

Play it again, SAMtools. Q&A with the SAMtools team on 12 years of providing bioinformatics “glue”

Scott Edmunds 

Published February 17, 2021

Citation

Edmunds, S. (n.d.). In *GigaBlog*. GigaBlog. <https://doi.org/10.59350/56qqm-9mv21>

Keywords

Technology, Genome Assembly, Genomics, Open Source, Q&A
Feature Image

Copyright

Copyright © Scott Edmunds 2021. Distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

The [SAMtools suite](#) of tools for manipulating sequencing data one of the most ubiquitous tools in bioinformatics, as the "glue" holding together much of bioinformatics we see it used in pretty much every genomics pipeline we are submitted. Originally published by Heng Li and colleagues in the 1000 Genome Project Data Processing Subgroup, the project is now maintained by a [team at the Wellcome Sanger Institute](#) that includes Andrew Whitwham, Petr Danecek, James Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Thomas Keane, Shane McCarthy, and Robert Davies, alongside contributors to the open codebase from across the world. Despite the* >1M downloads, >2,200 commits to the code, and many changes, there hasn't been a paper describing these updates since Heng Li's [original 2009 publication](#).*

Today in GigaScience we publish the first update in 12 years, with two papers on what's [new in SAMtools](#), and describing for the first time the associated [BCFtools](#) and [HTSlib software library](#). Providing some under-the-hood insight for users, as well as citeable credit for the authors. As recent signatories of the new Software Citation Guide [published in F1000Research](#) we've used these papers as an example of how to provide this type of credit according to best practice. And as fans of under-hood-insight we provide one in our long running series of [author Q&A's](#) with the team behind the paper. Andrew Whitwham from Sanger wrangling the comments from the team below.

SAMtools paper

How did Sequence Alignment/Map (SAM) format files and SAMtools come into being?

This started when the 1000 Genomes Project wanted to move away from the MAQ mapper format. Someone suggested designing a new format during a conference call which immediately gained traction. The overall TAB-delimited flavor of SAM came from an earlier format inspired by BLAT's PSL format. The name of SAM came from Gabor Marth. His group had a format under the same name but with a completely different syntax (more like the standard BLAST output).

The tags in SAM were proposed by Bob Handsaker and either Richard Durbin or Bob proposed the headers. The text-binary dual SAM-BAM format idea came from MAQ. MAQ uses generic gzip compression. Jue Ruan developed a library for random access in generic gzip. When Heng Li suggested that he would like to use the same binary and indexing for a binary SAM, Jue said he had a better idea, which was RAZF. Bob didn't like RAZF because RAZF has to use low-level zlib functions not available in other languages. Then during a conference call, Gerton Lunter said gzip blocks can be concatenated. This and RAZF motivated Bob to develop BGZF. Heng proposed the binning index, learned from UCSC, and Bob proposed the linear index.

The CIGAR string, which is used to represent details of the alignment once a read is mapped, was adapted from Exonerate developed by Guy Slater. Exonerate only had M, I and D operations, Heng added the rest and called the SAM CIGAR as extended CIGAR at the time. Commands like sort, merge and rmdup came from MAQ, while tview was inspired by James Bonfield's text alignment viewer.

An early history of the format is given in [Heng's old blog post](#).

Publishing a lot of genomics papers we see SAMtools near ubiquitously used in these, so what has made it such a crucial and fundamental tool in bioinformatics?

The 1000 Genomes Project published all of its alignment data in SAM/BAM format. Numerous aligners and other programs were written that produced and analyzed them, so the formats rapidly gained momentum. SAMtools was immediately available as a way of working with the formats, so it also rapidly became popular. The program was written in a very compact way, it was fast, easy to install and easy to use. Its permissive license enabled its code to be reused in any way, no questions asked. All these points probably contributed to its long lasting popularity.

How did BCFtools and HTSlib evolve and where do they fit in the SAMtools family?

Both BCFtools and HTSlib projects were originally just parts of the SAMtools codebase. However, as people started reusing the code and putting fragments of the code into their own projects, it became apparent that we needed a clear separation of the application code from the application programming interfaces (API) for parsing the formats, to guarantee long-term stability and to make it easier to interface with other programs. So, in 2013 the decision was made to separate the API from the command line tools and HTSlib was born. The earliest releases of SAMtools included the BCFtools variant caller. When HTSlib was split off, BCFtools was moved into its own package to focus on the processing and analysis of variant data.

SAMtools

BCF is a faster binary representation of the VCF format, with BCFtools designed for the reading and writing of these files.

Since the original Sequence Alignment/Map format and SAMtools paper was [published in 2009](#) this is the first update. What has changed in that time, and what are the most crucial things you've tried to get across in these new papers?

A lot has changed as described in the papers!

Two key changes in the field over the lifetime have been the enormous growth in sequencing data and the need for more advanced analyses. This has directly led to an emphasis on performance and adding new functions and features, including new file formats, [CRAM](#) and BCFv2. This has been coupled with improved robustness and security as we have seen the rise of inclusion of the software in web services.

The software can do much more than it could when the first paper came out. BCFtools in particular has gained an enormous collection of new functionality for analysing variant data.

A project lasting 12 years in the fast-paced world of bioinformatics is amazing, so how have you maintained longevity, stability and backwards compatibility? How many people have been involved in these efforts over the years?

We have, where possible, tried to make changes in a backwards-compatible way. While that means we occasionally have to jump through some hoops to get the results we want, it makes upgrading to the latest version much easier for our users.

The governance of the file formats have moved from being purely an internal 1000 Genomes format to an international standard within [GA4GH's remit](#). This has given safety and longevity to the format, which in turn has aided the implementations.

The main people behind the projects are included as authors on these papers, however there have been many more collectively improving the software. To date there are 146 people listed in the git history across the 3 projects, and we are indebted to the many, many others who have helped to improve the software by reporting bugs, asking questions and making suggestions.

As developers and maintainers of software working under-the-hood to power so many other people's research these efforts may sometimes be taken for granted. We've been involved in the development and [promotion of new guidelines](#) for citing and crediting software. Do you find it difficult to get credit for this work, and how important and useful is it for software developers to get credit in this way?

The huge number of downloads and the active community is a form of recognition. However papers are the generally recognised way of tracking credit, via citations, and can directly impact on grant funding. Without a DOI it can be hard to track use of software and the impact it has.

We are very fortunate that the Wellcome Sanger Institute funds this work as part of its core infrastructure, but we have witnessed funding elsewhere of some software projects drying up due to lack of recognition of their impact to the scientific community. Receiving appropriate credit can therefore be vital for software developers everywhere.

This is an issue not only for software developers, but for those actively involved in maintaining the GA4GH file format specifications too.

Further Reading

Li et al., The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, Volume 25, Issue 16, 15 August 2009, Pages 2078–2079, <https://doi.org/10.1093/bioinformatics/btp352>

Danecek et al., Twelve years of SAMtools and BCFtools, *GigaScience*, Volume 10, Issue 2, February 2021, giab008, <https://doi.org/10.1093/gigascience/giab008>

Bonfield et al., HTSLib: C library for reading/writing high-throughput sequencing data, *GigaScience*, Volume 10, Issue 2, February 2021, giab007, <https://doi.org/10.1093/gigascience/giab007>

SAMtools developers (2020). SAMtools (Version 1.11) <https://www.htslib.org/>

For more on using these resources, Simon Cockell has a [useful online tutorial](#) on using SAMtools.

Ein Fehler ist aufgetreten.

Sieh dir dieses Video auf www.youtube.com an oder aktiviere JavaScript, falls es in deinem Browser deaktiviert sein sollte.

The post [Play it again, SAMtools. Q&A with the SAMtools team on 12 years of providing bioinformatics "glue"](#) appeared first on [GigaBlog](#).