GENOMICS AND COMPUTATIONAL BIOLOGY Vol. 4 No. 2 (2018): e100042

Resource

Automated Processing of NGS Data from Raw Sequencing Files to Ready-To-Use Information Tables for Genome Modelling

Robert Deelen[1], Martin Wieland[1], Susanne Gerber[1], David Fournier[1],

[1]Faculty of Biology and Centre for Computational Sciences, Institute for Developmental Biology and Neurobiology. Johannes Gutenberg-Universität Mainz, Staudingerweg 9, 55128 Mainz, Germany

*Correspondence: dfournie@uni-mainz.de

Received 2017-10-27; Accepted 2018-03-02

ABSTRACT

Epigenetic features such as histone and DNA modifications are important mechanisms for the regulation of gene expression and for cell and tissue development. As a result, extensive efforts are currently undertaken using next-generation sequencing (NGS) to generate vast amounts of data regarding the epigenetic regulation of genomes. Several tools and frameworks for the processing of these NGS data have been developed in the last decade. Nevertheless, each user still bares the challenge to integrate all these tasks to perform the analysis. This procedure is not only tedious but also resource-intensive due to the putative large processing power involved. To automate, standardize and speed up the handling of NGS data, with focus on ChIP-seq data, we present a user-friendly pipeline that automatically processes a list of sequencing data files and returns a ready-to-use purified table for subsequent modelling or analysis attempts.

KEYWORDS

Epigenetics; Histone modification; Big data; Pipeline; Genome modelling; ChIP-seq; NGS;

AVAILABILITY AND REQUIREMENTS

- Project name: Next Gen Bender
- https://github.com/SusanneGerber/Epigenetics-Pipeline
- Systems: Linux, Slurm and LSF
- License: GNU

INTRODUCTION

Epigenetics is a strong component of living systems, that comprise all the mechanisms helping to convert a genotype (sum of all genetic information) of an organism into phenotypic traits (for instance, color of hair or number of digits) [1]. On the molecular side, epigenetic features regulate genes and usually maintain their regulation at a sustained pace through the different cell divisions during development, a mechanism called epigenetic inheritance [2]. Before the genome era, the molecular relationship between genetics and epigenetics was poorly understood. The situation changed with the development of molecular methods to study genome function regulation and structure, such as ChIP-seq (Chromatin Immunoprecipitation sequencing, to study binding of features to DNA), RNA-seq (to study gene expression) or bisulfite sequencing (to study DNA methylation, a modification that affects gene expression and DNA folding). Data generated with these methods (epitomized by the term "genomics") have recently shown emerging evidence for a role of epigenetics in gene regulation, gene expression and pathologies such as cancer and neurodegenerative diseases [3–5]. Genomics has also been crucial in revealing the influence of epigenetics in embryonic development [6] and aging [7]. Because of these new insights, extensive efforts are currently undertaken to study and map epigenetic features in various tissues and organisms [8, 9].

The standard methods for the analysis of epigenetic features, such as ChIP-seq or RNA-seq require next-generation sequencing (NGS). NGS is a cost and time efficient way to acquire knowledge about features related to a given genomic position. These features can be roughly divided into three categories: the functional features, such as binding of transcription factors or regulators, performed via the ChIP-seq technology and affiliated methods; structural features, via direct structural information (Hi-C) or by probing the accessibility of DNA to binding factors (ATAC-seq, FAIRE-seq); finally, regulatory features, mainly via histone modifications, which directly or indirectly influence the binding of factors to DNA and so play a major role in epigenetic regulation of the genome. Combinations or changes of these three types of features are usually associated to different status of chromatin. Chromatin can roughly be divided into active, inactive and repressed state, with more complex states that remain to be investigated.

In order to understand the intertwined mechanics of the different epigenetic features available (such as histone modifications and DNA methylation sites), several modeling approaches have been used previously: Markov Models [10, 11], Bayesian methods [12] or Boolean networks [13]. Polymer

Publisher: Kernel Press. Copyright © (2018) the Author(s). This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

models derived from Hi-C data, have also contributed to understanding the relationship between epigenetic features and the structure of the DNA [14]. As these methods usually require the acquisition of information (in the definition of information theory [15]) by the description of "peaks" (a cluster of sequencing events associated to a genomic position or region that form one unit of information), it would be useful to have an automated procedure to generate ready-to-use information tables from raw NGS data. Several pipelines have been released to process NGS data, either as standalone software (KNIME [16] and SeqAn [17]), or online (Galaxy, Mobyle [18]). Nevertheless, these tools provide output files listing peaks that require further processing before modelling can be performed (SeqAn, also requires skills in C++ scripting), or deliver information which is difficul to process further, such as genomic functional annotation (case of KNIME). Pipelines can be also rather sophisticated and require investment in the computational design of the pipeline from the user (SeqAn, already mentioned, or Conveyor [19]).

To simplify procedures and give a comprehensible output ready for genome modelling and analysis, we developed a pipeline that takes raw sequencing files generated from ChIP-seq experiments as input and extracts all the non-redundant epigenetic information they contain via a peak calling step. Overlap between the peaks of the different features studied is then assessed to compare co-occurence of features at any place on the genome. Finally, the different peaks are presented in a table where each column is a feature and each row a peak. For each peak detected in any ChIP-seq experiment, the pipeline thus computes the status of the other samples at the exact same position via feature-to-feature comparisons such as correlations or clustering. As a result, this table contains exhaustive epigenetic information located in the raw input sequencing files. The pipeline handles all standard tasks used in NGS processing such as downloading of sequences, quality control step, alignment to reference genomes and peak calling. Compared to other tools for automated ChIP-seq processing such as [20], who focus on functional peaks related to genes and overall visualization, our contribution lies in the presentation of the results in an information format ready to use for modellers and analyst. The pipeline can be easily adapted to include additional types of sequencing data, for instance DNA methylation states or gene expression data.

METHODS

Pipeline Overview

An overview of the different steps covered by the pipeline is presented in Figure 1. The code is written in Python 3 and uses a CSV file as input. A first step of the analysis is to set the different global variables: (i) name of the input file, (ii) number of processors available, (iii) number of parallel processes to use in order to

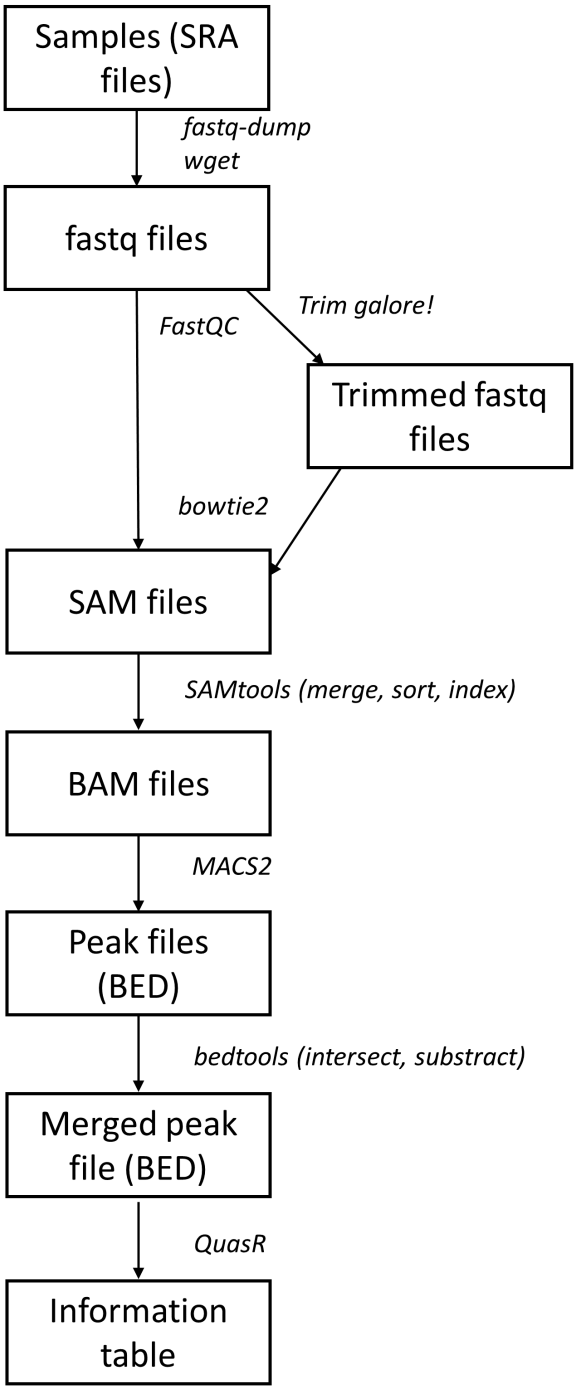

Figure 1: Overview of the pipeline.

enhance processing speed. The input file contains the names of the different epigenetic marks to consider, the associated file names containing the Chip-Seq data and eventually the different URL necessary to download them. Accepted file-formats are the standard formats fastq (usually compressed) or Sequence Read Archive (SRA). Both formats contain sequences of reads along with metadata such as sequencing quality. URL of files are not mandatory and the user has the liberty to either use his own data or to refer to public SRA files via their identifiers. By reading the CSV input file, the pipeline creates an array of epigenetic marks to be downloaded for each mark and a URL for each file.

The pipeline follows a protocol of tasks launched for each feature to be analyzed, and are parallelized whenever possible. Firstly, the SRA files are processed with fastq-dump to generate fastq files. Afterwards, a quality control (QC) test is performed to skip or trim sequence files with low quality. For this purpose, fastq files are tested with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for different measures such as quality of sequencing, and sequence bias. If the quality is too low and providing that the user allows for it, sequences of poor quality can be automatically removed from the analysis, or trimmed with custom parameter value. The fastq files which have passed the QC step are then aligned to the reference genome (using Bowtie 2 [21]) and saved as SAM files. The SAM files from biological replicates are then merged with SAMtools [22] and compressed as BAM files. The BAM files become sorted and indexed with SAMtools for faster access. Afterwards the peak calling step is performed by MACS2 [23] for each file and peaks are saved as a BED file.

To create the bed file containing the information merged from the different ChIP-seq sequence files, the first file in the list of marks is compared with the following peak bed file in the list. The intersection procedure of the BEDtools software suite finds the overlap region of two peaks and leaves out the non-overlapping part (command "BEDtools subtract"). Peaks are merged if they overlap by at least 50% of their width, a value that we selected to efficiently merge peaks from two ChIP-seq which appear by sight to be at the same position (for instance at a promoter start site). The procedure guarantees that only unique peaks are saved in the final output bed file. This process is repeated until all unique peaks are added, sorted and saved in the master file. Finally, the pipeline calls the R package QuasR [24]. This package provides tools to quantify and analyse sequence reads from NGS platforms, in order to find the enrichment values for each mark at the positions of the peaks, and normalize data in a single step. At the position of a given peak, we define the value of enrichment above input of normalized counts q_i as follows:

$$q_i = \frac{\frac{\text{Feature counts}}{\text{Feature total counts}}}{\frac{\text{Input counts}}{\text{Input total counts}}}, \quad (1)$$

where "total counts" refer to the amount of reads found in the observed sample, either the epigenetic feature considered or the input. The division by the total amount of counts ensures to take size factors into account, such as variations in the overall number of reads during sequencing. Division by input ensures that the values of the features are not biased by background noise but true signals. Prior to the computing of q_i, we add a pseudocount of 1 to the values of counts equal to zero. This avoids division by zero when input is equal to zero and allows the comparisons between samples even in cases where one of them might be equal to zero.

Finally, the pipeline generates a table where the columns are associated to marks and the rows to unique peaks. The table is ready to be processed by the user for modelling or genome analysis. The Python code of the pipeline is provided for two query system: LSF (https://www.ibm.com/support/knowledgecenter/en/SSETD4_9.1.3/QSG/lsf9.1.3_quick_start.html) and Slurm [25].

The software is available at https://github.com/SusanneGerber/Epigenetics-Pipeline

APPLICATION

We tested the pipeline on 12 histone modifications published by Pope et al., 2014 [26], and investigated the relationships between epigenetic features during early development. Embryonic stem cells (ESC) used in the study are a prominent model for epigenetics, as very important epigenetic remodelling events occur during development, thus providing a perfect platform to study epigenetics. Besides, the dataset from Pope et al. is a good source for ChIP-seq data as they provide about 50 different features for ESC, including histone modifications and transcription factors. To execute this example analysis, we run the pipeline on Mogon, the supercomputer of the university of Mainz, delivering the complete analysis for 21 sample fastq files in a wallclock-time of approximately 5 hours. The job was running on 32 cores of a AMD Opteron 6272 node (Cores frequency: 2.1 Ghz). Number of threads was critical. For instance, duration of the bowtie2 step was 191 minutes for 2 threads versus 1 minute for 32 threads.

In the resulting heatmap (Figure 2) the correlation between different marks is visible, associated to three distinct regions of different functionality, corresponding to active, inactive and deeply condensed chromatin. The correlations found seem to agree with rules of epigenetic associations published in previous major genomic studies (e.g. the H3K9 and H3K14 [27]). Several studies [11, 26, 28–30] classified the mammalian genome into three major epigenetic groups: (i) enriched in mark H3K4me3 and associated to active genes, (ii) H3K27me3 associated to rather inactive genes that are still activatable, (iii) H3K9m3 is mostly associated to deeply repressed chromatin. Here, we have shown that we can retrieve those clusters using our custom procedure (Figure 2, right panel). This rather simple example demonstrates the potential of the pipeline to help modellers and computational biologists to convert genomic data into information tables immediately usable for modelling and genome analysis.

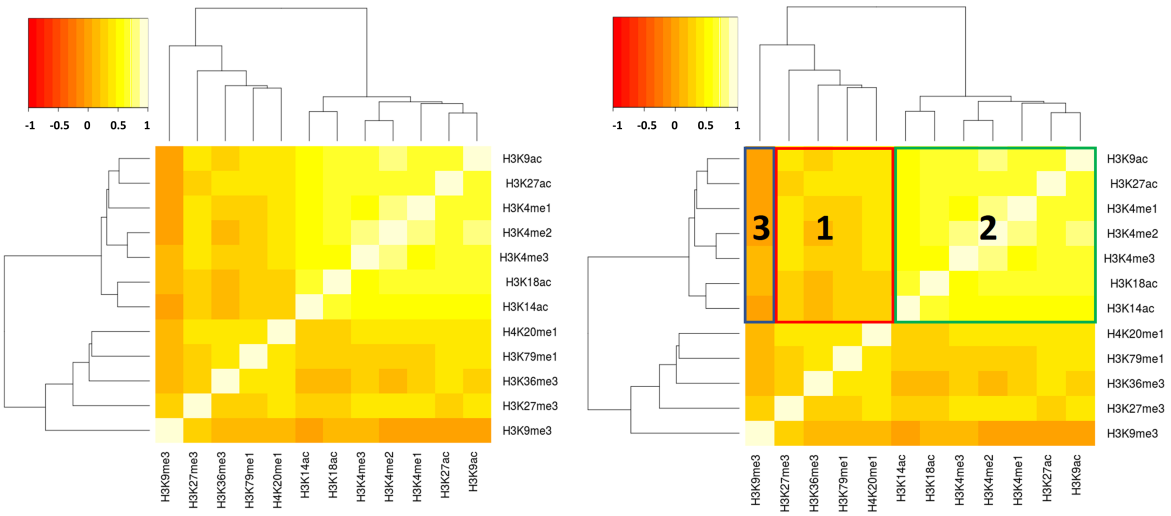

Figure 2: Correlation plot between epigenetic marks generated with the information table produced by the pipeline (using R package QuasR to calculate enrichment values and heatmap.2 command of package gplots). Left: heatmap with no clustering. Right: Three different regions are distinguished using hierarchical clustering: (1) Marks usually associated with gene inactivity; (2) marks associated with gene activity; (3) cluster associated with deeply condensed chromatin.

CONCLUSION

The pipeline presented here utilizes several well-known genomic tools and combines their outputs into one information table summarizing all the peaks found in one or multiple samples. This table is for further investigations of correlations between multiple epigenetic marks. Overall, we believe that the pipeline will be very valuable for all scientists working in the area of computational genomics as it delivers userfriendly information tables that can feed new algorithms or models, and is easy to implement on any workstation or server.

ACKNOWLEDGEMENTS

The project and the work of RD was partly funded by the Impulsfonds Forschungsinitiative Rheinland-Pfalz from the JGU “Deciphering cell identity and function using single-cell data analysis”. The work of SG and DF was funded by the Center for Computational Sciences in Mainz (CSM).

Parts of this research were conducted using the supercomputer Mogon and/or advisory services offered by Johannes Gutenberg University Mainz (hpc.uni-mainz.de), which is a member of the AHRP and the Gauss Alliance e.V.

The authors gratefully acknowledge the computing time granted on the supercomputer Mogon at Johannes Gutenberg University Mainz (hpc.uni-mainz.de).

AUTHOR CONTRIBUTIONS

RD implemented the Slurm pipeline and wrote the article. MW implemented the LSF pipeline. DF designed the study, supervised the project, implemented the classical pipeline and edited the text. SG supervised the project and edited the text.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ABBREVIATIONS

BAM: Binary SAM file

BED: Browser Extensible Data

ChIP-seq: Chromatin Immunoprecipitation Sequencing

H3K27me3: Histone 3 Lysine 27 Trimethylaton

H3K4me3: Histone 3 Lysine 4 Trimethylaton

H3K9me3: Histone 3 Lysine 9 Trimethylaton

NGS: Next-Generation Sequencing

SAM: Sequence Alignment Map

REFERENCES

1. Waddington C. **The epigenotype. 1942.** Int J Epidemiol. 2012;41(1):10–3.
2. Bonasio R, Tu S, Reinberg D. **Molecular signals of epigenetic states**. Science. 2010;330(6004):612–6. doi:10.1126/science.1191078.
3. Guenther M, Levine S, Boyer L, Jaenisch R, Young R. **A chromatin landmark and transcription initiation at most promoters in human cells**. Cell. 2007;130(1):77–88. doi:10.1016/j.cell.2007.05.042.
4. Oldridge DA, Wood AC, Weichert-leahey N, Crimmins I, Winter C, Mcdaniel LD, et al. **Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism**. Nature. 2016;528(7582):418–421. doi:10.1038/nature15540.
5. Soldner F, Stelzer Y, Shivalila CS, Abraham BJ, Jeanne C, Barrasa MI, et al. **Parkinson-associated risk variant in enhancer element produces subtle effect on target gene expression**. Nature. 2016;533(7601):95–99. doi:10.1038/nature17939.

6. Takahashi K, Yamanaka S. **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors**. Cell. 2006;126(4):663–76. doi:10.1016/j.cell.2006.07.024.
7. Sidler C, Kovalchuk O, I K. **Epigenetic Regulation of Cellular Senescence and Aging**. Front Genet. 2017;8:138. doi:10.3389/fgene.2017.00138.
8. Sarda S, Hannenhalli S. **Next-generation sequencing and epigenomics research: a hammer in search of nails**. Genomics Inform. 2014;12(1):2–11. doi:10.5808/GI.2014.12.1.2.
9. Yen A, Kheradpour P, Zhang Z, Heravi-moussavi A, Liu Y, Amin V, et al. **Integrative analysis of 111 reference human epigenomes**. Nature. 2015;518(7539):317–330. doi:10.1038/nature14248.
10. Ernst J, Manolis K. **Discovery and characterization of chromatin states for systematic annotation of the human genome**. Nature Biotechnology. 2010;28(8):817–825. doi:10.1038/nbt.1662.
11. Prakash K, Fournier D. **Histone code and higher-order chromatin folding: A hypothesis**. Genomics and Computational Biology. 2017;3(2):41. doi:10.18547/gcb.2017.vol3.iss2.e41.
12. Love MI, Huska MR, Jurk M, Schoepflin R, Starick SR, Schwahn K, et al. **Role of the chromatin landscape and sequence in determining cell type-specific genomic glucocorticoid receptor binding and gene regulation**. Nucleic acids research. 2017;45(4):1805–1819. doi:10.1093/nar/gkw1163.
13. Prakash K. **A binary combinatorial histone code**. Master thesis. 2012;Aalto University.
14. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping**. Cell. 2014;159(7):1665–1680. doi:10.1016/j.cell.2014.11.021.
15. Shannon C. **The Mathematical Theory of Communication**. Bell System Technical Journal. 1948;27:379–423.
16. Jagla B, Wiswedel B, Coppée J. **Extending KNIME for next-generation sequencing data analysis**. Bioinformatics. 2011;27(20):2907–2909. doi:10.1093/bioinformatics/btr478.
17. Reinert K, Dadi T, Ehrhardt M, Hauswedell H, Mehringer S, Rahn R, et al. **The SeqAn C++ template library for efficient sequence analysis: A resource for programmers**. J Biotechnol. 2017;261:157–168. doi:10.1016/j.jbiotec.2017.07.017.
18. Néron B, Ménager H, Maufrais C, Joly N, Maupetit J, Letort S, et al. **Mobyle: a new full web bioinformatics framework**. Bioinformatics. 2009;25(22):3005–3011. doi:10.1093/bioinformatics/btp493.
19. Linke B, Giegerich R, Goesmann A. **Conveyor: a workflow engine for bioinformatic analyses**. Bioinformatics. 2011;27(7):903–911. doi:10.1093/bioinformatics/btr040.
20. Park S, Kim J, Yoon B, Kim S. **A ChIP-Seq Data Analysis Pipeline Based on Bioconductor Packages**. Genomics Inform. 2017;15(1):11–18. doi:10.5808/GI.2017.15.1.11.
21. Langmead B, Salzberg S. **Fast gapped-read alignment with Bowtie 2**. Nat Methods. 2012;9(4):357–9. doi:10.1038/nmeth.1923.
22. 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. doi:10.1093/bioinformatics/btp352.
23. Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, et al. **Model-based analysis of ChIP-Seq (MACS)**. Genome Biol. 2008;9(9):R137. doi:10.1186/gb-2008-9-9-r137.
24. Gaidatzis D, Lerch A, Hahne F, Stadler MB. **QuasR: Quantification and annotation of short reads in R**. Bioinformatics. 2015;31(7):1130–1132. doi:10.1093/bioinformatics/btu781.
25. Yoo A, Jette M, M G. **SLURM: Simple Linux Utility for Resource Management**. In: Job Scheduling Strategies for Parallel Processing. vol. 2862. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. p. 44–60. doi:10.1007/10968987_3.
26. Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, et al. **Topologically associating domains are stable units of replication-timing regulation**. Nature. 2014;515(7527):402–405. doi:10.1038/nature13986.
27. Karmodiya K, Krebs AR, Oulad-Abdelghani M, Kimura H, Tora L. **H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells**. BMC Genomics. 2012;13(1):424. doi:10.1186/1471-2164-13-424.
28. Boettiger AN, Bintu B, Moffitt JR, Wang S, Beliveau BJ, Fudenberg G, et al. **Super-resolution imaging reveals distinct chromatin folding for different epigenetic states**. Nature. 2016;529(7586):418–422. doi:10.1038/nature16496.
29. Prakash K, Fournier D, Redl S, Best G, Borsos M, Tiwari VK, et al. **Superresolution imaging reveals structurally distinct periodic patterns of chromatin along pachytene chromosomes**. Proceedings of the National Academy of Sciences. 2015;112(47):14635–14640. doi:10.1073/pnas.1516928112.
30. Prakash K, Fournier D. **Evidence for the implication of the histone code in building the genome structure**. Biosystems. 2018;164:49–59. doi:10.1016/j.biosystems.2017.11.005.