# 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

## Summary

Mapping transcription factor binding events and histone modifications on a genome-wide scale is key to understanding the regulatory dynamics controlling gene expression. CUT&RUN, CUT&Tag and TIP-seq protocols provide new avenues to map chromatin-associated proteins, 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 bioinformatics pipeline for CUT&RUN, CUT&Tag and TIP-seq data. In contrast to existing alternative pipelines, nf-core/cutandrun was 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 peak callers, and provides extensive quality control metrics reporting. The pipeline supports a wide range of execution environments and operating systems, and is usable 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 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 binding events and 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 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 first expose 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 enables local DNA cleavage to generate protein-associated chromatin fragments, which is followed by DNA extraction, library preparation, amplification and sequencing. Most recently, the TIP-seq protocol was additionally developed, which combines CUT&Tag with *in vitro* transcription to provide a further 10-fold increase in sensitivity (Bartlett et al., 2021).

Here, we present nf-core/cutandrun, a best-practice bioinformatics pipeline for CUT&RUN, CUT&Tag and TIP-seq datasets based on the original data analysis workflow by Zheng et al. (2020). To meet the modern requirements for bioinformatics analysis, which we define to be reproducibility, flexibility, scalability, portability and robust reporting, we developed the pipeline in line with the nf-core community framework (Di Tommaso et al., 2017; Ewels et al., 2020). nf-core/cutandrun 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 can introduce considerable experimental error, leading to low alignment rates that are insufficient for normalisation. Consequently, we included additional read count-based normalisation modes and support for user-defined spike-in genomes. The pipeline additionally provides three distinct peak calling methods, including SEACR, MACS2 narrow, and MACS2 broad peak modes. We selected SEACR as the default peak caller, as it has been demonstrated to provide greater specificity and efficiency of peak calling on sparse data (Meers et al., 2019). Finally, nf-core/cutandrun provides extensive quality control (QC) 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 experiments (Figure 1).
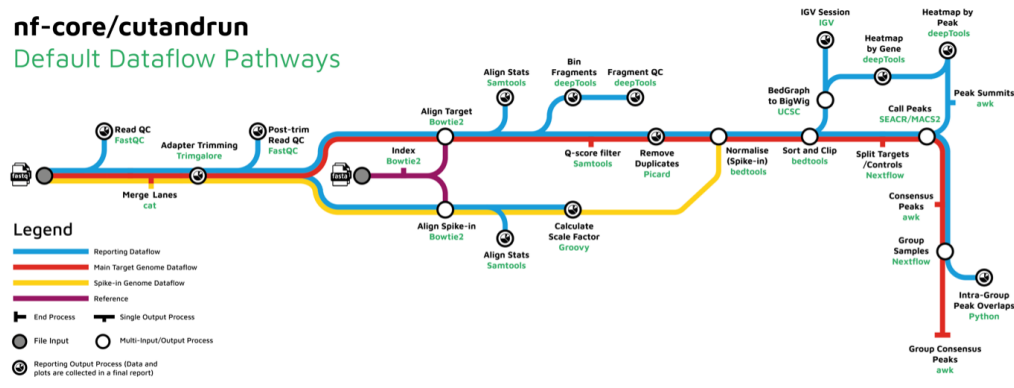


**Figure 1:** A schematic representation of the nf-core/cutandrun pipeline. Icons were adapted from nf-core/sarek; under a CC-BY 4.0 licence.

The workflow initialises by performing checks on input files and determining the experimental design in terms of sample grouping, potential IgG controls, and the required normalisation mode. 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 to collect QC metrics, before undergoing trimming by TrimGalore to remove sequencing adaptors from reads (Andrews, 2010; Krueger, 2012). FastQC is re-executed post-trimming, consistent with the workflow's emphasis on transparency and interpretability. The reads are next aligned to both the target reference genome and the spike-in genome using Bowtie2 (Langmead & Salzberg, 2012). Alignment statistics, fragment distributions and binned sample correlation metrics are calculated for each sample using SAMtools, deepTools, and custom Python scripts, aiding users in identifying potential issues

with the alignment process (Li et al., 2009; Ramírez et al., 2016). Reads which aligned to the target genome are subjected to q-score filtering using SAMtools, and deduplication using Picard (IgG controls only by default) (Broad Institute, 2019; Li et al., 2009). For TIP-seq data, linear amplification duplicates are additionally removed with custom Python scripts. Scaling factors for BAM-to-BedGraph conversion are calculated per sample, based on the number of reads that aligned to the spike-in genome. BedGraph files are also converted to a BigWig format to auto-generate IGV sessions, and generate heatmaps using deepTools (Ramírez et al., 2016; Robinson et al., 2011). By default, peak calling is performed using SEACR, which was specifically developed for protocols with low background noise (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 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, comprehensive summaries of the workflow execution, parameter settings, software versions, and QC metrics are aggregated into a single MultiQC report (Ewels et al., 2016).

## Pipeline installation and operation

nf-core/cutandrun offers minimal execution requirements, requiring only the installation of Nextflow (version $\geq$ 21.10.3) on the host system and a containerisation platform such as Docker (see for a review Rad et al., 2017) or Singularity (Kurtzer et al., 2017). The workflow is automatically downloaded by Nextflow when the keywords 'nf-core/cutandrun' are included within a run statement, rendering manual installation unnecessary. Pipeline installation and operation are easily verified by running pre-defined test routines with public data hosted in the cloud, allowing users to stress test the host system before data analysis. nf-core/cutandrun supports a wide range of execution environments including local, cluster and cloud instances that run on macOS, Linux or Windows operating systems. A rich set of parameters supports workflow customisation and flow-switching, and all results are provided in a single, intuitively-structured folder to facilitate troubleshooting. Readers are referred to the official nf-core/cutandrun documentation for further details regarding pipeline installation, operation, and parameter options.

## 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 TIP-seq data analysis. We provide additional command line parameters for workflow customisation, including the specification of technical replicates, flow-switching options, and pipeline execution reporting. Collectively, we conclude that these properties of nf-core/cutandrun 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

# 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 Bisulfite-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*, *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

Zheng, Y., Ahmad, K., & Henikoff, S. (2020). *CUT&Tag Data Processing and Analysis Tutorial*. https://www.protocols.io/view/cut-amp-tag-data-processing-and-analysis-tutorial-e6nvw93x7gmk/v1

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