

# Detecting genetic engineering in the wild

## Abstract

We present *k-mer model* for detecting the source lab of an engineered DNA sequence. The key advantages of this model are high speed of inference, ease of modification, few assumptions about the sample, and ability to work with short sequences. This allows evaluating raw sequencing datasets for traces of genetic engineering without additional complicated processing steps. Previously, one could only determine the source lab of a plasmid whose exact sequence is known, which limited potential application scenarios. The k-mer model enables *fully automatic* detection and attribution of engineered DNA in any biological sample.

In this report, we first validate the approach using simulations, showing that it can detect the presence of a single engineered plasmid in raw sequencing data. We then showcase it with a study that discovers signs of genetic engineering among 874 *Cannabis* genomes from a public dataset.

## Introduction

The task of genetic engineering attribution is, given a complete engineered DNA sequence, to identify its the most likely source lab. The attribution task is defined this way in recent publications [1, 2] and in the Prediction Track of the DrivenData challenge. The modern solution to this problem is a neural network, e.g. a CNN [2] or RNN [1]. They are precise, but time consuming to train, generally require re-training when training set changes (e.g. when a new lab needs to be recognized), and require expensive hardware to run. It is also difficult to extract concrete evidence of attribution from their predictions: the similarities between the input sequence and sequence(s) from the alleged source remain implicit.

This report is based on the submission to the Prediction Track which is an average of 8 models: 7 large CNNs, and a k-mer scoring model. This blend of models predicts attribution very accurately, almost perfectly for labs with >20 training samples. The focus of this report, however, is the k-mer model.

k-mer model works with DNA sequences only. It stores k-mers (continuous subsequences of length  $k$ , here  $19 \leq k \leq 27$ ) from the training samples. When predicting, it assigns scores to the labs based on how many stored k-mers from the lab are also present in the given input sequence. k-mers are weighted: the lab gets higher score if the k-mer is found in its training samples often, but is rarely found in other labs. The k-mer storage uses a hash table data structure, which makes lookups very fast, even in a large database. The method is inspired by Kraken[3], which uses a k-mer database to detect species in genomic data, a task which is similar to attribution.

k-mer model is fast, processing  $\sim 10$  million base pairs (Mbases) of input sequences per second. Its scores are derived from matching subsequences between training and test samples, which is transparent and simple, giving a way to recheck and explain the results. The database can be modified easily, by adding new k-mers from additional training samples, or removing k-mers that are considered irrelevant. The accuracy of the model is also high, only 3-4% below the accuracy of the full blended model. This is still more accurate than the current state of the art as reported in [1], even though that model uses additional biological data about the samples. All these advantages open up a new possibility: we can now scan raw sequencing datasets for engineered DNA, just like Kraken does for species identification.

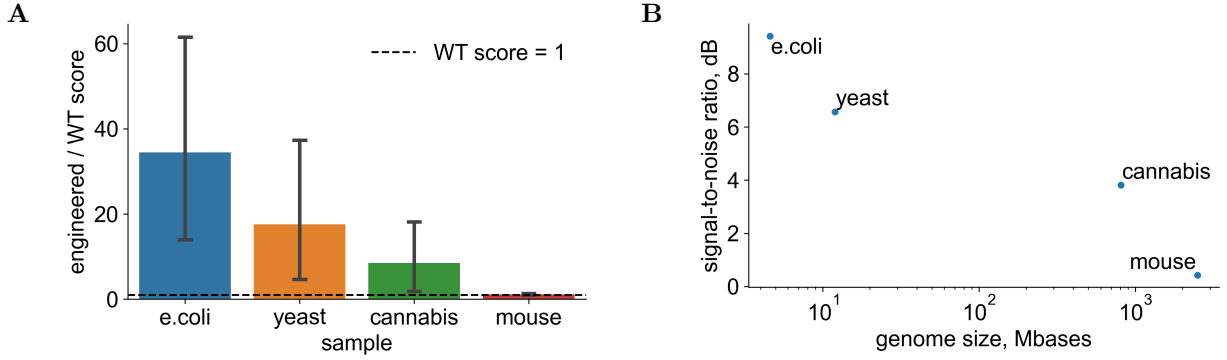


Figure 1: Detection of simulated genetic engineering. **A.** Detection score of samples with simulated plasmid reads, divided by the score of corresponding wildtype samples. Bar height: average score of 20 random simulations. Error bar: 95% confidence interval. Dashed line: WT score level. **B.** Dependency of detection score's signal-to-noise ratio on genome size. X: Size of species' reference genome, log scale. Y: Estimated signal-to-noise ratio, dB

## Validation

To evaluate whether k-mer model can detect genetic engineering in raw data, we simulated addition of a single plasmid to genomes. Public whole genome sequence (WGS) datasets of four species were downloaded and each was randomly split into two train and test parts. For each species separately, we take the k-mer database built from the competition training data, and then remove from it the k-mers found in the train part of species' data. The resulting database contains k-mers from engineered plasmids which were not observed in the wildtype (WT) sample. Presence of these k-mers in a dataset indicates genetic engineering. The test part of the WGS data is the negative control, providing the reference score of a non-engineered sample. The positive sample is created by simulating reads from a randomly chosen plasmid from the competition's test set. The number of simulated reads is such that would be observed if there is a single plasmid in each cell, according to the WGS dataset sequencing depth, species' genome size, and plasmid size. We sum the labs' scores calculated for all reads in the dataset, and take the maximum over all labs as the main *engineering detection score*.

Fig. 1A shows the scores of 20 random trials (each time selecting another test plasmid), relative to the negative control WT score. For *E. coli*, yeast, and *Cannabis*, the score of samples with simulated plasmids are reliably higher than the negative control, proving that the method can distinguish between WT and engineered samples. For mouse, however, the scores are very low.

The difference in detection power for different species is explained in Fig. 1B. Signal-to-noise ratio of the score, computed using the data above, falls to near zero with increased genome size. In large genomes, like that of mouse, k-mers from engineered plasmids occur randomly, drowning the signal from the plasmid.

To conclude, k-mer model works when applied to raw data, unless the proportion of engineered DNA in the sample is very small. Importantly, it works in relevant organisms.

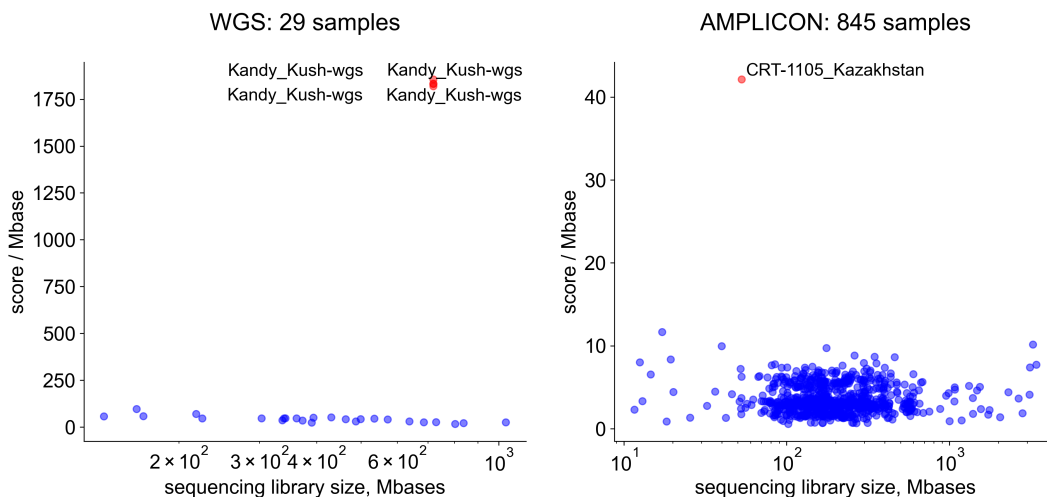


Figure 2: Detection of genetic engineering in *Cannabis* samples. Left: WGS samples from various sources. Right: AMPLICON samples from Phylos Bioscience. X: Size of the dataset, log scale. Y: detection score divided by dataset size. Outliers (score/Mbase > median + 5  $\times$  Interquartile range) are marked in red, and their sample names are indicated.

## Is *Cannabis* being genetically engineered?

*Cannabis* is an agriculturally important plant, one of the most widely cultivated worldwide[4]. Many strains exist, and many sequencing datasets are available[5]. Due to great commercial value, it is conceivable that some strains are genetically engineered, but there is little public information about it.

To demonstrate the possibilities opened up by the k-mer model, we will scan the public *Cannabis* genomes for signs of genetic engineering. The *Cannabis* WGS dataset from validation is used to remove WT k-mers from the training database. Then, 874 public datasets from [5] are scanned against this database (Fig. 2). The scores were normalized to the corresponding library size. Most samples in these datasets come from Phylos Bioscience, which used a custom AMPLICON targeted sequencing strategy. The other samples are WGS. AMPLICON, by targeting certain regions of the genome and ignoring others, makes a biased library, which results in a different range of scores; thus we analyze it separately from unbiased WGS. In both cases, natural variation of WT scores is stable across a wide range of sequencing depths. Outliers with much higher scores are possibly genetically engineered.

To prove that these samples are indeed genetically engineered, one may follow these findings up with checking which k-mers gave high scores, what training samples they correspond to, in which genomic context they are found. This is out of scope of this report. The results presented here show feasibility of this application scenario for k-mer model. It took  $\sim 36$  hours to process 1.8Tb of raw sequencing data on a single 8-core machine.

## Applications

Requiring a clean engineered sequence as input for attribution models is a major limitation of previous methods. Extracting an unknown plasmid sequence from a biological sample, as suggested in [1], is practically challenging. To employ targeted sequencing, we first have to

know what to target. *De novo* discovery of new DNA constructs in WGS requires enormous sequencing depth, and it is still difficult even with that: e.g. plant genomes like *Cannabis* have highly repetitive DNA that hinders assembly [4]. Doing this before we even know whether there *is* artificial DNA in the sample is infeasible.

We propose to run k-mer model to detect unusually high percentage of engineering-related sequence content in normal sequencing data, as we did with *Cannabis* above. This is fast and doesn't require very deep sequencing. If needed, this can be followed up with ultra-deep re-sequencing, or re-sequencing targeted on suspicious k-mers, of samples that cross the detection threshold. Then the detected plasmids or other DNA modifications can be studied in detail.

This method makes new useful application scenarios possible. To name a few:

- Monitoring of runaway genetic modifications. GMO escaping from farms and labs can be problematic. High-throughput detection and attribution of engineered DNA can be used to survey samples from the wild and trace artificial DNA contamination to its source.
- Verifying non-GMO status. It becomes easy for anyone to test any organism for presence of artificial DNA.
- Revealing competitors' genetic engineering efforts. An agricultural business can scan competing products in search for new genetic engineering ideas.
- As a biological warfare early warning system, new pathogens can be routinely checked for likelihood of being artificial.

Generally, genetic engineering will become more transparent and accountable, which brings about many benefits.

## Limitations

We can only detect genetic modifications which are similar to training samples in some way. The method is not currently applicable to large (e.g. mammalian) genomes with a small fraction of engineered DNA (e.g. a single short plasmid). Also, the score has to be calibrated using control (WT) samples of the same kind as the samples being tested.

## References

1. Alley, E. C. *et al.* Attribution of genetic engineering: A practical and accurate machine-learning toolkit for biosecurity. *bioRxiv* (2020).
2. Nielsen, A. A. K. & Voigt, C. A. Deep learning to predict the lab-of-origin of engineered DNA. *Nature Communications* **9**, 3135 (Aug. 2018).
3. Wood, D. E. *et al.* Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 257 (Nov. 2019).
4. Gao, S. *et al.* A high-quality reference genome of wild *Cannabis sativa*. *Horticulture Research* **7**, 73 (May 2020).
5. Day, A. *DNA Sequencing of 1000 Cannabis Strains publicly available in Google BigQuery* Mar. 2017.