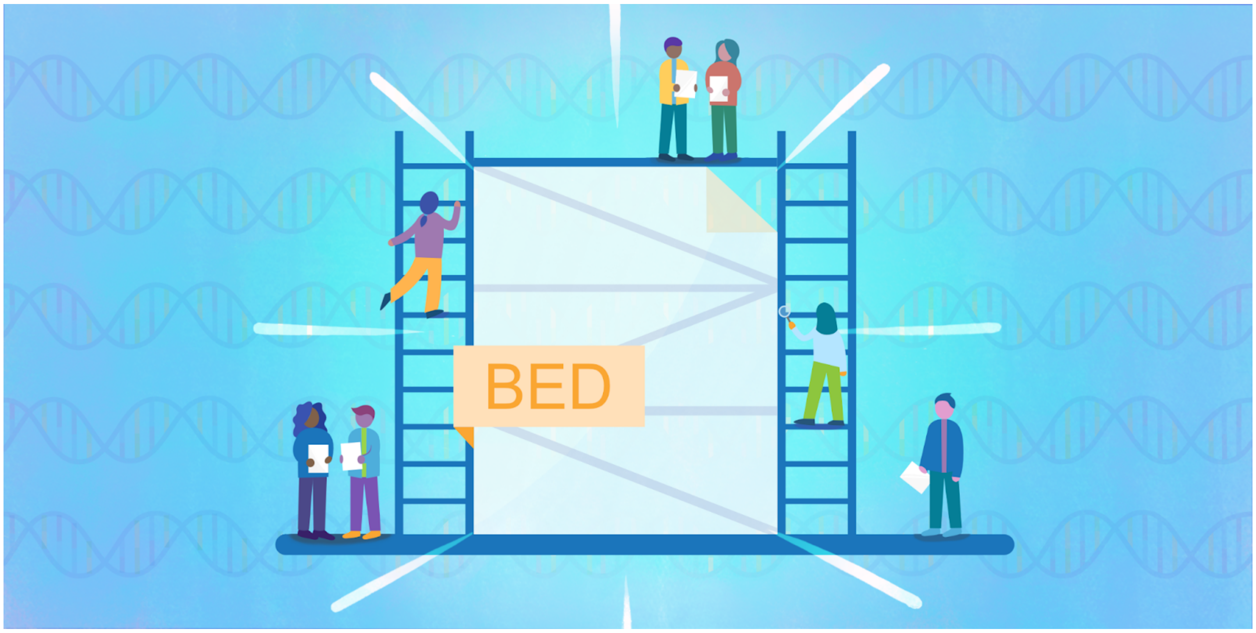
GA4GH BED v1.0: A formal standard sets ground rules for genomic features



Genomic features — such as genes, regulatory elements and repeated sequences, as well as RNA — can have consequences for human health and disease. To better understand disease-causing genes, we must clearly document these features. Investigators use a process called genome annotation to identify what genomic features are present in a DNA sequence, where they are located and what they do.

Over the past two decades, the Browser Extensible Data (BED) file format has become a popular method of capturing the location of genomic features and associated annotations.

Established by Jim Kent at the University of California Santa Cruz (UCSC) Genomics Institute, the BED file format was first used during the Human Genome Project. Since then, numerous genomics projects, analysis software and visualisation tools — including UCSC's Genome Browser application — have adopted the format as well.

"The BED format was developed in service of making genomic annotations visualisable on a graphical viewer, or browser," said Robert Kuhn, associate director of the UCSC Genome Browser. "Over time, due to the format's simplicity, flexibility and conciseness, it became the de facto standard."

Despite its widespread use, the BED file format has lacked a formal specification. The Large Scale Genomics Work Stream at the Global Alliance for Genomics & Health (GA4GH) set about addressing this need.

With guidance from UCSC, the Work Stream built upon the UCSC BED description to produce GA4GH BED v1.0.

"Thousands of users worldwide work with the BED format on a daily basis," said Aaron Quinlan, co-maintainer of the new specification and professor of human genetics and biomedical informatics at the University of Utah. “Formalising the intended use and structure of this fundamental format — from how to name chromosomes to what whitespace delimiter is preferred — will facilitate more reproducible research, thus saving time and improving accuracy.”

Approved in 2021 by the GA4GH Standards Steering Committee (SSC), GA4GH BED v1.0 fills in the gaps in the existing documentation and establishes a concrete set of guidelines for utilising the format.

“The BED format is an interesting project since we’re trying to formalise a standard that has a variety of ‘flavours’ and is widely used today,” said Oliver Hofmann, co-lead of the Large Scale Genomics Work Stream and professor at the University of Melbourne. “By bringing the format to GA4GH, we’re able to leverage community input to tighten up the ambiguous aspects and ensure interoperability.”

In essence, BED is a simple, plain text file format consisting of a series of fields. The first three mandatory fields capture the physical start and end positions of a genomic feature on a linear chromosome. Nine more specified optional fields provide additional information, such as gene name and aesthetic features. Custom fields allow addition of many other data types.

While seemingly straightforward, the lack of conventions has led to a plethora of ways to fill in and structure the fields. The new specification aims to define a numerical range for each specified BED field and provide semantics for whitespace, sorting, default values, and other missing details.

A formal specification is essential for developing software that works with the file format. Otherwise, various tools may read elements differently or simply reject the file altogether.

“Interoperability of tools would enable output from one tool to be used as input into other tools — if the formats are well-defined and predictable,” said Kuhn.

“By standardising the BED format, we can reduce any misinterpretation when using the format, minimise issues when interoperating between software tools, and ultimately avoid errors and inconsistencies in scientific results,” according to Michael Hoffman, co-maintainer of the specification, senior scientist at University Health Network and associate professor of medical biophysics and computer science at the University of Toronto.

To test and verify the performance of software packages that analyse BED files, contributors have developed quality assurance tools, such as the UCSC Genome Browser’s Kent tools, which can validate the output of tools that write BED files. Additionally, Hoffman’s team built a tool to screen for correct behaviour across software tools that read BED files. These efforts are reported in the preprint, “Assessing and assuring interoperability of a genomics file format,” available on bioRxiv.

In the future, the Large Scale Genomics Work Stream plans to continue iterating on the GA4GH BED specification.

“The 1.0 specification was a formalisation effort,” said Hoffman. “For future specifications, the next steps are gathering stakeholders together to determine the best way to encode new metadata and genome annotations within the format itself. Important metadata include the version of the genome assembly used, and the definition of custom data types that might differ from BED file to BED file.”