The Journal of Open Source Software

Bam-readcount - rapid generation of basepair-resolution sequence metrics

Ajay Khanna[*1], **David E. Larson**[†2,3], **Sridhar Nonavinkere Srivatsan**[1], **Matthew Mosior**[1,4], **Travis E. Abbott**[2,5], **Susanna Kiwala**[2], **Timothy J. Ley**[1,6], **Eric J. Duncavage**[7], **Matthew J. Walter**[1,6], **Jason R. Walker**[2], **Obi L. Griffith**[1,2,6,8], **Malachi Griffith**[1,2,6,8], **and Christopher A. Miller**[‡1,6]

1 Division of Oncology, Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO **2** McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO **3** Current Affiliation: Benson Hill, Inc. St. Louis, MO **4** Current Affiliation: Moffitt Cancer Center, Tampa, FL **5** Current Affiliation: Google, Inc. Mountain View, CA **6** Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO **7** Department of Pathology, Washington University School of Medicine, St. Louis, MO **8** Department of Genetics, Washington University School of Medicine, St. Louis, MO

DOI: 10.21105/joss.03722

Software

- Review
- Repository
- Archive

Editor: Lorena Pantano

Reviewers:

- @friedue
- @bebatut

Submitted: 15 August 2021
Published: 29 January 2022

Summary

`Bam-readcount` is a utility for generating low-level information about sequencing data at specific nucleotide positions. Originally designed to help filter genomic mutation calls, the metrics it outputs are useful as input for variant detection tools and for resolving ambiguity between variant callers (Koboldt et al., 2013a; Kothen-Hill et al., 2018). In addition, it has found broad applicability in diverse fields including tumor evolution, single-cell genomics, climate change ecology, and tracking community spread of SARS-CoV-2 (Miller et al., 2018; Müller et al., 2018; Paiva et al., 2020; Sun et al., 2020).

Statement of need

Bam-readcount is designed to meet two related needs related to genomic sequence analysis. The first is rapid genotyping of specific locations from a bam file, reporting not just the dominant bases, but counts of all bases. One context in which this is important is residual disease monitoring, where base changes with frequency below the sensitivity of standard genomic variant callers may still be informative. The second is reporting 15 key metrics for each reported base, including summarized mapping and base qualities, strandedness information, mismatch counts, and position within the reads. This information can be useful in a large number of contexts, with one frequent application being variant filtering, to remove false-positive calls, either with straightforward application of heuristic cutoffs or with semi-automated machine-learning approaches (Ainscough et al., 2018; Koboldt et al., 2013b). Another common use case is in ensemble variant calling situations where there is disagreement about base counts or key metrics at particular sites. Bam-readcount can be used to produce consistent, tool-agnostic metrics that are helpful in resolving such ambiguity (Anzar et al., 2019; Kockan et al., 2017; Kothen-Hill et al., 2018).

[*] co-first author
[†] co-first author
[‡] corresponding author

Implementation and results

The ongoing adoption of compressed data formats has necessitated additions to the code, and the version 1.0 release that we report on here utilizes an updated version of HTSlib to support rapid CRAM file access (Bonfield et al., 2021). This has also improved performance, and `bam-readcount` can report on 100,000 randomly selected sites from a 30x whole-genome sequencing (WGS) BAM in around 5 minutes (Griffith, Miller, et al., 2015). Its performance scales nearly linearly with the number of genomic sites queried and average sequencing depth (Figure 1). Querying the same 100,000 sites from a BAM with 300x WGS takes 48 minutes, roughly 10x as long.

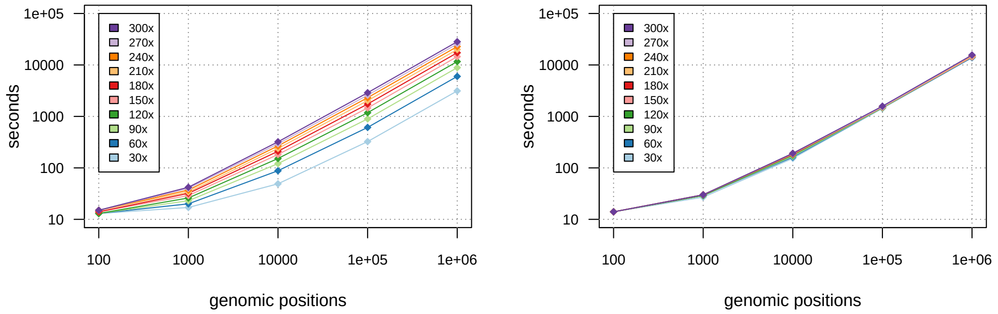

Figure 1: Performance of `bam-readcount` when querying randomly selected genomic positions from BAMs (left) or corresponding CRAMs (right) of varying sequencing depth. Colors correspond to average sequencing depth of the downsampled BAM/CRAM file.

Memory usage likewise is dependent on depth of sequencing, but still requires less than 1 GB of RAM for a 300x WGS BAM. Processing small CRAM files is somewhat slower than BAMs with comparable amounts of data, due to the increased CPU usage for decompression, but as depth increases, retrieval from disk becomes the bottleneck and operations on CRAMs exceed the speed of BAM. In our testing, on a fast SSD tier of networked disks, this transition occurs at a depth of about 180x. The problem is also embarrassingly parallel, so assuming adequate disk I/O, a roughly linear increase in speed can be achieved with a scatter/gather approach.

To lower barriers to adoption, we provide docker images for containerized workflows, and have developed a python wrapper that annotates a VCF file with read counts produced from this tool, available as part of the VAtools package (http://vatools.org).

Conclusions

`bam-readcount` provides fast and accurate genomic readcounts and associated metrics, which allow it to fill a key niche in many genomic workflows. It has been adopted as a lightweight variant caller, finding known mutations in pre-leukemic phenotypes and used for detecting therapy-altering mutations from cell-free DNA (Wyatt et al., 2016; Xie et al., 2014). Viral researchers have tracked nucleotide changes across samples to understand diversity in Varicella Zoster Virus Encephalitis and to perform epidemiological surveillance in wastewater of SARS-CoV-2 (Depledge et al., 2018; Mondal et al., 2021). Those with RNA-sequencing data have found it useful for identifying allele-specific expression in cancer, or for enabling copy-number detection in single-cell RNA sequencing by retrieving allele frequencies (Cancer Genome Atlas Research Network et al., 2013; Müller et al., 2018). Its feature-rich output has also enabled deep learning approaches to variant calling and filtering (Ainscough et al., 2018; Anzar et al., 2019). In these roles, and other related ones, `bam-readcount` has served as key infrastructure that supports groups of all sizes, from exploratory analyses to core facility pipelines to large multi-institution workflows (Griffith, Griffith, et al., 2015; Jensen et al., 2017; Sandmann et al., 2018). In the NCI's Genomic Data Commons pipelines alone, its use in variant filtering means that it has been run on tens of thousands of cancer genomes.

Looking forward, we anticipate that as machine learning makes deeper inroads into genomics, the ability to extract highly informative features from large cohorts in a rapid manner will continue to make `bam-readcount` useful for the next generation of genomics research.

The `bam-readcount` tool is available at https://github.com/genome/bam-readcount and is shared under a MIT license to enable broad re-use.

Data availability

The WGS data used for benchmarking is available through dbGaP study phs000159, under sample id `452198/AML31`. The summary data and scripts used to generate the figure are available at https://github.com/genome/bam-readcount/tree/joss-paper/figures. An archived snapshot of this 1.0 release is available at https://doi.org/10.5281/zenodo.5142454

Authors' contributions

Software Development: AK, DEL, SNS, MM, TEA, SK, CAM. Validation: AK, SNS, MM, CAM. Visualization: CAM. Supervision: CAM, MG, OLG, TJL, EJD, JRW, MJW Writing, review, and editing: AK, DEL, SNS, MM, TEA, SK, TJL, EJD, MJW, JRW, OLG, MG, CAM

Acknowledgements

This work was supported by the National Cancer Institute [R50CA211782 to CAM, P01CA101937 to TJL, K22CA188163 to OLG, 1U01CA209936 to OLG, U24CA237719 to OLG], the Edward P. Evans Foundation (to MJW), and the National Human Genome Research Institute [R00 HG007940 to MG]

References

Ainscough, B. J., Barnell, E. K., Ronning, P., Campbell, K. M., Wagner, A. H., Fehniger, T. A., Dunn, G. P., Uppaluri, R., Govindan, R., Rohan, T. E., Griffith, M., Mardis, E. R., Swamidass, S. J., & Griffith, O. L. (2018). A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nat. Genet.*, *50*(12), 1735–1743. https://doi.org/10.1038/s41588-018-0257-y

Anzar, I., Sverchkova, A., Stratford, R., & Clancy, T. (2019). NeoMutate: An ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC Med. Genomics*, *12*(1), 63. https://doi.org/10.1186/s12920-019-0508-5

Bonfield, J. K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., & Davies, R. M. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience*, *10*(2). https://doi.org/10.1093/gigascience/giab007

Cancer Genome Atlas Research Network, Ley, T. J., Miller, C., Ding, L., Raphael, B. J., Mungall, A. J., Robertson, A. G., Hoadley, K., Triche, T. J., Jr, Laird, P. W., Baty, J. D., Fulton, L. L., Fulton, R., Heath, S. E., Kalicki-Veizer, J., Kandoth, C., Klco, J. M., Koboldt, D. C., Kanchi, K.-L., … Eley, G. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.*, *368*(22), 2059–2074. https://doi.org/10.1056/NEJMoa1301689

Depledge, D. P., Cudini, J., Kundu, S., Atkinson, C., Brown, J. R., Haque, T., Houldcroft, C. J., Koay, E. S., McGill, F., Milne, R., Whitfield, T., Tang, J. W., Underhill, G., Bergstrom, T., Norberg, P., Goldstein, R., Solomon, T., & Breuer, J. (2018). High viral diversity and mixed infections in cerebral spinal fluid from cases of varicella zoster virus encephalitis. *J. Infect. Dis.*, *218*(10), 1592–1601. https://doi.org/10.1093/infdis/jiy358

Griffith, M., Griffith, O. L., Smith, S. M., Ramu, A., Callaway, M. B., Brummett, A. M., Kiwala, M. J., Coffman, A. C., Regier, A. A., Oberkfell, B. J., Sanderson, G. E., Mooney, T. P., Nutter, N. G., Belter, E. A., Du, F., Long, R. L., Abbott, T. E., Ferguson, I. T., Morton, D. L., … Wilson, R. K. (2015). Genome modeling system: A knowledge management platform for genomics. *PLoS Comput. Biol.*, *11*(7), e1004274. https://doi.org/10.1371/journal.pcbi.1004274

Griffith, M., Miller, C. A., Griffith, O. L., Krysiak, K., Skidmore, Z. L., Ramu, A., Walker, J. R., Dang, H. X., Trani, L., Larson, D. E., Demeter, R. T., Wendl, M. C., McMichael, J. F., Austin, R. E., Magrini, V., McGrath, S. D., Ly, A., Kulkarni, S., Cordes, M. G., … Wilson, R. K. (2015). Optimizing cancer genome sequencing and analysis. *Cell Syst*, *1*(3), 210–223. https://doi.org/10.1016/j.cels.2015.08.015

Jensen, M. A., Ferretti, V., Grossman, R. L., & Staudt, L. M. (2017). The NCI genomic data commons as an engine for precision medicine. *Blood*, *130*(4), 453–459. https://doi.org/10.1182/blood-2017-03-735654

Koboldt, D. C., Larson, D. E., & Wilson, R. K. (2013a). Using VarScan 2 for germline variant calling and somatic mutation detection. *Curr. Protoc. Bioinformatics*, *44*, 15.4.1–17. https://doi.org/10.1002/0471250953.bi1504s44

Koboldt, D. C., Larson, D. E., & Wilson, R. K. (2013b). Using VarScan 2 for germline variant calling and somatic mutation detection. *Current Protocols in Bioinformatics*, *44*(1), 15–14. https://doi.org/10.1093/bioinformatics/bty316

Kockan, C., Hach, F., Sarrafi, I., Bell, R. H., McConeghy, B., Beja, K., Haegert, A., Wyatt, A. W., Volik, S. V., Chi, K. N., Collins, C. C., & Sahinalp, S. C. (2017). SiNVICT: Ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA. *Bioinformatics*, *33*(1), 26–34. https://doi.org/10.1093/bioinformatics/btw536

Kothen-Hill, S. T., Zviran, A., Schulman, R. C., Deochand, S., Gaiti, F., Maloney, D., Huang, K. Y., Liao, W., Robine, N., Omans, N. D., & Landau, D. A. (2018, February). *Deep learning mutation prediction enables early stage lung cancer detection in liquid biopsy*.

Miller, C. A., Dahiya, S., Li, T., Fulton, R. S., Smyth, M. D., Dunn, G. P., Rubin, J. B., & Mardis, E. R. (2018). Resistance-promoting effects of ependymoma treatment revealed through genomic analysis of multiple recurrences in a single patient. *Cold Spring Harb Mol Case Stud*, *4*(2). https://doi.org/10.1101/mcs.a002444

Mondal, S., Feirer, N., Brockman, M., Preston, M. A., Teter, S. J., Ma, D., Goueli, S. A., Moorji, S., Saul, B., & Cali, J. J. (2021). A direct capture method for purification and detection of viral nucleic acid enables epidemiological surveillance of SARS-CoV-2. *Sci. Total Environ.*, *795*, 148834. https://doi.org/10.1101/2021.05.06.21256753

Müller, S., Cho, A., Liu, S. J., Lim, D. A., & Diaz, A. (2018). CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. *Bioinformatics*, *34*(18), 3217–3219. https://doi.org/10.1093/bioinformatics/bty316

Paiva, M. H. S., Guedes, D. R. D., Docena, C., Bezerra, M. F., Dezordi, F. Z., Machado, L. C., Krokovsky, L., Helvecio, E., Silva, A. F. da, Vasconcelos, L. R. S., Rezende, A. M., Silva, S. J. R. da, Sales, K. G. da S., Sá, B. S. L. F. de, Cruz, D. L. da, Cavalcanti, C. E., Neto, A. de M., Silva, C. T. A. da, Mendes, R. P. G., … Wallau, G. L. (2020). Multiple introductions followed by ongoing community spread of SARS-CoV-2 at one of the largest metropolitan areas of northeast brazil. *Viruses*, *12*(12). https://doi.org/10.3390/v12121414

Sandmann, S., Karimi, M., Graaf, A. O. de, Rohde, C., Göllner, S., Varghese, J., Ernsting, J., Walldin, G., Reijden, B. A. van der, Müller-Tidow, C., Malcovati, L., Hellström-Lindberg, E., Jansen, J. H., & Dugas, M. (2018). appreci8: A pipeline for precise variant calling integrating 8 tools. *Bioinformatics*, *34*(24), 4205–4212. https://doi.org/10.1093/bioinformatics/bty518

Sun, Y., Bossdorf, O., Grados, R. D., Liao, Z., & Müller-Schärer, H. (2020). Rapid genomic and phenotypic change in response to climate warming in a widespread plant invader. *Glob. Chang. Biol.*, *26*(11), 6511–6522. https://doi.org/10.1111/gcb.15291

Wyatt, A. W., Azad, A. A., Volik, S. V., Annala, M., Beja, K., McConeghy, B., Haegert, A., Warner, E. W., Mo, F., Brahmbhatt, S., Shukin, R., Le Bihan, S., Gleave, M. E., Nykter, M., Collins, C. C., & Chi, K. N. (2016). Genomic alterations in Cell-Free DNA and enzalutamide resistance in Castration-Resistant prostate cancer. *JAMA Oncol*, *2*(12), 1598–1606. https://doi.org/10.1001/jamaoncol.2016.0494

Xie, M., Lu, C., Wang, J., McLellan, M. D., Johnson, K. J., Wendl, M. C., McMichael, J. F., Schmidt, H. K., Yellapantula, V., Miller, C. A., Ozenberger, B. A., Welch, J. S., Link, D. C., Walter, M. J., Mardis, E. R., Dipersio, J. F., Chen, F., Wilson, R. K., Ley, T. J., & Ding, L. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.*, *20*(12), 1472–1478. https://doi.org/10.1038/nm.3733