorithm for Drug Positioning & Rescue (RADR) platform

Other key players in the market include: AI Therapeutics, Inc., Freenome Holdings, Inc., BioSymetrics, Engine Biosciences Pte. Ltd., Clover Therapeutics, Coral Genomics, Cyclica Inc., Desktop Genetics Ltd., DNAnexus, Data4Cure, Inc, Empiric Logic, Cambridge Cancer Genomics Limited, PrecisionLife Ltd, SOPHiA GENETICS, Inc., Verge Genomics, Fabric Genomics, Inc., LifebitAI, Genoox, Ltd., Congenica Ltd, Predictive Oncology, Ares Genetics GmbH, Emedgene, Microsoft (Project Hanover), CureMatch, Inc., Trace Genomics, WhiteLab Genomics, and WuXi Nextcode Genomics among others.

Table A2. Non-exhaustive list of established companies within the genomics market.

Bayer, **Pfizer**, **AstraZeneca**, Takeda, GSK, Janssen, Boehringer Ingelheim, Roche, Sanofi, **Merck**, **Amgen**, **Bristol-Myers Squibb**, Abbvie, Novo Nordisk, **Novartis**, **Johnson & Johnson**, Celgene, Evotec, Astellas, Aum Biosciences, IVA, OncXerna, Immuno Precise, Yuhan, Almirall, Hansoh Pharma, GC, Illumina

Company	Founded	Size / Funding	Description / Activity
Abbott Laboratories	1888 Illinois, USA	US$3.19 bn in revenue, 2019	Multinational medical devices and healthcare company engaged in pharmaceuticals and manufacturing healthcare products.
Almac Group Ltd.	2001 Ireland	US$368 mn net worth, 2018	Provides medical and wide range of pharmaceutical products -- delivers to markets throughout the U.K.
Amgen Inc.	1980 California, USA	US$7.84 bn net income, 2019	Multinational biopharmaceutical company -- one of the world's largest independent biotech companies
ANGLE plc	1994, United Kingdom	US$2.67 mn market cap, 2021	World leading liquid biopsy company with sample-to-answer solutions. ANGLE's patent-protected platforms include an epitope-independent circulating tumor cell (CTC) harvesting technology and a downstream analysis system for cost effective, highly multiplexed analysis of nucleic acids and proteins.
Astellas Pharma Inc			
AstraZeneca PLC			
ASURAGEN INC.			
Bayer			
Bio-Rad Laboratories, Inc.			
bioMérieux SA.			
Bristol-Myers Squibb Company			
Cardiff Oncology			
CETICS Healthcare Technologies GmbH			

Danaher Corporation			
Eli Lilly and Company Limited			
Epic Sciences, Inc.			
F. Hoffmann-La Roche Ltd			
GE Corporation			
Gilead Sciences, Inc.			
GlaxoSmithKline Plc			
Illumina, Inc.			
Intomics A/S			
Johnson & Johnson Company			
Konica Minolta, Inc.			
Laboratory Corporation of America MDxHealth, Inc.			
Menarini Silicon Biosystems, Inc.			
Merck & Co., Inc.			
Myriad Genetics, Inc.			
Novartis AG.			
Oracle Corporation			
Partek, Inc.			
Pfizer, Inc.			
QIAGEN N.V.			
Quest Diagnostics Inc.			

Randox Laboratories Ltd.			
Sanofi SA			
Sysmex Corporation			
Teva Pharmaceuticals Industries Ltd.			
Thermo Fisher Scientific, Inc.			

APPENDIX B. Key Terminology, Definitions, and Abbreviations.

Artificial Intelligence (AI)

Artificial intelligence refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving. There are four types of AI: Reactive Machines, Limited Memory, Theory of Mind, and Self-Awareness. AI can be helpful when implemented in conjunction with enormous datasets, which is why it pairs well with genome sequencing data, which can be billions of data points. See Machine Learning for more information.

Further reading: Understanding the Four Types of Artificial Intelligence, The State of AI in 2020

Bioinformatics

An interdisciplinary field of biology and computer science that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. The biological data is most often DNA and amino acid sequences – this data is acquired, stored, analyzed, and disseminated.

Bioinformatics utilizes computer programs for a variety of purposes, e.g. determining gene and protein functions, establishing involuntary relationships, and predicting 3-D shapes of proteins.

Further reading: Bioinformatics (NIH)

DNA base pair (bp)

The four nucleotide bases that comprise DNA bind together in a specific pattern: Adenine always binds to Thymine (A–T) and Cytosine always binds to Guanine (C–G). These two pairs of bases form the double helical structure of DNA.

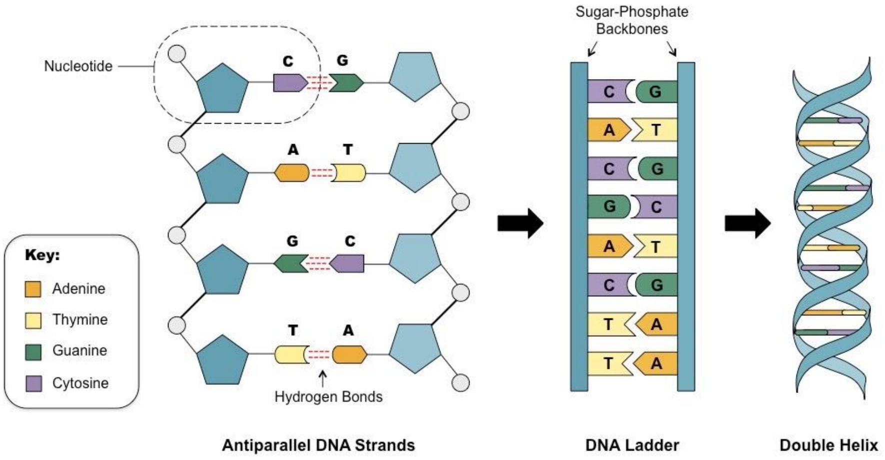


Figure B1. Make-up of DNA, including base pairing and double-helical structure.

DNA Microarray
Also known as a **gene chip**, **DNA chip**, or **biochip**, these are a collection of microscopic DNA samples on a small surface, no larger than a postage stamp. They are used to measure the expression level of a large number of genes simultaneously or genotype multiple regions of a genome. Each DNA spot contains picamoles of a specific DNA sequence, which are known as probes.

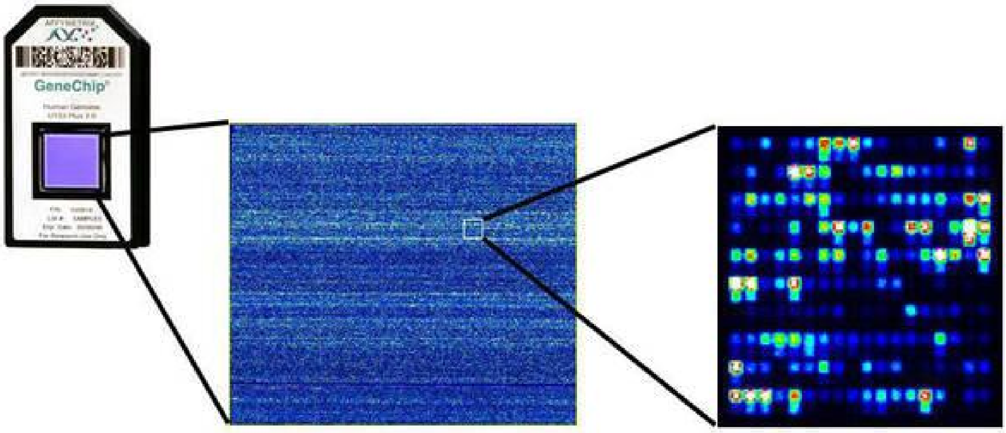


Figure B2. Gene chip and stain.

Gene chips can be used in lieu of full genome sequencing for more pointed genetic testing (a whole genome sequence is often unnecessary) due to the lower cost. These chips allow for a quick panel of genomic testing against common variants with known phenotypes.

Further reading: DNA Microarray Technology Fact Sheet (NIH)

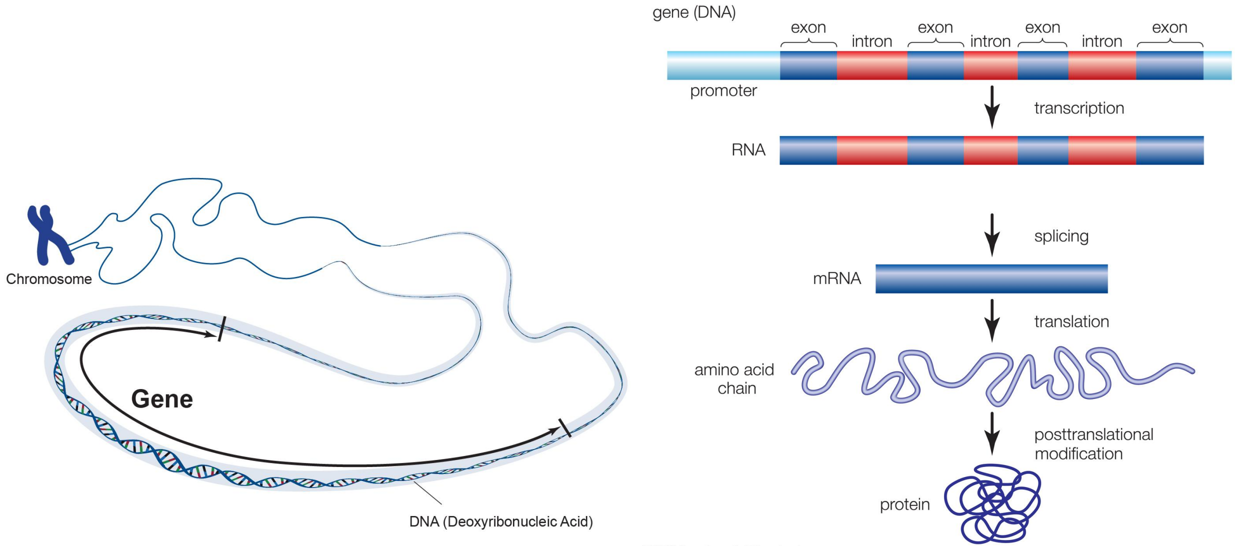
DNA Sequencing
The process of determining the nucleic acid sequence (the order of nucleotides in DNA) of a biological sample. It includes any method or technology that is used to determine the order of the four bases: adenine, guanine, cytosine, and thymine. The human genome contains about 3 billion base pairs that spell out the instructions for making and maintaining a human being.

The sequence shows the type of genetic information that is stored in a particular DNA segment. For example, scientists can use sequence information to determine which segments of DNA contain genes and which segments carry regulatory instructions, turning genes on or off. In addition, and importantly, sequence data can highlight changes in a gene that may cause disease.

Further reading: DNA Sequencing Fact Sheet (NIH), The Cost Of Sequencing a Human Genome (NIH)

Gene

Genes are units of inheritance or strings of genetic information that are found within our chromosomes, of which humans have 46. More specifically, they are a sequence of nucleotides in DNA or RNA, that encodes the synthesis of a gene product, either RNA or protein. Gene expression occurs when the cell reads and interprets the genetic code and adds the appropriate amino acids indicated in the DNA to build specific proteins.



© 2013 Encyclopædia Britannica, Inc.

Figure B3. Gene expression.

Phenotype

A set of observable characteristics of an individual resulting from the interaction of its genotype with its environment. This includes traits as benign as eye color or height to diseased phenotypes such as a genetic predisposition to breast cancer or Parkinson's disease.

It is important to distinguish phenotype from *genotype*, which is the genetic cause of the observable trait. Notably, phenotype does not always directly correlate with a given genotype, as environmental factors can have equal or even more influence on the phenotype.

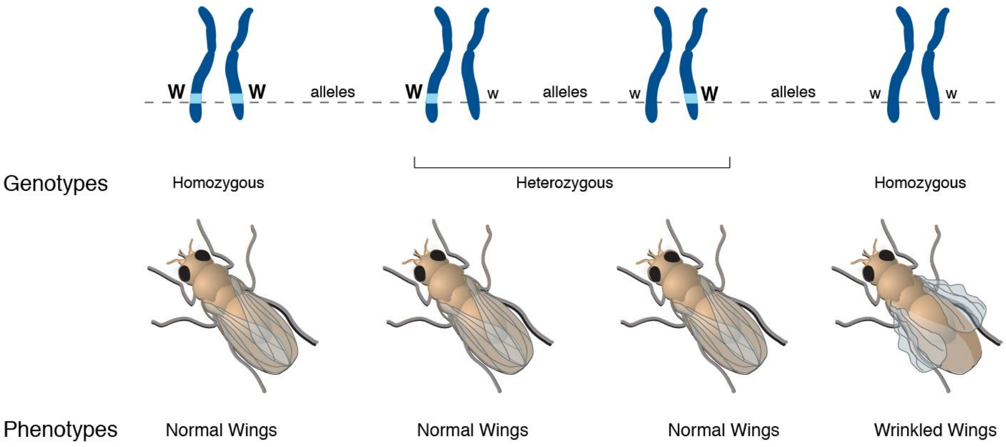


Figure B4. Example of various genotypes and the resulting phenotypes of fruit flies.

Genetic Variants

Changes in the DNA sequence compared to the reference sequence. A disease-causing variant is called a **mutation**. Genetic variants are classified according to size, nature and location of varying segments of DNA relative to a reference genome:

- *Single Nucleotide Variants (SNVs)*, *Single Nucleotide Polymorphisms (SNPs)*, and *Point Mutations* are discrete single base pair substitutions. (See SNP for more info)
- *Multi-nucleotide variants (MNVs)* are multiple SNVs of a few base-pairs in length.
- *Indels* are small insertions or deletions in the variant genome, typically between 1-50 bp in length.
- *Frameshift mutations* are indels whose lengths are not multiples of 3.

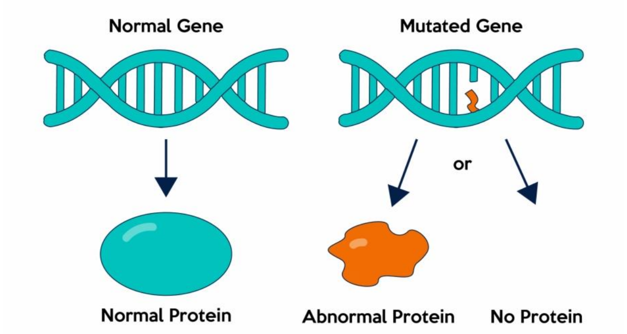


Figure B5. Comparison of normal versus mutated gene outcomes.

Genome-Wide Association Studies (GWAS)

An approach used in genetics research to associate specific genetic variations with particular diseases. The method involves scanning genomes from many different people and looking for genetic markers that can be used to predict the presence of a disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease.

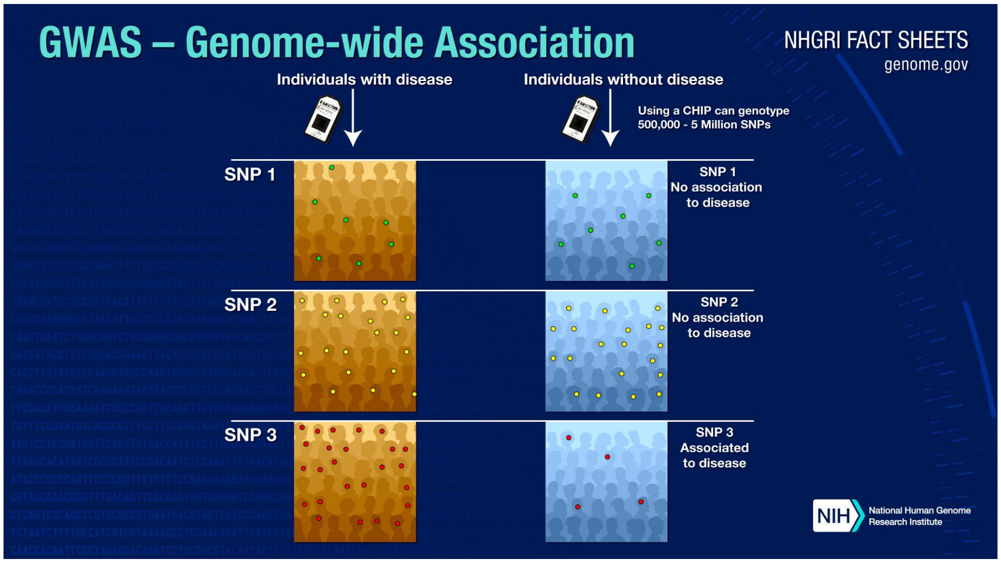

GWAS – Genome-wide Association
NHGRI FACT SHEETS
genome.gov
Individuals with disease
Individuals without disease
Using a CHIP can genotype 500,000 - 5 Million SNPs
SNP 1
SNP 1 No association to disease
SNP 2
SNP 2 No association to disease
SNP 3
SNP 3 Associated to disease
NIH National Human Genome Research Institute

Figure B6. How GWAS are conducted.

Further Reading: Genome-Wide Association Studies Fact Sheet (NIH)

Genomics
Genomics is an interdisciplinary field of biology focusing on the structure, function, evolution, mapping, and editing of genomes. A major part of genomics is determining the sequence of molecules that make up the genomic deoxyribonucleic acid (DNA) content of an organism. A genome, consequently, is an organism's complete set of DNA, including all of its genes.

Genomics includes the scientific study of complex diseases such as heart disease, asthma, diabetes, and cancer because these diseases are typically caused more by a combination of genetic and environmental factors than by individual genes.

Further Reading: [A Brief Guide to Genomics (NIH)]()

Machine Learning (ML)
ML is when a computer uses data to learn a model for predicting a value, where the relationship between the data and the value is not explicitly provided.

The data is composed of instances (i.e. samples) and features (i.e. independent variables) that describe those instances. For example, if our instances are *genes*, features describing those genes could be the GC content, the presence or absence of a specific functional domain, or its level of conservation across species. If the values being predicted are not known a priori for any instance, then unsupervised ML approaches (e.g. clustering) can be applied to extract previously unknown patterns. If the values being predicted are known for some instances, these values are referred to as labels and one can learn from these labels using a supervised ML approach.

- *Classification problem*: if the known labels are categorical (e.g. is the gene upregulated or downregulated?)
- *Regression problem*: if the labels are continuous (e.g. gene expression levels)

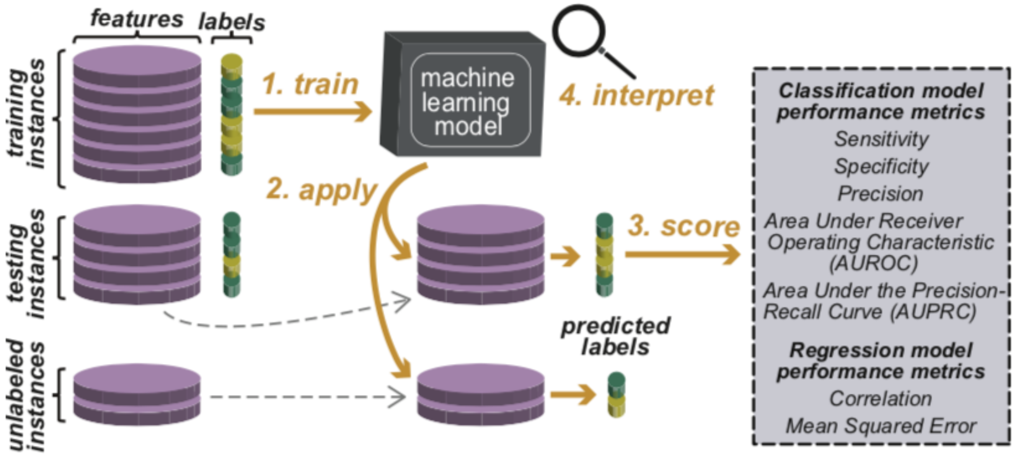


Figure B7. Overview of Machine Learning.

Further reading: [Machine Learning in Genomics](), [7 Applications of ML in Pharma and Medicine]()

Next Generation Sequencing (NGS)
A technology for determining the sequence of DNA or RNA to study genetic variant association with diseases or other biological phenomena. What makes it unique is it enables scientists to analyze the entire human genome in a single sequencing experiment. It has ultrahigh throughput, scalability, and speed, making sequencing, e.g. large-scale whole genome sequencing (WGS), accessible and practical.

Further reading: Introduction to NGS (Illumina), What is Next-Generation Sequencing (NGS)? (Thermo Fisher)

Pharmacogenomic testing
The practice of utilizing genetic testing (normally via microarrays) to determine the most effective drug for patients based on their genomic makeup. It is considered a vital component of precision medicine as a way to compound custom medication.

A good illustration of its utility would be a cancer treatment example – depending on the cancer, there are often multiple treatment routes, but some are more effective than others depending on the individual. Normally patients receive a treatment that performs the best across large populations of people, however, because cancer cells multiply rapidly, it is detrimental to the patient if they first receive a treatment that ultimately proves ineffective.

Further reading: What is Pharmacogenomics? (NIH)

Reference Genome
A digital nucleic acid sequence database, assembled by scientists as a representative example of the set of genes in one idealized individual organism of a species. They are typically used as a guide on which new genomes are built, enabling them to be assembled more quickly.

Because they are assembled from the DNA sequences of various donors, they do not accurately represent the set of genes of any single individual organism. The most accurate and up-to-date version of the human genome is CRCh38, released in 2014 – Build 38 is updated four times a year.

Further reading: Reference genome: defining human difference (Genomics Education Programme)

Single Nucleotide Polymorphism (SNP)
See “Genetic Variants.” SNPs - pronounced "snips" - are common, but minute, variations that occur in the human genome at a frequency of one in every 300 bases. That means 10 million positions out of the 3 billion base-pair human genome have common variations. These variations can be used to track inheritance in families and susceptibility to disease, so many

scientists are working to develop a catalogue of SNPs as a tool to use in their efforts to uncover the causes of common illness like diabetes or heart disease.