**ppsPCP: A Plant Presence/absence Variants Scanner and Pan-genome Construction Pipeline**

Muhammad Tahir ul Qamar1,2, Xitong Zhu1,2, Feng Xing1,2, Ling-Ling Chen1,2,\*

**1**National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, P. R. China

**2**Hubei Key Laboratory of Agricultural Informatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, P. R. China

**\* Correspondence:**

Ling-Ling Chen (llchen@mail.hzau.edu.cn)

****

**Supplementary Figure 1: An example of merging overlapped PAVs and their boundaries correction on chromosome-1 of Shuhui498 genome when compared as query with Nipponbare reference genome using ppsPCP.** PAV (shown in green color) was screened by MUMmer at step-1, but it was not covering the whole gene region of OsR498G0101472900.01 (shown in orange color). This PAV is extended, and its boundaries are corrected to cover whole gene region at step-5, but after extending PAV-Draft overlapped with nearby gene region OsR498G0101473200.01 (shown in blue color). Thus, this PAV is again extended at step-10, so it covers all the genes region.

**Supplementary note 1: ppsPCP usage and available options for users**

**Usage:**

|  |
| --- |
| make\_pan.pl [options] --ref [reference\_genome] --ref\_anno [refernece\_anno] --query query1\_genome[query2...] --query\_anno query1\_anno[query2...] &> [job\_name].log |

**Help:**

|  |  |
| --- | --- |
| --help|-h | Print the help message and exit.  |

**Required parameters:**

|  |  |
| --- | --- |
| --ref | Reference sequence file, usually a fasta file  |
| --ref\_anno | The gff3 annotation file for the reference sequence  |
| --query | The query sequence files, can be one or more, separated with space  |
| --query\_anno | The gff3 annotation files corresponding to the query sequence files, must have the same order with the query sequence files  |

**Filtering parameters:**

|  |  |
| --- | --- |
| --coverage | The coverage used to filter similar PAVs. Can be any number between 0 and 1. Default: 0.9  |
| --sim\_pav | The similarity used to filter similar PAVs. Can be any number between 0 and 1. Default: 0.95  |
| --sim\_gene | Then similarity used to filter mapped genes in blat mapping. Can be any number between 0 and 1. Default: 0.8  |

**Other parameters:**

|  |  |
| --- | --- |
| --tmp | The temporary directory where you want to save the temporary files. Default: ./tmp  |
| --no\_tmp | Delete tmp file when job finished  |
| --thread | The number of threads used for mummer and blastn. Remember not all the phases of ppsPCP are parallelized. Default: 1 |

**Supplementary note 2: Resources needed to run ppsPCP**

ppsPCP currently available for Linux based platforms. To scan three rice sized genomes (~400 MB each) and make their pan-genome, ppsPCP only takes about 25 CPU hours.

|  |  |
| --- | --- |
| **Software**  | **Availability**  |
| MUMmer | <https://github.com/mummer4/mummer/releases/download/v4.0.0beta2/mummer-4.0.0beta2.tar.gz>  |
| Blast+ | ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.7.1+-x64-linux.tar.gz  |
| Bedtools | <https://github.com/arq5x/bedtools2/releases/download/v2.25.0/bedtools-2.25.0.tar.gz>  |
| Blat | <http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/>  |
| gffread | <http://cole-trapnell-lab.github.io/cufflinks/assets/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz>  |
| Bioperl | <https://bioperl.org/INSTALL.html>  |

\*complete ppsPCP details are available at <https://doi.org/10.5281/zenodo.2567390> and <http://cbi.hzau.edu.cn/ppsPCP/>.

**Supplementary note 3: Construction of rice and *arabidopsis thaliana* pan-genome to demonstrate ppsPCP pipeline**

Rice [*Oryza sativa*] is model cereal and one of the most important staple food crops globally. We selected 3 different elite species of rice (1 *japonica* and 2 *indica*) which have the most complete genomes to benchmark ppsPCP pipeline. We selected rice reference species Nipponbare (International Rice Genome Sequencing Project, 2005), nearly complete sequenced species Shuhui498 (Du et al., 2017) and the most complete sequenced specie by our research group Minghui63 (Zhang et al., 2016). Non-TE Genomes of Nipponbare, Shuhui498 and Minghui63 were downloaded from <http://rice.plantbiology.msu.edu/>, <http://www.mbkbase.org/R498/> and <http://rice.hzau.edu.cn/rice/> respectively. We selected Nipponbare as reference and Shuhui498, Minghui63 as query genomes respectively. We run ppsPCP using default parameters and 24 threads. All the input and output data can be retrieved from ppsPCP webpage using following link: <http://cbi.hzau.edu.cn/ppsPCP/files/rice_ppsPCP.tar.gz>.

*Arabidopsis thaliana* is a model dicot plant and has most complete/compact genome. We tested ppsPCP with whole genome sequence assemblies of 19 *A. thaliana* ecotypes (Gan et al., 2011). WGS assemblies corresponding to 19 annotated *A. thaliana* ecotypes were retrieved from <http://mtweb.cs.ucl.ac.uk/mus/www/19genomes/>. We selected Col-0 as reference and Bur-0, Can-0, Ct-1, Edi-0, Hi-0, Kn-0, Ler-0, Mt-0, No-0, Oy-0, Po-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0 and Zu-0 as query genomes respectively. We run ppsPCP using default parameters and 24 threads. All the input and output data can be retrieved from ppsPCP webpage using following link: <http://cbi.hzau.edu.cn/ppsPCP/files/arabidopsis_ppsPCP.tar.gz>.

**References**

Du, H. *et al*. (2017) Sequencing and de novo assembly of a near complete *indica* rice genome. *Nat. Commun.,* 8, 15324.

Gan, X. *et al*. (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, 477, 419.

International Rice Genome Sequencing Project. (2005) The map-based sequence of the rice genome. *Nature*, 436, 793-800.

Zhang, J. *et al*. (2016) Extensive sequence divergence between the reference genomes of two elite *indica* rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl Acad. Sci. USA* 113, E5163-E5171.