JHIT TÜRKİYE SAĞLIK ENSTİTÜLERİ BAŞKANLIĞI DERGİSİ
JOURNAL OF HEALTH INSTITUTES OF TÜRKİYE

REVIEW / DERLEME

J Health Inst Turk 2025;8(3):62-69
DOI: 10.4274/jhit.galenos.2025.87597

The Türkiye National Genome and Bioinformatics Project: An Overview

Türkiye Ulusal Genom ve Biyoinformatik Projesi: Genel Bir Bakış

Fatma Duygu Özel Demiralp[1], Emine Altun[1], Salih Berkay Berkcan[1], Ayhan Demir[2], Melike Efeer[1], Ezgi Göksoy Oruç[1], Tuğçe Gültan[1], Mehmet Ali Kök[1], Tuba Özbay[1], Adem Özleyen[1], Saniye Elvan Öztürk[1], Tunç Tuncel[1], Büşra Ahata[2], Gizem Turaç Karakurt[2], Tuğçe Kan Mutlu[2], Hatice Cemre Ünver[1], Rabia Yılmaz Öztürk[2]

[1]Health Institutes of Türkiye, Türkiye Biotechnology Institute, Aziz Sancar Research Center, Ankara, Türkiye
[2]Health Institutes of Türkiye, Türkiye Biotechnology Institute, İstanbul, Türkiye

ABSTRACT

The Turkish National Genome and Bioinformatics Project is a large-scale initiative aimed at advancing genomic research and precision medicine in Türkiye. This project focuses on whole genome sequencing of diverse population samples, conforming to international standards and utilizing validated next-generation sequencing technologies. Genomic DNA extraction and sequencing are performed using automated high-throughput platforms to ensure accuracy and scalability.

The project, following the completion of the Türkiye Genome Project phase in 2018, continues with the COVID-19 Genome Project, covering healthy individuals, cancer, and undiagnosed diseases. As of March 2025, biological samples from a total of 2.500 individuals from the Turkish population have been collected, with half of the samples sequenced and secondary bioinformatics analyses completed. Variants and frequencies obtained from the sequenced cohorts have been systematically structured and are being shared securely through the Türkiye Genome Project data sharing portal. This portal allows researchers to analyze genetic data within ethical guidelines and facilitates the enhancement of global scientific collaborations.

By integrating big data analytics and advanced bioinformatics pipelines, the project enhances the understanding of population-specific genetic variations, disease associations, and potential therapeutic targets. This initiative marks a significant step toward the implementation of genomic medicine in Türkiye and strengthens the nation's contribution to global advancements in personalized healthcare and precision diagnostics.

Keywords: Genome project, Turkish genome, whole genome sequencing (WGS), next-generation sequencing (NGS)

ÖZ

Türkiye Ulusal Genom ve Biyoinformatik Projesi, Türkiye'de genom araştırmalarını ve tıbbi ilerlemeyi hedefleyen geniş ölçekli bir girişimdir. Proje, farklı kohortların tüm genom dizilemesi üzerine odaklanmakta olup uluslararası standartlara uygun şekilde doğrulanmış yeni nesil dizileme teknolojilerinden faydalanmaktadır. Genomik DNA izolasyonu ve dizileme süreçleri, doğruluk ve ölçeklenebilirliği sağlamak amacıyla otomatik yüksek verimli platformlar kullanılarak gerçekleştirilmektedir.

Proje; 2018 yılında tamamlanan Türkiye Genom Projesi fazını takiben, COVID-19 Genom Projesi, sağlıklı bireyler, kanser ve tanısız hastalıklar kapsamında devam etmektedir. Proje kapsamında Mart 2025 itibarıyla, Türkiye popülasyonuna ait toplamda 2,500 bireyin biyolojik örnekleri toplanmış olup toplam örneklemin yarısı dizilenmiş ve ikincil biyoinformatik analizleri tamamlanmıştır. Dizilenen kohortlardan elde edilen varyantlar ve frekansları, sistematik olarak yapılandırılmıştır ve güvenli bir şekilde Türkiye Genom Projesi veri paylaşım portalı aracılığıyla paylaşılmaktadır. Bu portal, etik kurallar çerçevesinde araştırmacıların genetik verileri analiz etmesine ve küresel düzeyde bilimsel iş birliklerini artırmasına olanak tanımaktadır.

Büyük veri analitiği ve ileri biyoinformatik altyapıları ile entegre edilen bu proje, popülasyona özgü genetik varyasyonların, hastalık ilişkilerinin ve potansiyel terapötik hedeflerin daha iyi anlaşılmasına katkı sağlayacaktır. Bu girişim, Türkiye'de genombilimin uygulanmasına yönelik önemli bir adım niteliğinde olup ülkenin kişiselleştirilmiş sağlık hizmetleri ve hassas tıp alanındaki küresel ilerlemelere katkısını güçlendirmektedir.

Anahtar Kelimeler: Genom projesi, Türkiye genomu, tüm genom dizileme (WGS), yeni-nesil dizileme (NGS)

Corresponding Author/Sorumlu Yazar: Prof. Fatma Duygu Özel Demiralp,
Health Institutes of Türkiye, Türkiye Biotechnology Institute, Aziz Sancar Research Center, Ankara, Türkiye
E-mail: Duygu.Ozeldemiralp@tuseb.gov.tr

ORCID ID: orcid.org/0000-0002-1798-7951

Received/Geliş Tarihi: 09.04.2025 **Accepted/Kabul Tarihi:** 26.12.2025 **Publication Date/Yayınlanma Tarihi:** 31.12.2025

Cite this article as/Atıf: Özel Demiralp FD, Altun E, Berkcan SB, Demir A, Efeer M, Göksoy Oruç E, et al. The Türkiye national genome and bioinformatics project: an overview. J Health Inst Turk. 2025;8(3):62-69

INTRODUCTION

The question that arose following the discovery of DNA in 1953 was whether the alignment of these consecutively arranged nucleotides could be sequenced. Fortunately, the sequencing of small DNA fragments commenced during the 1970s, facilitated by the advent of the Sanger sequencing technique. After small-scale sequencing such as bacteriophage genome, *Haemophilus influenzae* was the first organism to have its entire genome comprehensively sequenced (1). Following the success achieved in genome sequencing, attention subsequently shifted toward the sequencing of the human genome.

Genome projects have profound significance in various fields, particularly in medicine, agriculture, and environmental science. Currently, significant number of genome projects covering various regions and targets around the world have been completed or are ongoing (Table 1). The aims of these projects, which are programmed according to needs, can be summarized as advancing personalized medicine, improving healthcare, optimizing agricultural practices, and improving our understanding of biodiversity. From large-scale international initiatives such as the Human Genome Project (HGP) to more region-specific programs such as the Saudi Human Genome Program and the Turkish National Genome and Bioinformatics Project, each initiative contributes uniquely to the scientific community. These projects aim to gain valuable insights into genetic variation, disease mechanisms, and environmental interactions by mapping and sequencing genomes across different populations and species. Through these efforts, genomic research continues to shape the future of medicine, agriculture, and biodiversity conservation.

The HGP started in the 1990s. The first draft was released in 2001 (2) and completed in 2003 (3). The HGP provided a comprehensive map of the human genome, enabling breakthroughs in personalized medicine, gene therapy, and enhanced understanding of hereditary diseases (4). This project has also fostered global cooperation, making genomic data freely accessible to researchers worldwide, which is crucial for ongoing research and development (5).

Rare Disease Genome Projects

Among rare diseases, which are important public health problems, 10.000 different rare diseases have been identified to date, and it is known that they affect more than 300 million people worldwide (6). It has been observed that these diseases, which show a high rate of genetic transmission (approximately 80%), mostly emerge in childhood. It is understood that this rate increases especially in countries with a history of consanguineous marriage. The reason for this is known to be the increase in the prevalence of diseases with autosomal recessive transmission.

The rare disease genome projects aim to enhance the understanding and diagnosis of rare genetic disorders through advanced genomic technologies and collaborative efforts. These projects focus on identifying disease-causing genetic variants, improving diagnostic success rates, and addressing inequities in genomic research. They employ innovative bioinformatics strategies, variant prioritization methods, and large-scale data analysis to achieve these goals.

The Rare Genomes Project (USA) employs genome sequencing to identify causal variants, using computational models to prioritize variants based on quality scores, allele frequency, and phenotype. This approach has led to the discovery of novel diagnostic variants and disease-gene candidates (7). Another rare disease project is the Solve-RD project which is a major European initiative. This project has centralized and reanalyzed genetic datasets, identifying disease-causing variants in over 700 rare disease families. This project has developed new methods to detect unknown genetic variants and utilized long-read sequencing to diagnose previously undiagnosed families (8).

Data related to rare diseases are also emerging from large-scale genome projects that do not directly focus on rare diseases. For instance, the 100.000 Genomes Project (100kGP) has applied an analytical gene burden framework to discover 88 novel rare disease-gene associations, potentially diagnosing 456 previously undiagnosed cases. This highlights the clinical impact of large-scale statistical approaches in discovery of novel variants that are responsible for rare diseases (9). Structural variants, including inversions, have been analyzed in 33.924 families, revealing their role in rare diseases and resolving complex diagnostic cases (10). Genome projects that have been conducted and are ongoing have shown that with a unified data infrastructure, collaborative data analysis, and long-term storage of genomic data, it becomes easier to improve diagnostic methods for undiagnosed patients and develop therapeutic drugs for rare diseases.

Cancer Genome Projects

Cancer genome projects, especially the 100kGP, have significantly advanced the understanding and application of WGS in oncology. The aim of 100kGP is to integrate genomic data into clinical practice to provide and develop more specific recommendations for cancer patients and develop treatment strategies. With the 100kGP, it was found that there were distinct pathogenic variants (PV) between European and non-European patients. In particular, 4.6% of South Asian patients

Table 1. Comprehensive overview of global human genome projects

Project name	Status	Objectives	Country	Websites
Human Genome Project	Completed (2003)	To map and sequence the entire human genome, providing foundational knowledge for genetic research and personalized medicine.	International	https://www.genome.gov/human-genome-project
Human Microbiome Project	Completed (2007-2016)	Study microbial communities and their health implications.	United States	https://hmpdacc.org/
1000 Genomes Project	Completed (2015)	To create a detailed catalog of human genetic variation by sequencing genomes from diverse global populations.	International	https://www.internationalgenome.org/
Cancer Genome Atlas	Completed (2006-2018)	Characterize genomic alterations in over 30 cancer types.	United States	https://portal.gdc.cancer.gov/
deCODE Genetics Project	Ongoing (since 1996)	Explore genetic variations and their implications for health.	Iceland	https://www.decode.com/
Canadian Genomics Enterprise	Ongoing (since 2000)	Support genomic research for advancements in health, agriculture, and environment.	Canada	https://genomecanada.ca/
Personal Genome Project China	Ongoing (since 2005)	To provide ethical alternatives for problematic human subject consent and to test novel technologies to collect data on genomes, environments and traits.	China	http://pgpchina.org/
UK Biobank	Ongoing (since 2006)	To compile extensive genetic, health, and lifestyle data to investigate disease determinants and promote personalized medicine.	United Kingdom	https://www.ukbiobank.ac.uk/
International Cancer Genome Consortium	Ongoing (since 2008)	Map genomic abnormalities in diverse cancer types.	Global	https://www.icgc-argo.org/
The African Genome Variation Project	Ongoing (since 2010)	Map genetic variation across African populations.	Africa	https://www.sanger.ac.uk/collaboration/african-genome-variation-project/
Tohoku Medical Megabank Project	Ongoing (since 2011)	To support personalized healthcare by analyzing genetic and environmental data from populations affected by the 2011 disaster.	Japan	https://www.megabank.tohoku.ac.jp/english/
IRDiRC	Ongoing (since 2011)	Promote global research collaboration for rare disease diagnosis and treatment.	Global	https://irdirc.org/
Saudi Human Genome Program	Ongoing (since 2013)	Map genetic mutations prevalent in the Saudi population.	Saudi Arabia	https://www.vision2030.gov.sa/en/explore/projects/the-saudi-genome-program
Genome Russia Project	Ongoing (since 2013)	Map genetic diversity among Russia's ethnic groups.	Russia	(-)
Genomics England	Ongoing (since 2013)	Sequence 100.000 genomes for rare diseases and cancer.	United Kingdom	https://www.genomicsengland.co.uk/
Korean Genome Project	Ongoing (since 2015)	To create a reference genome database for the Korean population to support precision medicine.	South Korea	https://kbds.re.kr/
China Precision Medicine Initiative	Ongoing (since 2016)	To sequence 10 million genomes to advance precision medicine tailored to the Chinese population.	China	(-)
GenomeAsia 100K	Ongoing (since 2016)	Sequence 100.000 genomes from diverse Asian populations.	Asia	https://www.genomeasia100k.org/
Australian Genomics Health Alliance	Ongoing (since 2016)	Integrate genomics into clinical practice for rare diseases and cancer.	Australia	https://www.australiangenomics.org.au/
Rare Genomes Project	Ongoing (since 2016)	Provide genome sequencing for individuals with undiagnosed rare diseases.	United States	https://raregenomes.org/

Table 1. Continued

Project name	Status	Objectives	Country	Websites
All of Us Research Program	Ongoing (since 2018)	To gather health data from over one million Americans to facilitate personalized medicine and health equity.	United States	https://allofus.nih.gov/
Türkiye National Genome and Bioinformatics Project	Ongoing (since 2018)	Analyze genetic diversity to understand diseases and develop personalized medicine strategies.	Türkiye	https://tgd.tuseb.gov.tr/en/
Earth BioGenome Project	Ongoing (since 2018)	Sequence genomes of all known eukaryotic species.	Global	https://www.earthbiogenome.org/
Solve-RD	Ongoing (since 2018)	Resolve diagnostic gaps for rare diseases using genomics.	Europe	https://solve-rd.eu/
Darwin Tree of Life Project	Ongoing (since 2019)	Sequence genomes of all eukaryotic species in the UK and Ireland.	United Kingdom	https://www.darwintreeoflife.org/
IndiGen Genome Project	Ongoing (since 2019)	Catalog genetic diversity to support personalized healthcare.	India	https://indigen.igib.in/

IRDiRC: International Rare Diseases Research Consortium

and 5.3% of African patients had PV, indicating the need for improved variant classification in various populations (11).

The use of whole genome sequencing (WGS) has enabled not only the detection of pathogenic variants related to lineage in cancer patients but also the implementation of personalized applications. In a study conducted by Leung et al. (12) it was shown that 59.2% of the participants received personalized clinical recommendations based on their genomic data. With the 100kGP, genomic data can be linked to understanding real health information to understand treatment outcomes and examine the long-term impact of genomic testing on patient care (13), enabling better survival rates analysis also allowing for patient-based clinical recommendations. Additionally, Shibata emphasizes the discovery of non-coding drivers and structural abnormalities through cancer genome sequencing, which may be surprising given their less understood role in cancer compared to coding regions (14).

Türkiye National Genome and Bioinformatics Project

The Türkiye National Genome and Bioinformatics Project has been established with the aim of elucidating molecular biological underpinnings of diseases that pose both significant social and economic burdens, including cancer, rare diseases, and coronavirus disease-2019. Additionally, comprehensive genome, transcriptome, and metagenome sequencing studies are being conducted with healthy volunteers to identify genomic variants specific to the Turkish population and to determine the frequencies of these variants.

The Türkiye National Genome and Bioinformatics Project adopts a vision parallel to large-scale genome initiatives conducted worldwide. For instance, the UK launched the 100kGP in 2013, aiming to sequence the entire genomes of 100.000 individuals. This project has been instrumental in uncovering the genetic basis of diseases, particularly rare diseases and cancer, achieving a diagnostic success rate of 25-35% for undiagnosed cases (15). Similarly, the United States' All of Us Research Program integrates personal health records with genomic information to advance personalized medicine approaches and has collected data from over one million volunteers (16). In Japan, the initiative on rare and undiagnosed diseases has developed a national model integrating genomic analysis for cases of rare diseases with undiagnosed conditions (17).

Rare diseases, identified as a global public health priority, are reported to occur more frequently in our country due to the high prevalence of consanguineous marriages (approximately 20-25%) (18). Therefore, the importance of conducting studies specific to our country, including analyses of variants unique to Türkiye, is increasingly recognized. Inspired by global examples, the Türkiye National Genome and Bioinformatics Project aims to increase diagnostic rates for rare diseases in Türkiye's population, characterized by its unique demographic and genetic structure. It also seeks to discover novel genetic variants and produces reliable, large-scale genomic data that contribute to international genomic research. Consequently, the project positions itself as a strategic initiative, providing a scientific foundation for advancing early diagnosis and the widespread implementation of personalized medicine within Türkiye's healthcare system.

In Türkiye, the diagnosis rate in the first 6 months was 69%, and almost 10% of the patients remained undiagnosed in terms of rare diseases (19). One of the most poignant aspects of undiagnosed rare diseases is their emergence during childhood. This high rate of undiagnosed cases underscores the necessity of incorporating national genome projects and

advanced genomic techniques into routine practices at a systemic level.

As emphasized within the 11th development plan of the Republic of Türkiye (2019-2023), which was published by the Presidency of the Republic of Türkiye in 2018, establishing a national genomic database for the early diagnosis and treatment of genetic diseases, as well as advancing personalized medicine applications, are among the primary priorities. Similarly, research projects focusing on rare diseases are being encouraged, and the development of biotechnological solutions is being targeted with the support of institutions such as the Health Institutes of Türkiye (TÜSEB) and the Scientific and Technological Research Council of Türkiye.

The Workflow of Türkiye Genome Project

WGS is a procedure that involves sequencing the DNA of collected samples while adhering to international standards, ensuring appropriate quality and depth in accordance with the principles of economic scale. This process is carried out using established and validated next-generation sequencing (NGS) technologies. The WGS process at the National Genome Center of Türkiye (TUGEM) encompasses the following key stages (Figure 1).

1. Accepting the Samples Taken According to the Project Acceptance Criteria at the Sequencing Center

Within the scope of the Türkiye National Genome and Bioinformatics Project, blood and tissue samples were collected from distinct cohorts during different phases of the study (Table 2). The peripheric blood samples were collected from patient groups and healthy volunteers at city hospitals, family health centers countrywide and General Directorate of Public Health of Türkiye. As of March 2025, biological samples (blood, tissue, serum, swab, etc.) have been collected from 2500 healthy volunteers and WGS process has been completed for a total of 1067 individuals from the Turkish population (Table 3).

2. Verifying the Quality of the Accepted Samples for Sequencing

Genomic DNA was isolated from blood and tissue samples using the QiaSymphony automation system (Qiagen), and the concentrations of the elutions were determined with Qubit Flex Fluorometer (Thermo Fisher) using Qubit dsDNA BR analysis kit (Invitrogen).

3. Conducting the Library Preparation (Wet Lab) Process by Using Robotic Technology

Illumina DNA PCR-Free Prep tagmentation beads and buffers, IDT for Illumina DNA/RNA UD indexes and Illumina DNA PCR-Free Prep purification beads and buffers kits were used for tagmentation, dual-index ligation and purification of the single stranded DNA libraries and this process was conducted using the Hamilton ML STAR automation system (Hamilton Company). Library quantification was carried out with Qubit Flex Fluorometer (Thermo Fisher Scientific) by Qubit ssDNA assay kit (Invitrogen).

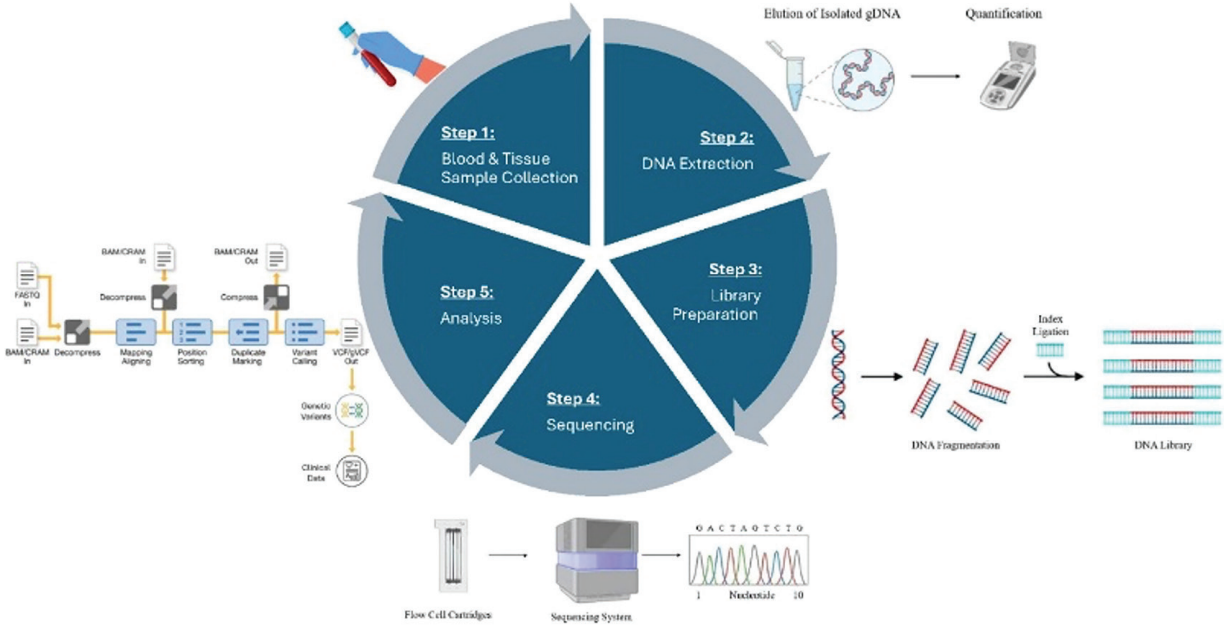

Figure 1. The workflow of WGS process at the TUGEM

WGS: Whole genome sequencing, TUGEM: National Genome Center of Türkiye

4. Sequencing the Prepared DNA Libraries by Using the Appropriate NGS System

Sequencing was performed using the S4 reagent kit v1.5 (300 cycles) on the NovaSeq 6000 sequencing system (Illumina) following the manufacturer's standard protocol, 151 base pair (bp) paired-end reads with an average insert size of approximately 450 bp were generated and an average coverage of 30x was targetted.

5. Processing Raw Sequencing Data and Secondary Bioinformatics Pipeline

Sequencing raw data obtained from TUGEM were recorded in data storage units operating on a local network and processed on the Illumina Dragen Bio-IT platform. Dragen DNA pipeline is used to perform bioinformatic analysis of the sequenced samples including mapping, aligning, QC check, sorting, small variant calling, copy number calling, and structural variant calling. DNA pipeline with Dragen targeted callers like HLA caller, star allele caller, HBA caller, RH caller, and LPA caller were also combined. Population-specific multi sample joint genotyping using genomic VCF files was assessed in healthy population. In rare diseases cohorts, pedigree based joint genotyping using trio data to better discover diseases related genomic variants, integrated with Dragen Expansion Hunter to identify disease related nucleotide repeats was determined.

To facilitate data accessibility, TÜSEB has developed the Türkiye Genome Data Sharing Portal, an intuitive online platform that allows researchers, clinicians, and patients to explore genomic data without requiring advanced bioinformatics expertise. The portal, accessible at https://tgd.tuseb.gov.tr provides interactive tools for genomic analysis, helping users visualize and query large-scale variant datasets in real time.

6. Storing the Obtained Raw Data in the Databank to be Analyzed with Backup

Efficient data storage and backup are essential for preserving the integrity of sequencing data. Raw sequencing data are securely stored in a local, internet-isolated data center, ensuring controlled access and data security. Redundant on-premises backup systems are implemented to prevent data loss and enable long-term retrieval. Adhering to standardized data management practices ensures the security, reproducibility, and scalability of large-scale genomic studies within a secure infrastructure.

Privacy and Data Security of Türkiye National Genome and Bioinformatics Project

Over the past decade, large-scale international consortia have leveraged NGS technologies to characterize the human genome, including its variations, dynamics, and associated pathologies. For example, the ongoing 100kGP, initiated by the British National Health Service, aims to sequence the genomes of 100.000 individuals within a clinical framework to establish a comprehensive population-scale genomic database with clinical annotations (20). The Cancer Genome Atlas research network has conducted extensive multi-omics analyses across major human cancer types, encompassing more than 11.000 patient samples (21).

Similarly, initiatives such as the Encyclopedia of DNA Elements and the Roadmap Epigenomics Project seek to construct high-resolution maps of chromatin organization and its variability. The 4D nucleome initiative further extends these efforts by investigating the spatiotemporal organization of the cellular nucleus in both physiological and pathological states.

The volume of data generated by these collaborative projects surpasses previous benchmarks by several orders of magnitude. Beyond large-scale consortia, an increasing number of smaller research initiatives are contributing to the accumulation of genomic data. For instance, the ArrayExpress database currently hosts over 10.000 records of research projects involving RNA sequencing data (22). Collectively, these datasets provide insights into the complex and heterogeneous nature of the human genome. In the future, the development of advanced computational approaches will be essential for structuring these vast datasets and integrating them with locally generated experimental data.

Table 2. Phases of Türkiye National Genome and Bioinformatics Project

Phase name	Status	Dates	Cohort
Pilot phase	Completed	2018-2019	Healthy volunteers
COVID-19 Genome Project	Completed	2022-2024	COVID-19 patients healthy volunteers
1000 Genomes Project	Ongoing	2024-	Healthy volunteers
Cancer Genome Project	Ongoing	2023-	Cancer patients
National Genome and Bioinformatics Project for Rare Diseases	Ongoing	2024-	Rare disease patients and their families
COVID-19: Coronavirus disease-2019			

Table 3. Current status in Türkiye National Genome and Bioinformatics Project

Donor type	Sample received	Sequenced
Healthy Turkish population	2500	1067
Cancer patients	77	67
Rare disease patients and family members	75	75

The potential for genomic datasets to enable individual identification has been previously demonstrated, underscoring the critical importance of privacy and confidentiality in genomic data management (23). In the case of the Türkiye National Genome and Bioinformatics Project, implementing stringent access controls and utilizing pseudonymization techniques, such as encoding individual metadata within barcodes, offer partial protection of subject identities. Furthermore, local system administrators, researchers and project staff are meticulously informed about data security principles and storage locations of sensitive data. TUGEM, located at the Aziz Sancar Research Center, Ankara, is protected 7/24 by fingerprint access and the individuals authorized to access are assigned by TÜSEB. All personnel are informed about the legally binding confidentiality of the patient data. Continuous security assessments are carried out to mitigate risks associated with potential data breaches and cyber threats. The closed circuit and fully-equipped technologic infrastructure enables the entire genome sequencing and data analysis of the Turkish population to be executed without the need for any patient sample or data output to leave the borders of TUGEM. The variant frequency data obtained in the Türkiye National Genome and Bioinformatics Project are accessible to researchers from all around the world via the "Türkiye Genome Project data sharing portal" (24).

Over the past ten years, most countries have made considerable investments in laboratory facilities, information technologies, and software infrastructure, leading to the widespread implementation of fundamental procedures at major genomic research centers worldwide. In the light of these technological achievements, in 2018, TÜSEB has undertaken the task of implementing the Türkiye Genome and Bioinformatics Project to analyze the molecular mechanisms of diseases, develop new diagnostic and therapeutic methods, and initiate individual-specific medicine studies within the scope of the tasks assigned to it in the 11th development plan. Accordingly, the TUGEM was included in the 2022 investment plan of the Presidency of the Republic of Türkiye, the Presidency of Strategy and Budget, and was established and put into operation within TÜSEB. Türkiye National Biobank, where biological samples obtained within the scope of the Türkiye National Genome and Bioinformatics Project is being stored, was established in 2020, and state of the art technological infrastructure have been allocated.

All types of donated biological samples are barcoded and labeled according to ISBT 128 standard (25). A donation identification number (DIN), which consists of the facility number, sample admission year, sample number, type and status, is appointed to the sample as soon as it is accepted. All biological samples are stored in Türkiye National Biobank with these labels which include a unique barcode and the DIN of the sample. This ensures the traceability for all types of biological samples and side-products related to the Türkiye National Genome and Bioinformatics Project, as well as the vigilance and surveillance tools to assist with data sharing and protection.

CONCLUSION

Despite significant advancements in the genomics field over the past three decades, the establishment of NGS data analysis workflows remains a complex challenge, particularly in core facility environments where computational infrastructure must accommodate the processing of data from thousands of samples annually. Although standardized protocols for fundamental data processing steps have emerged, numerous parameters still require optimization, imposing a substantial workload on researchers managing these pipelines.

TÜSEB continues to carry out research activities with high-level equipment using multiple methods, especially WGS. As the focus shifts from data acquisition to data utilization, there is an increasing demand for efficient data exploitation strategies. Global initiatives aiming to integrate genomic data from multiple sources necessitate substantial efforts in data organization and interconnectivity, yet these areas remain in their early developmental stages. Over the next decade, the integration of big data paradigms into genomic medicine is expected to drive substantial progress, ultimately enhancing medical outcomes such as developing diagnostic and therapeutic tools for cancer and rare diseases.

Acknowledgements

The Türkiye National Genome and Bioinformatics Project, led by Health Institutes of Türkiye, is funded by the Presidency of the Republic of Türkiye through the Presidency of Strategy and Budget under project number 2022K12-187826.

Footnotes

Authorship Contributions

Concept: F.D.Ö.D., T.Ö., A.Ö., T.T., B.A., G.T.K., T.K.M., R.Y.Ö., Design: F.D.Ö.D., T.Ö., T.T., Data Collection or Processing: E.A., S.B.B., A.D., M.E., E.G.O., T.G., M.A.K., T.Ö., A.Ö., S.E.Ö., T.T., H.C.Ü., Analysis or Interpretation: F.D.Ö.D., E.A.,

S.B.B., A.D., M.E., E.G.O., T.G., M.A.K., T.Ö., A.Ö., S.E.Ö., T.T., H.C.Ü., Literature Search: F.D.Ö.D., E.A., S.B.B., A.D., M.E., E.G.O., T.G., M.A.K., T.Ö., A.Ö., S.E.Ö., T.T., B.A., G.T.K., T.K.M., H.C.Ü., R.Y.Ö., Writing: F.D.Ö.D., E.A., S.B.B., A.D., M.E., E.G.O., T.G., M.A.K., T.Ö., A.Ö., S.E.Ö., T.T., H.C.Ü.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Adachi T, Kawamura K, Furusawa Y, Nishizaki Y, Imanishi N, Umehara S, et al. Japan's initiative on rare and undiagnosed diseases (IRUD): towards an end to the diagnostic odyssey. Eur J Hum Genet. 2017;25(9):1025-8.
2. All of Us Research Program Investigators. The "All of Us" research program. N Engl J Med. 2019;381:668-76.
3. 100,000 Genomes Project Pilot Investigators; Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, et al. 100,000 genomes pilot on rare-disease diagnosis in health care - preliminary report. N Engl J Med. 2021;385(20):1868-80.
4. Birney E. The International Human Genome Project. Hum Mol Genet. 2021;30(R2):R161-3.
5. Cipriani V, Vestito L, Magavern EF, Jacobsen JO, Arno G, Behr ER, et al. Rare disease gene association discovery from burden analysis of the 100,000 Genomes Project data. medRxiv [Preprint]. 2023:2023.12.20.23300294.
6. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995;269(5223):496-512.
7. García-Sancho M, Lowe J. The human genome project(s). In: A history of genomics across species, communities and projects. Medicine and Biomedical Sciences in Modern History. Cham: Palgrave Macmillan, 2023:79-116.
8. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013;339(6117):321-4.
9. Leung EYL, Robbins HL, Zaman S, Lal N, Morton D, Dew L, et al. The potential clinical utility of whole genome sequencing for patients with cancer: evaluation of a regional implementation of the 100,000 Genomes Project. Br J Cancer. 2024;131(11):1805-13.
10. Murugaesu N, Sosinsky A, Ambrose J, Cross W, Turnbull C, Henderson S, et al. Insights for precision healthcare from the 100,000 genomes cancer programme. Research Square. 2022.
11. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28(2):165-73.
12. Steward C. International human genome sequencing consortium nature 409, 860-921. International Human Genome Sequencing Consortium. 2001.
13. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45.
14. Nguyen T, Tallman S, Cho Y, Sosinsky A, Ambrose J, Thorn S, et al. Disparities in cancer genomics by ancestry in the 100,000 Genomes Project. medRxiv. 2024.
15. Pagnamenta AT, Yu J, Walker S, Noble AJ, Lord J, Dutta P, et al. The impact of inversions across 33,924 families with rare disease from a national genome sequencing project. Am J Hum Genet. 2024;111:1140-64.
16. Peplow M. The 100,000 genomes project. BMJ. 2016;353:i1757.
17. Shibata T. [Prospects on the whole cancer genome sequence project]. Gan To Kagaku Ryoho. 2022;49(7):713-8.
18. Stenton SL, O'Leary MC, Lemire G, VanNoy GE, DiTroia S, Ganesh VS, et al. Critical assessment of variant prioritization methods for rare disease diagnosis within the rare genomes project. Hum Genomics. 2024;18(1):44.
19. Steyaert WAR. From data to diagnosis: innovative bioinformatics strategies for diagnosing rare genetic diseases. Netherlands: Radboud University Press, 2025.
20. Tunçbilek E. Clinical outcomes of consanguineous marriages in Turkey. Turk J Pediatr. 2001;43(4):277-9.
21. Türkiye Genom Projesi Veri Paylaşım Portalı [Internet]. Erişim adresi: https://tgd.tuseb.gov.tr/tr/ (Erişim tarihi: 14 Mart 2025).
22. Sarkans U, Parkinson H, Lara GG, Oezcimen A, Sharma A, Abeygunawardena N, et al. The ArrayExpress gene expression database: a software engineering and implementation perspective. Bioinformatics. 2005;21(8):1495-501.
23. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546-58.