The Journal of Open Source Software

nf-core/cutandrun: A Nextflow pipeline for the analysis of CUT&RUN, CUT&Tag and TIP-seq datasets

Tamara L. Hodgetts [1], **Charlotte West** [2], **Nicholas M. Luscombe** [3], **Jernej Ule** [1], **James Briscoe** [1], **and Chris Cheshire** [1¶]

1 The Francis Crick Institute, London, UK **2** European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK **3** Okinawa Institute of Science and Technology, Okinawa, Japan ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 12 August 2025
Published: unpublished

Summary

Mapping transcription factor binding events and histone modifications on a genome-wide scale is a central goal for gaining a better understanding of the regulatory dynamics controlling gene expression. The recent development of CUT&RUN, CUT&Tag, and TIP-seq protocols provides new avenues to characterise protein-DNA interactions and histone modifications, offering advantages compared to ChIP-seq in terms of sensitivity, resolution, and sample consumption. Despite differences in biochemistry, the core computational analysis of CUT&RUN, CUT&Tag, and TIP-seq data exhibit notable similarities, representing an opportunity for a unified bioinformatics solution.

Here, we present nf-core/cutandrun, a best-practice bioinformatic analysis workflow for CUT&RUN, CUT&Tag and TIP-seq data. In contrast to existing alternative pipelines, nf-core/cutandrun was written using the Nextflow programming language and developed based on the nf-core community framework to provide enhanced reproducibility, flexibility, scalability, portability and robustness. nf-core/cutandrun additionally enables the specification of spike-in genomes with different normalisation options, supports multiple peakcallers, and provides extensive quality control metrics reporting. The pipeline supports a wide range of execution environments, operating systems and platforms and is designed to be useable with minimal technical knowledge. We encourage user engagement with the wider nf-core community through the official nf-core Github repository and Slack channels and aim to incorporate user feedback in subsequent releases to ensure that the pipeline remains dynamic and up-to-date.

Statement of need

Characterising the genome-wide distribution of transcription factor (TF) binding events and epigenetic histone modifications, is key to elucidating the mechanics and dynamics of gene expression regulation. In recent years, the CUT&RUN and CUT&Tag protocols have emerged as alternatives to chromatin immunoprecipitation (ChIP), providing enhanced accuracy, specificity, and signal-to-noise ratios, as well as reduced costs and input sample requirements (Kaya-Okur et al., 2019; Skene & Henikoff, 2017; see for a review Furey, 2012; Park, 2009). The CUT&RUN and CUT&Tag protocols involve the exposure of permeabilised, unfixed cells to primary antibodies, which bind a particular DNA-associated transcription factor or histone modification of interest. Secondary antibodies are added for signal amplification, and to facilitate the binding of either Protein A-MNase (CUT&RUN) or Protein A-Tn5 transposase (CUT&Tag) fusion constructs. Enzymatic activation subsequently results in local DNA cleavage to generate protein-associated chromatin fragments, which is followed by DNA extraction,

library preparation, amplification and sequencing. The TIP-seq protocol was additionally developed most recently, which combines CUT&Tag with *in vitro* transcription to allow multiple copies of the same insert site to be generated. TIP-seq provides a 10-fold increase in sensitivity relative to traditional CUT&Tag protocols and, like CUT&RUN and CUT&Tag, is amenable to single-cell assays (Bartlett et al., 2021).

Here, we present nf-core/cutandrun, a best-practice bioinformatic analysis pipeline for CUT&RUN, CUT&Tag and TIP-seq datasets. To meet the modern requirements for bioinformatic analysis, which we define to be reproducibility, flexibility, scalability, portability and robust reporting, the pipeline has been developed in line with the nf-core community framework. nf-core is an international, community-driven project which aims to develop a curated set of gold-standard bioinformatic analysis pipelines written in the scientific workflow language Nextflow (Di Tommaso et al., 2017; Ewels et al., 2020). The nf-core community provides various modules and sub-workflows that can be used for the development of modularised, robust workflows, all of which must adhere to the established nf-core guidelines. nf-core/cutandrun is fully containerised and supports a wide range of execution environments including local, cluster and cloud instances that run on macOS, Linux or Windows operating systems. All software used in nf-core/cutandrun was selected using a stringent process that ensured the pipeline to be open-source, high-quality, and actively well-maintained.

The nf-core/cutandrun pipeline provides several advantages relative to CUT&RUNTools and ePeak, which are the only current alternative pipelines for the end-to-end analysis of CUT&RUN and CUT&Tag data (Daunesse et al., 2022; Yu et al., 2022; Zhu et al., 2019). First, as a result of the protein production process in original CUT&RUN and CUT&Tag protocols, there is residual *E. coli* DNA that can be used as a natural spike-in calibrator. This *E. coli* DNA is increasingly purified out of commercial kits, with spike-in DNA being manually added during the protocol. Manual addition of spike-in DNA is a considerable source of experimental error, which can lead to low alignment rates that are insufficient for normalisation. Consequently, we included additional support for user-defined spike-in genomes, and additional modes for normalisation based on read counts. The pipeline additionally provides three distinct peak calling methods, which include SEACR, MACS2 narrow, and MACS2 broad peak modes. We selected SEACR as the default peakcaller, as it has been demonstrated to provide greater specificity and efficiency of peak calling on sparse data, such as CUT&RUN, CUT&Tag and TIP-seq data (Meers et al., 2019). Finally, nf-core/cutandrun provides extensive quality control metrics reporting to maximise confidence prior to downstream analysis, and uniquely supports the analysis of TIP-seq data.

Workflow overview

nf-core/cutandrun offers a robust data analysis workflow that is designed for Illumina-based sequencing data obtained from CUT&RUN, CUT&Tag and TIP-seq experimental protocols (Figure 1).

The Journal of Open Source Software

Figure 1: A schematic representation of the nf-core/cutandrun pipeline. Icons were adapted from nf-core/sarek; under a CC-BY 4.0 licence.

The workflow initialises by performing checks on input files and determining the experimental design in terms of sample grouping, the presence of IgG controls and the required type of spike-in normalisation. During pre-processing, the raw FASTQ files are merged, considering multi-lane sequencing or technical replicates as defined in the sample sheet. The resultant merged FASTQ files are initially screened using FASTQC before undergoing trimming by TrimGalore to remove sequencing adaptors from reads (Andrews, 2010; Krueger, 2012). FASTQC is subsequently re-executed post-trimming, in line with the workflow's strong emphasis on transparency and interpretability. The raw reads are next aligned to both the target reference genome and the spike-in genome using Bowtie2 (Langmead & Salzberg, 2012). Several quality control metrics are computed following alignment, aiding users in identifying potential issues with the alignment process. Alignment statistics, fragment distributions and binned sample correlation metrics are calculated on a per-sample basis using Samtools and custom Python scripts (Li et al., 2009). Reads which were aligned on the target genome are subject to q-score filtering using Samtools and duplicate removal (IgG controls only by default) using Picard (Broad Institute, 2019; Li et al., 2009). For TIP-seq data, the removal of linear amplification duplicates is additionally supported with custom Python scripts. Scaling factors for BAM to BedGraph conversion are calculated on a per-sample basis, based on the number of reads that aligned to the spike-in genome. BedGraph files are also converted to a BigWig format for the auto-generation of IgV sessions and for heatmap generation using DeepTools (Ramírez et al., 2016; Robinson et al., 2011). By default, peak calling is performed using SEACR, which was specifically developed for experiments with low background noise and thus detects a larger number of high-accuracy peaks compared to more generalised peak callers (Meers et al., 2019). However, the pipeline also supports peak calling using MACS2 either instead of or in addition to SEACR. At this stage, any IgG controls specified in the experimental design are split out and used to normalise peak calling against background levels of non-specific antibody binding. Peaks are merged into consensus peak sets based on sample groupings, and intra-group peak reproducibility is assessed using both the number of reproducible peaks and FRiP scores (Fraction of Reads in Peaks). Finally, a comprehensive summary of the workflow execution, parameter settings, software versions, and quality metrics are aggregated into a single MultiQC report (Ewels et al., 2016).

Pipeline installation and operation

nf-core/cutandrun provides minimal execution requirements, which include only the installation of Nextflow (version $\geq$ 21.10.3) on the host system, and a software execution environment that is container-based, for example Docker (see for a review Rad et al., 2017) or Singularity (Kurtzer et al., 2017). Nextflow additionally supports a large range of workflow execution

environments including Slurm, Sun Grid Engine, LSF and Kubernetes and cloud-based computing environments such as Google Cloud and AWS. Switching between execution locations is trivial, often requiring just a single parameter adjustment on execution. The workflow is automatically downloaded by Nextflow when the keywords nf-core/cutandrun are included within a run statement, rendering manual installation unnecessary. Installation and pipeline operation can be verified easily by running pre-defined test routines with public data hosted in the cloud that can stress test the host system before any user-provided data is analysed. Resource requirements and run times may vary as they are dependent on the target genome size, number of samples, and number of reads per sample. Thus, although the pipeline is designed for optimal portability and scalability by facilitating execution on local computers, high-performance computing clusters (HPCs) and cloud computing environments, the choice of execution system should depend on the memory and CPU requirements of the dataset in question. A rich set of parameters is provided to support workflow customisation and flow-switching, and all results are provided in a single folder that is intuitively structured to facilitate troubleshooting. For further details regarding pipeline installation, operation, and parameter options, readers are referred to the official nf-core/cutandrun documentation.

Conclusion

Here we have presented nf-core/cutandrun, a best-practice bioinformatics pipeline for the end-to-end analysis of CUT&RUN, CUT&Tag and TIP-seq data. By adhering to nf-core community guidelines, the pipeline fulfils our defined prerequisites for modern bioinformatics analysis, and supports a range of execution platforms, operating systems, and resource-limited scenarios. Our pipeline supports user-defined spike-in genomes, offers additional read count-based normalisation modes, supports three distinct peak calling methods, and uniquely facilitates the analysis of TIP-seq data. We provide additional command line parameters for workflow customisation, including the specification technical replicates, flow-switching options, and pipeline execution reporting. Collectively, we conclude that these properties of the nf-core/cutandrun pipeline provide significant advantages relative to current alternative pipelines. Readers are encouraged to engage with the nf-core community through the official nf-core Github repository and Slack channels.

Acknowledgements

This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (CC001051, CC0102, FC010110), the UK Medical Research Council (CC001051, CC0102, FC010110), and the Wellcome Trust (CC001051, CC0102, FC010110). It was also supported by the Wellcome Trust (215593/Z/19/Z to J.U. and N.M.L.). N.M.L. was a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of the Francis Crick Institute. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. We would like to thank all members of the Francis Crick Institute and the nf-core community for using the pipeline and providing structured feedback throughout its development and release.

References

Andrews, S. (2010). *Babraham bioinformatics-FastQC: a quality control tool for high throughput sequence data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bartlett, DA., Dileep, V., Handa, T., Ohkawa, Y., Kimura, H., Henikoff, S., & Gilbert, DM. (2021). High-throughput single-cell epigenomic profiling by targeted insertion of

promoters (TIP-seq). *Journal of Cell Biology*, *220*, e202103078. https://doi.org/10.1083/jcb.202103078

Broad Institute. (2019). *Picard Toolkit, GitHub Repository*. https://broadinstitute.github.io/picard/

Daunesse, M., Legendre, R., Varet, H., Pain, A., & Chica, C. (2022). ePeak: from replicated chromatin profiling data to epigenomic dynamics. *NAR Genomics and Bioinformatics*, *4*, lqac041. https://doi.org/10.1093/nargab/lqac041

Di Tommaso, P., Chatzou, M., Floden, EW., Barja, PP., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*, 316–319. https://doi.org/10.1038/nbt.3820

Ewels, PA., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Ewels, PA., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, MU., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*, 276–278. https://doi.org/10.1038/s41587-020-0439-x

Furey, TS. (2012). ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, *13*, 840–852. https://doi.org/10.1038/nrg3306

Kaya-Okur, HS., Wu, SJ., Codomo, CA., Pledger, ES., Bryson, TD., Henikoff, JG., Ahmad, K., & Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, *10*, 1930. https://doi.org/10.1038/s41467-019-09982-5

Krueger, F. (2012). *Trim Galore: a wrapper tool around Cutadapt and FastQC for quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries*. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Kurtzer, GM., Sochat, V., & Bauer, MW. (2017). Singularity: Scientific containers for mobility of compute. *PloS One*, *12*, e0177459. https://doi.org/10.1371/journal.pone.0177459

Langmead, B., & Salzberg, SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359. https://doi.org/10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Meers, MP., Bryson, TD., Henikoff, JG., & Henikoff, S. (2019). Improved CUT&RUN chromatin profiling tools. *eLife*, *8*, e46314. https://doi.org/10.7554/eLife.46314

Park, PJ. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, *10*, 669–680. https://doi.org/10.1038/nrg2641

Rad, BB., Bhatti, HJ., & Ahmadi, M. (2017). An introduction to docker and analysis of its performance. *International Journal of Computer Science and Network Security (IJCSNS)*, *17*, 228–235.

Ramírez, F., Ryan, DP., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, AS., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, *44*(Web Server issue), W160–5. https://doi.org/10.1093/nar/gkw257

Robinson, JT., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, ES., Getz, G., &

Mesirov, JP. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*, 24–26. https://doi.org/10.1038/nbt.1754

Skene, PJ., & Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*, *6*, e21856. https://doi.org/10.7554/eLife.21856

Yu, F., Sankaran, VG., & Yuan, GC. (2022). CUT&RUNTools 2.0: a pipeline for single-cell and bulk-level CUT&RUN and CUT&Tag data analysis. *Bioinformatics*, *38*, 252–254. https://doi.org/10.1093/bioinformatics/btab507

Zhu, Q., Liu, N., Orkin, SH., & Yuan, GC. (2019). CUT&RUNTools: a flexible pipeline for CUT&RUN processing and footprint analysis. *Genome Biology*, *20*, 1–12. https://doi.org/10.1186/s13059-019-1802-4