Genomic Data Integrity: Setting the Standard

Outcomes of the Genomic Data Integrity Summit
Hosted by BIO-ISAC and HudsonAlpha Institute for Biotechnology

bioisac
Creating Collaborative Threat Intelligence for the Bioeconomy

Mitigating data risk

Ensuring the accuracy, reliability, and security of genomic data is essential to make informed decisions based on that data in many areas, from agriculture science to personalized medicine. Our ability to generate and use genomic data has become more available and swiftly shareable, while our computational analyses advance.

Verifying a genomic dataset is free from manipulation, however, is limited and relies heavily on trust rather than calculation. An individual's personal decision process or an organization's operating procedures are often all that stands between safe, secure data and/or the compromise of volumes of findings. Current frameworks for the use, protection, and privacy of genomic data in the United States need to advance to address this issue and already compromised volumes of data. In 2023, it is estimated that more than 10,000 peer-reviewed articles were removed after publication due to issues with the integrity of the data examined in the studies.

In May 2024, the Bioeconomy Information Sharing and Analysis Center (BIO-ISAC) convened a select group of 20 individuals in genomic science, cybersecurity, and academic research, as well as 6 United States Government agencies and law enforcement to craft a data collection and use standard that addresses genomic data integrity.

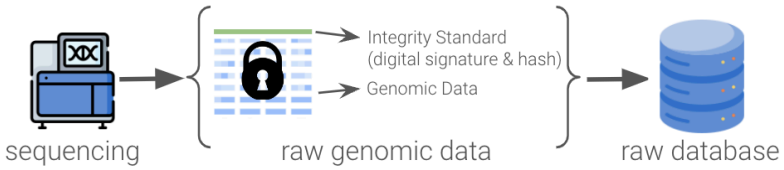
- ATCC (American Type Culture Collection)
- Air Force Research Laboratory
- Cultivarium
- Geneinfosec
- HudsonAlpha Institute for Biotechnology
- Johns Hopkins University Applied Physics Laboratory
- Lincoln Laboratory, Massachusetts Institute of Technology
- Pacific Northwest National Laboratory
- Sandia National Laboratories
- University of Alabama at Huntsville
- U.S. Department of Agriculture
- U.S. Department of Veterans Affairs
- U.S. Federal Bureau of Investigation
- U.S. Senate, National Security Commission on Emerging Biotechnology

This standard includes attribution, compatibility (forwards and backwards), and data signatures in order to allow researchers a way to verify the integrity of a genomic dataset was created. Advancing the models of FAST-A and FAST-Q, this standard identifies where the data came from, how it was collected and on what equipment, and which other parties have had what type of access to the data, prior to its receipt and use.

Implementing this standard will protect economic investments and defend against false findings by safeguarding the raw material our nation's researchers use to advance new science and discoveries.

Standard for Genomic Data Integrity

The Standard for Genomic Data Integrity includes attribution, compatibility (forwards and backwards), and data signatures in order to allow researchers a way to verify the quality and integrity of a genomic dataset.



Attribution refers to the process of identifying the party responsible for a specific action or incident, such as the creation, extraction, or manipulation of data. Attribution notes the provenance of the data as well as any additions or deletions in all prior instances of the data. Accurate attribution establishes the evidence for any malicious activity within the dataset, as well, strengthening overall security of the data.

Compatibility refers to the ability of different systems, software, and protocols to work together seamlessly without compromising security measures or data accuracy. This involves ensuring that various security tools, encryption methods, and data storage formats can interact and function cohesively while maintaining the integrity and confidentiality of genomic data. To ensure backwards compatibility, this standard was confirmed to work with current FAST-Q methods, this adds approximately $x/xB of data in costs while helping maintain robust security frameworks, safeguarding against data manipulation, and protecting the genomic data during its lifecycle.

A **Digital Signature** is a cryptographic technique used to validate the authenticity and integrity of digital documents, data, and communications. Using public and private key pairs, a unique signature is created that can be attached to a digital data file. This signature ensures that the data has not been altered since it was signed and verifies the identity of the sender. Digital signatures are crucial for maintaining trust in the exchange of data, as they provide a reliable means to detect tampering and confirm the source of the data, enhancing overall security and integrity.

Genomic Data Integrity Standard		
Attribution	**Compatibility**	**Digital Signature**
Requires the following: - (legal) originating party and institution identifier (and digital signature) - subsequent user identifiers - RefSeq reannotated - data users must alert originating party when a datafile is changed (optional)	Requires the following: - adhere attribution to data file - include ORCID - repeat extraction (zip + unzip) - adoption - invest on checker - trimmed vs raw submissions	Requires the following: - verifiable decentralized attestation method (ie: DANE) - root per vendor, lower intermediate CAs Encourages: - allowing verification (optional)

Call to Action

This genomic data integrity standard is in implementation, including a workgroup at the Research Data Alliance. To join this implementation effort email **tips@isac.bio**.

bioisac
Creating Collaborative Threat Intelligence for the Bioeconomy