Improving rigor and reproducibility in chromatin immunoprecipitation assay data analysis workflows with Rocketchip

Viktoria Haghani[1,2]

Aditi Goyal[2]

Alan Zhang[2]

Osman Sharifi[1,2]

Natasha Mariano[2]

Dag Yasui[1]

Ian Korf[2]

Janine LaSalle[1*]

[*] Corresponding author, jmlasalle@ucdavis.edu, (530) 754-7598

1. Department of Medical Microbiology and Immunology, Genome Center, University of California, Davis. Davis, CA, USA.
2. Department of Molecular and Cellular Biology, Genome Center, University of California, Davis. Davis, CA, USA.

Abstract

As genome sequencing technologies advance, the accumulation of sequencing data in public databases necessitates more robust and adaptable data analysis workflows. Here, we present Rocketchip, which aims to offer a solution to this problem by allowing researchers to easily compare and swap out different components of ChIP-seq, CUT&RUN, and CUT&Tag data analysis, thereby facilitating the identification of reliable analysis methodologies. Rocketchip enables researchers to efficiently process large datasets while ensuring reproducibility and allowing for the reanalysis of existing data. By supporting comparative analyses across different datasets and methodologies, Rocketchip contributes to the rigor and reproducibility of scientific findings. Furthermore, Rocketchip serves as a platform for benchmarking algorithms, allowing researchers to identify the most accurate and efficient analytical approaches to be applied to their data. In emphasizing reproducibility and adaptability, Rocketchip represents a significant step towards fostering robust scientific research practices.

Keywords

ChIP-seq, CUT&RUN, CUT&Tag, bioinformatics, workflow, Snakemake

Description of the Authors

Viktoria Haghani is a PhD Candidate in the UC Davis Integrative Genetics and Genomics Graduate Group. She is interested in the application of bioinformatics and automation in big data analysis.

Aditi Goyal is an MS candidate in the Stanford Biomedical Data Science Program. She was previously an undergraduate at UC Davis and graduated with a double bachelor’s degree in Statistics and Genetics & Genomics. She is interested in personalized medicine and cancer genomics.

Alan Zhang was an undergraduate student in the UC Davis Genetics and Genomics Program and is currently a PhD student in the UC Santa Cruz Biomolecular Engineering and Bioinformatics Program. He is interested in computational genomics and population genetics.

Osman Sharifi is a PhD Candidate in the Biochemistry, Molecular, Cellular, and Developmental Biology program at UC Davis. He is interested in utilizing bioinformatics for NGS big data.

Natasha Mariano, a former UC Davis undergraduate, is currently a graduate student in the labs of Dr. Cedric Feschotte and Dr. Eirene Markenscoff-Papadimitriou at Cornell University. She is studying how a transposon derived autism risk gene affects embryonic neurodevelopment.

Dag Yasui, Ph.D. is a Project Scientist at the UC Davis Department of Medical Microbiology and Immunology with expertise in Rett, Prader-Willi syndrome research, and chromatin organization.

Ian Korf, Ph.D. is a Professor in the Department of Molecular Cellular Biology at UC Davis and Associate Director for Bioinformatics at the UC Davis Genome Center, with expertise in bioinformatics and genomics.

Janine LaSalle, Ph.D. is a Professor of Medical Microbiology and Immunology, Co-Director of the Perinatal Origins of Disparities Center, and Deputy Director of the Environmental Health Sciences Center at UC Davis, with expertise in epigenomics and neurodevelopmental disorders.

Introduction

As genome sequencing technologies and their applications continue to rapidly evolve to include epigenomic information, a vast amount of sequencing data is accumulating (1). Journals now commonly mandate the deposition of raw sequence data to public databases, such as the International Nucleotide Sequence Database Collaboration Sequence Read Archive (SRA), generating a substantial volume of sequencing data (2). These mandates are further supported by funding agencies, such as the National Institutes of Health (NIH), which requires NIH-funded research to publish sequence data as part of their Genomic Data Sharing Policy (3). Therefore, there is an increasing need for more comprehensive and biologically relevant data analysis workflows that promote reproducibility of results and allow for increased leverage and comparisons of publicly available data. Although mandated availability of data is beneficial to science, data analysis pipelines are often complicated, with divergent results due to variation in analysis steps, parameter usage, software used, and software version. These problems can partly be solved by workflow managers, which control for analysis step order, software versions via virtual environments, software parameters, etc. This is especially important for workflows requiring multiple analysis steps, such as sequence data produced by chromatin immunoprecipitation assays, where small differences in analytical steps can yield different results.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a technique commonly used to identify protein binding sites in the genome (4,5). Briefly, DNA associated proteins are chemically fixed onto DNA via a crosslinking agent. The resulting chromatin is then fragmented by either sonication or enzymatic digestion into 100-500 base pair fragments. Next, chromatin fragments bearing the protein of interest are immunoprecipitated using protein-specific antibodies. Typically, ChIP-seq experiments utilize two controls: (1) an input control to correct for differences in sonication and genomic DNA sequence bias and (2) a mock IP to account for nonspecific interactions of the antibody used (6–8). Finally, the chemical cross-link is reversed, allowing the DNA bound by the protein to be purified and then sequenced. In addition to ChIP-seq, other chromatin immunoprecipitation assays, such as cleavage under targets and release using nuclease (CUT&RUN) and cleavage under targets and tagmentation (CUT&Tag) are similarly utilized to assess protein binding sites (9,10). In CUT&RUN, antibody binding occurs directly on protein-bound DNA fragments within intact nuclei, with DNA fragmentation accomplished enzymatically using a fusion protein containing Protein A and micrococcal nuclease (MNase), contrasting with ChIP-seq's reliance on sonication for DNA fragmentation. In CUT&Tag, a fusion protein combines Protein A and Tn5 transposase, allowing for simultaneous tagging and binding of sequencing adapters to protein-bound DNA fragments.

Following the unique approaches of ChIP-seq, CUT&RUN, and CUT&Tag in isolating protein-bound DNA, the generated DNA sequence data is aligned to a reference genome, and a peak caller is typically used to identify regions of interest. These peaks are “broad” for proteins with large areas of interactions with DNA, such as histones and DNA methylation interacting proteins. Conversely, proteins like transcription factors such as CTCF, that have a sequence specific region of interaction, produce “narrow” peaks. Analyzing these different types of DNA binding proteins can sometimes call for different computational controls. Given the variation in how peaks are generated, there is a significant amount of statistical noise in these experiments, which can hinder the efficacy of peak calling algorithms. This is further complicated by sequencing issues associated with strand bias, GC content, PCR amplification, library preparation, primer choice, sequencing platform, and antibody choice (8,11–20). Consequently, it is increasingly difficult to reproduce or replicate experimental findings. From this point forward, “reproducibility” refers to an individual’s or group’s ability to generate the same findings as another study, whereas “replicability” represents an individual’s or group’s ability to recreate findings multiple times using the same data input. Overall, this highlights the need for consistent and controlled data analysis practices, ensuring reproducibility and replicability of results.

Here, we present Rocketchip, available at https://github.com/vhaghani26/rocketchip, to address key aspects of these problems. Rocketchip reduces variation in data analysis methodologies, increase reproducibility and replicability of experimental results, and encourage greater usage of publicly available sequence data.

Methods and Results

Implementation

Rocketchip is an automated bioinformatics workflow written in the Python-based (v3.10.12) workflow manager, Snakemake (v7.32.4) (21) (Figure 1). Rocketchip downloads ChIP-seq data directly from the SRA, the largest publicly available sequence database, using the SRA Toolkit (sra-tools, 3.0.9) (2,22). The SRA Toolkit downloads data via the prefetch and fasterq-dump functions, then splits the raw DNA sequence read file into respective paired-end (PE) read files using the flag --split-files and converts them into FASTQ file formats, whereas for single-end (SE) read files, the SRA Toolkit is utilized solely for the download and conversion of raw read data into FASTQ format. Rocketchip also presents the option for users to use local (i.e. non-SRA sourced) ChIP-seq data by providing their own FASTQ files. In parallel, Rocketchip downloads and processes a reference genome of the user's choice from the UCSC Genome Browser (23). After downloading and converting files, files are stored in the local file system in FASTA and FASTQ formats. Users also have the option of storing their own custom genomes and sequencing reads that are not yet publicly available.

In Rocketchip, raw sequence data undergoes a quality control step using FastQC (v0.12.1) with default parameters to assess levels of data duplication and sequence quality (24). Raw sequence data is then aligned to the reference genome using the user's choice of alignment software from BWA-MEM (v0.7.17), Bowtie2 (v2.5.2), or STAR (v2.7.11a), each with the default parameters (25–27). Intermediate data files are processed as necessary for deduplication. PCR duplicates are removed from sequence data using a user's choice of deduplication software from Samtools (v1.18) with --mode s (i.e. standard PCR duplicate detection), Picard (v3.1.1) using MarkDuplicates, Sambamba (v1.0.0) with default parameters, or no deduplication (28–30). Deduplicated data files are subject to another quality control step using FastQC to ensure data integrity. Next, Deeptools' (v3.5.4) bamCoverage function with default parameters is used to convert data from the BAM file format to the bigwig file format, which can be used for visualization of ChIP-seq data in the UCSC Genome Browser or other visualization tools (31). Finally, in Rocketchip, peaks are called with a user's choice of software from MACS3 (v3.0.0b3) with the --bdg option, Genrich (v0.6.1) using default parameters, PePr (v1.1.24) using default parameters, or CisGenome (v2.0) (32–35). For CisGenome, if a control was used, the seqpeak command was employed with the default options. Without a control, CisGenome ran two rounds of peak calling using the hts_peakdetectorv2 command with options "-w 100 -s 25 -c 10 -br 1 -brl 30 -ssf 1."

Peak-calling options in Rocketchip include narrow versus broad peak calling and use of a control where appropriate. Software version control is handled using Conda to ensure reproducibility of results (36). Software options were chosen based on options for command line use, ease of installation, and their standard use in the field.

Validation of all Software Options

In order to ensure that all combinations of software can be integrated seamlessly, we selected a deeply sequenced experimental ChIP-seq study targeting a transcription factor, MeCP2, in mouse main olfactory epithelium conducted by Rube et al. 2016, hereafter referred to as "Rube", (37) and a ChIP-seq study targeting NRF2 in human non-small lung cells conducted by Namani et al. 2019, hereafter referred to as "Namani" (38). To facilitate this comprehensive assessment, these data sets were run through Rocketchip using all four peak callers, namely MACS3, CisGenome, Genrich, and PePr. This was used in combination with each of the three aligners, BWA-MEM, Bowtie2, and STAR and four deduplication techniques, Samtools, Sambamba, Picard, and no deduplication. Additionally, to validate that Rocketchip can be run with or without a control (i.e. input or IgG control), this analysis was conducted with both the usage and omission of the corresponding control for each data set, with the exceptions of CisGenome and PePr, which can only be run if a control is used in the analysis. Each test was run three times to assess Rocketchip's ability to replicate experimental results.

Each algorithm demonstrated varying peak-calling efficiency, influenced by several key factors (Figure 2). Most notably, the source of the data (i.e, Namani vs. Rube) revealed dramatic differences in the performance of peak-calling algorithms. For the Rube data, MACS3 consistently yielded the highest number of called peaks. All methods of deduplication produced comparable peak counts. Bowtie2 and BWA-MEM performed similarly, identifying slightly more peaks compared to STAR. CisGenome called the fewest peaks, significantly influenced by the deduplicator and aligner used. Without deduplication, CisGenome identified the fewest peaks across any software combination, while peak counts increased with any deduplication method. When STAR was used as the aligner with CisGenome, peak counts were the lowest. BWA-MEM yielded a higher peak count than Bowtie2 but exhibited more non-deterministic behavior compared to Bowtie2 and STAR. For Genrich, peak counts increased when the control was omitted during peak calling. In contrast to CisGenome, Genrich had the highest peak counts when no deduplication was used and when STAR was used for alignment. The other deduplicators and aligners showed negligible differences in peak counts. When using PePr for peak calling on the Rube data, no deduplication yielded the highest peak count, with other deduplication methods yielding similar results. Unlike the other algorithms, each aligner performed differently with PePr, with STAR identifying the most peaks, followed by BWA-MEM and then Bowtie2. Both the narrow- and broad-peak-calling algorithms yielded negligible differences, except for CisGenome, which had less deterministic results when peaks were defined as narrow and run through Rocketchip.

For the Namani data, notable contrasts in peak-calling outcomes were observed across various algorithms. Genrich consistently yielded the highest number of peaks, contrasting with the Rube data where MACS3 showed higher peak counts. Interestingly, PePr consistently produced the lowest peak counts for the Namani dataset despite performing second best for the Rube data. The choice of aligner did not significantly impact peak counts overall; however, omitting deduplication generally resulted in slightly higher peak counts across all peak-callers. Unlike the Rube data, where the distinction between narrow- and broad-peak calling showed minimal differences across algorithms, the Namani data exhibited variability based on this distinction. Specifically, MACS3 identified higher peak counts under the assumption of broad peaks compared to narrow peaks.

In addition to assessing the results of the different algorithms, we also assessed Rocketchip's ability to replicate experimental findings. A total of 288 combinations were tested, accounting for both data sets and software combinations. Among the 288 combinations tested, 274 trials (95.14%) demonstrated perfect replication of peak counts across all three trials. Surprisingly, despite identical software versions, computational resources, and inputs, certain combinations of data and software were non-deterministic, with 4.86% (14 out of 288) exhibiting variability in peak counts (Table 1). The greatest variability in peak counts occurred when CisGenome was used as the peak-caller, where the range between peak counts in these cases were 20,356 and 159,544 peaks. Cases with lower variation (i.e. differences in peak counts ranging from 1-6 peaks) occurred when the STAR aligner was used in the workflow.

Overall, these findings underscore four critical points. First, the choice of software combination, including the algorithm, deduplication method, and aligner, has a significant impact on peak-calling outcomes. Second, even with strict control of factors impacting the analysis, we still observe some variability in peak calling, albeit in a small subset of cases. Third, dataset-specific nuances significantly impact the performance of different software combinations, resulting in a differing consensus on what the "best" software combination is. Finally, the results validate that all available software combinations can be successfully executed within Rocketchip, ensuring flexibility and robustness in ChIP-seq analyses.

Assessing Run Times for Experimental ChIP-seq Data from Varying Genomes and Read Sizes

The UCSC Genome Browser has seven genomes displayed by default in the “Genomes” tab: human (hg38), mouse (mm10), rat (rn6), zebrafish (danRer11), fruitfly (dm6), worm (ce11), and yeast (sacCer3). Three published ChIP-seq data sets on the SRA with varying read coverage were selected per genome and run through Rocketchip (Figure 3) (39–50). These experiments were run on an HPC with 64 CPUs and 250 GB of memory available. However, jobs were run without being parallelized (i.e. one job at a time with one thread). Additionally, genome copies were deleted between runs using the same genome to ensure that the run time accounts for the full workflow. All data selected was run on Rocketchip for narrow-peak calling using PE data. The software used was BWA-MEM for alignment, Samtools for deduplication, and MACS3 for peak-calling. The results of this experiment validated the use of the seven major genomes from the UCSC Genome Browser and use of sequence data hosted on the SRA. Additionally, it provides users with estimates of how long Rocketchip should run for different genomes to allow for appropriate computational resource requests. Run times ranged from 0.55 hours to 20.54 hours depending on the genome size and read data used.

Validating use of CUT&RUN and CUT&Tag with Rocketchip

As of June 18, 2024, there are 363,213 ChIP-seq, 494,128 CUT&RUN, and 19,492 CUT&Tag data sets available on the SRA. Due to the increasing use of CUT&RUN and CUT&Tag, we wanted to assess Rocketchip’s ability to effectively process data generated by these mapping techniques. Therefore, we applied Rocketchip to CUT&RUN and CUT&Tag data generated by Akdogan-Ozdilek et al. that sought to characterize the zebrafish epigenome during embryogenesis (51). This data set was chosen due to the thorough documentation of the results and high sequence data quality. In evaluating the performance of Rocketchip for CUT&RUN and CUT&Tag data analysis, we assessed the percentage of reads aligned. The alignment percentage was chosen as a metric due to its significance in assessing the overall data processing efficiency and alignment accuracy. Alignment percentage serves as a key indicator of how effectively Rocketchip handles the unique characteristics of CUT&RUN and CUT&Tag data sets, ensuring that a substantial proportion of reads are appropriately mapped to the reference genome and available for further analysis.

The CUT&RUN data set consisted of nine samples. Six samples were SE reads and corresponded to two replicates each for detection of H3K4me3, H3K27me3, and H3K9me3 (SRR14850825 and SRR14850826, SRR14850827 and SRR14850828, and SRR14850829 and SRR14850830, respectively). The remaining three sets were PE read samples that corresponded to two replicates for the detection of RNA polymerase II and mock IP control using the IgG antibody (SRR14850831 and SRR14850832, and SRR14850833, respectively). The study that originally produced these data sets employed Bowtie2 to align sequences, Samtools to filter aligned sequences, and HOMER for peak-calling. Rocketchip was run using Bowtie2, Samtools, and MACS3 for broad peak-calling. As the data was partially SE and partially PE reads, analyses were conducted separately, with the control being used for the PE analysis due to compatibility. Original alignment percentages were compared to those obtained via Rocketchip (Figure 4). A paired t-test was conducted using each SRA input as an observation, yielding a *p*-value of 0.00302, with Rocketchip consistently producing better alignment percentages for the CUT&RUN data compared to the original analysis. The mean difference in the alignment percentages was 32.66%. ‘We hypothesize that the alignment accuracy differences are related to the use of different Bowtie2 versions, as the original study used Bowtie 2.4.1 to align their CUT&RUN data, whereas Rocketchip used version 2.5.2. The GitHub change log for Bowtie2 version 2.5.1, one version earlier than the one Rocketchip uses, notes: “fixed an issue affecting bowtie2 alignment accuracy” and “fixed a segmentation fault that would occur while aligning SRA data." This highlights the need to revisit and utilize publicly available data, as software updates can improve data processing and thus yield more accurate results. This also highlights Rocketchip's ability to accurately and effectively process CUT&RUN data.

The CUT&Tag data set used for testing in Rocketchip was comprised of six PE read samples. There are three replicates each of H2A.Z at 6 hours and 24 hours post fertilization (SRR14870792, SRR14870793, and SRR14870794 and SRR14870795, SRR14870796, and SRR14870797, respectively). The original study used Bowtie2 for alignment, Samtools for filtering, Picard for deduplication, and MACS2 for peak-calling. We ran Rocketchip using Bowtie2, Samtools, and MACS3. Original alignment percentages were compared to those obtained via Rocketchip (Figure 5). A paired t-test was conducted using each SRA input as an observation, yielding a *p*-value of 0.00015, with Rocketchip resulting in lower alignment percentages for the CUT&Tag data compared to the original analysis. The magnitude of difference in alignment percentages, however, was less than that of the CUT&RUN data, as the mean difference by Rocketchip was -8.41% for CUT&Tag as opposed to +32.66% for CUT&RUN. However, the alignment percentages achieved in our Rocketchip CUT&Tag analysis still surpassed every alignment percentage reported in the original study for their CUT&RUN data, suggesting that Rocketchip is suitable for analyzing CUT&Tag data. We hypothesize that the alignment difference may be due to the usage of both Samtools and Picard in the original CUT&Tag analysis as opposed to just using Samtools in Rocketchip.

Discussion

Rocketchip is distinct as a novel and innovative tool due to its unique approach to automating and allowing for flexibility in ChIP-seq, CUT&RUN, and CUT&Tag data analysis workflows. Unlike traditional methods that often require manual intervention and lack reproducibility, Rocketchip provides a straightforward solution by integrating existing software to automatically run analyses for large-scale datasets. Researchers can easily interchange analysis components and rerun their analysis to identify the most appropriate software options for their data. Additionally, Rocketchip was designed to be user-friendly, making it more accessible to researchers with limited bioinformatics expertise compared to traditional methods that require users to navigate software installation, parameter determination, and command inputs from scratch, promoting broader utilization of publicly available sequence data.

In our analyses of Rocketchip using published experimental ChIP-seq, CUT&RUN, and CUT&Tag data, we demonstrated Rocketchip's ability to handle diverse software combinations seamlessly, which is critical given the variability observed in peak-calling efficiency across different datasets and software tools. We found that the choice of peak caller, aligner, and deduplication method significantly influenced peak-calling outcomes in a data-specific manor. These disparities persisted even when identical software combinations were employed, highlighting the significant impact of dataset-specific factors on the results.

These variations underscore the importance of selecting the appropriate software combination tailored to specific experimental contexts. Moreover, Rocketchip's ability to replicate results across multiple runs (95.14% perfect replication rate) is noteworthy, demonstrating its reliability in ensuring reproducibility in peak calling. While minor variability (4.86%) was observed in peak counts across some combinations, particularly with CisGenome, this was mitigated with other software tools like MACS3 and Genrich, which exhibited more consistent performance. This is consistent with another study that found that CisGenome performed significantly worse with peak-calling compared to MACS, a prior version of MACS3, as well as having lower consistency in peak-calling (52). Among the other cases where variation was observed in peak-calling, the only consistent variable was that STAR was used as the aligner. There is little documentation to suggest that STAR may yield non-deterministic results, with the exception being a Google Groups thread titled “Reproducibility of alt/ref counts w/STAR alignment (through RSEM)” (53), that suggests that the seed-searching portion of the algorithm is deterministic, but the parallelization (i.e. multithreading) may be the cause of the minor differences observed. Ultimately, the persistence of any variation in results using Rocketchip, which provides controlled software versions and parameters, highlights the need for the further investigation of determinism in algorithms commonly used for analyzing genomic data. Furthermore, there is a seemingly constant push to standardize pipelines and tools, but these analyses demonstrate that standardization is likely not possible due to differences in genomic data. There is, therefore, an increasing need for flexible workflows that provide a streamlined approach to facilitate robust data analysis.

For the CUT&RUN analysis, Rocketchip significantly improved the percentage of mapped reads, likely due to the use of the updated Bowtie2 version. However, for CUT&Tag, Rocketchip resulted in lower read alignment that could not be easily explained by differences in the aligner or deduplicator used prior to peak-calling. This unsolved discrepancy highlights the necessity of documenting software versions and parameters used in analyses to enable replication of results.

When using Rocketchip, a few possible limitations to Rocketchip should be considered. First, it should be noted that broad and narrow-peak-calling must be done in separate Rocketchip runs. This is due to the inherent variability in how peaks are represented via read counts. Similarly, SE and PE data sets must be run separately, as these data types are processed differently at the start of the analysis. Additionally, updating software may yield incompatibilities between dependencies; however, Rocketchip ensures version control via Conda to eliminate potential problems with version incompatibilities.

Future goals for Rocketchip include packaging it to be directly pip or conda installable rather than having a user clone the environment from a YAML file. We are also interested in making a Nextflow adaptation to facilitate cloud computing options. We also look forward to expanding Rocketchip’s selection of software, including other peak-calling algorithms, to provide further user customization options. For instance, WASP (54) has recently become a leading method for deduplication of sequence data. It takes an allele-aware approach to mapping reads back to the reference genome, and discards reads that fail to map to the same region of the genome when the complementary read is considered. This approach reduces false positives for allele imbalance and can help improve peak quality in ChIP-seq data. This algorithm has already been incorporated into STAR alignment. Thus, we aim to release a future version of Rocketchip with an option of using WASP, which will thereby circumvent the following deduplication step in the pipeline. Furthermore, because ChIP-seq can be used for various reasons, such as motif finding or differential binding analysis, automation of further analysis is particularly difficult. For future updates of Rocketchip, we hope to include options for motif analysis via HOMER (55) and differential binding analysis via BEDTools (56) followed by gene ontology enrichments.

Future goals for Rocketchip analyses include using Rocketchip to conduct a meta-analysis of all published data sets for a specific transcription factor. With increased sample sizes and varying coverage, this may yield improved accuracy of transcription factor motifs and *in vivo* binding properties. Additionally, we hope to conduct tests on simulated ChIP-seq data to better understand what factors impact ChIP-seq data and how they do so. This includes modeling narrow vs. broad peak regions, as well as varying peak density and coverage, GC-rich regions, overlapping and bimodal peak regions, and levels of PCR duplication. Synthetic data with these modeled characteristics can be run through the various software combinations within Rocketchip to better understand which tools are better suited for different types of data. Ultimately, this would further researchers’ ability to better tailor specific analysis tools to their data.

Key Points

- We've developed Rocketchip, a Python-based command-line tool that integrates existing software, enabling automated ChIP-seq, CUT&RUN, and CUT&Tag data analysis with ease and efficiency.
- Rocketchip allows for increased leverage of publicly available ChIP-seq, CUT&Tag, and CUT&RUN sequence data.
- Rocketchip is designed to facilitate replicability and reproducibility in molecular analyses of protein-DNA interactions by enabling head-to-head comparisons across published data and software.

Data Availability

The synthetic sequence data sets, all analysis scripts, and usage instructions can be found in this GitHub repository: https://github.com/vhaghani26/rocketchip_tests. The Rocketchip source code, installation instructions, and usage instructions can be found in the main Rocketchip GitHub repository: https://github.com/vhaghani26/rocketchip.

Funding

This work was supported by the National Institutes of Health [R01 AA027075 to J.M.L.]; and the Autism Science Foundation and Rett Syndrome Research Trust [22-001 to V.H].

Acknowledgements

The authors would like to thank C. Titus Brown, Katie Lee, Brandon Hom, and Aryss Hearne for their computational expertise, statistical expertise, and guidance.

Conflict-of-Interest Declaration

The authors declare no conflicts of interest.

References

1. Bansal V, Boucher C. Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going? Vol. 18, iScience. Elsevier Inc.; 2019. p. 37–41.
2. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. 2011;39(November 2010):2010–2.
3. National Institutes of Health. Genomics Data Sharing Policy Overview [Internet]. [cited 2023 Sep 18]. Available from: https://sharing.nih.gov/genomic-data-sharing-policy/about-genomic-data-sharing/gds-policy-overview
4. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science (1979). 2007;316(5830):1497–503.
5. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007;4(8):651–7.
6. Xu J, Kudron MM, Victorsen A, Gao J, Ammouri HN, Navarro FCP, et al. To mock or not: A comprehensive comparison of mock IP and DNA input for ChIP-seq. Nucleic Acids Res. 2021;49(3):1–13.
7. Kidder BL, Hu G, Zhao K. ChIP-Seq: Technical considerations for obtaining high-quality data. Vol. 12, Nature Immunology. 2011. p. 918–22.
8. Liang K, Keles undüz. Normalization of ChIP-seq data with control [Internet]. Vol. 13. 2012. Available from: http://www.biomedcentral.com/1471-2105/13/199
9. Skene PJ, Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites.
10. Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Commun. 2019 Dec 1;10(1).
11. Guo Y, Li J, Li CI, Long J, Samuels DC, Shyr Y. The effect of strand bias in Illumina short-read sequencing data. BMC Genomics. 2012;13(1):1–11.
12. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40(10):1–14.
13. Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, et al. Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. DNA Research. 2019;26(5):391–8.
14. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics [Internet]. 2012;13(1):1. Available from: BMC Genomics
15. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol [Internet]. 2011;12(2):R18. Available from: http://genomebiology.com/2011/12/2/R18
16. Gohl DM, Magli A, Garbe J, Becker A, Johnson DM, Anderson S, et al. Measuring sequencer size bias using REcount: A novel method for highly accurate Illumina sequencing-based quantification. Genome Biol. 2019;20(1):1–17.
17. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol [Internet]. 2013;02(5):1. Available from: http://genomebiology.com/2013/14/5/R51%5Cnhttp://www.biomedcentral.com/1471-2164/13/1%5Cnhttp://genomebiology.com/2011/12/2/R18
18. Schirmer M, Ijaz UZ, D’Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 2015;43(6).

19. Wardle FC, Tan H. A ChIP on the shoulder? Chromatin immunoprecipitation and validation strategies for ChIP antibodies. Vol. 4, F1000Research. F1000 Research Ltd; 2015.
20. Teng M, Irizarry RA. Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data. Genome Res. 2017 Nov 1;27(11):1930–8.
21. Koster J, Rahmann S. Snakemake — a scalable bioinformatics workflow engine. Bioinformatics. 2012;28(19):2520–2.
22. National Center for Biotechnology Information. Sequence Read Archive Toolkit [Internet]. [cited 2023 Sep 18]. Available from: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software
23. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. 2002;996–1006.
24. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequencing Data [Internet]. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
25. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
26. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357–9.
27. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan;29(1):15–21.
28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
29. Broad Institute. Picard Toolkit [Internet]. 2019 [cited 2023 Sep 18]. Available from: https://github.com/broadinstitute/picard
30. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Available from: http://picard.sourceforge.net/.
31. Ram F, Friederike D, Diehl S. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42(187–191).
32. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9).
33. Gaspar J. Genrich: Detecting Sites of Genomic Enrichment [Internet]. 2021 [cited 2023 Sep 18]. Available from: https://github.com/jsh58/Genrich
34. Zhang Y, Lin YH, Johnson TD, Rozek LS, Sartor MA. PePr: A peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. Bioinformatics. 2014 Sep 15;30(18):2568–75.
35. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol. 2008 Nov;26(11):1293–300.
36. Anaconda Software Distribution. Conda Computer Software [Internet]. 2023 [cited 2023 Sep 18]. Available from: https://anaconda.com
37. Rube HT, Lee W, Hejna M, Chen H, Yasui DH, Hess JF, et al. Sequence features accurately predict genome-wide MeCP2 binding in vivo. Nat Commun. 2016;7.
38. Namani A, Liu K, Wang S, Zhou X, Liao Y, Wang H, et al. Genome-wide global identification of NRF2 binding sites in A549 non-small cell lung cancer cells by ChIP-Seq reveals NRF2 regulation of genes involved in focal adhesion pathways. Aging. 2019;11(24):12600–23.
39. Richart L, Picod-Chedotel ML, Wassef M, Macario M, Aflaki S, Salvador MA, et al. XIST loss impairs mammary stem cell differentiation and increases tumorigenicity through Mediator hyperactivation. Cell. 2022 Jun 9;185(12):2164-2183.e25.
40. Morita T, Hayashi K. Actin-related protein 5 functions as a novel modulator of MyoD and MyoG in skeletal muscle and in rhabdomyosarcoma. Elife. 2022 Mar 1;11.

41. Pradhan SJ, Reddy PC, Smutny M, Sharma A, Sako K, Oak MS, et al. Satb2 acts as a gatekeeper for major developmental transitions during early vertebrate embryogenesis. Nat Commun. 2021 Dec 1;12(1).
42. Sidlowski P, Czerwinski A, Liu Y, Liu P, Teng RJ, Kumar S, et al. OLA1 Phosphorylation Governs the Mitochondrial Bioenergetic Function of Pulmonary Vascular Cells. Am J Respir Cell Mol Biol. 2023 Apr 1;68(4):395–405.
43. Jaura R, Yeh SY, Montanera KN, Ialongo A, Anwar Z, Lu Y, et al. Extended intergenic DNA contributes to neuron-specific expression of neighboring genes in the mammalian nervous system. Nat Commun. 2022 Dec 1;13(1).
44. Edwards SL, Erdenebat P, Morphis AC, Kumar L, Wang L, Chamera T, et al. Insulin/IGF-1 signaling and heat stress differentially regulate HSF1 activities in germline development. Cell Rep. 2021 Aug 31;36(9).
45. Rawal Y, Qiu H, Hinnebusch AG. Distinct functions of three chromatin remodelers in activator binding and preinitiation complex assembly. PLoS Genet. 2022 Jul 6;18(7).
46. Bellec M, Dufourt J, Hunt G, Lenden-Hasse H, Trullo A, Zine El Aabidine A, et al. The control of transcriptional memory by stable mitotic bookmarking. Nat Commun. 2022 Dec 1;13(1).
47. Wu M, Xu Y, Li J, Lian J, Chen Q, Meng P, et al. Genetic and epigenetic orchestration of Gfi1aa-Lsd1-cebpa in zebrafish neutrophil development. Development. 2021 Sep 1;148(17).
48. Wei W, Liu Y, Qiu Y, Chen M, Wang Y, Han Z, et al. Characterization of Acetylation of Histone H3 at Lysine 9 in the Trigeminal Ganglion of a Rat Trigeminal Neuralgia Model. Oxid Med Cell Longev. 2022;2022.
49. Xu W, He C, Kaye EG, Li J, Mu M, Nelson GM, et al. Dynamic control of chromatin-associated m6A methylation regulates nascent RNA synthesis. Mol Cell. 2022 Mar 17;82(6):1156-1168.e7.
50. Pelletier A, Mayran A, Gouhier A, Omichinski JG, Balsalobre A, Drouin J. Pax7 pioneer factor action requires both paired and homeo DNA binding domains. Nucleic Acids Res. 2021 Jul 21;49(13):7424–36.
51. Akdogan-Ozdilek B, Duval KL, Meng FW, Murphy PJ, Goll MG. Identification of chromatin states during zebrafish gastrulation using CUT&RUN and CUT&Tag. Developmental Dynamics. 2022 Apr 1;251(4):729–42.
52. De Boer BA, Van Duijvenboden K, Van Den Boogaard M, Christoffels VM, Barnett P, Ruijter JM. OccuPeak: ChIP-seq peak calling based on internal background modelling. PLoS One. 2014 Jun 17;9(6).
53. Hoff A, Dobin A. Reproducibility of alt/ref counts w/STAR alignment (through RSEM) [Internet]. 2020 [cited 2024 Jun 17]. Available from: https://groups.google.com/g/rna-star/c/kQGfbQhezsU/m/1RpoIq4gBgAJ
54. Van De Geijn B, Mcvicker G, Gilad Y, Pritchard JK. WASP: Allele-specific software for robust molecular quantitative trait locus discovery. Vol. 12, Nature Methods. Nature Publishing Group; 2015. p. 1061–3.
55. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell. 2010 May 28;38(4):576–89.
56. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.

Figures and Tables

Figure 1. Rocketchip Pipeline for ChIP-seq Data Analysis. Raw sequencing data is aligned to a reference genome and processed to generate bigwig files for data visualization and delineate ChIP-seq peaks.

Figure 2. Peak Counts for all Software Combinations. ChIP-seq data from Namani et al. 2019 and Rube et al. 2016 were run through all software combinations in Rocketchip three times each. Raw peak counts were log2 transformed and plotted in the heatmap. Darker red corresponds to higher peak counts while darker blue corresponds to lower peak counts. Gray corresponds to “NA” values, as PePr and CisGenome cannot be run without a control. The heatmap was created using R (v4.2.3) with a kernel (r-irkernel v1.3.2) in Jupyter Notebook (v1.0.0) using the following packages: ComplexHeatmap (v2.14.0), dplyr (v1.1.4), tidyr (v1.3.1), reshape2 (v1.4.4), stringr (v1.5.1), and MASS (v7.3.60.0.1).

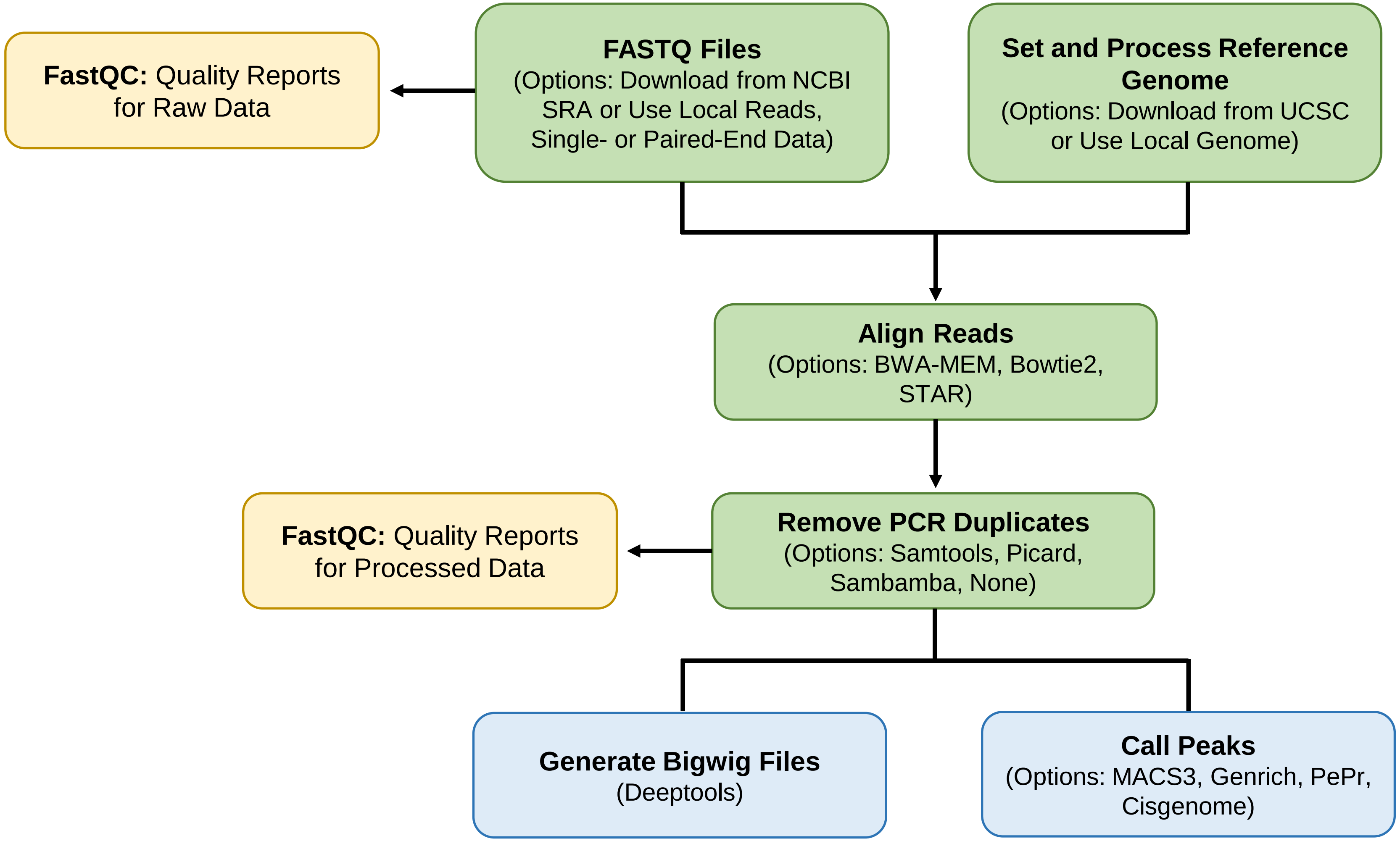
Table 1. Variation in Called Peaks. ChIP-seq data from Namani et al. 2019 and Rube et al. 2016 were run through all software combinations in Rocketchip three times each. This table depicts all combinations of software and data in which peak counts were not replicated perfectly each of the three runs, exhibiting variation in peak-calling. “Project” details the source of the data set. “Control” refers to whether a control was used or excluded during peak-calling. “Aligner”, “Peak Caller”, and “Deduplicator” correspond to the sequence aligner, peak caller, and deduplicator tool used for the Rocketchip run, respectively. “Peak Count Range” represents the minimum and maximum peak counts for the Rocketchip run. “Difference in Called Peaks” represents the range between the minimum and maximum peak counts, highlighting the magnitude of variation in peak-calling across each of the three trial runs.

Figure 3. Rocketchip Execution Times per Genome. This is a horizontal bar plot representing how long (in hours) each sample per genome took to run. The label on the bars represents strictly hours (i.e. not hours and minutes). The color of the bars corresponds to which organism the sample comes from and which genome it was run with. The bar plot was created using R

(v4.2.3) with a kernel (r-irkernel v1.3.2) in Jupyter Notebook (v1.0.0) using the following packages: ggplot2 (v3.4.4) and dplyr (v1.1.4).

Figure 4. Rocketchip Alignment Percentages for CUT&RUN Data. This is a grouped bar plot comparing the percent of raw reads aligned by Akdogan-Ozilek et al. 2023 (red) compared to Rocketchip (blue) for the CUT&RUN data. The bar plot was created using R (v4.2.3) with a kernel (r-irkernel v1.3.2) in Jupyter Notebook (v1.0.0) using the following packages: ggplot2 (v3.4.4), dplyr (v1.1.4), and tidyr (v1.3.1).

Figure 5. Rocketchip Alignment Percentages for CUT&Tag Data. This is a grouped bar plot comparing the percent of raw reads aligned by Akdogan-Ozilek et al. 2023 (red) compared to Rocketchip (blue) for the CUT&Tag data. The bar plot was created using R (v4.2.3) with a kernel (r-irkernel v1.3.2) in Jupyter Notebook (v1.0.0) using the following packages: ggplot2 (v3.4.4), dplyr (v1.1.4), and tidyr (v1.3.1).



FastQC: Quality Reports for Raw Data

FASTQ Files
(Options: Download from NCBI SRA or Use Local Reads, Single- or Paired-End Data)

Set and Process Reference Genome
(Options: Download from UCSC or Use Local Genome)

Align Reads
(Options: BWA-MEM, Bowtie2, STAR)

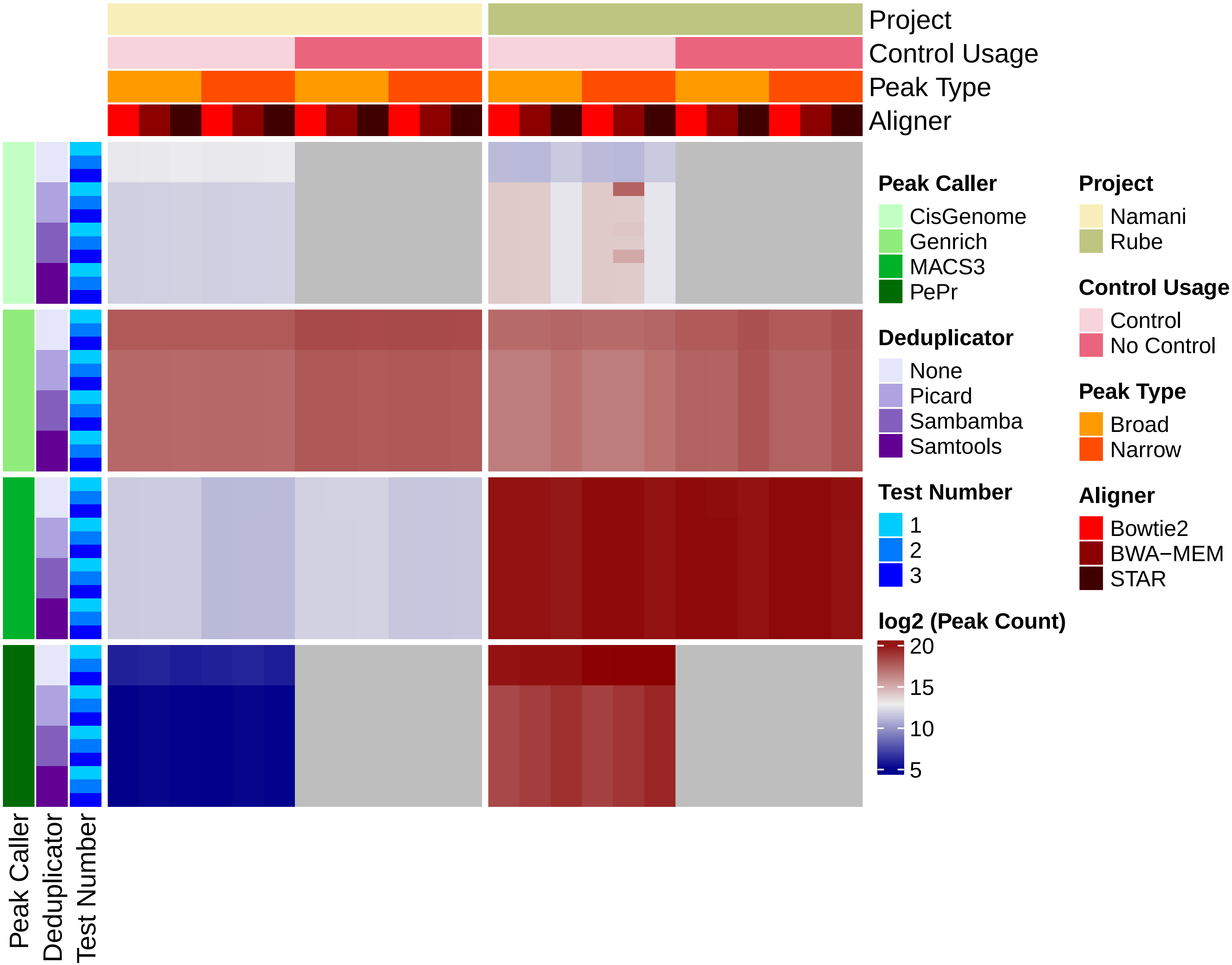
FastQC: Quality Reports for Processed Data

Remove PCR Duplicates
(Options: Samtools, Picard, Sambamba, None)

Generate Bigwig Files
(Deeptools)

Call Peaks
(Options: MACS3, Genrich, PePr, Cisgenome)

Figure 1



Project
Control Usage
Peak Type
Aligner
Peak Caller
CisGenome
Genrich
MACS3
PePr
Deduplicator
None
Picard
Sambamba
Samtools
Test Number
1
2
3
log2 (Peak Count)
20
15
10
5
Project
Namani
Rube
Control Usage
Control
No Control
Peak Type
Broad
Narrow
Aligner
Bowtie2
BWA−MEM
STAR
Peak Caller
Deduplicator
Test Number

Figure 2

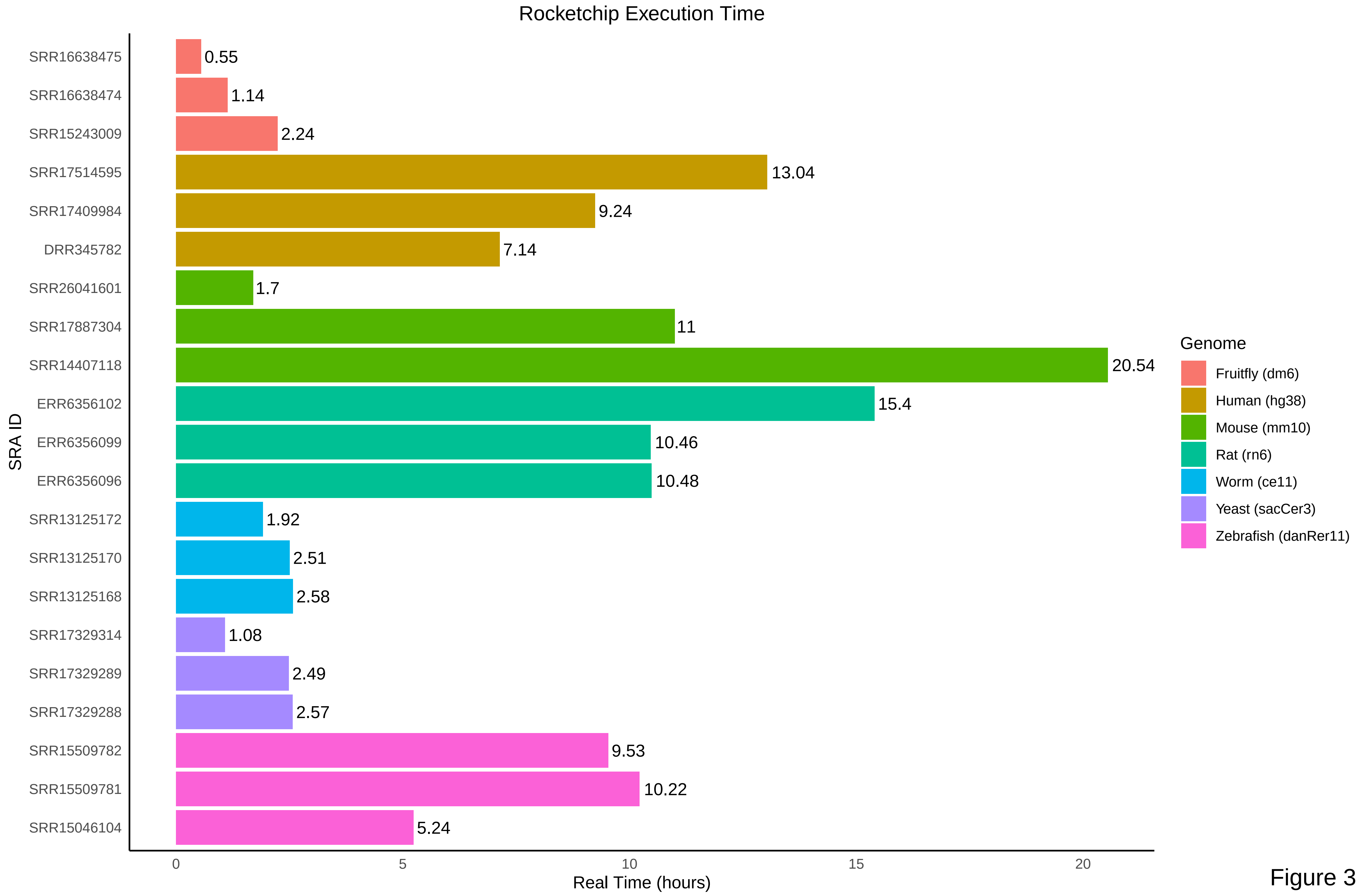

Rocketchip Execution Time
SRA ID
SRR16638475
0.55
SRR16638474
1.14
SRR15243009
2.24
SRR17514595
13.04
SRR17409984
9.24
DRR345782
7.14
SRR26041601
1.7
SRR17887304
11
SRR14407118
20.54
ERR6356102
15.4
ERR6356099
10.46
ERR6356096
10.48
SRR13125172
1.92
SRR13125170
2.51
SRR13125168
2.58
SRR17329314
1.08
SRR17329289
2.49
SRR17329288
2.57
SRR15509782
9.53
SRR15509781
10.22
SRR15046104
5.24
0
5
10
15
20
Real Time (hours)
Genome
Fruitfly (dm6)
Human (hg38)
Mouse (mm10)
Rat (rn6)
Worm (ce11)
Yeast (sacCer3)
Zebrafish (danRer11)

Figure 3

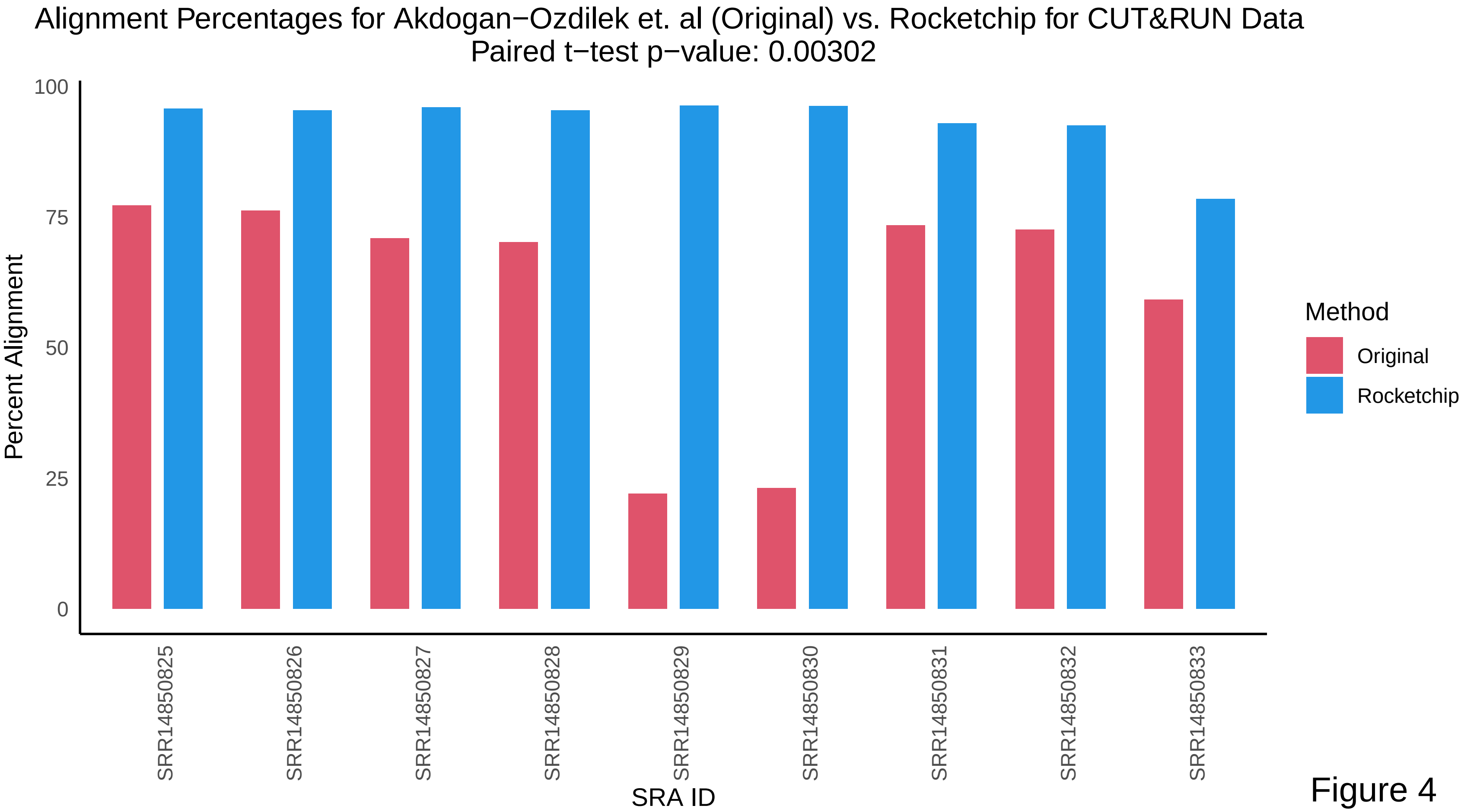

Alignment Percentages for Akdogan−Ozdilek et. al (Original) vs. Rocketchip for CUT&RUN Data
Paired t−test p−value: 0.00302
Percent Alignment
100
75
50
25
0
SRR14850825
SRR14850826
SRR14850827
SRR14850828
SRR14850829
SRR14850830
SRR14850831
SRR14850832
SRR14850833
SRA ID
Method
Original
Rocketchip

Figure 4

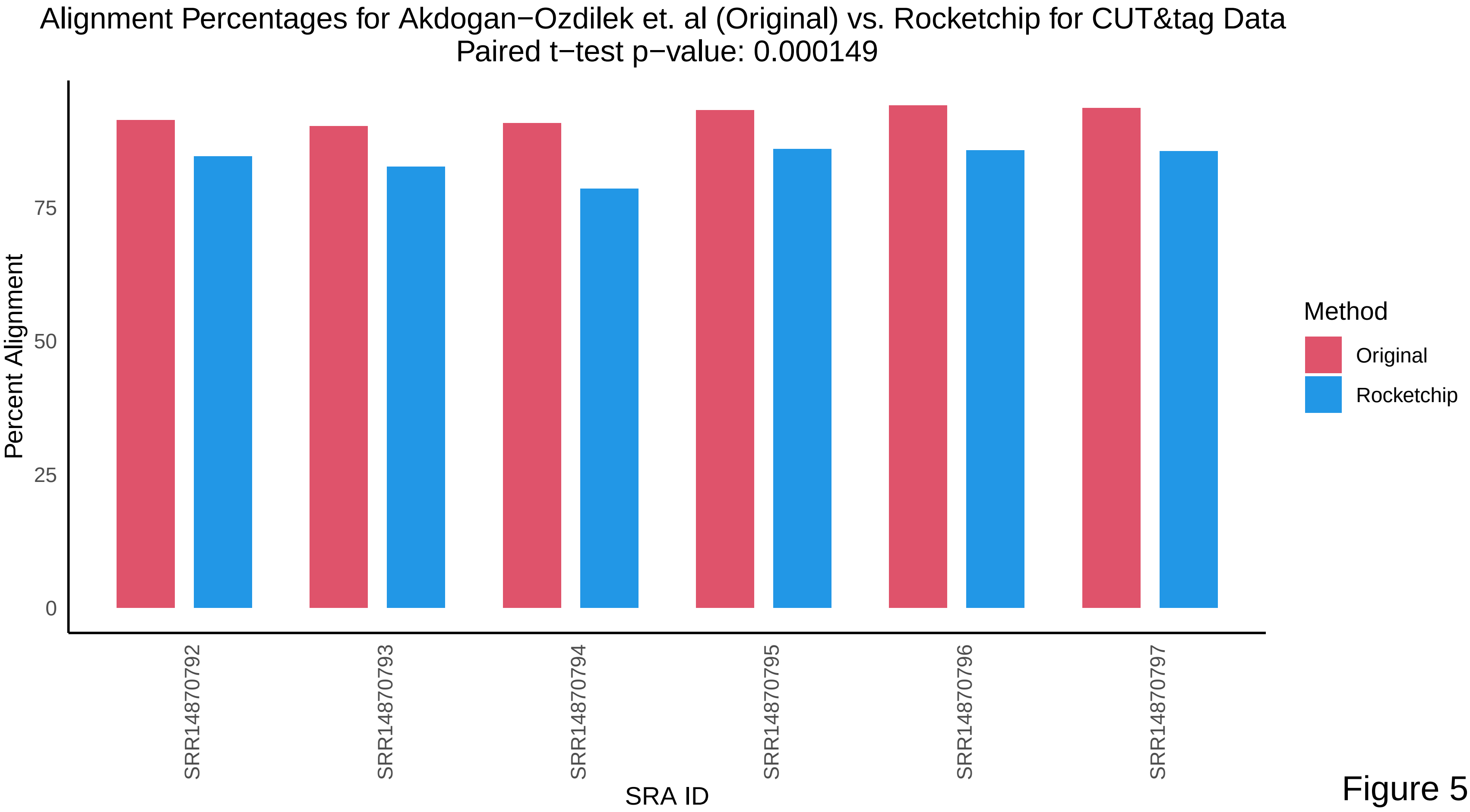

Alignment Percentages for Akdogan−Ozdilek et. al (Original) vs. Rocketchip for CUT&tag Data
Paired t−test p−value: 0.000149
Percent Alignment
0
25
50
75
SRR14870792
SRR14870793
SRR14870794
SRR14870795
SRR14870796
SRR14870797
SRA ID
Method
Original
Rocketchip

Figure 5

Project	Control	Peak Type	Aligner	Peak Caller	Deduplicator	Peak Count Range	Difference in Called Peaks
Rube	No Control	Broad	STAR	Genrich	Sambamba	252158-252160	2
				MACS3	Sambamba	1157876-1157882	6
		Narrow	STAR	Genrich	Sambamba	252158-252160	2
				MACS3	Sambamba	1168345-1168346	1
	With Control	Broad	STAR	Genrich	Sambamba	127542-127543	1
				MACS3	No Deduplication	1002670-1002671	1
					Sambamba	1004044-1004045	1
				PePr	Sambamba	568554-568556	2
		Narrow	STAR	Genrich	Sambamba	127542-127543	1
				MACS3	Sambamba	1143541-1143542	1
				PePr	No Deduplication	1713483-1713484	1
					Sambamba	721307-721309	2
			BWA-MEM	Cisgenome	Picard	15553-175097	159544
					Sambamba	15553-35909	20356

Table 1