Yuan et al. *Zool. Res.* 2022, 43(1): 75−77
https://doi.org/10.24272/j.issn.2095-8137.2021.250

ZR | Zoological Research

Letter to the editor Open Access

COSINE: A web server for clonal and subclonal structure inference and evolution in cancer genomics

Cancer cell genomes originate from single-cell mutation with sequential clonal and subclonal expansion of somatic mutation acquisition during pathogenesis, thus exhibiting a Darwinian evolutionary process (Gerstung et al., 2020; Nik-Zainal et al., 2012). Through next-generation sequencing of tumor tissue, this evolutionary process can be characterized by statistical modelling, which can identify the clonal state, somatic mutation order, and evolutionary process (Gerstung et al., 2020; Mcgranahan & Swanton, 2017). Inference of clonal and subclonal structure from bulk or single-cell tumor genomic sequencing data has a huge impact on studying cancer evolution. Clonal state and mutation order can provide detailed insight into tumor origin and future development. In the past decade, various methods for subclonal reconstruction using bulk tumor sequencing data have been developed. However, these methods had different programming languages and data input formats, which limited their use and comparison. Therefore, we established a web server for Clonal and Subclonal Structure Inference and Evolution (COSINE) of cancer genomic data, which incorporated twelve popular subclonal reconstruction methods. We deconstructed each method to provide a detailed workflow of single processing steps with a user-friendly interface. To the best of our knowledge, this is the first web server providing online subclonal inference based on the integration of most popular subclonal reconstruction methods. COSINE is freely accessible at www.clab-cosine.net.

Inference of subclonal structure using tumor-based bulk genomic sequencing data is an important part of tumor evolution research and provides a new way to study the relative sequence of mutations and mutation processes in tumorigenesis. Cancer evolution can be inferred from next-generation sequencing data based on the "most recent common ancestor (MRCA)", as applied in classical population genetics. Mutations that occur before the MRCA and are found in all tumor cells in a sample can be used as markers of clonal populations (Gerstung et al., 2020; Salcedo et al., 2020).

In the past decade, a lots subclonal reconstruction methods have been developed for a single or multiple sample(s) tumor's bulk or single cell genomic data over time and/or multiple sites (Cun et al., 2018; Malikic et al., 2015; Miller et al., 2014; Miura et al., 2018; Nik-Zainal et al., 2012; Salcedo et al., 2020; Strino et al., 2013; Xiao et al., 2020). Generally, subclonal reconstruction involves three steps: first, calculate the fraction of variant alleles of somatic mutations with relevant copy number changes and tumor purity; second, calculate the cancer cell fraction (CCF) in the tumor (using structural variation information correction); third, cluster the CCFs to identify subclonal structures and construct related phylogenetic trees. Thus, the accuracy and resolution of each subclonal inference method depends on the experimental design and mutation characteristics of the specific tumor being reconstructed. Among these methods, most employ non-parametric Bayesian approaches for clustering, e.g., Dirichlet process with stick-breaking representation (Cmero et al., 2020; Nik-Zainal et al., 2012), which require Markov chain Monte Carlo (MCMC) resampling and incur high computational costs, especially with increasing mutation number. A more economical computation way is to use a variational Bayesian mixture model, such as SciClone (Miller et al., 2014). Combinational phylogenetic approaches are also applied for clustering, e.g., TrAP (Strino et al., 2013), CITUP (Malikic et al., 2015), and CloneFinder (Miura et al., 2018). The deconvolution of single-nucleotide variant (SNV) density of cancer cells is computationally efficient for subclonal inference, as applied in Sclust (Cun et al., 2018) and FastClone (Xiao et al., 2020).

However, since the above-mentioned subclonal

Received: 25 September 2021; Accepted: 25 November 2021; Online: 26 November 2021

Foundation items: This work was supported by the CAS Pioneer Hundred Talents Program and National Natural Science Foundation of China (32070683) to Y.P.C.; the Science and Technology Planning Project of XI'AN (GXYD6.2) and National Natural Science Foundation of China (61771369) to X.G.Y.

reconstruction methods are developed using different programming languages and implemented under the Linux platform, most users may find it difficult to run and compare them. In this paper, we established a web server for subclonal inference in cancer genomics with the incorporation of 12 popular subclonal reconstruction methods (Cun et al., 2018; Malikic et al., 2015; Miller et al., 2014; Miura et al., 2018; Salcedo et al., 2020; Strino et al., 2013; Xiao et al., 2020), including three popular used methods: DPclust, PyClone, PhyloWGS and our own Sclust. Each method was deconstructed into detailed operational steps and implemented through a relevant operational interface, allowing easy and convenient comparison of methods when running data. Although in the DREAM challenge project on subclonal inferencing, Salcedo et al. (2020) reviewed current major approaches in subconal inference and compared the performance of DPclust, PyClone, PhyloWGS in the real genomic data. But a lots non-Dirichlet type method did not include in their review and comparison, it is still required to include more subclonal inference methods to model comparison. Our new online tool for subclonal inference, which integrates the 12 most popular subclonal inference methods, will help resolve model-to-user gaps and give user more choice for subclonal inferencing.

To facilitate the use of our previously developed Sclust method and 11 other approaches, we developed an online web server for subclonal inference called COSINE. Of the 12 selected methods, all are run under the Linux system, seven use only one programming language (Sclust developed in C++; PyClone, FastClone, and CloneFinder developed in Python; DPclust, SciClone developed in R; TrAp developed in Java), and five others use more than two programming languages. All those information was summarized in Supplementary Table S1. These differences may hinder their application by non-professionals wishing to perform rapid or comparative subclonal inference. Figure 1A, B show the general workflow for the inference of clonal and subclonal structure, which includes five steps: (1) somatic mutation calling from matched normal-tumor tissue samples based on next-generation sequencing (NGS) data; (2) gene copy number calling using NGS data; (3) CCF estimation; (4) clonal and subclonal structure inference via CCF clustering; and (5) clonal and subclonal evolutionary tree construction. A step-by-step pipeline for mapping raw data to reference genome, base calibrating and PCR duplication filtering, mutation and copy number calling were given in Supplementary Text and cunlab.org/cosine.

In the COSINE, we added all methods to a high-performance computing cluster, thus allowing the user to directly call each subclonal inference method via their web interface using 1 to 5 commands in the method's frame box, and then download the results when finished. Figure 1C showed an example of how to run the Sclust in the COSINE. With SNVs and copy number variation information (structure variation needed for some methods), user can employ method of those twelve methods for subclonal inferencing on the COSINE. As each method had their own input file format, we made some Python scripts to change the same somatic mutation variant call format (VCF) file and copy number alteration file to the format of each method, which were

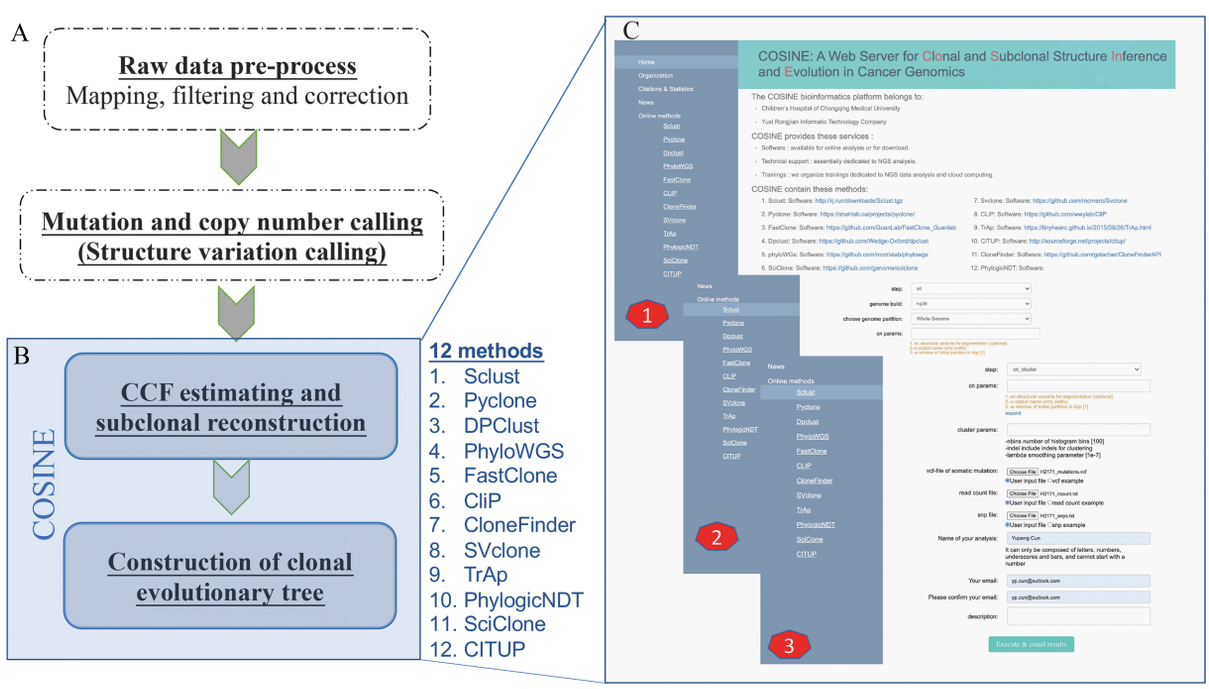

Figure 1 Raw data pre-processing, CCF estimation and subclonal reconstruction

A, B: Typical pipeline for mapping, bam filtering, and mutation and copy number/structure variation calling from raw clean data. Information on somatic mutation VCF, copy number alteration, and structure variation (optimal) is used for subclonal reconstruction. C: COSINE web server containing 12 subclonal reconstruction methods with detailed operational step(s).

available at: cunlab.org/cosine.

As shown in Figure 1C, users can follow the following steps for subclonal inference: (1) visit the COSINE website (www.clab-cosine.net/cun-web/) and click the relevant method (Figure 1C1); (2) choose a new task on the method page (Figure 1C2); and (3) upload and run the program (Figure 1C3), and an e-mail will send to user when the job is finished.

The COSINE is an online computational platform for subclonal structure inference in the cancer genome. It integrates twelve popular subclonal inference methods and provides an easy-to-access and user-friendly interface. Although various subclonal inference models have been proposed in recent years, many contain inherent difficulties for researchers regarding method selection, installation, and program operation. The COSINE not only helps to bridge the gap between model developer to normal user, but also allows easier and more convenient subclonal inference method comparison. In the future, we will develop additional functions and methods for online subclonal evolutionary tree plotting and adjustment, and include subclonal reconstruction methods from single-cell genomic sequencing data.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

Y.P.C. and X.G.Y. conceived and designed the study. Y.P.C., M.P., Z.D.L., W.L., S.Y.W., and T.G. developed the program and wrote the computer codes for the web server. Y.P.C., L.M.G., Q.L., Z.B.W., and P.N.Z. designed the web interface. Y.P.C., Y.Z., and Y.G. wrote the supplementary practical guideline. Y.P.C. and X.G.Y. wrote and edited the manuscript. All authors read and approved the final version of the manuscript.

Xi-Guo Yuan[1,#], Yuan Zhao[1,#], Yang Guo[1], Lin-Mei Ge[2], Wei Liu[3], Shi-Yu Wen[3], Qi Li[1], Zhang-Bo Wan[1], Pei-Na Zheng[1], Tao Guo[3], Zhi-Da Li[3], Martin Peifer[4], Yu-Peng Cun[5,2,*]

[1] *School of Computer Science and Technology*, *Xidian University*, *Xi'an*, *Shaanxi* 710071, *China*

[2] *iFlora Bioinformatics Center*, *Germplasm Bank of Wild Species*, *Kunming Institute of Botany*, *Chinese Academy of Sciences*, *Kunming*, *Yunan* 650201, *China*

[3] *Yuxi Rongjian Information Technology Co.*, *Ltd.*, *Yuxi*, *Yunan* 653100, *China*

[4] *Center for Molecular Medicine Cologne (CMMC)*, *University of Cologne*, *Cologne* 50931, *Germany*

[5] *Pediatric Research Institute*, *Ministry of Education Key Laboratory of Child Development and Disorders*, *National Clinical Research Center for Child Health and Disorders*, *China International Science and Technology Cooperation Base of Child Development and Critical Disorders*, *Chongqing Key Laboratory of Translational Medical Research in Cognitive Development and Learning and Memory Disorders*, *Children's Hospital of Chongqing Medical University*, *Chongqing* 400014, *China*

[#]Authors contributed equally to this work

*Corresponding author, E-mail: cunyp@cqmu.edu.cn

REFERENCES

Cmero M, Yuan K, Ong CS, Schröder J, Adams DJ, Anur P, et al. 2020. Inferring structural variant cancer cell fraction. *Nature Communications*, **11**(1): 730.

Cun YP, Yang TP, Achter V, Lang U, Peifer M. 2018. Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust. *Nature Protocols*, **13**(6): 1488−1501.

Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. 2020. The evolutionary history of 2, 658 cancers. *Nature*, **578**(7793): 122−128.

Malikic S, Mcpherson AW, Donmez N, Sahinalp CS. 2015. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, **31**(9): 1349−1356.

Mcgranahan N, Swanton C. 2017. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, **168**(4): 613−628.

Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, et al. 2014. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Computational Biology*, **10**(8): e1003665.

Miura S, Gomez K, Murillo O, Huuki LA, Vu T, Buturla T, et al. 2018. Predicting clone genotypes from tumor bulk sequencing of multiple samples. *Bioinformatics*, **34**(23): 4017−4026.

Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. 2012. The life history of 21 breast cancers. *Cell*, **149**(5): 994−1007.

Salcedo A, Tarabichi M, Espiritu SMG, Deshwar AG, David M, Wilson NM, et al. 2020. A community effort to create standards for evaluating tumor subclonal reconstruction. *Nature Biotechnology*, **38**(1): 97−107.

Strino F, Parisi F, Micsinai M, Kluger Y. 2013. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Research*, **41**(17): e165.

Xiao Y, Wang XQ, Zhang HJ, Ulintz PJ, Li HY, Guan YF. 2020. FastClone is a probabilistic tool for deconvoluting tumor heterogeneity in bulk-sequencing samples. *Nature Communications*, **11**(1): 4469.