Genetics and population analysis

EUPAN enables pan-genome studies of a large number of eukaryotic genomes

Zhiqiang Hu¹,²,†, Chen Sun¹,²,†, Kuang-chen Lu¹, Xixia Chu³, Yue Zhao¹, Jinyuan Lu¹,², Jianxin Shi⁴,* and Chaochun Wei¹,²,*

¹Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China, ²Shanghai Center for Bioinformation Technology, 1278 Keyuan Road, Pudong District, Shanghai 201203, China, ³Bio-X Institutes, Shanghai Jiao Tong University, Shanghai 200240, China and ⁴Department of Genetic and Developmental Science, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate editor: Oliver Stegle

Received on December 21, 2016; revised on March 1, 2017; editorial decision on March 20, 2017; accepted on April 6, 2017

Abstract

Summary: Pan-genome analyses are routinely carried out for bacteria to interpret the within-species gene presence/absence variations (PAVs). However, pan-genome analyses are rare for eukaryotes due to the large sizes and higher complexities of their genomes. Here we proposed EUPAN, a eukaryotic pan-genome analysis toolkit, enabling automatic large-scale eukaryotic pan-genome analyses and detection of gene PAVs at a relatively low sequencing depth. In the previous studies, we demonstrated the effectiveness and high accuracy of EUPAN in the pan-genome analysis of 453 rice genomes, in which we also revealed widespread gene PAVs among individual rice genomes. Moreover, EUPAN can be directly applied to the current re-sequencing projects primarily focusing on single nucleotide polymorphisms.

Availability and Implementation: EUPAN is implemented in Perl, R and C++. It is supported under Linux and preferred for a computer cluster with LSF and SLURM job scheduling system. EUPAN together with its standard operating procedure (SOP) is freely available for non-commercial use (CC BY-NC 4.0) at http://cgm.sjtu.edu.cn/eupan/index.html.

Contact: ccwei@sjtu.edu.cn or jianxin.shi@sjtu.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Thanks to the rapid decrease of sequencing cost, pan-genome studies are routinely carried out for bacteria currently aiming to reveal gene PAVs within a species. However, there are only a handful of pan-genome studies of eukaryotes with large genomes (Supplementary Table S1) (Hirsch et al., 2014; Li et al., 2010, 2014; Schatz et al., 2014; Yao et al., 2015). These studies demonstrated that gene PAVs are of great importance and have unique roles in within-species differentiation, especially for plants/crops. Most of these studies only focused on the pan-genome of the species, exploring the novel sequences and novel genes missed in a reference genome instead of revealing the gene PAVs (Hirsch et al., 2014; Li et al., 2010; Yao et al., 2015). Only two of them studied the gene PAVs and revealed their widespread existence (Li et al., 2014; Schatz et al., 2014). These two studies followed the traditional analysis strategy, in which individual genomes were first de novo assembled and annotated, followed by determination of gene PAVs by comparison of protein sequences among individuals (left panel of Figure 1). However, assembly of a relatively complex eukaryotic genome is of high cost, requiring high sequencing depth and multiple DNA libraries with various insertion sizes due to its large size and high level of repeats. Therefore, the individual numbers
EUPAN can be installed easily and it is user-friendly, though it integrated many independent tools, including FastQC and Trimmomatic (Bolger et al., 2014) for read quality operation, BWA (Li and Durbin, 2009), Bowtie2 (Langmead and Salzberg, 2012) and SAMtools (Li et al., 2009) for mapping, SOAPdenovo2 (Luo et al., 2012) and QUAST (Gurevich et al., 2013) for assembly, BLAST and CD-HIT (Fu et al., 2012) for alignments and clustering. Besides, though EUPAN support any Unix-like machine, we highly recommend running EUPAN on a computer cluster due to the massive computation and high storage involved in the analyses.

3 Conclusion

Besides SNP and structural variation, gene PAV is another variation form playing an important role in subspecies differentiation for bacteria and plants and its potential in animals was poorly explored. We introduced the map-to-pan strategy and EUPAN toolbox, enabling the analyses to be involved in the pan-genome studies of hundreds or even thousands of individuals for higher eukaryotes with large-sized genomes.

Acknowledgements

The authors would thank Center for High Performance Computing at Shanghai Jiao Tong University for computing.

References