Bioinformatics, 37(16), 2021, 2502–2503
doi: 10.1093/bioinformatics/btaa999
Advance Access Publication Date: 1 December 2020
Applications Note

Databases and ontologies

FinaleDB: a browser and database of cell-free DNA fragmentation patterns

Haizi Zheng[1], Michelle S. Zhu[2] and Yaping Liu [1,3,4,5,*]

[1]Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA, [2]Department of Computer Science, University of Texas at Austin, Austin, TX 78712, USA, [3]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA, [4]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA and [5]Department of Electrical Engineering and Computing Sciences, University of Cincinnati College of Engineering and Applied Science, Cincinnati, OH 45229, USA

*To whom correspondence should be addressed.

Associate Editor: Robinson Peter

Received on August 18, 2020; revised on October 8, 2020; editorial decision on November 16, 2020; accepted on November 17, 2020

Abstract

Summary: Circulating cell-free DNA (cfDNA) is a promising biomarker for the diagnosis and prognosis of many diseases, including cancer. The genome-wide non-random fragmentation patterns of cfDNA are associated with the nucleosomal protection, epigenetic environment and gene expression in the cell types that contributed to cfDNA. However, current progress on the development of computational methods and understanding of molecular mechanisms behind cfDNA fragmentation patterns is significantly limited by the controlled-access of cfDNA whole-genome sequencing (WGS) dataset. Here, we present FinaleDB (**F**ragmentat**I**o**N** **A**naLysis of c**E**ll-free DNA **D**ata**B**ase), a comprehensive database to host thousands of uniformly processed and curated de-identified cfDNA WGS datasets across different pathological conditions. Furthermore, FinaleDB comes with a fragmentation genome browser, from which users can seamlessly integrate thousands of other omics data in different cell types to experience a comprehensive view of both gene-regulatory landscape and cfDNA fragmentation patterns.

Availability and implementation: FinaleDB service: http://finaledb.research.cchmc.org/. FinaleDB source code: https://github.com/epifluidlab/finaledb_portal, https://github.com/epifluidlab/finaledb_workflow.

Contact: lyping1986@gmail.com

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

1 Introduction

Circulating cell-free DNA (cfDNA) in peripheral blood and urine have recently been shown as a promising biomarker for disease diagnosis and prognosis (Phallen *et al.*, 2017). The fragment lengths of cfDNA are not uniform across the genome and are influenced by the local epigenetic environment and different physiological conditions (Ivanov *et al.*, 2015; Snyder *et al.*, 2016). cfDNA fragment length has revealed a predominant 167-base pair (bp) peak and 10-bp periodicity pattern in cfDNA, which is highly correlated with local nucleosomal structure and histone modifications. The cfDNA fragmentation patterns and their derived patterns from whole-genome sequencing (WGS), such as nucleosome positions, patterns near transcription start sites or transcription factor binding sites, ended position of cfDNA and large-scale fragmentation changes at mega-base level, offer extensive signals from the diseased tissues, as well as possible alterations from peripheral immune cell deaths, which can significantly increase the sensitivity for disease diagnosis (Cristiano *et al.*, 2019; Jiang *et al.*, 2018; Snyder *et al.*, 2016; Sun *et al.*, 2019; Ulz *et al.*, 2016; 2019).

Due to the protection of genotype information from the patients, which is not needed for the fragmentation analysis, most cfDNA WGS datasets are deposited in the controlled-access repositories. The data access in these repositories requires special and lengthy application processes and sometimes even data transfer agreements that may take several months between the two organizations' legal departments. Moreover, the cfDNA fragmentation patterns are inferred from the mapping locations of paired-end short-read sequencing, which are highly affected by the reads' quality, length and choices of the mapping strategy. These 'batch effects' will significantly affect the downstream computational inference and data analysis. Currently, a centralized database with uniformly processed cfDNA datasets from a variety of physiological conditions is still not publicly available for the community.

To address these challenges, we developed FinaleDB, a comprehensive and interactive cfDNA fragmentation pattern genome

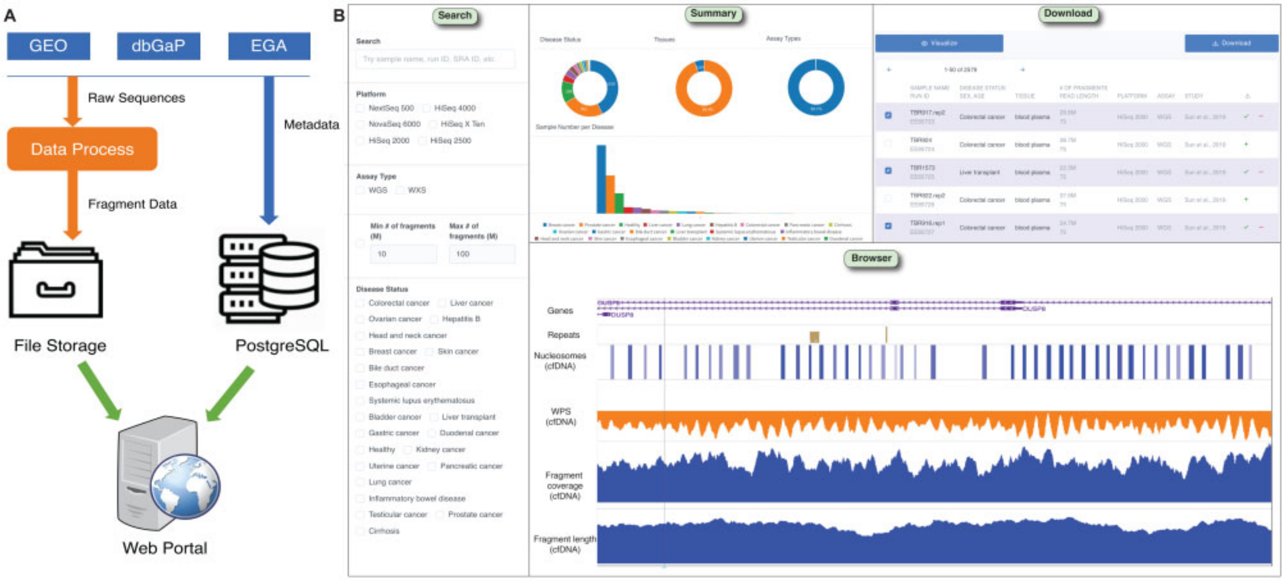

Fig. 1. (**A**) Overview of system design. (**B**) Web portal and fragmentation browser

browser and database that collected thousands of publicly available cfDNA WGS datasets (Fig. 1A).

2 The database

In the current version of FinaleDB, we collected 2579 paired-end cfDNA WGS datasets across 23 different pathological conditions from GEO, EGA and dbGaP (Supplementary Tables S1 and S2). We processed the raw sequencing datasets by an in-house workflow. The workflow is managed by snakemake v5.19 and is tailored for a Kubernetes cluster with AWS Spot Instances, which ensures optimal cost-effectiveness.

In the back-end, we built a database powered by Amazon RDS for PostgreSQL. The database stored the essential metadata, including the sample information, sequencing platform and study design. The fragmentation data itself is served by an HTTP static file server.

3 The application programming interface

The application programming interface (API) serves as an intermediate between the database and the front-end web portal. The API, based on the RESTful standard, can be accessed directly using any common programming language (Supplementary Table S4).

4 The front-end web portal and fragmentation browser

We developed a web portal for the database based on React.js, with the source code publicly available (Fig. 1B). At the query page, users can search with a number of criteria, such as GEO/dbGaP/EGA ID and pathological condition. The visualization page comes with a modified WashU Epigenome Browser embedded within. Users can visualize fragmentation pattern tracks of selected datasets, along with any other tracks that can be either local, remote, or those natively provided by WashU Epigenome Browser. In addition, the web portal allows users to download fragmentation data files such as the coverage, fragment size profile, etc. (Supplementary Section S3.3).

Acknowledgements

The authors acknowledged the helps from all the research groups for their cfDNA data.

Funding

This work was supported by the CCHMC start-up grant and Trustee Award to Y.L. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation [ACI-1548562]. This work used the XSEDE at the Pittsburgh Supercomputing Center (PSC) through allocation MCB190124P and MCB190006P.

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.

Ivanov,M. *et al.* (2015) Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics*, **16**, S1.

Jiang,P. *et al.* (2018) Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. *Proc. Natl. Acad. Sci. USA*, **115**, E10925–E10933.

Phallen,J. *et al.* (2017) Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med*, **9**, eaan2415.

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.

Ulz,P. *et al.* (2019) Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat. Commun.*, **10**, 4666.

Ulz,P. *et al.* (2016) Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat. Genet.*, **48**, 1273–1278.