Bioinformatics, 2017, 1–2
doi: 10.1093/bioinformatics/btx170
Advance Access Publication Date: 7 April 2017
Applications Note

Genetics and population analysis

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

Zhiqiang Hu[1,2,†], Chen Sun[1,2,†], Kuang-chen Lu[1], Xixia Chu[3], Yue Zhao[1], Jinyuan Lu[1,2], Jianxin Shi[4,*] and Chaochun Wei[1,2,*]

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

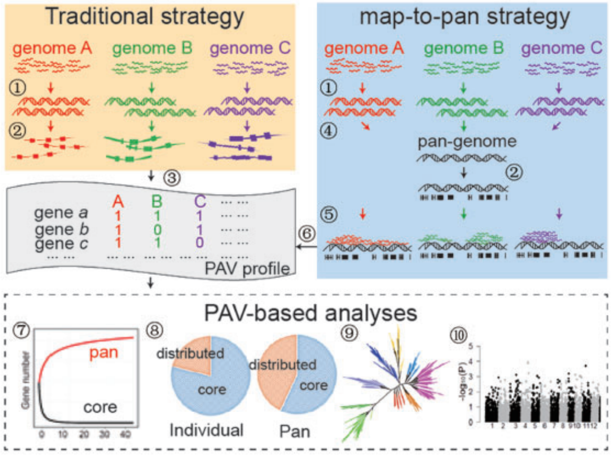

Fig. 1. Comparison of methods for gene PAV-based pan-genome analyses. The processes include ① *de novo* assembly; ② gene annotation; ③ gene PAV determination by protein sequence comparison; ④ construction of pan-genome sequences; ⑤ read mapping; ⑥ gene PAV determination by read coverage; ⑦ simulation of pan-genome; ⑧ genome composition; ⑨ population structure; and ⑩ PAV-based GWAS

involved in the two studies were very limited (3 and 7, respectively). Unfortunately, for most cases, limited individuals cannot represent the whole species, as we demonstrated in recent reports of the 3000 Rice Genomes Project (Rice Genomes Project, 2014; Sun *et al.*, 2016). We also demonstrated that gene-PAV-based genome-wide association studies (GWAS), on which are effective to detect phenotype-associated genes, could serve as an important complement to traditional SNP-based GWAS. Therefore, tools/methods for large-scale eukaryotic pan-genome study are of great importance and of urgent demand.

In this paper, we present the ‘map-to-pan’ strategy to determine gene PAVs (right panel of Figure 1) which includes the following steps: (1) *de novo* assembly of individual genomes; (2) construction of pan-genome sequences based on the assemblies and available reference genomes; (3) gene annotation of the pan-genome sequences; and (4) determination of PAVs based on gene coverage of mapped reads against pan-genome sequences. This strategy was primarily used in the 3000 Rice Genomes Project (Rice Genomes Project, 2014; Sun *et al.*, 2016), in which we observed that the genome can be fully covered by read mapping at sequencing depth >20×, though it is poorly assembled (Supplementary Figure S1). In order to enable and accelerate large-scale pan-genome studies of higher eukaryotes, we reorganized and refined the codes and built the EUPAN toolbox, rendering it to be a highly configurable set of command-line tools.

2 Implementation and application

The detailed pipeline of EUPAN is shown in Supplementary Figure S2. EUPAN toolkit can perform the following operations for thousands of samples in parallel on a computer cluster with LSF/SLURM system or sequentially on a single machine:

- check and plot the overall sequencing qualities;
- extract high-quality reads with both filtering and trimming methods;
- conduct *de novo* assembly with automatically selected best Kmer;
- evaluate *de novo* assembly;
- align contigs to a reference genome, extract non-redundant novel sequences and build pan-genome sequence set;
- map reads to a pan-genome or a reference genome;
- evaluate read mapping;
- determine gene PAVs and gene family PAVs by mapping reads to the reference pan-genome.

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.

Funding

This work has been supported by the grants from the National Natural Science Foundation of China (61272250, 61472246 and J1210047), the National Basic Research Program of China (2013CB956103), and the National High-Tech R&D Program (863) (2014AA021502). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of interest: none declared.

References

Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Fu,L.M. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Gurevich,A. *et al.* (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Hirsch,C.N. *et al.* (2014) Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*, **26**, 121–135.

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

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,R. *et al.* (2010) Building the sequence map of the human pan-genome. *Nat. Biotechnol.*, **28**, 57–63.

Li,Y.H. *et al.* (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.*, **32**, 1045–1052.

Luo,R. *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, **1**, 18.

Rice Genomes Project. (2014) The 3,000 rice genomes project. *Gigascience*, **3**, 7.

Schatz,M.C. *et al.* (2014) Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. *Genome Biol.*, **15**, 506.

Sun,C. *et al.* (2016) RPAN: rice pan-genome browser for approximately 3000 rice genomes. *Nucleic Acids Res.*, **45**, 597–605.

Yao,W. *et al.* (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol.*, **16**, 187.