BMC Bioinformatics

RESEARCH

Open Access

DNA sequence contamination analyzer (DNASCAN): a supervised analysis toolkit for detecting and removing DNA contaminants

John Stephen Malamon[1,2*]

*Correspondence:
John Stephen Malamon
john.malamon@cuanschutz.edu
[1]Division of Transplant Surgery, Department of Surgery, University of Colorado, Anschutz Medical Campus, 1635 Aurora Court, Aurora, CO 80045, USA
[2]Division of Transplant Surgery, Colorado Center for Transplantation Care, Research and Education (CCTCARE), Aurora, CO 80045, USA

Abstract

DNA N-gram analysis methodologies have been successfully deployed to provide a wide range of elegant solutions to a variety of complex problems in bioinformatics, such as sequence alignment, DNA barcoding, the identification of functional gene elements, the analysis of microbial genomes, the characterization of protein structures, and the detection of DNA sequence artifacts. Because biological DNA contamination is ubiquitous and therefore unavoidable, it has a significant impact on genomics and genetics research, posing substantial challenges in population genotype calling quality, model organism research, proteomics, and clinical gene therapies such as recombinant adeno-associated viral vector preparations. To this end, I present DNASCAN, DNA Sequence Contamination Analyzer, a scalable and efficient algorithm designed for the high-resolution detection of biological contaminants within source DNA sequences. DNASCAN leverages N-gram analysis, supervised random forest classification, and Bayesian change-point detection to provide precise breakpoints and highly accurate impurity estimates. In summary, DNASCAN yielded 100% purity estimates and breakpoint detection accuracy rates at bacterial contamination levels above 0.1%. Using long-read sequencing data, DNASCAN detected all impurity levels above 0.025%, making it ideal for the removal and reconstitution of contaminated DNA sequences. The software, documentation, and vignettes with detailed code demonstrations are available at https://github.com/jmal0403/DNASCAN/wiki.

Introduction

I recently introduced DNAnamer, a novel and generalized stochastic modeling and N-gram frequency analysis framework designed for the supervised classification and identification of cross-species DNA sequences. Because DNAnamer was able to accurately (>99%) and reliably identify species-specific DNA segments as small as 40 kb in length, I hypothesized that this framework would be ideal for developing an efficient and sensitive analysis toolkit for detecting biological DNA contaminants [1]. Thus, DNASCAN was built using DNAnamer but with many methodological advances such as Bayesian change-point detection, impurity estimates, and breakpoint detection. The

sources and downstream effects of biological DNA contamination are widespread, ubiquitous, and well-documented, presenting a host of challenges in the analysis of human and non-human DNA. Bacterial contaminants have been noted to significantly affect a variety of human DNA studies [2–4], resulting in the miscalling of population variants and genotypes, false estimations of genetic variability, and the detection of thousands of spurious protein structures [5–7], which makes genetic research significantly more costly and time-consuming. Despite these acknowledged concerns, there remains a lack of original methodologies for identifying biological DNA contaminants.

First, it is critical to consider the sources and scope of biological DNA contamination to recognize the impact it has on critical research in molecular genetics, genomics, bioinformatics, and gene-based therapeutics. Potential sources include the presence of bacterial DNA contamination in ultrapure biology-grade water [8, 9], PCR reagents [10–12], and DNA extraction kits [13–16]. Common bacterial sources of contamination in human DNA include, but are not limited to, *Streptococcus, Bradyrhizobium, Microbacterium, Staphylococcus, Acinetobacter,* and *Pseudomonas* [17–19]. These biological artifacts are ubiquitous and persistent, causing genotype heterogeneity and propagating errors throughout the many steps required to sequence and analyze human DNA. Of course, biological contaminants and their effects are not limited to handling human DNA.

Caenorhabditis elegans is a one-millimeter, transparent nematode commonly used in genetics research. Its transparent body and short life cycle make it an ideal model organism for a wide range of important research domains, including developmental biology, memory, aging, and neurodegeneration [20–23]. In the laboratory environment, *C. elegans* is cultured on a diet of *Escherichia coli* [24]. As a result, *E. coli* contamination is highly prevalent in DNA experiments that use this inexpensive and indispensable model organism. Thus, *E. coli* contamination poses widespread and significant limitations to analyzing and interpreting the results of these genetic experiments, with one experiment finding a 4 kb insertion of *E. coli* DNA in the *C. elegans* reference genome [25–27].

Another burgeoning field of research plagued by sequence contamination is the study of the human microbiome. Microbiome-based research has revolutionized our understanding of the gut-brain axis, cancer development and progression, and the modulation of the immune system, offering many new and promising therapeutics [28–30]. However, much concern remains due to the frequency and variety of microbial contaminants, highlighting the need for improvements in the standardization of quality control processes and novel methodologies to accurately detect and remove sequence contaminants [31–34]. As a result, several complex and expensive technologies, such as metagenomic sequencing, 16S rRNA gene sequencing, and other PCR-based methods, have been employed to better identify and describe natural sequence contaminants [35–37]. They are assays that can only identify contaminants but cannot remove them. One goal of this work is to provide a more cost-effective and accurate solution that requires only genotype information. Another advantage to this approach is the potential to remove these sequencing artifacts and improve the quality and interpretability of downstream genomic analyses.

To this end, I present DNASCAN, DNA Sequence Contamination Analyzer, an efficient and scalable toolkit for accurately detecting bacterial and biological DNA sequences within human and model organism DNA. Because DNASCAN is highly

sensitive, it provides precise breakpoints for the removal and reconstitution of affected DNA sequences. DNASCAN is tested and available as an R package with code examples and user tutorials.

Materials and methods

Data source and description

Two host organisms, *C. elegans* and *Homo sapiens*, and one bacterial species, *E. coli*, were selected to provide specific use cases, a proof of concept, and bootstrapped performance metrics. The three reference genomes were downloaded from the National Center for Biotechnology Information's genome resource database using the 'biomartr' package [38]. The reference genomes totaled approximately 3300 Mb (human), 100 Mb (*C. elegans*), and 186 kb (*E. coli*). To demonstrate DNASCAN on individually sequenced organisms, 10 high-quality, long-read *C. elegans* whole-genome assemblies were downloaded from the Sequence Read Archive. Each genome contained approximately 400 Mb of DNA sequence. Pacific Biosciences' RSII platform was used to construct high-quality reads at 80 × coverage, averaging approximately 760 kb in length [39].

Experimental design

A graphical illustration of the DNASCAN methodology and the in silico experimental design is provided in Fig. 1, where the left and middle panels describe the three primary analytical components: N-gram frequency analysis, supervised random forest (RF) classification, and Bayesian change-point detection. This study design consists of two main experiments. The first experiment was a proof of concept and was performed by randomly selecting and spiking in *E. coli* DNA segments at four impurity percentages [0.5, 1, 2.5, and 5%] into the entire reference genome of *C. elegans*. These spiked-in fragment sizes of *E. coli* DNA ranged from 100 to 500 kb. Next, fifth-order N-gram frequency tables were calculated for each segment with the contaminated sequence inserted. The

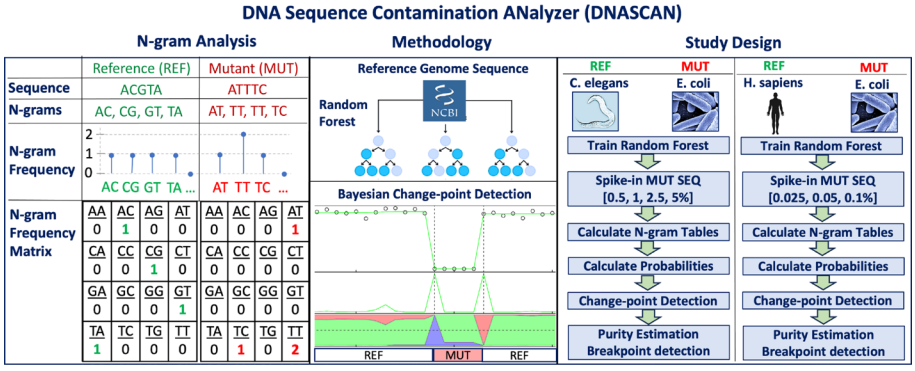

Fig. 1 A graphical abstract of the methodology and experimental design. An illustration of the DNASCAN (DNA Sequence Contamination Analyzer) methodology is provided, where two simple examples of second-order DNA N-gram sequences and frequency tables are given (left panel). REF denotes the sequence of the targeted reference species, whereas MUT denotes the contaminated sequence. The NCBI reference genome for each species was divided into equal-length training and test sets. The training set was used to train a random forest model and then calculate the probability of unique, contaminated sequence segments spiked in across the entire reference genome. Bayesian change-point analysis with a variable sliding window provides high detection accuracy, precise breakpoints, and impurity estimations (middle panel). As proof of concept, two experiments were performed to validate this methodology. First, E. coli sequences were spiked into the reference organism, *C. elegans*, at impurity levels of 0.5, 1, 2, and 5%. Second, E. coli sequences were spiked into 50Mb chunks of the human genome (right panel). 200 bootstraps were performed with E. coli segments ranging from 2.5 to 50 kb in length

RF model was then used to predict the probability distribution of each N-gram frequency matrix. The probability is the cumulative proportion of votes in the forest, or ensemble [40]. Bayesian change-point detection was employed to estimate impurity rates and determine precise breakpoint locations. To gauge the accuracy and resolution of this methodology at scale, a bootstrap analysis was by randomly selecting segments of the human reference genome and spiking in *E. coli* segments at impurity rates of 0.05, 0.1, 0.5, and 1%. Fifth-order N-gram analysis was used for all experiments, including the *C. elegans*' reference genome and ten long-read samples. All analyses were performed using the R statistical language version 4.4.0 [41].

Model training and tuning

DNA sequences were randomly sampled across the entire reference genome of each species to create evenly balanced training sets, each consisting of continuous sequence segments. Specifically, all training and test segments were equal in length, did not overlap, and ranged from 2.5 to 50 kb in length. N-gram frequency distribution priors were calculated using the RF method on only the training data. N-gram frequencies were neither filtered nor normalized. All 1360 5th-order N-grams were used for training and testing. The validation, or test, sets consisted of continuous DNA segments and were used to classify segments using N-gram frequency probability tables. A sensitivity analysis was conducted to optimize this model by varying the number of trees from 100 to 5000 while holding all other parameters constant. In summary, 2000 trees provided the most stable performance without drastically increasing the processing time. The number of permutations represents the number of out-of-bag (OOB) recursions performed per tree, which was used to assess variable importance and perform bootstrapping to estimate in-sample accuracy and error. The mtry parameter specifies the maximum number of input features or variables available to a decision tree. 222 input features per split were optimal. An mtry value of 222 was determined using R's 'caret' package. 'Caret' provides a bootstrapping procedure that gives the optimal mtry value [42]. Variable importance was computed using OOB sample bootstrapping to determine the mean decrease accuracy of each independent variable, or sequence N-gram, which is based on how much the accuracy decreased when the variable was excluded throughout the bootstrapping process.

Classification using the random forest methodology

The random forest methodology [43] is a highly versatile and effective regression and classification model for a variety of biomedical and genomic applications [44–47] which has been successfully employed to perform DNA barcoding [48–50]. The RF model was used to calculate N-gram frequency priors using only training sequence sets. For prediction, the RF model was leveraged to provide probability estimates for all N-gram frequency tables. All experiments were performed under the same parameters, or initial conditions. Namely, a maximum of 2000 decision trees was specified with 100 permutations and an mtry of 222. The number of decision trees represents the number of decision nodes produced for each model, which presents a trade-off between prediction accuracy and computational efficiency [51].

In silico spike-in analysis

For proof of concept, five random locations were selected along the *C. elegans* and human reference genomes. Five adjoining segments of *E. coli* DNA were inserted at the center of each randomly selected spike-in location, with segment lengths ranging from 100 to 500 kb. For example, at 1% impurity, 100 kb of *E. coli.* sequence was inserted into 100 Mb of *C. elegans* sequence. For the second experiment, the human genome was divided into 50 Mb segments. Spike-in segments ranged from 2.5 to 50 kb at impurity levels of 0.1, 0.05, and 0.025%.

Bayesian change-point analysis

Bayesian change-point detection and time-series decomposition were performed on the probability estimates given by the RF model using 'Rbeast,' or Bayesian Estimator of Abrupt change, Seasonality, and Trend [52]. 'Rbeast' is based on Bayes' theorem (Eq. 1) and is fast and extensible, offering accurate posterior probabilities using the Bayesian model averaging (BMA) equation (Eq. 2). Here, $P(\theta \mid y)$ is the posterior distribution for the parameter θ given the observed data y, where θ represents the probability of contamination for a given DNA sequence segment. $\mathbf{P}(\theta \mid M_j, y)$ is the posterior distribution for the parameter θ given the model M_j and the observed data, y. $P(M_j \mid y)$ is the posterior probability of model M_j given the observed data *y*, which represents the weight assigned to each model. BMA doesn't assume a best model but rather accounts for uncertainty by averaging all models.

$$Bayes' \, Theorum = P\left(A|B\right) = \frac{P\left(B|A\right) * P\left(A\right)}{P\left(B\right)} \tag{1}$$

$$BMA = P\left(\theta|y\right) = \sum_{j=1}^{J} P(\theta|M_j, y) * p(M_j|y) \tag{2}$$

Impurity estimation

Sequence impurity was estimated using Eq. 3 to determine the number of impure sequence segments, where *No. Impure Segments* was equal to the sum of segments (θ) with a posterior probability greater than 0.5 ($P\left(\theta|y\right) > 0.5$). Finally, the *Percent Impurity* (Eq. 4) was equal to the *No. Impure Segments* multiplied by the segment length of the MUT sequence divided by the total REF sequence and multiplied by 100. For example, given a total of 1 Mb of reference sequence length and a contaminant sequence segment size of 10 kb, an additional 1% impurity is added for each additional impure segment.

$$No. \, Impure \, Segments = P\left(\theta|y\right) > 0.5 = \sum_{j=1}^{J} \left\{ \begin{array}{ll} 1, & \theta = \theta + \theta_j \\ 0, & 0 \end{array} \right\} \tag{3}$$

$$Percent \, Impurity = \frac{No. \, Impure \, Segments * Segment \, Length_{MUT}}{Total \, Sequence_{REF}} * 100 \tag{4}$$

Breakpoint detection and accuracy

Left and right breakpoints were provided for all bacterial DNA sequence fragments identified, or those with a posterior probability below 0.5. Breakpoint detection accuracy

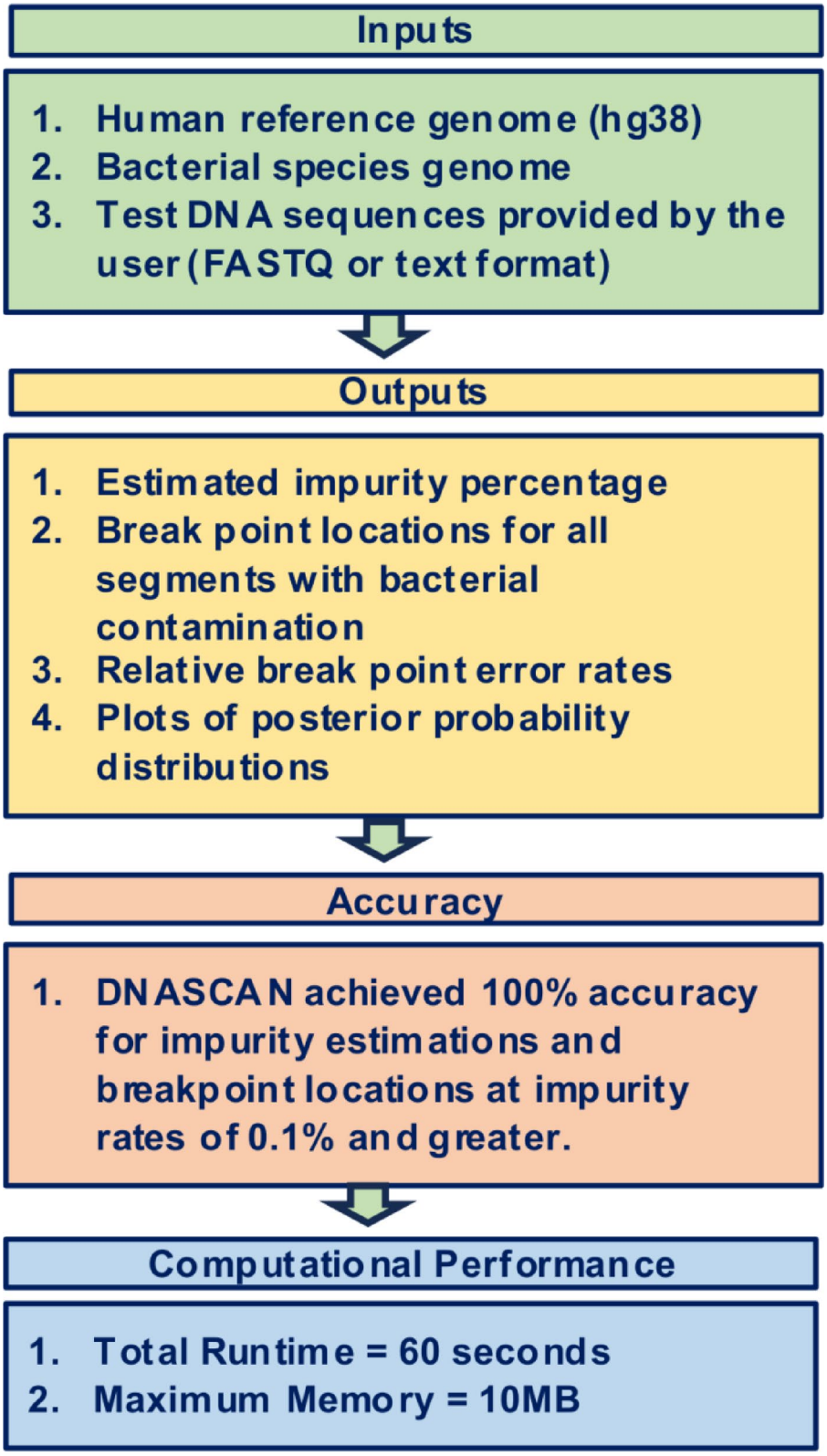

Fig. 2 (See legend on next page.)

(See figure on previous page.)

Fig. 2 Software workflow diagram of the DNASCAN methodology. To summarize DNASCAN's required inputs, outputs, results, and performance, a workflow diagram is provided, with four high-level components: inputs, outputs, accuracy, and computational performance. DNASCAN requires a reference genome or WGS sequence. It can be used to detect contaminants in any well-genotyped data. These genome assemblies were then used to construct N-gram frequency tables and train the random forest model. The user provides DNA sequences as input in FASTQ or text format. DNASCAN provides the estimated bacterial impurity percentage, breakpoint locations, and posterior probability plots. In summary, DNASCAN detected 100% of spiked-in bacterial DNA sequences at impurity rates of 0.5%. Finally, DNASCAN is relatively fast and efficient, with a total runtime of approximately one minute

was defined as the absolute value of the location of the mutant sequence subtracted from the location of the target reference sequence for both left and right breakpoints. Thus, accuracy represents the absolute distance of the overlap of both the left and right boundaries, where 100% overlap is equal to 100% accuracy.

Benchmark analysis

Finally, computational performance benchmarking was performed on the human reference genome. Total runtimes (seconds) and the total memory allocated (GB) were recorded on a 2.4 GHz 8-Core Intel Core i9 processor with 64GB of available memory. Performance benchmarks were provided for all necessary functions and the five DNA segment lengths. 100 Mb of DNA sequences were processed for each of the benchmark experiments. Fifth-order N-gram analysis was performed to provide a realistic estimation of the computational resources required to replicate these experiments.

Results

DNASCAN

The DNA Sequence Contamination Analyzer (DNASCAN) is a novel N-gram frequency analysis framework for the supervised classification of cross-contaminated DNA sequences and is available as an R software package. Documentation and vignettes with detailed code demonstrations are available at https://github.com/jmal0403/DNASCAN/wiki. All experiments performed herein can be reproduced. In summary, DNASCAN provides an extensible methodology and efficient analysis toolkit to accurately detect contaminants, estimate impurity rates, and locate precise breakpoints. Figure 2 provides a high-level summary of DNASCAN's required inputs, outputs, results, and computational benchmarks.

Proof of concept using the *C. elegans* reference genome

Figure 3 provides a summary of the results for each in silico spike-in experiment, with the experimental parameters, randomly selected spike-in locations, and impurity estimations for each test provided in the left panel. The right panel provides the trend plots [23] of the probability of detecting a sequence from the *C. elegans* genome, provided by the random forest model. The posterior probability of detecting a change-point in the N-gram frequencies, or Pr(tcp), is provided along the bottom axis in light green. Spike-in segments ranged from 100 to 500 kb in length at impurity levels of 0.5, 1, 2.5, and 5%. In summary, DNASCAN was able to identify all twenty *E. coli* sequences that were spiked into the *C. elegans* genome.

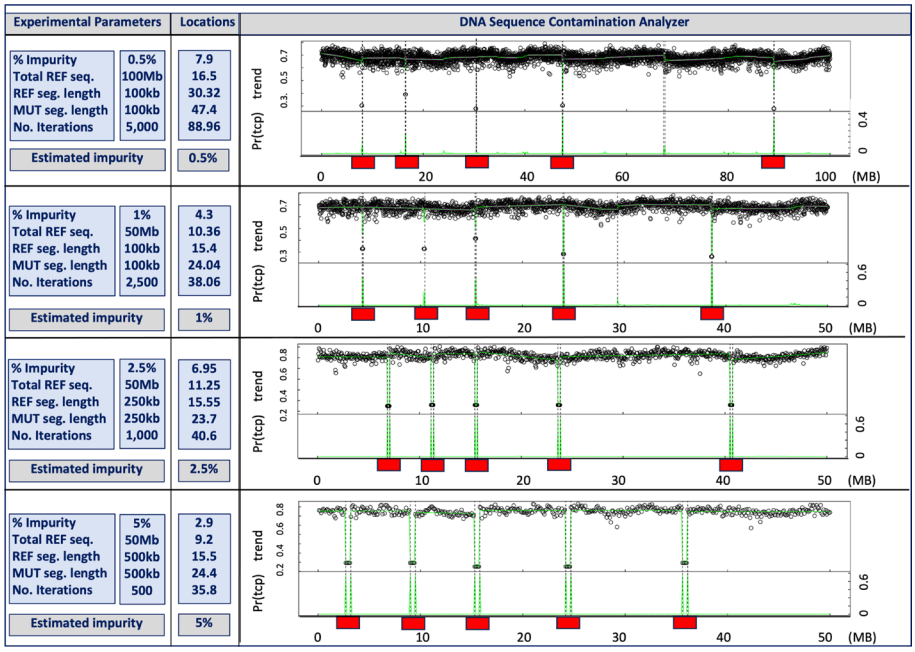

Fig. 3 DNASCAN for *E. coli* contamination in the *C. elegans* reference genome. Four impurity levels (horizontal panels) were tested, with in silico spike-in segments inserted in the C. elegans reference genome, ranging from 100 to 500 kb in length at impurity levels of 0.5, 1, 2.5, and 5%. 200 bootstraps were performed for all impurity levels. Five spike-in locations (shown as red boxes) were randomly chosen and inserted into the C. elegans genome. Experimental parameters, randomly selected spike-in locations (x-axis), and impurity estimations for each test are provided in the left panel. The right panel provides trend plots of the probability of detecting (y-axis) a sequence from the *C. elegans* genome, based on a priori N-gram frequencies provided by the random forest model. The posterior probabilities of detecting a change-point in the N-gram frequencies, or Pr(tcp), are provided along the x-axis and are colored in light green. A posterior probability of less than 0.5 was used to classify a contaminated sequence

Impurity and breakpoint accuracy in the human genome

Figure 4 provides a summary of the results for each in silico spike-in experiment, with the experimental parameters, impurity estimations, and breakpoint detection error metrics. The mean impurity estimations and breakpoint detection accuracy metrics are provided (left panel). 200 bootstraps, which were run for each impurity rate, were performed by randomly selecting and spiking impurities into each 50 Mb reference sequence, totaling 10 billion base pairs of human DNA sequence analyzed. In summary, all contamination levels greater than 0.1% yielded an estimated impurity and breakpoint detection accuracy of 100%. Performance for 0.05% targeted impurity varied. Mean impurity estimates ranged from 2 to 11%, with a mean estimate of 7.83% (SD = 0.32%). Notably, all 200 estimates showed at least 2% contamination. The mean absolute breakpoint detection error was excellent at 0.29%. Finally, at a targeted impurity of 0.025%, impurity estimates ranged from 0.6 to 11%, with a mean of 6.96% (SD = 0.0375). Again, all 200 runs produced an impurity estimate of at least 0.6%. Breakpoint detection error rates were significantly higher.

Impurity and breakpoint accuracy in *C. elegans* genome

Figure 5 provides a summary of the results for each in silico spike-in experiment, with the experimental parameters, impurity estimations, and breakpoint detection error metrics. All performance metrics reported are the average of the ten WGS samples. All prior experiments were repeated for ten long-read *C. elegans* whole-genome de novo

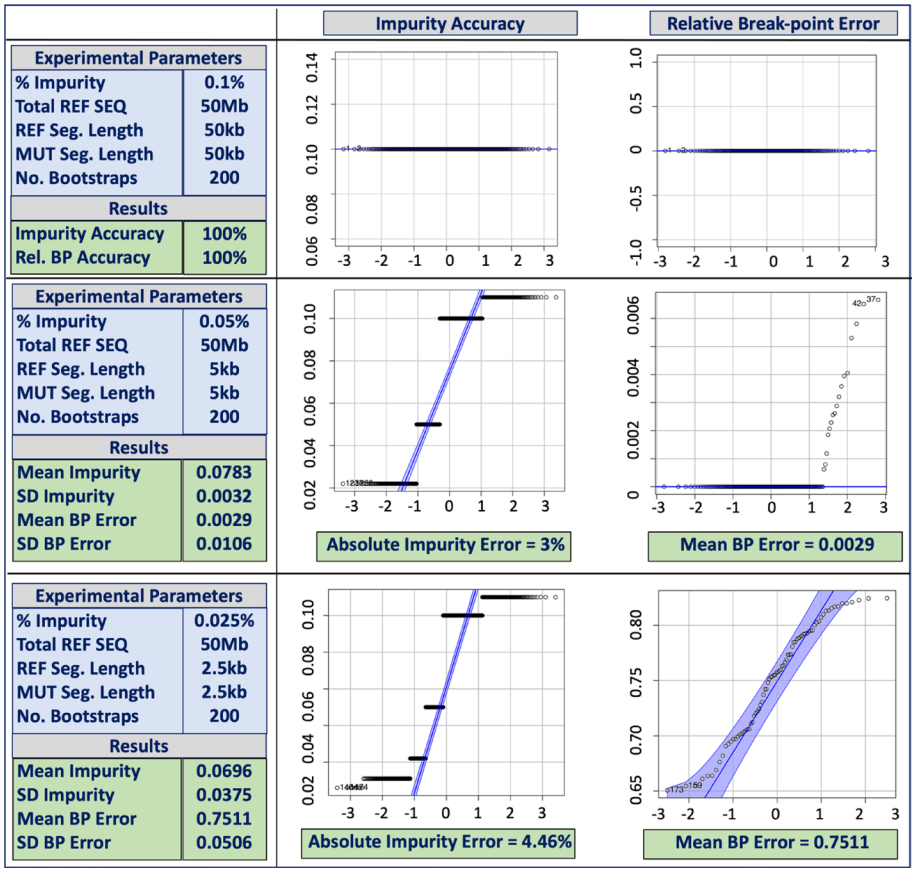

Fig. 4 DNASCAN for *E. coli* contamination in the human reference genome. Three impurity levels (horizontal panels) are shown, with in silico spike-in segments ranging from 2.5 to 50 kb in length at impurity levels of 0.1, 0.05, and 0.025%. 200 bootstrap simulations per experiment were executed, and five spike-in locations were randomly selected and inserted into the human reference genome. Means and standard deviations for the impurity estimation and breakpoint accuracy were provided. The right panel shows quantile–quantile plots for impurity and breakpoint estimations. A posterior probability of less than 0.5 was used to classify a contaminated sequence

assemblies. The mean impurity estimations and breakpoint detection accuracy metrics are provided (left panel). 200 bootstraps across ten samples, which were run for each impurity rate, were performed by randomly selecting and spiking impurities into each 50 Mb reference sequence, totaling 10 billion base pairs of DNA sequence analyzed. In summary, all contamination levels greater than 0.025% yielded an estimated impurity and breakpoint detection accuracy of 100%. Performance for a targeted impurity of 0.025% varied. Mean impurity estimates ranged from 0 to 0.005%, with a mean estimate of 0.005% (SD = 0.0001%). The mean absolute breakpoint detection error was excellent at 8%. The absolute impurity error rate was 2.45%.

Computational performance benchmarking

Table 1 summarizes the computational performance benchmarks for fifth-order N-gram analysis. Fifth-order N-grams were used to provide an approximation of this model's computational performance using a total of 100 Mb of DNA sequence for each benchmark. Three functions were required to perform supervised classification. They were *getSequence()*, *getFreqDF()*, and *runDNASCAN()*. The *getSequence()* function randomly selected and constructed the contiguous DNA segments. On average, this function

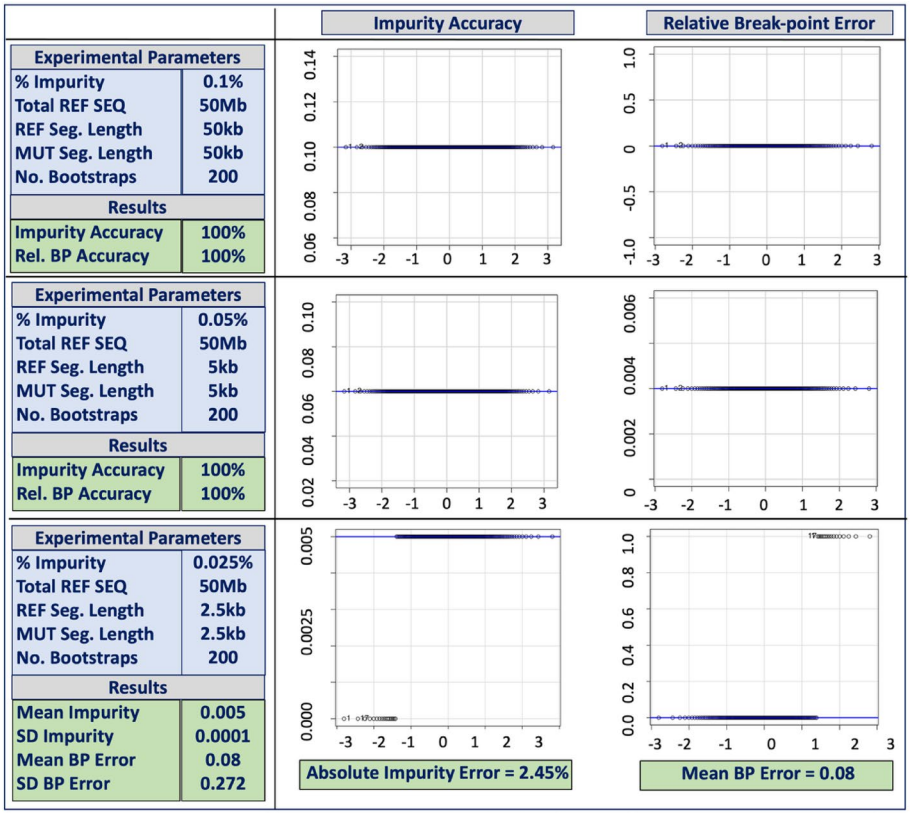

Fig. 5 DNASCAN for *E. coli* contamination in *C. elegans* whole genome. All experiments were repeated using 10 high-quality, long-read *C. elegans* whole-genome assemblies downloaded from the Sequence Read Archive. All performance metrics reported are the average of the ten samples. Specifically, three impurity levels (horizontal panels) are shown, with in silico spike-in segments ranging from 2.5 to 50 kb in length at impurity levels of 0.1, 0.05, and 0.025%. 200 bootstrap simulations per experiment were executed, and five spike-in locations were randomly selected and inserted into these genomes. Means and standard deviations for the impurity estimation and breakpoint accuracy were provided. The right panel provides quantile–quantile plots for impurity and breakpoint estimations

Table 1 Computational performance benchmarks for fifth-order N-gram analysis

Function	Segment length (kB)	Total runtime (seconds)	Core seconds	Maximum memory (GB)
getSequence	10	19.2	2.40	4.75
getSequence	20	18.7	2.34	4.77
getSequence	40	18.8	2.35	4.74
getSequence	80	18.5	2.31	4.77
getSequence	100	18.9	2.36	4.76
getFreqDF	10	69.6	8.70	32.3
getFreqDF	20	45.4	5.68	16.7
getFreqDF	40	34.4	4.30	12.7
getFreqDF	80	32.6	4.08	11.7
getFreqDF	100	30.8	3.85	11.5
runDNASCAN	10	0.1	5.53	0.01
runDNASCAN	20	0.08	0.01	0.01
runDNASCAN	40	0.08	0.01	0.01
runDNASCAN	80	0.04	0.005	0.01
runDNASCAN	100	0.01	0.0013	0.01

*Core Seconds are equal to the Total Runtime divided by the number of processing cores (n = 8)

required 19 s to process 20 MB of DNA using 4.7GB of memory. The *getFreqDF()* function calculated the N-gram frequency tables and required between 30 and 50 s to complete while using 12 to 32 GB of memory. This function's runtime and memory increased with the number of tests, as expected with a larger matrix. Finally, the *runDNASCAN()* function calculated impurity estimations and required less than 1 s to process using less than 10 MB of memory. The average combined processing time for a fifth-order binary classification experiment was only 60 s. Core Seconds are equal to the Total Runtime divided by the number of processing cores (n = 8).

Discussion

DNASCAN provides an efficient, accurate, and extensible methodology and statistical analysis toolkit designed to tackle a broad range of complex and computationally expensive regression and classification problems applicable to biomedical sciences, genomics, and genetics. This methodology achieves high levels of classification accuracy using only genotype information as the input. This innovative toolkit for stochastic modeling and machine learning has been validated and documented and is available as a complete and well-tested R package.

To establish theoretical feasibility, a number of fundamental assumptions and hypotheses have been addressed. Specifically, it has been demonstrated that relatively small training sets of bacterial DNA sequences can be efficiently leveraged to optimize the detection of biological sequence contaminants across a wide range of organisms of varying taxonomic distances, offering flexibility and extensibility to this approach. It is important to note that the *E. coli* reference genome totals 186 kb in length. Because N-gram analysis offers a rich variable space by providing a numerical representation of natural sequence structures and complex syntax patterns encoded in DNA, it is ideal for the identification of cross-species DNA artifacts. Further, Bayesian change-point detection complements this probabilistic classification methodology and offers precise and reliable breakpoint detection and accuracy at relatively low contamination rates. Throughout the next several paragraphs, I will further emphasize the three most important functional design aspects of DNASCAN, with a particular emphasis on continued application and the improvement of DNA sequence quality and its impact on science. To provide additional feasibility and a practical example using WGS data, *E. coli* sequences were spiked into ten WGS C. elegans assemblies. DNASCAN performed slightly better using WGS as compared to the reference genome.

The primary design consideration of DNASCAN was extensibility to the types of species it can analyze, making the scope of its use much broader. Using the DNAnamer algorithm, five mammalian species were efficiently barcoded and accurately classified using (> 99%) 100 kb DNA fragments. In this experiment, the increased taxonomic distance between *C. elegans* and *E. coli* appears to have increased statistical discrimination, made evident by the small amount of training data required to detect contaminants. Second, DNASCAN was designed for flexibility in terms of the data it can analyze. DNASCAN can be employed to detect bacterial sequence contamination in paired or unpaired and unaligned sequencing reads, such as in FASTQ files. Given that DNASCAN detected all contaminants in WGS data using sequence lengths of 5 kb and greater, it is recommended that the user specify segment lengths of 5 kb or greater. This allows for the filtering of sequence reads before sequence alignment, since misalignments are very

common and thus making this tool useful. Finally, DNASCAN works seamlessly with existing variant callers and genotype quality control. DNASCAN's ability to work in parallel with existing genotype QC software tools also provides a consensus quality control approach, proven to be ideal for genotype quality control [53–56]. This extends the scope of DNASCAN's biological applications by allowing users to easily train models and classify DNA sequences for any species with sufficient genotype information.

The next major design consideration of DNASCAN was classification accuracy. In summary, for the proof-of-concept experiment, DNASCAN correctly identified all twenty variants spiked into the C. elegans reference genome at impurity levels of 0.5, 1, 2.5, and 5%. This provided a 100% detection accuracy across a total of 100 tests. When examining the human reference genome, all impurity levels ranging from 0.1 to 1% yielded 100% content and breakpoint detection accuracy. 200 bootstrap simulations were performed for each impurity level to determine the estimated impurity and breakpoint accuracy metrics. At impurity rates of 0.05 and 0.025%, all 400 sequences with mutant sequences spiked in were correctly identified, despite the fact that impurity estimates and breakpoint predictions varied. For example, at 0.05% impurity, the mean breakpoint error was only 0.29%. However, this increased to 75% at 0.025% impurity. Because human genomes are broken down into many thousands of segments so they can be efficiently processed, DNASCAN's segment analysis window can be tailored to the specific needs of the user. For example, if a user needs to analyze 500 Mb of total source DNA sequence, then, given an assumed impurity rate of 0.1%, a segment window of 50 kb should provide 100% detection accuracy for a given bacterial species.

It is also important to contrast DNASCAN against similar software tools and methodologies that are currently available. FASTQC is a popular software toolkit for detecting a number of genotype quality checks, including read duplication, adapter contamination, and over-represented sequences, which are all forms of potential DNA sequence contamination [57]. Popular DNA sequence quality control tools such as Falco [58] and 'fastp' [59] are fast and reliable; however, their algorithms rely on indirect methods to identify potential sequence contamination, and they do not provide precise breakpoints. Namely, the FASTQC methodology aggregates summary statistics related to GC content and static repeat structures to formulate its heuristics. Whereas DNASCAN leverages high-dimension priors, Bayesian change-point detection, and machine learning to maximize detection accuracy. All of the aforementioned software tools provide impurity estimates. Another distinct advantage of DNASCAN over traditional methods is its ability to provide precise breakpoints. DeconSeq is a software tool that can detect and remove contaminated DNA sequences by querying each sequence read against many large genomics and metagenomics databases [60]. Although sometimes useful, DeconSeq methodology is not computationally efficient, as it relies on recursive queries of large databases. Whereas DNASCAN was designed to use minimal information entropy. When dealing with vast amounts of information such as whole genome assemblies, it is preferred that the minimum information be used to achieve a satisfactory and scalable result, as has been demonstrated by information theorists and mathematicians.

Finally, application and integration were the practical motivations for this work. As previously discussed, unavoidable biological sequence artifacts have a significant impact on genomics and genetics research which impedes the development of safe and effective biologics and therapeutics. I have highlighted two such domains of research that are

significantly curtailed by sequence contamination. In the analysis of the human microbiome, sequence contamination has led to skewed results, flawed estimations of microbial community compositions, and contradictory findings. To date, the Food and Drug Administration has approved two microbiome-based therapies, Rebyota and Vowst, federally approved fecal microbiota treatments that are used to prevent the recurrence of the *Clostridioides difficile* infection. Microbiota treatments aim to restore the balance of the gut microbiome, improving patient survival in cases where antibiotics have not been successful in stopping infection. Further, as many bacterial species become resistant to traditional antibiotics, new approaches to restoring immune health will be needed to prevent excess mortality, especially in older populations. Despite the potential for the propagation of experimental errors, removing contaminated sequences remains a value-based judgement call.

Finally, DNA contamination continues to impede some of the most promising and novel gene therapies, recombinant adeno-associated virus (rAAV) preparations. While these have been successful as delivery vectors for many gene therapies, the contamination poses significant challenges to advancing the application of gene therapies, particularly regarding patient safety and efficacy. To date, the FDA has approved at least seven unique rAAV gene therapies designed to treat a range of medical conditions from retinal dystrophy to hemophilia. In 2024, the FDA approved Kebilidi, a gene therapy for the treatment of Aromatic L-amino Acid Decarboxylase deficiency. In summary, DNA contamination is unavoidable and must be given greater priority to facilitate the adoption of novel and lifesaving gene therapies [61–63]. It is also important to remember that these technologies for identifying and removing DNA contaminants are still advancing and will take considerable time and effort to improve to a higher standard. DNASCAN is only one step to doing so. Nonetheless, these concerns continue to hamper the adoption of this promising and potentially lifesaving treatment [61–63].

Like all models, this one has both strengths and limitations. A major limitation is the lack of WGS with a known, or contrived, level of bacterial sequence contamination. Also, this experiment did not evaluate DNASCAN's performance using multiple sources of bacterial contamination. Another limitation of this toolkit is that it used 32GB of memory to calculate the largest N-gram frequency table. While not ideal, it can easily run on a modern laptop computer. The computational performance of DNASCAN can be improved by utilizing parallel processing and data reduction techniques. Furthermore, this algorithm's detection limit can be improved by increasing the size of the training data sets and exploring more advanced model optimization and parameter tuning options. However, these considerations are not necessary for excellent detection accuracy. Among this model's main strengths are efficiency, accuracy, and scalability. In an era of 'big data,' it is critical to remember that more data does not necessarily imply more meaning. Rather, we must continue to drill down into the marvelous complexity of DNA by deconstructing its highly organized syntax structures so that we may sift through its many layers of information to increase our unders.

Abbreviations

AAV	Adeno-associated virus
ANOVA	Analysis of variance
AUC	Area under the curve
BMA	Bayesian model averaging
C. elegans	*Caenorhabditis elegans*
DNA	Deoxyribonucleic acid

DNASCAN	DNA Sequence Contamination Analyzer
E. coli	*Escherichia coli*
FASTQ	FASTA quality-based sequencing format
hg38	Human reference genome version 38
human	Homo sapiens
Gb	Gigabyte
kb	Kilobase
Mb	Megabase
MB	Megabyte
MUT	Bacterial sequence mutant
NCBI	National Center for Biotechnology Information
OOB	Out-of-bag
PCR	Polymerase chain reaction
REF	Reference genome
RF	Random forest
SD	Standard deviation

Acknowledgements
The author has no acknowledgments.

Author contributions
John Stephen Malamon designed, drafted, and wrote this manuscript and performed the analyses contained herein.

Funding
No funding was used.

Availability of data and materials
All data used to conduct this study can easily be downloaded from the National Center for Biotechnology Information's genome resource database. The software, documentation, and vignettes with detailed code demonstrations are available at https://github.com/jmal0403/DNASCAN/wiki and knowledge of the mysterious and beautiful language of DNA.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Links to NCBI Reference Genomes
Human (hg38): https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/. E. coli (ASM584v2): https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000005845.2/. C. elegans (WBcel235): https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002985.6/. Links to NCBI Whole Genomes: C. elegans (CB4856): https://trace.ncbi.nlm.nih.gov/Traces/index.html?view=study&acc=SRP186435.

Received: 23 October 2025 / Accepted: 24 December 2025
Published online: 03 March 2026

References

1. Malamon JS. DNA N-gram analysis framework (DNAnamer): a generalized N-gram frequency analysis framework for the supervised classification of DNA sequences. Heliyon. 2024;10(17):e36914. https://doi.org/10.1016/j.heliyon.2024.e36914.
2. Korlevic P, Gerber T, Gansauge MT, Hajdinjak M, Nagel S, Aximu-Petri A, et al. Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. Biotechniques. 2015;59(2):87–93. https://doi.org/10.2144/000114320.
3. Lee JE, Hong EJ, Shim SM, Kim JW, Bae GR, Cho YS, et al. Bacterial contamination of blood DNA samples is associated with donor's health condition. Biopreserv Biobank. 2010;8(3):127–31. https://doi.org/10.1089/bio.2010.0012.
4. Yoon CJ, Kim SY, Nam CH, Lee J, Park JW, Mun J, et al. Estimation of intrafamilial DNA contamination in family trio genome sequencing using deviation from Mendelian inheritance. Genome Res. 2022;32(11–12):2134–44. https://doi.org/10.1101/gr.276794.122.
5. Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. Human contamination in bacterial genomes has created thousands of spurious proteins. Genome Res. 2019;29(6):954–60. https://doi.org/10.1101/gr.245373.118.
6. Goig GA, Blanco S, Garcia-Basteiro AL, Comas I. Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. BMC Biol. 2020;18(1):24. https://doi.org/10.1186/s12915-020-0748-z.
7. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet. 2012;91(5):839–48. https://doi.org/10.1016/j.ajhg.2012.09.004.
8. Nogami T, Ohto T, Kawaguchi O, Zaitsu Y, Sasaki S. Estimation of bacterial contamination in ultrapure water: application of the anti-DNA antibody. Anal Chem. 1998;70(24):5296–301. https://doi.org/10.1021/ac9805854. (PubMed PMID: 9868920).

9. Shen H, Rogelj S, Kieft TL. Sensitive, real-time PCR detects low-levels of contamination by *Legionella pneumophila* in commercial reagents. Mol Cell Probes. 2006;20(3–4):147–53. https://doi.org/10.1016/j.mcp.2005.09.007.
10. Maiwald M, Ditton HJ, Sonntag HG, von Knebel Doeberitz M. Characterization of contaminating DNA in Taq polymerase which occurs during amplification with a primer set for *Legionella* 5S ribosomal RNA. Mol Cell Probes. 1994;8(1):11–4. https://doi.org/10.1006/mcpr.1994.1002.
11. Newsome T, Li BJ, Zou N, Lo SC. Presence of bacterial phage-like DNA sequences in commercial Taq DNA polymerase reagents. J Clin Microbiol. 2004;42(5):2264–7. https://doi.org/10.1128/JCM.42.5.2264-2267.2004.
12. Rand KH, Houck H. Taq polymerase contains bacterial DNA of unknown origin. Mol Cell Probes. 1990;4(6):445–50. https://doi.org/10.1016/0890-8508(90)90003-i.
13. Charif D, Lobry JR. SeqinR 1.0–2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural approaches to sequence evolution: molecules, networks, populations. Berlin: Springer; 2007. p. 207–32.
14. Jue E, Witters D, Ismagilov RF. Two-phase wash to solve the ubiquitous contaminant-carryover problem in commercial nucleic-acid extraction kits. Sci Rep. 2020;10(1):1940. https://doi.org/10.1038/s41598-020-58586-3.
15. Mohammadi T, Reesink HW, Vandenbroucke-Grauls CM, Savelkoul PH. Removal of contaminating DNA from commercial nucleic acid extraction kit reagents. J Microbiol Methods. 2005;61(2):285–8. https://doi.org/10.1016/j.mimet.2004.11.018.
16. Toole K, Roffey P, Young E, Cho K, Shaw T, Smith M, et al. Evaluation of commercial forensic DNA extraction kits for decontamination and extraction of DNA from biological samples contaminated with radionuclides. For Sci Int. 2019;302:109867. https://doi.org/10.1016/j.forsciint.2019.06.025.
17. Chrisman B, He C, Jung JY, Stockham N, Paskov K, Washington P, et al. The human "contaminome": bacterial, viral, and computational contamination in whole genome sequences from 1000 families. Sci Rep. 2022;12(1):9863. https://doi.org/10.1038/s41598-022-13269-z.
18. Laurence M, Hatzis C, Brash DE. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS ONE. 2014;9(5):e97876. https://doi.org/10.1371/journal.pone.0097876.
19. Peters RP, Mohammadi T, Vandenbroucke-Grauls CM, Danner SA, van Agtmael MA, Savelkoul PH. Detection of bacterial DNA in blood samples from febrile patients: underestimated infection or emerging contamination? FEMS Immunol Med Microbiol. 2004;42(2):249–53. https://doi.org/10.1016/j.femsim.2004.05.009.
20. Chen F, Mackerell AD Jr., Luo Y, Shapiro P. Using *Caenorhabditis elegans* as a model organism for evaluating extracellular signal-regulated kinase docking domain inhibitors. J Cell Commun Signal. 2008;2(3–4):81–92. https://doi.org/10.1007/s12079-008-0034-2.
21. Marsh EK, May RC. *Caenorhabditis elegans*, a model organism for investigating immunity. Appl Environ Microbiol. 2012;78(7):2075–81. https://doi.org/10.1128/AEM.07486-11.
22. Murakami S. *Caenorhabditis elegans* as a model system to study aging of learning and memory. Mol Neurobiol. 2007;35(1):85–94. https://doi.org/10.1007/BF02700625. (PubMed PMID: 17519507).
23. Roussos A, Kitopoulou K, Borbolis F, Palikaras K. *Caenorhabditis elegans* as a model system to study human neurodegenerative disorders. Biomolecules. 2023. https://doi.org/10.3390/biom13030478.
24. Zecic A, Dhondt I, Braeckman BP. The nutritional requirements of *Caenorhabditis elegans*. Genes Nutr. 2019;14:15. https://doi.org/10.1186/s12263-019-0637-7.
25. Bush ZD, Naftaly AFS, Dinwiddie D, Albers C, Hillers KJ, Libuda DE. Comprehensive detection of structural variation and transposable element differences between wild type laboratory lineages of C. elegans. bioRxiv. 2023. https://doi.org/10.1101/2023.01.13.523974.
26. Liu H, Wang X, Wang HD, Wu J, Ren J, Meng L, et al. *Escherichia coli* noncoding RNAs can affect gene expression and physiology of *Caenorhabditis elegans*. Nat Commun. 2012;3(1):1073. https://doi.org/10.1038/ncomms2071.
27. Steinegger M, Salzberg SL. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biol. 2020;21(1):115. https://doi.org/10.1186/s13059-020-02023-1.
28. Hou K, Wu ZX, Chen XY, Wang JQ, Zhang D, Xiao C, et al. Microbiota in health and diseases. Signal Transduct Target Ther. 2022;7(1):135. https://doi.org/10.1038/s41392-022-00974-4.
29. Rodriguez J, Cordaillat-Simmons M, Pot B, Druart C. The regulatory framework for microbiome-based therapies: insights into European regulatory developments. NPJ Biofilms Microbiomes. 2025;11(1):53. https://doi.org/10.1038/s41522-025-00683-0.
30. Sorbara MT, Pamer EG. Microbiome-based therapeutics. Nat Rev Microbiol. 2022;20(6):365–80. https://doi.org/10.1038/s41579-021-00667-9.
31. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog. 2016;8(1):24. https://doi.org/10.1186/s13099-016-0103-7.
32. Greathouse KL, Sinha R, Vogtmann E. DNA extraction for human microbiome studies: the issue of standardization. Genome Biol. 2019;20(1):212. https://doi.org/10.1186/s13059-019-1843-8.
33. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12(1):87. https://doi.org/10.1186/s12915-014-0087-z.
34. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. Tracking down the sources of experimental contamination in microbiome studies. Genome Biol. 2014;15(12):564. https://doi.org/10.1186/s13059-014-0564-2.
35. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019;10(1):5029. https://doi.org/10.1038/s41467-019-13036-1.
36. Kupfer DM, Chaturvedi AK, Canfield DV, Roe BA. PCR-based identification of postmortem microbial contaminants–a preliminary study. J Forensic Sci. 1999;44(3):592–6 (PubMed PMID: 10408116).
37. Liu Y, Elworth RAL, Jochum MD, Aagaard KM, Treangen TJ. De novo identification of microbial contaminants in low microbial biomass microbiomes with Squeegee. Nat Commun. 2022;13(1):6799. https://doi.org/10.1038/s41467-022-34409-z.
38. Drost HG, Paszkowski J. Biomartr: genomic data retrieval with R. Bioinformatics. 2017;33(8):1216–7. https://doi.org/10.1093/bioinformatics/btw821.
39. Kim C, Kim J, Kim S, Cook DE, Evans KS, Andersen EC, et al. Long-read sequencing reveals intra-species tolerance of substantial structural variations and new subtelomere formation in C. elegans. Genome Res. 2019;29(6):1023–35. https://doi.org/10.1101/gr.246082.118.

40. Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability machines: consistent probability estimation using nonparametric learning machines. Methods Inf Med. 2012;51(1):74–81. https://doi.org/10.3414/ME00-01-0052.
41. Team RC. R: A language and environment for statistical computing (Version 4.0. 2). R Foundation for Statistical Computing. 2020.
42. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28(5):1–26. https://doi.org/10.18637/jss.v028.i05.
43. Liaw AaW, Matthew. Classification and Regression by randomForest. R News. 2002;2:18–22.
44. Diaz-Uriarte R, de Alvarez Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinform. 2006;7:3. https://doi.org/10.1186/1471-2105-7-3.
45. Goldstein BA, Polley EC, Briggs FB. Random forests for genetic association studies. Stat Appl Genet Mol Biol. 2011;10(1):32. https://doi.org/10.2202/1544-6115.1691.
46. Pellegrino E, Jacques C, Beaufils N, Nanni I, Carlioz A, Metellus P, et al. Machine learning random forest for predicting oncosomatic variant NGS analysis. Sci Rep. 2021;11(1):21820. https://doi.org/10.1038/s41598-021-01253-y.
47. Toth R, Schiffmann H, Hube-Magg C, Buscheck F, Hoflmayer D, Weidemann S, et al. Random forest-based modelling to detect biomarkers for prostate cancer progression. Clin Epigenetics. 2019;11(1):148. https://doi.org/10.1186/s13148-019-0736-8.
48. Meher PK, Sahu TK, Gahoi S, Tomar R, Rao AR. funbarRF: DNA barcode-based fungal species prediction using multiclass Random Forest supervised learning model. BMC Genet. 2019;20(1):2. https://doi.org/10.1186/s12863-018-0710-z.
49. Meher PK, Sahu TK, Rao AR. Identification of species based on DNA barcode using k-mer feature vector and Random forest classifier. Gene. 2016;592(2):316–24. https://doi.org/10.1016/j.gene.2016.07.010.
50. Riza LS, Zain MI, Izzuddin A, Prasetyo Y, Hidayat T, Abu Samah KAF. Implementation of machine learning in DNA barcoding for determining the plant family taxonomy. Heliyon. 2023;9(10):e20161. https://doi.org/10.1016/j.heliyon.2023.e20161.
51. Oshiro TM, Perez PS, Baranauskas JA, editors. How Many Trees in a Random Forest?2012; Berlin, Heidelberg: Springer Berlin Heidelberg.
52. Zhao K, Wulder MA, Hu T, Bright R, Wu Q, Qin H, et al. Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: a Bayesian ensemble algorithm. Remote Sens Environ. 2019;232:111181. https://doi.org/10.1016/j.rse.2019.04.034.
53. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020;12(1):91. https://doi.org/10.1186/s13073-020-00791-w.
54. Malamon JS, Farrell JJ, Xia LC, Dombroski BA, Das RG, Way J, et al. A comparative study of structural variant calling in WGS from Alzheimer's disease families. Life Sci Alliance. 2024. https://doi.org/10.26508/lsa.202302181.
55. Naj AC, Lin H, Vardarajan BN, White S, Lancour D, Ma Y, et al. Quality control and integration of genotypes from two calling pipelines for whole genome sequence data in the Alzheimer's disease sequencing project. Genomics. 2019;111(4):808–18. https://doi.org/10.1016/j.ygeno.2018.05.004.
56. Trubetskoy V, Rodriguez A, Dave U, Campbell N, Crawford EL, Cook EH, et al. Consensus genotyper for exome sequencing (CGES): improving the quality of exome variant genotypes. Bioinformatics. 2015;31(2):187–93. https://doi.org/10.1093/bioinformatics/btu591.
57. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. . Scientific Research. 2010.
58. de Sena Brandine G, Smith AD. Falco: high-speed FastQC emulation for quality control of sequencing data. F1000Res. 2019;8:1874. https://doi.org/10.12688/f1000research.21142.2.
59. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560.
60. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS ONE. 2011;6(3):e17288. https://doi.org/10.1371/journal.pone.0017288.
61. Brimble MA, Winston SM, Davidoff AM. Stowaways in the cargo: contaminating nucleic acids in rAAV preparations for gene therapy. Mol Ther. 2023;31(10):2826–38. https://doi.org/10.1016/j.ymthe.2023.07.025.
62. Higashiyama K, Yuan Y, Hashiba N, Masumi-Koizumi K, Yusa K, Uchida K. Quantitation of residual host cell DNA in recombinant adeno-associated virus using droplet digital polymerase chain reaction. Hum Gene Ther. 2023;34(11–12):578–85. https://doi.org/10.1089/hum.2023.006.
63. Luthers CR, Ha SM, Mittelhauser A, Morselli M, Long JD, Kuo CY, et al. DNA contamination within recombinant adeno-associated virus preparations correlates with decreased CD34(+) cell clonogenic potential. Mol Ther Methods Clin Dev. 2024;32(4):101334. https://doi.org/10.1016/j.omtm.2024.101334.

Publisher's note