The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types

AI-powered analysis · Generated June 14, 2026 · simplify-research.org

■ Paper Summary

This paper presents the Genome Sequence Archive (GSA) family, a comprehensive data repository system developed by the China National Center for Bioinformation. The GSA family consists of three main components: the updated GSA for general raw sequence data, GSA-Human for human genetics data with controlled access, and OMIX for miscellaneous data types. The system addresses the exponential growth in genomic data volume and diversity, growing from 200 TB to over 8.5 PB by 2021. GSA has been significantly enhanced with improved data models, batch submission capabilities, quality control pipelines, and support for 66 sequencing platforms. GSA-Human provides secure, controlled-access management for sensitive human genetic data with user-defined Data Access Committees, while OMIX serves as an open archive for various data types including transcriptome, epigenome, proteome, and clinical information. The family has served 1,574 users from 391 institutions, archived 359,017 experiments and 395,977 runs, and been cited in 711 research articles across 252 journals. The system provides bilingual support, high-speed data transfer via FTP and Aspera, and integrates with other CNCB-NGDC resources through unified access portals.

■ Simple Explanation (ELI15)

Main Concept	Scientists created a massive digital library system to store and share genetic data from around the world, handling the explosion of genomic information safely and efficiently
Key Takeaway	The GSA family grew to store over 8.5 petabytes of genetic data (equivalent to millions of movies) and serves thousands of researchers globally with secure access controls
Analogies	• Like a massive digital warehouse system with different sections: one for general items (GSA), one for sensitive medical records (GSA-Human), and one for miscellaneous stuff (OMIX) • Similar to how a library expanded from a single building to a campus with specialized wings, each designed for different types of books and access levels

■ Technical Deep-Dive

Aspect	Details
Objective	To develop a comprehensive family of data repositories capable of handling explosive growth in genomic data volume and diversity while providing secure, controlled access for sensitive human genetic information

Methodology	Development of three integrated database systems: updated GSA with enhanced data models and batch submission capabilities, GSA-Human with controlled-access mechanisms and Data Access Committees, and OMIX for miscellaneous data types, all built on INSDC standards with quality control pipelines
Results	Successfully archived 8.5+ PB of data from 1,574 users across 391 institutions, supporting 66 sequencing platforms and 13 data formats, with GSA-Human managing 68,241 individuals and processing 743 access requests, while maintaining high-speed transfer capabilities and bilingual interfaces
Conclusion	The GSA family effectively addresses the challenges of massive genomic data management through specialized repositories, providing a scalable solution for global data sharing while ensuring security and privacy protection for human genetic data

■ Impact Scorecard

Industry Relevance

Critical infrastructure for global genomic research, precision medicine, and biodiversity studies, supporting major sequencing projects and enabling data-driven scientific discoveries

Real-World Applications

- Supporting large-scale genomic projects like Earth BioGenome Project and SARS-CoV-2 sequencing efforts
- Enabling precision medicine research through secure management of human genetic data with clinical information

■■ Research Roadmap

Future Prospects

- Enhancement through national big data infrastructure with upgraded storage, computing, and network resources
- Development of cloud infrastructure and high-speed data transfer tools for massive dataset management

Limitations

- Explosive data growth continues to pose significant storage and management challenges
- Complex regulatory and ethical requirements for human genetic data sharing across international boundaries

Recommended Next Steps

- Optimize data models and curation processes based on evolving user needs and technological advances
- Enhance security protection measures and develop standardized approaches for global biodiversity and health data sharing

■■ Research Methodology Flow

1

System Architecture Design

Developed three-component GSA family architecture with specialized repositories for different data types and access requirements

→ Built specialized database architecture for diverse genomic data types

▼

2

Data Model Enhancement

Updated GSA with improved data structures, separated BioProject and BioSample as independent metadata databases, and implemented batch submission capabilities

→ Enhanced data organization and submission workflows

▼

3

Quality Control Implementation

Integrated automated quality control pipelines and expert curation processes to ensure high-quality data archiving

→ Implemented comprehensive quality assurance systems

▼

4

Performance Analysis

Analyzed system growth from 200 TB to 8.5+ PB, tracking user engagement, data submissions, and access patterns across global research communities

→ Evaluated system scalability and global adoption metrics

▼

5

Security and Access Management

Successfully implemented controlled-access mechanisms for GSA-Human with Data Access Committees, processing 743 access requests while maintaining data privacy and security

→ Established secure framework for sensitive human genetic data