Faster Pan-Genome Construction for Efficient Differentiation of Naturally Occurring and Engineered Plasmids with Plaster

Qi Wang
Systems, Synthetic, and Physical Biology (SSPB) Graduate Program,
Rice University, Houston, TX 77005, USA
qw17@rice.edu

R. A. Leo Elworth
Department of Computer Science, Rice University, Houston, TX 77005, USA
r.a.leo.elworth@rice.edu

Tian Rui Liu
Department of Computer Science, Rice University, Houston, TX 77005, USA
tl62@rice.edu

Todd J. Treangen
Department of Computer Science, Rice University, Houston, TX 77005, USA
treangen@rice.edu

Abstract

As sequence databases grow, characterizing diversity across extremely large collections of genomes requires the development of efficient methods that avoid costly all-vs-all comparisons [17]. In addition to exponential increases in the amount of natural genomes being sequenced, improved techniques for the creation of human engineered sequences is ushering in a new wave of synthetic genome sequence databases that grow alongside naturally occurring genome databases. In this paper, we analyze the full diversity of available sequenced natural and synthetic plasmid genome sequences. This diversity can be represented by a data structure that captures all presently available nucleotide sequences, known as a pan-genome. In our case, we construct a single linear pan-genome nucleotide sequence that captures this diversity. To process such a large number of sequences, we introduce the *plaster* algorithmic pipeline. Using *plaster* we are able to construct the full synthetic plasmid pan-genome from 51,047 synthetic plasmid sequences as well as a natural pan-genome from 6,642 natural plasmid sequences. We demonstrate the efficacy of *plaster* by comparing its speed against another pan-genome construction method as well as demonstrating that nearly all plasmids align well to their corresponding pan-genome. Finally, we explore the use of pan-genome sequence alignment to distinguish between naturally occurring and synthetic plasmids. We believe this approach will lead to new techniques for rapid characterization of engineered plasmids. Applications for this work include detection of genome editing, tracking an unknown plasmid back to its lab of origin, and identifying naturally occurring sequences that may be of use to the synthetic biology community. The source code for fully reconstructing the natural and synthetic plasmid pan-genomes as well for *plaster* are publicly available and can be downloaded at `https://gitlab.com/qiwangrice/plaster.git`.

2012 ACM Subject Classification Applied computing → Bioinformatics; Applied computing → Molecular sequence analysis; Applied computing → Computational genomics

Keywords and phrases comparative genomics, sequence alignment, pan-genome, engineered plasmids

Digital Object Identifier 10.4230/LIPIcs.WABI.2019.19

Funding *Qi Wang*: Q. W. was supported by funds from Rice University and by funds from the National Institute for Neurological Disorders and Stroke (NINDS) of the National Institutes of Health under award number R21NS106640.
R. A. Leo Elworth: R. A. L. E. was supported by the National Science Foundation (DMS-1547433) and was partially supported by the FunGCAT program from the Office of the Director of National

Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the Army Research Office (ARO) under Federal Award No. W911NF-17-2-0089.
Tian Rui Liu: T.R.L. was supported by funds from Rice University.
Todd J. Treangen: T.J.T was supported by startup funds from Rice University and was partially supported by the FunGCAT program from the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the Army Research Office (ARO) under Federal Award No. W911NF-17-2-0089.

Acknowledgements The authors would like to thank Dr. Caleb Bashor for critical discussion and feedback, and Dr. Joanne Kamens from Addgene for providing full access to the synthetic plasmids utilized in this study. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, ARO, or the US Government.

1 Introduction

Thanks to the advancement of sequencing and genome-editing technologies, over the past decade genome engineering has become more affordable and accessible. Plasmids are commonly found as double-stranded DNA, which can replicate independently from chromosomal DNA [12]. They are ubiquitous in bacteria and provide benefits such as enhanced host range [4], antibiotic resistance [18], and even forced cooperation [20]. As plasmids are easy to engineer and can confer functions in a broad range of species, they are widely used in biology labs for understanding and modifying genetic elements. To date, more than 65 thousand engineered plasmids have been deposited in the Addgene repository [13]. With the benefits of genetic engineering microbial sequences also come risks [3], some with significant dual use of research concerns [22]. In response, methods have been designed both to detect signatures of engineering [1] and also trace back synthetic plasmids to a lab of origin using deep learning methods [19]. Although deep learning shows potential to characterize engineered plasmids and identify the lab-of-origin, there is a need for explainable, white-box approaches that can provide a full list of sequence features specific to a biological function or lab. We thus propose a novel method that enables the detection of a plasmid's lab-of-origin with precise characterization the diversity within and among all engineered plasmid sequences.

Traditionally, a single genome is chosen as a reference to describe genome diversity within a group. Unfortunately, a single genome usually fails to reveal the full picture of similarities and discrepancies among all individuals. The "pan-genome" concept was introduced to reflect the diversity of all strains within a specific clade [31, 27]. This concept is similar to the problem of finding a smaller set of founder sequences, which can map to a given sequence in the group, called the *founder sequence reconstruction* problem [30]. The pan-genome is designed to describe both the core genome shared by all the individuals and the accesory genome contained by only some strains [27]. This concept can be applied to different domains. In metagenomic studies, a pan-genome can highlight essential genetic elements responsible for adaptations to the environment and the co-evolution interactions among the microorganisms [17]. In terms of building a phylogenetic tree, pan-genomes can detect weak evolutionary signals, which may be omitted by multiple sequence alignment. It also opens up the possibilities for discovering the origin of unknown organisms [33], pathogen transmission history [6], or the inference of cancer cell evolution [10].

In microbes, the pan-genome is commonly defined as the union set of genes that exist in all the genomes in a selected clade [27]. Many bioinformatic tools have been developed to build gene-based pan-genomes, such as PanOCT [8], PGAP [35], and Roary [24]. Given

their gene-centric construction of the pan-genome, they are vulnerable to missing genes [34] and do not include intergenic regions, which can have substantial impact on phenotypes [28]. To address these limitations, methods such as Piggy [29] build pan-genomes from intergenic regions. In this paper we move away from a gene and intergenic centric approach; the term pan-genome refers to the sum of all core and accessory DNA sequence fragments which are longer than a minimal sequence fragment length l (*plaster* default $l = 50bp$).

In order to efficiently analyze vast numbers of DNA sequences, there is a need to create a pan-genome sequence that encodes all existing variations in the minimal possible size [21]. To solve this problem efficiently via pan-genomics we introduce *plaster*, a new pan-genome construction algorithm inspired by fast DNA clustering techniques [5]. *plaster*'s speed exceeds the fastest existing tool seq-seq-pan and can be easily scaled to handle large data sizes. Seq-seq-pan [14] is a recently introduced method for efficient pan-genome construction that was shown to be an order of magnitude faster than the fastest available methods for pan genome construction [7]. In brief, it relies on progressiveMauve to catalog all of the differences (including insertions, deletions, substitutions, inversions, and rearrangements) within a set of genomes through multiple sequence alignment. Then, it applies majority vote to decide on consensus sequences from a set of segments of aligned sequences and merges those sequences with delimiter sequences. Our approach *plaster* employs a similar approach, but instead we calculate pairwise alignments with NUCmer and identify unaligned regions using dnadiff. We then append all unaligned regions along with delimiter sequences to the end of the reference sequence Fig. 1. We introduce the algorithm of *plaster* in detail and compare its performance with seq-seq-pan. In addition, we describe differences and similarities between the synthetic and natural plasmid pan-genome sequences based on pairwise alignment results.

2 Methods

2.1 *plaster* Workflow

The goal of our algorithm is to construct a pan-genome P from a set $S = [s_0, s_1, ..., s_n]$ of n genome sequences. Throughout this paper, we focus on plasmid sequences, though any arbitrary sequences can be used to construct the pan-genome P. A high level overview of our algorithmic pipeline is outlined in Fig. 1. Though not necessary to run the *plaster* software, in all analyses in this paper we first sort S so that $|s_0| \geq |s_1| \geq ... \geq |s_n|$ where $|s|$ denotes the length of a sequence s. This initialization offers optimal performance of the method and we recommend it to be used as a standard practice for users of *plaster*.

As shown in Fig. 1, the pipeline for our full *plaster* algorithm can be broken up into a series of discrete tasks. After ordering the sequences by length, the pan-genome P, also referred to as the reference sequence, is initialized as the longest sequence s_0 in S. We then proceed to build the full pan-genome by looping through all the remaining sequences, $[s_1, s_2, ..., s_n]$, and performing the steps detailed in Fig. 1. We refer to each of these sequences as a query sequence which we compare to the current pan-genome reference sequence.

For each query sequence, *plaster* begins by identifying the regions that do not align between the query and current reference sequence. For this, NUCmer and dnadiff [25] (from the MUMmer package [16], default dnadiff parameters were used). Pairwise genome alignment is performed by NUCmer; following alignment, dnadiff provides detailed analysis of all differences between the two sequences. Once the differences between the query and reference sequences have been calculated, we add unaligned query sequence regions to the pan-genome reference. The .report file output by dnadiff contains the total alignment percentage

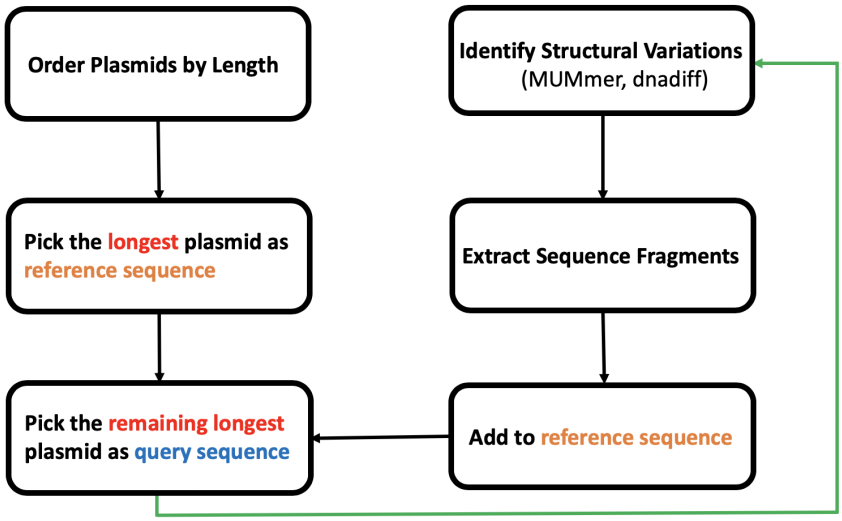


Figure 1 *plaster* Pipeline: To build a pan-genome sequence, first we order all of the input sequences by length. We next select the longest plasmid as an initial reference sequence and the second longest plasmid as the first query sequence. We perform pairwise local alignment to identify DNA sequences that are not contained in the current pan-genome. We then extract the novel sub-sequences from the query sequence and update the pan-genome with these sequences. We continue the pan-genome construction process by selecting the next longest sequence in the remaining data set as the next query sequence. We iterate over all input sequences until the entire set has been processed.

between the reference and query sequence. If the alignment percentage is zero, that is, the query sequence is totally unaligned to the reference, we add the entire query to the plasmid pan-genome by appending a delimiter sequence and the entire query sequence to the end.

Following this procedure, there are five different types of structural variants we capture: gaps, duplications, breaks, relocations, and inversions. Each of these represents a different type of sequence variant. For the purpose of capturing the minimal pan-genome, we only care about variations where there are nucleotides in the query sequence that are not described in the current reference. This includes: i) gaps, which identifies indels between two consistently ordered and oriented alignments, and ii) breaks, which refers to a fragment of query sequence not aligned to the reference sequence. Inversions, relocations, and duplications are sequence variants of interest, but do not require an update to the linear pan-genome. By default, if the length of a gap or break area is longer than l (default is 50bp), the corresponding unaligned query sequence area will be appended to the end of the reference sequence after an additional delimiter sequence. This process of finding and appending all differences between a query sequence and the current pan-genome reference sequence continues until every sequence in S has been iterated over. The final output of *plaster* is the resulting reference sequence which represents a linear pan-genome P for the genomes of S.

3 Results

Having developed a novel algorithm for efficiently constructing pan-genomes that can handle very large numbers of sequences, we set out to categorize the differences between all known naturally occurring and synthetic plasmids. To justify the efficacy and utility of our newly developed *plaster* algorithm, we perform experiments on the running time of our method.

The run-time for *plaster* also includes the time for sorting the input sequences based on length. We build the full pan-genomes of all natural and synthetic plasmids and we take a first look at a novel approach for categorizing plasmids as natural or synthetic based on percent alignment against these two pan-genomes. All the experiments in this paper are written in python and run on a server running Ubuntu 18.04 LTS with two Intel Xeon Gold 6138 2.0 GhZ processors at 1 terabyte of RAM.

3.1 Evaluation of sensitivity

A total of 43 *M. tuberculosis* genomes were used to build a pan-genome sequence in [14]. However, only 41 genomes were still available in [23]. We used these 41 *M. tuberculosis* genomes to build a pan-genome sequence using both seq-seq-pan and *plaster*. Seq-seq-pan requires 34 min to construct the pan-genome sequence. On the other hand, it only takes *plaster* around 4.2 min. The total length of the pan-genome sequence built by *plaster* is 385,335 bp shorter than the one from seq-seq-pan. However, the pan-genome sequence generated by *plaster* can align a total of 8,828 bp more (or 99.994% average alignment percentage compared to 99.989% for seq-seq-pan) when performing pairwise alignment of all 41 *M. tuberculosis* sequences as query sequences against the pan-genome built from these same 41 genomes as the reference sequence. This shows that *plaster* is an order of magnitude faster than seq-seq-pan while simultaneously producing a more compacted pan-genome sequence and capturing more detailed variations within it.

3.2 Evaluation of speed

Given the rapid increase in the number of natural and engineered plasmids, there is a need to develop a platform which can build pan-genomes for large data sets within a reasonable time and be able to update the pan-genome information quickly for newly sequenced genomes. To evaluate the pan-genome construction speed of *plaster*, we first performed experiments on subsets of synthetic plasmid sequences and compared the performance of *plaster* against the state of the art seq-seq-pan pan-genome construction method. We randomly selected subsets of 10, 100, and 1000 synthetic plasmids from the full set of all synthetic plasmids. Then, we recorded the wall clock time for building a pan-genome from these sequences from start to finish.

As seen in Table 2, the performance of *plaster* for building a pan-genome sequence ranges from 10 to 100 times faster than seq-seq-pan. *plaster* processes plasmids at a rate of 0.25 seconds/plasmid for 10 plasmids, 0.3 seconds/plasmid for 100 plasmids, and 0.337 seconds/plasmid for 1000 plasmids. For seq-seq-pan, the rate is 2.56 seconds/ plasmid, 4.7 seconds/plasmid, and 35.3 seconds/plasmid for 10, 100, and 1000 plasmids respectively. From these results we can see that seq-seq-pan will not scale to building pan-genome sequences for all known synthetic or natural plasmids. On the other hand, *plaster* is able to build a pan-genome for 51,047 complete engineered plasmid sequences in 8.9 hours.

Table 1 Run-time, sensitivity, and pan-genome length for 41 *M. tuberculosis* genomes.

	Seq-seq-pan	*plaster*
Run-time (s)	2058.64	252.215
Pan-genome length (bp)	4,874,793	4,489,458
Total Aligned Length (bp)	180,681,735	180,690,563
Average Alignment Percentage (%)	99.989±0.00298	99.994±0.00514

Table 2 Run-time for building a pan-genome sequence and updating a pan-genome sequence with one additional new sequence. We compare the speed of *plaster* to seq-seq-pan when building pan-genome sequences using 10, 100, and 1000 random plasmids. We repeated the 10 and 100 sequence runs 20 times and report their average and standard deviation.

No. of plasmids	Elapsed wall clock time (s) Pan-genome Construction		Elapsed wall clock time (s) Pan-genome Update	
	seq-seq-pan	*plaster*	seq-seq-pan	*plaster*
10	25.265±0.736	1.90±0.0717	3.5±0.163	0.22±0.029
100	495.5±40.926	22.189±0.401	7.84±1.89	0.239±0.0243
1000	35295.3	337.2	86.679	0.886

In addition to the times for full pan-genome construction in Table 2, we compared the timings for the pan-genome update speed. This update speed is the speed at which we can add a single new sequence to an already fully constructed pan-genome. For this experiment, we randomly selected one plasmid from our synthetic plasmid data set and added it to the pan-genome sequences created in the previous experiment for 10, 100 and 1000 plasmids using either seq-seq-pan or *plaster*.

As demonstrated by the results of Table 2, the update speed of seq-seq-pan grows intractably for larger pan-genomes when compared to *plaster*. *plaster* is about 15 times faster for the 10 plasmid pan-genome sequence, 30 times faster for the 100 plasmid pan-genome sequence, and 100 times faster for the 1000 plasmid pan-genome sequence. Given the speed of single sequence updates and the rapid growth of sequence databases, we envision that *plaster* could be used as an online algorithm which updates a pan-genome sequence every time new sequences are added to the database. A final timing experiment was performed showing the update speed for *plaster* for all of a set of 51,047 synthetic plasmid sequences. Fig. 2 shows these update times. The curvature of Fig. 2 suggests that we have created a closed pan-genome, where update times gradually converge to a nearly constant time as the pan-genome grows large.

3.3 Full Natural and Synthetic Plasmid Pan-Genomes

We built a full synthetic plasmid pan-genome and full natural plasmid pan-genome using *plaster* based on 51,047 synthetic plasmid sequences, 73,727 partial synthetic plasmid sequences, and 6,642 natural plasmids. All the synthetic plasmids came from the Addgene engineered plasmid database from January 2019 [13]. A JavaScript Object Notation (JSON) file containing these DNA sequences was obtained directly from Addgene. These sequences were grouped into four categories which were full plasmid sequences submitted by Addgene, partial sequences (segments of the plasmid) submitted by Addgene, full plasmid sequences submitted by a depositing lab, and partial sequences submitted by a depositing lab. There were a total of 51,047 plasmids with complete sequences, in which 23,875 were submitted by Addgene, and 73,727 partial sequences. The average size of a complete synthetic plasmid is 7,159 bp.

To construct the full synthetic plasmid pan-genome, we first ordered all the plasmids with full sequences based on their lengths. Then, we used those ordered plasmids as input to *plaster* to build a full pan-genome sequence. It took about 8.9 hours to build the pan-genome sequence and the final resulting linear pan-genome consisted of 8,307,070 bp. The total number of nucleotides of all the synthetic plasmids with complete sequences was 365,468,935 bp. After building the pan-genome sequence with only full plasmid sequences, we generated

a second pan-genome sequence using both full and partial sequences. We ranked the partial sequences based on their lengths and input those sequences into the *plaster* pipeline as query sequences. Given the online nature of *plaster*, we were able to use the already constructed pan-genome for the full synthetic plasmids as the initial reference sequence, iterating over all the partial sequences as new query sequences. *plaster* spent less than 35 hours to construct the final pan-genome for both full and partial sequences. The total number of nucleotides of this synthetic plasmid pan-genome sequence using both full and partial sequences is 18,163,933. The total number of nucleotides of all the sequences used to construct the pan-genome was 442,923,876. For natural plasmids, we obtained the DNA sequences from the latest database of plasmid sequences [2]. All of these natural plasmids are complete sequences that were curated from the entire NCBI database [23]. There are a total of 6,642 plasmids in this database with an average size of 128,953 bp. The total number of nucleotides of all natural plasmids was nearly one billion base pairs (856,388,404 bp total length). It took *plaster* around 50.2 hr (wall clock time) to build the full natural plasmid pan-genome sequence based on these sequences. The total number of nucleotides in the resulting pan-genome sequence was 205,605,679 bp. The size of the pan-genome sequence is around 24% of the total size of the natural plasmids.

3.4 Pan-genome Sequence Evaluation and Natural Versus Synthetic Differentiation

An exciting new area of research in bioinformatics is in characterizing engineered DNA and in differentiating engineered versus naturally occurring nucleotide sequences. To assess our synthetic and natural pan-genomes, we realigned all natural and synthetic plasmids against these two pan-genomes and report on the percent alignment for each of these plasmids. We also highlight how this technique unveils a novel approach for differentiating synthetic and natural plasmid sequences, given that each is expected to align well to its corresponding pan-genome but not align well to the other pan-genome. Aligning a given query sequence to both pan-genomes can therefore be used as an initial test for evidence of genomes having been subjected to genome-editing.

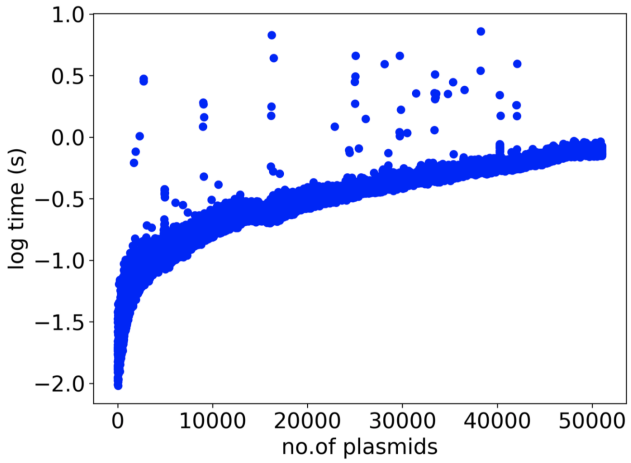


Figure 2 Cumulative run time per individual plasmid used when building the synthetic plasmid pan-genome sequence. x-axis is the cumulative number of plasmids. y-axis is the cumulative log scale time. As more plasmids are added to the pan-genome, the run time for each individual plasmid increases. The run-time growth rate decreases after a few thousand plasmids are appended as the sequence shifts from a more open pan-genome to a more closed pan-genome.

To assess the representation of each of the plasmids within the newly constructed pan-genome sequences, we did pairwise alignment using MUMmer. We used the plasmid pan-genome sequences as the reference and all the plasmids from the two sets of plasmids as query sequences. The percentage of a query sequence aligned to the pan-genome sequence is a good estimator for evaluating how well the query sequence was included as part of the corresponding pan-genome sequence. Figure 3a shows that, for most of the synthetic plasmids, more than 75% of the DNA sequence aligns to the synthetic plasmid pan-genome sequence. On the other hand, when natural plasmids are aligned to this same synthetic plasmid pan-genome in Fig. 3b, most of the natural plasmids have below 5% of their DNA aligning well. In Fig. 3c and Fig. 3d, we perform this same experiment but we align synthetic and natural plasmids against the natural plasmid pan-genome. Again, most of the natural plasmids are well aligned to the corresponding natural pan-genome sequence with more than 75% of the DNA aligning to the pan-genome in Fig. 3c. On the other hand, smaller portions of synthetic plasmids align well to the natural plasmid pan-genome sequence in Fig. 3d. The alignment percentages for synthetic plasmids range mainly from 25% to 75%

To evaluate the utility of natural and synthetic plasmid pan-genomes for differentiating natural and synthetic plasmids, we randomly selected 1000 synthetic plasmids (submitted to Addgene from January 2019 to June 2019) and 1000 natural plasmids from [9]. None of these plasmids were used in the pan-genome construction steps. As such, we used these 2000 total plasmids as a validation test data set. Our classification was performed as follows: Given an unclassified plasmid p, if its pairwise alignment percentage of p compared to the synthetic pan-genome SPG is above the threshold t AND its pairwise alignment percentage against the natural pan-genome NPG is below the threshold t, we classified p as a synthetic plasmid (and vice versa). Note: if p has alignment percentage above or below the threshold t for both synthetic and natural plasmid pan-genomes, we leave the sequence unclassified. Table 3 indicates that by mapping to the SPG and NPG, an unknown plasmid can be differentiated between being a synthetic or natural plasmid with high specificity. For synthetic plasmid classification, a high threshold yields high sensitivity. On the other hand, natural plasmid categorization has a high sensitivity with a low threshold.

To investigate the impact of the order of input sequences on building the pan-genome, we created a synthetic pan-genome using full synthetic plasmids with the input order starting from the shortest to the longest sequence. All the other parameters remained the same. The total number of nucleotides of the full synthetic plasmid pan-genome with this reverse input order is 6,585,872 bp. This is 1,721,198 bp shorter than the pan-genome built starting from the longest to the shortest sequence. The average pairwise alignment percentage of a plasmid against the initial pan-genome and the pan-genome with reversed input order are 91.52±11.63 % and 89.12±13.656 % respectively. Pairwise alignment results indicate that the alignments against the pan-genome with reversed input order are slightly worse than that against the initial synthetic plasmid pan-genome, but most of the synthetic plasmids still have more than 75% alignment against the full synthetic pan-genome.

In addition, we analyzed the number of synthetic plasmids that aligned to each nucleotide of the two pan-genome sequences. When aligning the plasmids against the synthetic pan-genome, most of the plasmids were mapped at the beginning of the synthetic plasmid pan-genome sequence (results not shown). In other words, the start of the synthetic plasmid pan-genome captures most of the nucleotide sequences shared by synthetic plasmids. The fragment with the highest number of aligned plasmids, which started at position 6077 in the synthetic plasmid pan-genome sequence, is annotated by *prokka* [26] as gene *bla* with protein product of "Beta-lactamase TEM precursor". The results of aligning synthetic plasmids

Table 3 Differentiation of natural and synthetic plasmids through pan-genome alignment. We classify 2000 new plasmids as either synthetic or natural plasmids based on alignment percentage against synthetic and natural plasmid pan-genomes. We calculated the sensitivity, specificity and F-score for each result.

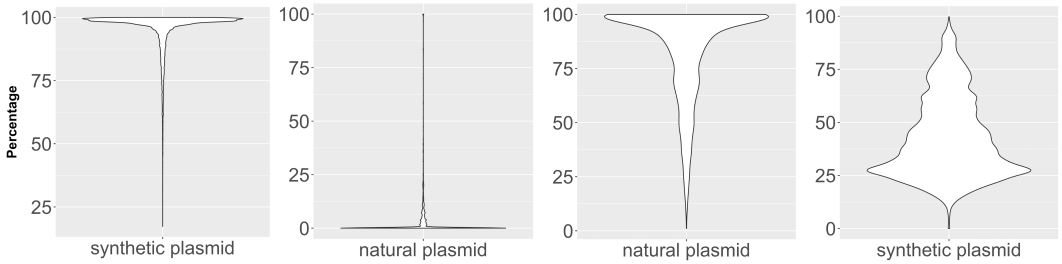
Threshold (t) (Alignment %)	Synthetic Plasmids			Natural Plasmids		
	Sensitivity	Specificity	F-score	Sensitivity	Specificity	F-score
40	0.579	0.998	0.73	0.767	1	0.868
60	0.775	0.999	0.87	0.683	0.999	0.811
80	0.821	0.999	0.90	0.536	0.998	0.697
85	0.773	0.999	0.872	0.502	0.995	0.667

and natural plasmids against the natural pan-genome are shown in Fig. 4. Specifically, the natural plasmid pan-genome sequence also includes some synthetic plasmid fragments (see Fig. 4a). Among the fragments with more than 10,000 plasmids mapped to them, two were assigned functions by *prokka*, with the rest having no results. Of these two fragments, one started at position 95,790,389 and was annotated as *lacZ* Beta-galactosidase enzyme; the other started at position 160,749,886 and was annotated as gene `bla` (Beta-lactamase). Both *lacZ* and *bla* were involved in one of the first minimal fully synthetic plasmids [15, 32], and known to be ubiquitous to the synthetic plasmidome.

4 Discussion and Conclusions

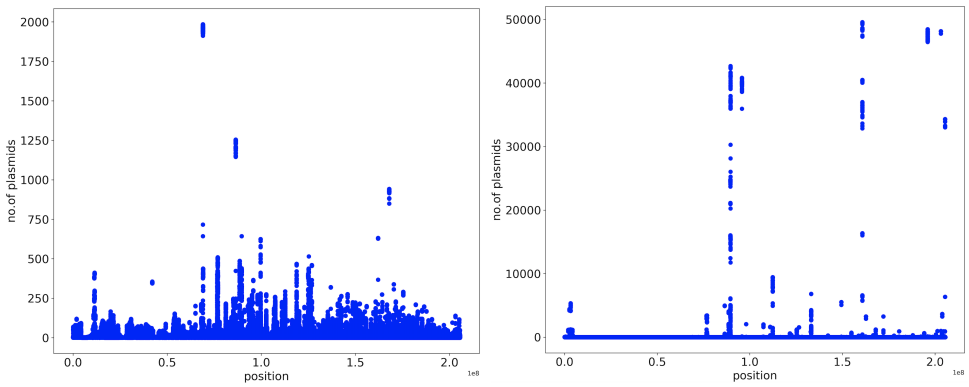
In this paper, we introduced a novel pipeline for building pan-genomes, called *plaster*, which takes advantage of closed nature of pan-genomes when possible and displays an order of magnitude improvements to the pan-genome construction speed when compared with the fastest existing method in seq-seq-pan. For instance, given a new sequence, it can rapidly update the pan-genome sequence within 0.01s for an already existing pan-genome built from 1000 plasmids. Using *plaster*, to the best of our knowledge we constructed the first ever pan-genomes for natural and synthetic plasmids. These pan-genomes reflect the core sequences as well as the sequence diversity that exists among sequenced natural and engineered plasmids.

Alongside these newly created pan-genomes, we presented a novel technique that serves as a first step in inferring whether a plasmid is natural or synthetic. The previous work of [19] used machine learning to attempt to infer labs of origin for synthetic DNA. One common



(a) SPG vs. SPP. **(b)** NPG vs. SPP. **(c)** NPG vs. NPP. **(d)** SPG vs NPP.

Figure 3 Alignment percentages for naturally occurring plasmids vs synthetic plasmids comparisons. (a) SPG=Synthetic plasmid genomes, (b) SPP=Synthetic plasmid pan-genome, (c) NPG=Natural plasmid genomes, NPP=Natural plasmid pan-genome, (d) SPG=Synthetic plasmid genome.



(a) NPG vs. NPP. **(b)** SPG vs. NPP.

Figure 4 Natural plasmid genome alignment (a) , and synthetic plasmid genome alignment (b), with respect to the Natural plasmid pan-genome. x-axis represents the position in the pan-genome (1 to 205 Mbp), y-axis represents the number of plasmids that align to a given position in the natural plasmid pan-genome.

disadvantage of machine learning approaches, in particular neural networks, suffer from poor explainability due to their black box nature [11]. Our approach presented here performs rapid pairwise sequence alignment of the query sequence against the natural and synthetic plasmid pan-genome and, as such, provides a white-box approached that allows the user to directly query the regions that are shared between natural and synthetic plasmids, specific to a give lab-of-origin, as well as what regions can be used to differentiate plasmids that have undergone human engineering [1]. Given a synthetic plasmid, our pipeline could be used to recover the full set of structural variations to fully categorize why it was determined to be synthetic as well as what engineering it has undergone.

There are several areas for future research left open by this work. As mentioned, alignment against pan-genomes can yield a full set of differences between a suspected synthetic plasmid and all natural and synthetic nucleotides contained in the pan-genomes. To investigate new ways to infer a possible lab of origin, the synthetic pan-genome could include labels for all of its contained variation and the lab(s) where that variation has been seen before. If a machine learning approach is ultimately preferred and is the most accurate for determining an origin lab or discriminating natural versus synthetic, the set of all mutations and structural variations for a given plasmid will still be crucial to have in the set of features for the final inferrence procedure. Functional annotations could also be added to determine what a particular lab was trying to achieve with the particular structural variations and mutations that they introduced into a plasmid. Investigations into fundamental improvements to the *plaster* method also remain as future work. For instance, an arbitrary choice was made to use variations greater than 50 bp when adding new nucleotides to the pan-genome. There will be pros and cons for varying this value to higher and lower values. This minimum alignment length could also be tuned depending on the final purpose of constructing the pan-genome, for instance for determining a lab of origin for synthetic plasmids. Tweaks to this value, as well as other algorithmic improvements, should push these alignments towards all being 100 percent alignment as well as aiding in the following step of differentiating natural and synthetic plasmids.

References

1 Jonathan E Allen, Shea N Gardner, and Tom R Slezak. DNA signatures for detecting genetic engineering in bacteria. *Genome biology*, 9(3):R56, 2008.

2 Lauren Brooks, Mo Kaze, and Mark Sistrom. A Curated, Comprehensive Database of Plasmid Sequences. *Microbiol Resour Announc*, 8(1):e01325–18, 2019.

3 Hans Bügl, John P Danner, Robert J Molinari, John T Mulligan, Han-Oh Park, Bas Reichert, David A Roth, Ralf Wagner, Bruce Budowle, Robert M Scripp, et al. DNA synthesis and biological security. *Nature biotechnology*, 25(6):627, 2007.

4 Jean Cury, Pedro H Oliveira, Fernando de la Cruz, and Eduardo PC Rocha. Host Range and Genetic Plasticity Explain the Coexistence of Integrative and Extrachromosomal Mobile Genetic Elements. *Molecular biology and evolution*, 35(9):2230–2239, 2018.

5 Robert C Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010.

6 Mark Eppinger, Talima Pearson, Sara SK Koenig, Ofori Pearson, Nathan Hicks, Sonia Agrawal, Fatemeh Sanjar, Kevin Galens, Sean Daugherty, Jonathan Crabtree, et al. Genomic epidemiology of the Haitian cholera outbreak: a single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. *MBio*, 5(6):e01721–14, 2014.

7 Corinna Ernst and Sven Rahmann. PanCake: a data structure for pangenomes. In *German Conference on Bioinformatics 2013*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.

8 Derrick E Fouts, Lauren Brinkac, Erin Beck, Jason Inman, and Granger Sutton. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic acids research*, 40(22):e172–e172, 2012.

9 Valentina Galata, Tobias Fehlmann, Christina Backes, and Andreas Keller. PLSDB: a resource of complete bacterial plasmids. *Nucleic acids research*, 47(D1):D195–D202, 2018.

10 Chris D Greenman, Erin D Pleasance, Scott Newman, Fengtang Yang, Beiyuan Fu, Serena Nik-Zainal, David Jones, King Wai Lau, Nigel Carter, Paul AW Edwards, et al. Estimation of rearrangement phylogeny for cancer genomes. *Genome research*, 22(2):346–361, 2012.

11 Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

12 Finbarr Hayes. The function and organization of plasmids. In *E. coli Plasmid Vectors*, pages 1–17. Springer, 2003.

13 Melanie Herscovitch, Eric Perkins, Andy Baltus, and Melina Fan. Addgene provides an open forum for plasmid sharing. *Nature biotechnology*, 30(4):316, 2012.

14 Christine Jandrasits, Piotr W Dabrowski, Stephan Fuchs, and Bernhard Y Renard. seq-seq-pan: Building a computational pan-genome data structure on whole genome alignment. *BMC genomics*, 19(1):47, 2018.

15 Wlodek Mandecki, Mark A Hayden, Mary Ann Shallcross, and Elizabeth Stotland. A totally synthetic plasmid for general cloning, gene expression and mutagenesis in Escherichia coli. *Gene*, 94(1):103–107, 1990.

16 Guillaume Marçais, Arthur L Delcher, Adam M Phillippy, Rachel Coston, Steven L Salzberg, and Aleksey Zimin. MUMmer4: a fast and versatile genome alignment system. *PLoS computational biology*, 14(1):e1005944, 2018.

17 Tobias Marschall, Manja Marz, Thomas Abeel, Louis Dijkstra, Bas E Dutilh, Ali Ghaffaari, Paul Kersey, Wigard P Kloosterman, Veli Makinen, Adam M Novak, et al. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 19(1):118–135, 2018.

18 Elizabeth Anne McMillan, Sushim K Gupta, Laura Williams, Thomas Jové, Lari M Hiott, Tiffanie A Woodley, John B Barrett, Charlene Renee Jackson, Jamie L Waslienko, Mustafa Simmons, et al. Antimicrobial Resistance Genes, Cassettes, and Plasmids present in Salmonella enterica associated with US Food Animals. *Frontiers in microbiology*, 10:832, 2019.

19 Alec AK Nielsen and Christopher A Voigt. Deep learning to predict the lab-of-origin of engineered DNA. *Nature communications*, 9(1):3135, 2018.

20 Teresa Nogueira, Daniel J Rankin, Marie Touchon, François Taddei, Sam P Brown, and Eduardo PC Rocha. Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Current Biology*, 19(20):1683–1691, 2009.

21 Tuukka Norri, Bastien Cazaux, Dmitry Kosolobov, Veli Mäkinen, et al. Minimum Segmentation for Pan-genomic Founder Reconstruction in Linear Time. In *18th International Workshop on Algorithms in Bioinformatics (WABI 2018)*. Schloss Dagstuhl Leibniz Center for Informatics, 2018.

22 Ryan S Noyce, Seth Lederman, and David H Evans. Construction of an infectious horsepox virus vaccine from chemically synthesized DNA fragments. *PloS one*, 13(1):e0188453, 2018.

23 Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2015.

24 Andrew J Page, Carla A Cummins, Martin Hunt, Vanessa K Wong, Sandra Reuter, Matthew TG Holden, Maria Fookes, Daniel Falush, Jacqueline A Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, 2015.

25 Adam M Phillippy, Michael C Schatz, and Mihai Pop. Genome assembly forensics: finding the elusive mis-assembly. *Genome biology*, 9(3):R55, 2008.

26 Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.

27 Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 102(39):13950–13955, 2005.

28 Harry A Thorpe, Sion C Bayliss, Laurence D Hurst, and Edward J Feil. Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics*, 206(1):363–376, 2017.

29 Harry A Thorpe, Sion C Bayliss, Samuel K Sheppard, and Edward J Feil. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *Gigascience*, 7(4):giy015, 2018.

30 Esko Ukkonen. Finding founder sequences from a set of recombinants. In *International Workshop on Algorithms in Bioinformatics*, pages 277–286. Springer, 2002.

31 George Vernikos, Duccio Medini, David R Riley, and Herve Tettelin. Ten years of pan-genome analyses. *Current opinion in microbiology*, 23:148–154, 2015.

32 Barry L Wanner. Molecular cloning of Mu d (bla lacZ) transcriptional and translational fusions. *Journal of bacteriology*, 169(5):2026–2030, 1987.

33 Tom A Williams, Peter G Foster, Cymon J Cox, and T Martin Embley. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, 504(7479):231, 2013.

34 Derrick E Wood, Henry Lin, Ami Levy-Moonshine, Rajiswari Swaminathan, Yi-Chien Chang, Brian P Anton, Lais Osmani, Martin Steffen, Simon Kasif, and Steven L Salzberg. Thousands of missed genes found in bacterial genomes and their analysis with COMBREX. *Biology direct*, 7(1):37, 2012.

35 Yongbing Zhao, Jiayan Wu, Junhui Yang, Shixiang Sun, Jingfa Xiao, and Jun Yu. PGAP: pan-genomes analysis pipeline. *Bioinformatics*, 28(3):416–418, 2011.