# PlasmidIdentifier: A multi-pronged approach to plasmid identification

Quinn Thomas, Tristan Kosciuch, Jared Tomei

**Abstract:** Bacterial plasmids have a large importance to ecology and human health. However identification of plasmids relies on being able to distinguish plasmid sequences from chromosomal sequences. Current tools for identifying plasmids are limited by biased databases and repetitive sequences (Schwengers et al., 2020). In order to mitigate the limitations of any single tool, we propose a multi-pronged approach. Our pipeline assembles fastq sequencing data and runs it through multiple plasmid identification tools. Output of this application is a single CSV file with contig names and whether each tool identified the contig as plasmid or chromosomal. All tools will be integrated into the pipeline using Docker in a Unix based operating system. A python file will run all the docker and pipeline based on flags given by the user . The tools used in the Docker container are SPAdes, Platon, and Plasforest. Running the pipeline on a 200Mb E. coli sequencing file, our tool took a total of 30 minutes to run and 18.5GB of memory. Results indicated that these three tools make unique plasmid predictions creating a comprehensive list of potential plasmid sequences.

**Introduction:** Plasmids are independent genetic elements in bacteria that replicate independently of the chromosome. Due to the mobility of their genetic elements, plasmids have a great impact on ecology and evolution of microbial communities and human health (Carattoli A. 2013; Crits-Christoph A Et. Al., 2022). Identifying plasmid sequences is essential to understanding their impact on human health. Plasmid DNA isolation and sequencing may be used to identify plasmid sequences however it is more common to sequence an entire bacterial genome and identify plasmid sequences (Orlek A et al., 2017). This requires being able to

distinguish plasmid sequences from chromosomal sequences. Several tools have been developed in order to identify these reads, but common issues such as databases that are biased to already known organisms, and short reads with repetitive elements causing issues with de Bruijn assemblers (Schwengers et al., 2020). Common features used for plasmid identification include coverage, replicon sequences, pentamer frequencies and homology databases (Antipov D. et al, 2016; Schwengers, O. et al., 2020; Carattoli A. et al., 2014; Zhou F. et al., 2010; Pradier, L. et al., 2021). In order to reduce biases caused by use of a single tool, we propose a multi-pronged approach by implementing multiple unique plasmid identification tools. The tools used in the pipeline for plasmid prediction are SPAdes (v3.15.4), plasmidSPAdes (v3.15.4), Platon (v1.6), and Plasforest (v1.0) (Bankevich A et al., 2012; Antipov D et al, 2016; Schwengers, O. et al., 2020; Pradier, L. et al., 2021). SPAdes is used to assemble contigs for Platon, and Plasforest as these tools need assembled reads. SPAdes also implements its own plasmid predictor, plasmidSPAdes which the raw data will also be run through. plasmidSPAdes separates chromosomal contigs from plasmid by checking the read coverage of contigs. plasmidSPades focuses on long contigs compared to all contigs in order to exclude repeats and low variance in coverage of long contigs. By removing long chromosomal edges and short dead-end edges from the assembly graph, the assembly graph is transformed into a plasmid graph output (Dmitry Antipov et al., 2016). Platon uses replicon distribution scores (RDS) and marker protein sequences (MPS) for plasmid identification. Platon depends on a custom database based on MPS, RDS, RefSeq Plasmid database, PlasmidFinder db as well as manually curated MOB HMM models from MOBscan, custom conjugation and replication HMM models and oriT sequences from MOB-suite (https://zenodo.org/record/4066768#.YnR5SvPMI-Q 2020). Contigs and their RDS are searched against this database in which RDS and MPS distinguish plasmid

and chromosomal sequences by representing the measured frequency biases of their protein coding sequence distributions (Schwengers et al., 2020). PlasForest is a homology-based random forest classifier used to identify plasmid sequences in partially assembled genomes. PlasForest first checks for plasmid annotation before analyzing contig features used for machine learning classification. Not only are contig features considered but also plasmid database matches are used to train the classifier. This database is from the NCBI RefSeq Genomes FTP server (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq; September 1st, 2019) and is composed of 36,450 sequences used as reference for homology. Contigs are then run through the classifier to determine whether they are of plasmid or chromosomal origin (Pradier et al., 2021).

**Implementation**: We have created a pipeline to automate contig assembly and plasmid prediction for plasmidSPAdes, platon and plasForest. All of the tools generate a fasta file of plasmid sequences or a csv file that indicates which contigs generated are plasmids. plasmidSPAdes assemble contigs that are different from SPAdes. In order to directly compare plasmidSPAdes to the other tools we created a local blast database of SPAdes contigs. plasmidSPAdes contigs were matched to their closest SPAdes contigs and the best match is reported. Results were then combined with platon and plasforest predictions in one comprehensive csv file. Our pipeline saves the output file of each tool to a results folder called “output” and the complete csv file called “results.csv” listing the contig ID and if each tool predicted the contig as a plasmid or chromosome. Also included is “results_summary.txt” with a summary of total counts and overlaps for each tool. A file called “plasmidSPAdes_closest_conitg.csv” lists the E and bit scores of the best match between plasmidSPAdes plasmids and SPAdes contigs.

Our pipeline is distributed as a docker image hosted on dockerhub at https://hub.docker.com/repository/docker/triskos/plasmid-id. The source code is available at https://github.com/qthomas612/CompBio_Group3. Our docker image is based on Ubuntu 20.04. Two scripts are called by our dockerfile while building our docker image; install.sh and download_db.sh. install.sh downloads and installs the latest version of SPAdes and plasforest from source. PlasForest and any python dependencies for tools are installed by conda. download_db.sh downloads local databases needed for Platon and Plasforest. The plasForest database comes from the NCBI RefSeq Genomes FTP server and contains 36,450 sequences. The platon database comes from RefSeq release 202 and contains 20,041 chromosomal sequences and 18,795 plasmid sequences. install.sh runs in approximately 15 minutes, but download_db.sh can take 2 hours or more, depending on internet speed. We provide a python wrapper, plasmid.py, to execute a run of our pipeline from the users local computer. This python command passes the input reads, the number of threads to use, read type, and output directory to a docker run command that executes our pipeline.

**Results:** To test our pipeline we used short read data coming from an *E. coli* Illumina MiSeq genome sequencing run (https://www.ncbi.nlm.nih.gov/sra/SRX13373163[accn]). Running the program on this 200Mb file with 16 threads took 30 minutes total. Both SPAdes and plasmidSPAdes used 7gb and took 11 minutes each,  plasforest used 1.5gb of memory and took 40 seconds, platon used 3gb of memory and took 7.5min. SPAdes assembled our raw reads into 5389 contigs. Of the 5389 contigs, 408 total contig sequences were identified as plasmid by one or more tools and 45 sequences were identified by 2 or more tools. Platon identified 49

sequences as plasmid, plasforest identified 208 sequences as plasmid and plasmidSPAdes identified 197 sequences as plasmid (Table 1).

| **Tool(s)** | **Number of plasmid contigs** |
|---|---:|
| platon | 49 |
| plasForest | 208 |
| plasmidSPAdes | 197 |
| platon & plasForest | 15 |
| platon & plasmidSPAdes | 16 |
| plasForest & plasmidSPAdes | 13 |
| platon, plasForest & plasmidSPAdes | 1 |

**Table 1.** Number of plasmid predictions for each tool and combination of tools.

Our results indicate that each tool identifies unique plasmid sequences with limited sequences overlapping. We utilize tools with a broad range of plasmid detection methods which likely explains the discrepancy between each program. This allows for the user to have the most comprehensive list of potential plasmids as our compiled list contains almost twice the amount of sequences (408 plasmids) as the tool with the largest amount of plasmids detected (208 sequences). Our pipeline implements 3 unique plasmid identification tools, two of which were released in the last 2 years. However, plasmid detection and identification methods are actively being developed, therefore we recommend that as new methods are released, they are integrated into this pipeline.

References:

Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-477. doi:10.1089/cmb.2012.0021
Carattoli A. Plasmids and the spread of resistance. *Int J Med Microbiol.* 2013;303:298–304. doi: 10.1016/j.ijmm.2013.02.001.
Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014 Jul;58(7):3895-903. doi: 10.1128/AAC.02412-14. Epub 2014 Apr 28. PMID: 24777092; PMCID: PMC4068535.
Crits-Christoph A, Hallowell HA, Koutouvalis K, Suez J. Good microbes, bad genes? The dissemination of antimicrobial resistance in the human microbiome. *Gut Microbes*. 2022;14(1):2055944. doi:10.1080/19490976.2022.2055944
Dmitry Antipov, Nolan Hartwick, Max Shen, Mikhail Raiko, Alla Lapidus, Pavel A. Pevzner, plasmidSPAdes: assembling plasmids from whole genome sequencing data, Bioinformatics, Volume 32, Issue 22, 15 November 2016, Pages 3380–3387
Orlek A, Stoesser N, Anjum MF, Doumith M, Ellington MJ, et al. Plasmid classification in an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. *Front Microbiol.* 2017;8:182. doi: 10.3389/fmicb.2017.00182.
Pradier, L., Tissot, T., Fiston-Lavier, AS. et al. PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. BMC Bioinformatics 22, 349 (2021). https://doi.org/10.1186/s12859-021-04270-w
Schwengers, O., Barth, P., Falgenhauer, L., Hain, T., Chakraborty, T., & Goesmann, A. (2020). Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. Microbial genomics, 6(10), mgen000398. https://doi.org/10.1099/mgen.0.000398
Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. Mobility of plasmids. *Microbiol Mol Biol Rev.* 2010;74:434–452. doi: 10.1128/MMBR.00020-10.
Zhou F, Xu Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics. 2010 Aug 15;26(16):2051-2. doi: 10.1093/bioinformatics/btq299. Epub 2010 Jun 10. PMID: 20538725; PMCID: PMC2916713.

platonDB: https://zenodo.org/record/4066768#.YnMbBPNuc-Q
plasForestDB: ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq