INTERNATIONAL STANDARD

ISO/IEC 23092-6

First edition
2023-11

Information technology — Genomic information representation —

Part 6: Coding of genomic annotations

Reference number
ISO/IEC 23092-6:2023(E)

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

A list of all parts in the ISO/IEC 23092 series can be found on the ISO and IEC websites.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

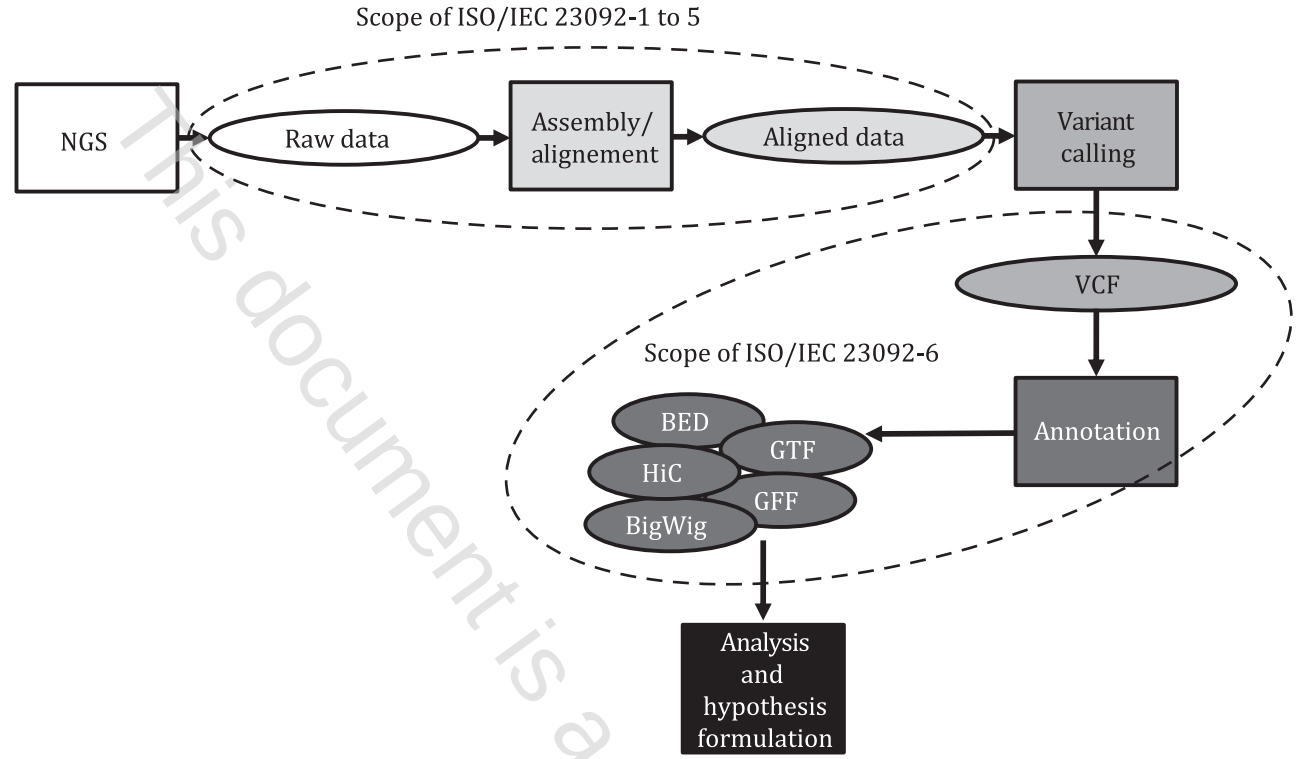
Introduction

While ISO/IEC 23092-1 to ISO/IEC 23092-5 (MPEG-G) deal with the representation of genomic information derived from the primary analysis of high-throughput sequencing (HTS) data – sequencing reads and qualities, and their alignment to a reference genome – which is only the first step in a long series. In particular, the results of primary analysis are usually processed further in order to obtain higher-level information. Such a process of aggregating information deduced from single reads and their alignments to the genome into more complex results is generally known as secondary analysis. In most HTS-based biological studies, the output of secondary analysis is usually represented as different types of annotations associated to one or more genomic intervals on the reference sequences.

Biological studies typically produce genomic annotation data such as mapping statistics, quantitative browser tracks, variants, genome functional annotations, gene expression data and Hi-C contact matrices. These diverse types of downstream genomic data are currently represented in different formats such as VCF, BED, WIG, etc., with loosely defined semantics, leading to issues with interoperability, the need for frequent conversions between formats, difficulty in the visualization of multi-modal data and complicated information exchange. Figure 1 depicts a typical pipeline for the primary and secondary analyses of HTS data, the file formats involved and the scopes of different parts of the ISO/IEC 23092 series.

Furthermore, the lack of a single format has stifled the work on compression algorithms and has led to the widespread use of general compression algorithms with suboptimum performance. These algorithms do not exploit the fact the annotation data typically comprises of multiple fields (attributes) with different statistical characteristics and instead compress them together. Therefore, while these algorithms support efficient random access with respect to genomic position, they do not allow extraction of specific fields without decompressing all the whole file.

In response to the aforementioned challenges, this document details a unified data format for the efficient representation and compression of diverse genomic annotation data for file storage or data transport. The benefits are manifold: reducing the cost of data storage, improving the speed of random data access and processing, providing support for data security and privacy in selective genomic regions, and creating linkages across different types of genomic annotation and sequencing data. The ultimate goal is to enable the secured and seamless sharing, processing and analysis of multi-modal genomic data in order to reduce the burden of data manipulation and management, so scientists can focus on biological interpretation and discovery.



Key

- sequencing generates raw reads
- read alignment
- variant calling
- variants annotations
- analysis

Figure 1 — Typical pipeline for the primary and secondary analyses of HTS data

Information technology — Genomic information representation —

Part 6: Coding of genomic annotations

1 Scope

This document provides specifications for the normative representation of the following types of genomic information:

— variants with genotyping information

— functional annotations

— tracks

— expression matrices

— contact matrices (from Hi-C experiments or similar).

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this specification. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 10646, *Information technology — Universal coded character set (UCS)*

ISO/IEC 11544, *Information technology — Coded representation of picture and audio information — Progressive bi-level image compression*

ISO/IEC 23092-1, *Information technology — Genomic information representation — Part 1: Transport and storage of genomic information*

ISO/IEC 23092-2, *Information technology — Genomic Information Representation — Part 2: Coding of Genomic Information*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
access unit
logical data structure containing a coded representation of genomic information to facilitate bit stream access and manipulation