Bioinformatics, 37(22), 2021, 4251–4252
doi: 10.1093/bioinformatics/btab413
Advance Access Publication Date: 27 May 2021
Applications Note

Sequence analysis

cfDNApipe: a comprehensive quality control and analysis pipeline for cell-free DNA high-throughput sequencing data

Wei Zhang[1], Lei Wei[1,*], Jiaqi Huang[2], Bixi Zhong[1], Jiaqi Li[1], Hanwen Xu[1], Shuying He[3], Yu Liu[3], Juhong Liu[3], Hairong Lv[1,3] and Xiaowo Wang [1,*]

[1]Ministry of Education Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, Bioinformatics Division, Beijing National Research Center for Information Science and Technology, Department of Automation, Tsinghua University, Beijing 100084, China, [2]Department of Physics, Tsinghua University, Beijing 100084, China and [3]Fuzhou Institute for Data Technology Co., Ltd., Fuzhou, Fujian 350207, China

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on January 16, 2021; revised on April 21, 2021; editorial decision on May 22, 2021; accepted on May 26, 2021

Abstract

Motivation: Cell-free DNA (cfDNA) is gaining substantial attention from both biological and clinical fields as a promising marker for liquid biopsy. Many aspects of disease-related features have been discovered from cfDNA high-throughput sequencing (HTS) data. However, there is still a lack of integrative and systematic tools for cfDNA HTS data analysis and quality control (QC).

Results: Here, we propose cfDNApipe, an easy-to-use and systematic python package for cfDNA whole-genome sequencing (WGS) and whole-genome bisulfite sequencing (WGBS) data analysis. It covers the entire analysis pipeline for the cfDNA data, including raw sequencing data processing, QC and sophisticated statistical analysis such as detecting copy number variations (CNVs), differentially methylated regions and DNA fragment size alterations. cfDNApipe provides one-command-line-execution pipelines and flexible application programming interfaces for customized analysis.

Availability and implementation: https://xwanglabthu.github.io/cfDNApipe/.

Contact: xwwang@tsinghua.edu.cn or weilei92@tsinghua.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recently, Cell-free DNA (cfDNA), which is present in body fluids like plasma and urine, has attracted significant attentions due to its non-invasive nature for disease diagnosis. The rapid development and application of high-throughput sequencing (HTS) technologies provide deep insight into the information contained by cfDNA. Researchers have found lots of signatures that can be derived from cfDNA HTS data, such as fragment size alterations (Cristiano *et al.*, 2019; Mouliere *et al.*, 2018), copy number variations (CNVs) (Jiang *et al.*, 2015) and DNA methylation changes (Li *et al.*, 2018). These features are promising to be applied in various scenarios such as prenatal diagnosis, early cancer detection and therapy response monitoring (Kilgour *et al.*, 2020; Oliveira *et al.*, 2020).

However, there still lacks an integrative and systematic tool for cfDNA HTS data processing and analysis. Researchers need to build their own pipeline with diverse tools and finely tuning the parameters to tackle cfDNA data, which is tedious and time-consuming. To fill this gap, we propose a highly integrated python package named cfDNApipe to deal with widely used cfDNA HTS data including both whole-genome sequencing (WGS) and whole-genome bisulfite sequencing (WGBS) data. cfDNApipe covers essential steps for cfDNA HTS data processing, including raw data pre-processing, quality control (QC) and state-of-the-art statistical analysis such as CNV detection and fragmentation profile analysis. The package is based on conda/bioconda and all its dependencies can be easily installed by one command line in shell.

2 Design and implementation

2.1 cfDNA data analysis workflow

The flowchart of cfDNApipe is shown in Figure 1.The entire workflow includes three major parts, data pre-processing, QC and statistical analysis.

In the pre-processing part, cfDNApipe takes raw fastq files of WGS or WGBS data directly as the input. It removes adapters and then employs bowtie2 (Langmead and Salzberg, 2012) or Bismark

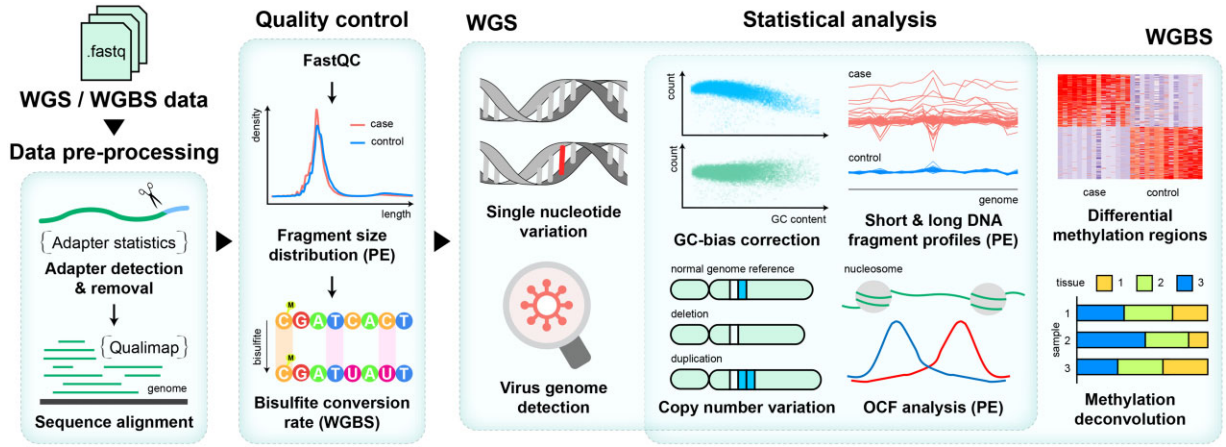

Fig. 1. The flowchart of cfDNApipe. PE, paired-end data

(Krueger and Andrews, 2011) for read alignment of different data types. Post-alignment operations such as sorting, filtering as well as removing duplicates are executed seamlessly.

Multi-level QC functions are provided in cfDNApipe. Quality reports for raw reads and post-alignment features such as coverage at different genomic regions will be generated. It executes fragment length QC to examine whether samples are contaminated. The bisulfite conversion rate can also be calculated to estimate the technical noise for WGBS data.

The statistical analysis step performs comprehensive analysis to obtain multi-aspect signatures. In terms of genomic features, cfDNApipe provides a method to detect unbiased CNVs accompanied with a GC-bias correction approach (Jiang *et al.*, 2015). Besides, it identifies single nucleotide variations as well as the genomic sequences from virus for WGS data. As for methylomics, cfDNApipe distinguishes differentially methylated regions between case and control samples. Deconvolution can be performed according to the methylome profiles to infer the fraction of each component as well as the tissue-of-origin. For fragmentomics, cfDNApipe integrates the nucleosome positioning inference method (Snyder *et al.*, 2016) and alterations of short and long fragmentation ratio are detected at whole-genome scale (Cristiano *et al.*, 2019). Orientation-aware cfDNA fragmentation (OCF) analysis (Sun *et al.*, 2019) is also performed to detect the tissue-specific open chromatin signals. Detailed functions and usage guidelines of cfDNApipe are shown in the Supplementary Materials.

2.2 Implementation

For the convenience of users, cfDNApipe provides automatic downloading and index building for genome references. The whole pipeline can be executed simply with one Python command. An HTML report will be generated for visualization of analysis results, and outputs of each step are well-arranged in their own folder separately.

The pipeline is organized by a built-in dataflow mechanism with a strictly defined up- and down-stream data interface. Therefore, users can execute particular modules or customize their own pipeline through the application programming interface easily.

cfDNApipe provides parallel computing options for maximizing its efficiency. It also contains breakpoint detection to continue the rest of computational tasks after interruption.

3 Conclusion

Here, we propose cfDNApipe as a highly integrated cfDNA HTS data analysis toolkit. Users without sophisticated programming background can start it easily to discover the information in cfDNA HTS data comprehensively and efficiently. This toolkit will make the analysis of cfDNA HTS data much easier in a wide range of scenarios to promote the development and application of cfDNA-related researches.

Acknowledgements

The authors thank Prof. Peiyong Jiang and Prof. Yuk Ming Dennis Lo for sharing cfDNA data. They also thank Zheng Wei and Xianglin Zhang for helpful discussion.

Funding

This work was supported by the National Natural Science Foundation of China [61773230, 61721003].

Conflict of Interest: none declared.

References

Cristiano,S. *et al.* (2019) Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, **570**, 385–389.

Jiang,P. *et al.* (2015) Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc. Natl. Acad. Sci. USA*, **112**, E1317–E1325.

Kilgour,E. *et al.* (2020) Liquid biopsy-based biomarkers of treatment response and resistance. *Cancer Cell*, **37**, 485–495.

Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Li,W. *et al.* (2018) CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Res.*, **46**, e89.

Mouliere,F. *et al.* (2018) Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med*, **10**, eaat4921.

Oliveira,K.C. *et al.* (2020) Current perspectives on circulating tumor DNA, precision medicine, and personalized clinical management of cancer. *Mol. Cancer Res.*, **18**, 517–528.

Snyder,M.W. *et al.* (2016) Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*, **164**, 57–68.

Sun,K. *et al.* (2019) Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome Res.*, **29**, 418–427.