




CRED: a rapid peak caller for Chem-seq data

Jason Lin^{1, 2}, Tony Kuo², Paul Horton^{3, 4}, and Hiroki Nagase¹

1 Laboratory of Cancer Genetics, Chiba Cancer Center Research Institute, Chuo-ku, Chiba, Japan **2** Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo, Japan **3** Institute of Medical Informatics, National Cheng Kung University, Tainan, Taiwan **4** Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

DOI: [10.21105/joss.01423](https://doi.org/10.21105/joss.01423)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Submitted: 10 April 2019

Published: 05 May 2019

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

Background

Chem-Seq Read Enrichment Discovery (CRED) is a rapid peak caller written in C for next-generation sequencing (NGS) data, particularly designed with analyzing affinity-enrichment sequencing experiments with pyrrole-imidazole polyamides. Pyrrole-imidazole (PI) polyamides are synthetic molecules, which have primary sequences composed of *N*-methylpyrrole and *N*-methylimidazole subunits with highly sequence-specific DNA minor-groove binding (Dervan & Edelson, 2003). Upon functionalization with other small molecules such as alkylating agents (Hiraoka et al., 2015) or histone deacetylase inhibitors (Pandian et al., 2014), PI polyamides provide a conduit for those small molecules to interact with specific regions of the genome. PI polyamides' relatively short recognition motif and molecular weight, however, can result in generally smaller binding surfaces in polyamide-DNA interactions compared to interactions of DNA with other biomolecules, for instance proteins and transcription factors. A direct consequence of this difference in interaction leads to a mixture of broad and narrow peaks in sequencing experiments conducted with PI polyamides (Chem-seq) that can be atypical of other NGS experiments (Lin et al., 2016). To properly analyze Chem-seq data necessitates the creation of a Chem-seq specific computational tool that can analyze general regions of enriched precipitation of polyamide-DNA ligands (a process known as *peak calling*) from sequencing reads in the genome. We previously designed and reported a workflow to characterize genomic sites enriched with PI polyamide-bound DNA fragments, but the approach required extended preprocessing to convert aligned reads (typically stored in BAM, a compressed binary standard) to BED files, a popular tab-delimited flat text format for storing positional data in the genome. As NGS data can reach upwards of tens to hundreds of gigabytes, this conversion step required a significant amount of time and buffer storage; in addition, the performance and post-processing of the output unnecessarily lengthened the workflow further and hindered throughput. Such shortcomings necessitated computational improvements that remain unmet in the field of Chem-seq research.

Overview

Despite that various peak callers for analyzing NGS data exist, most embrace the philosophy that peaks could be modeled as Poisson scattering events; while this approach works well in most ChIP-seq cases, various publications have reported deficiencies of this algorithm in applications such as DNase-seq.

In our research of pyrrole-imidazole polyamides, a class of DNA minor-groove binders, we have also noticed that MACS (Zhang et al., 2008), a popular model-based peak caller,

suffered from similar issues in the analysis of affinity-enriched DNA fragments sequenced by Ion Torrent systems (a method we hereafter will refer as “Chem-seq”). We previously proposed a coverage-based approach (Lin et al., 2016), in which we employed Perl-based diffReps (Shen et al., 2013) as the initial candidate selection component, followed by bootstrapped Kolmogorov-Smirnov comparisons to characterize Chem-seq peaks.

This workflow, however, required extensive pre- and post-processing of data, as Perl lacked a direct library to access sequencing data stored in BAM files, a format standard shared by various short-read aligners and NGS tools. While the popular BioPerl package included APIs for processing BAM files, the large amount of dependencies (most of which being unrelated to the workflow) and the size of the BioPerl library made it an undesirable for implementation. There were also performance issues associated with external system calls and the reliance on R to perform more complex mathematical computations.

CRED tries to address the aforementioned shortcomings by streamlining the existing workflow. We chose to develop CRED in C to take advantage of HTSLib, a native C library for processing NGS data (H. Li et al., 2009). Access to HTSLib led to time savings in data preprocessing, as the program could now accept BAM files as direct inputs. Writing the program in C also provided performance improvements in computation, and overall eliminated the need to rely on R, along with the associated need to save and retrieve intermediate output.



Figure: Example of a Chem-seq site by CRED and MACS in Integrated Genome Viewer (IGV). LS180 cells were treated with either a 9-bp biotinylated PI polyamide (“treatment”) or DMSO (“control”) and affinity precipitated with streptavidin. After Ion Torrent sequencing, reads were aligned with TMAP, followed by peak calling with either MACS 1.4.2 or CRED. Regions boxed with red dotted lines indicate putative regions of positive enrichment identified by CRED; Top track, treatment; bottom track,

control.

CRED accepts a pair of treatment ('pulldown') and control ('input') coordinate-sorted and indexed BAM files from Chem-seq experiments. The program then compiles a list of preliminary candidates and tests such regions against the hypothesis that there is significant enrichment compared to the same site in a control track, either via Welch's t (Majumder & Bhattacharjee, 1973) or Kolmogorov-Smirnov test (J. Durbin, 1973). The output is reported in a BED-like format to standard output, so they can be easily piped into a Perl array or R vector within a larger workflow script. This output format allows results to be easily visualized in genome browsers such as IGV and requires no additional reformatting. While designed with processing Chem-seq data in mind, CRED may also be compatible with other NGS applications in cases where reads may be too heterogeneous to fit a strict Poisson mixture model.

Acknowledgement

This work was supported by Grant-in-Aids for Scientific Research B and for Young Scientists B from Japan Society for the Promotion of Science (JP17H03602 to HN and JP17K15047 to JL) as well as Japan Agency for Medical Research and Development (AMED, JP17cm0106510, JP17ck0106263 and JP17ck0106356 to HN; JP18ck0106422 to HN and JL). High-performance computing was provided by the Institute of Medical Science at the University of Tokyo as well as AIST Artificial Intelligence Research Center, courtesy of PH. We would also like to thank Prof. Seiya Imoto at the Institute of Medical Science, the University of Tokyo, for providing supercomputing support and technical advice on Chem-seq. JL held a visiting appointment at AIST for the duration of this project.

References

- Dervan, P., & Edelson, B. (2003). Recognition of the dna minor groove by pyrrole-imidazole polyamides. *Current Opinion in Structural Biology*, 13(3), 284–299. doi:[10.1016/S0959-440X\(03\)00081-2](https://doi.org/10.1016/S0959-440X(03)00081-2)
- Durbin, J. (1973). *Distribution theory for tests based on the sample distribution function*. Society for Industrial & Applied Mathematics. doi:[10.1137/1.9781611970586](https://doi.org/10.1137/1.9781611970586)
- Hiraoka, K., Inoue, T., Taylor, R., Watanabe, T., Koshikawa, N., Yoda, H., Shinohara, K., et al. (2015). Inhibition of kras codon 12 mutants using a novel dna-alkylating pyrrole-imidazole polyamide conjugate. *Nature Communications*, 6(6706). doi:[10.1038/ncomms7706](https://doi.org/10.1038/ncomms7706)
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16), 2078–2079. doi:[10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352)
- Lin, J., Hiraoka, K., Watanabe, T., Kuo, T., Shinozaki, Y., Takatori, A., Koshikawa, N., et al. (2016). Identification of binding targets of a pyrrole-imidazole polyamide kr12 in the ls180 colorectal cancer genome. *PLoS ONE*, 11(10), 1–19. doi:[10.1371/journal.pone.0165581](https://doi.org/10.1371/journal.pone.0165581)
- Majumder, K., & Bhattacharjee, G. (1973). Algorithm as 63: The incomplete beta integral. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 22(3), 409–411. doi:[10.2307/2346797](https://doi.org/10.2307/2346797)

Pandian, G., Sato, S., Anandhakumar, C., Taniguchi, J., Takashima, K., Syed, J., Han, L., et al. (2014). Identification of a small molecule that turns on the pluripotency gene circuitry in human fibroblasts. *ACS Chemical Biology*, 9(12), 2729–2736. doi:[10.1021/cb500724t](https://doi.org/10.1021/cb500724t)

Shen, L., Shao, N.-Y., Liu, X., Maze, I., Feng, J., & Nestler, E. (2013). DiffReps: Detecting differential chromatin modification sites from chip-seq data with biological replicates. *PLoS ONE*, 8(6), 1–13. doi:[10.1371/journal.pone.0065598](https://doi.org/10.1371/journal.pone.0065598)

Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nusbaum, C., et al. (2008). Model-based analysis of chip-seq (macs). *Genome Biology*, 9(9), R137. doi:[10.1186/gb-2008-9-9-r137](https://doi.org/10.1186/gb-2008-9-9-r137)