Neptune: A Tool for Rapid Microbial Genomic Signature Discovery

Eric Marinier[1], Rahat Zaheer[1], Chrystal Berry[1], Kelly Weedmark[1], Michael Domaratzki[2], Philip Mabon[1], Natalie Knox[1], Aleisha Reimer[1], Morag Graham[1,3], The Canadian Listeria Detection and Surveillance using Next Generation Genomics (LiDS-NG) Consortium[†] and Gary Van Domselaar [*1,3]

[1] *National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB, Canada*
[2] *Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada*
[3] *Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, MB, Canada*

Abstract

Neptune locates genomic signatures using an exact k-mer matching strategy while accommodating k-mer mismatches. The software identifies sequences that are sufficiently represented within "inclusion targets" and sufficiently absent from "exclusion targets". The signature discovery process is accomplished using probabilistic models instead of heuristic strategies. We have evaluated Neptune on *Listeria monocytogenes* and *Escherichia coli* genome data sets and found that signatures identified from these experiments are sensitive and specific to their respective data sets. In addition, the identified loci provide a catalog of differential loci for research of group-specific traits. Neptune has broad implications in microbial characterization for public health applications due to its efficient ad hoc signature discovery based upon differential genomics.

1 Introduction

The ability to identify and respond to emergent infectious agents in a time-sensitive manner is critical for ensuring public health safety [23]. The advancement of high-throughput next generation sequencing (NGS) has necessitated computational approaches for effective, real-time, comprehensive outbreak investigation and response. An important component of public health response is rapid characterization of infectious agents, including the discovery of discriminatory signature sequences that may be leveraged to uniquely identify a group of organisms, such as those associated with a disease cluster.

This work defines a genomic signature as a string of characters, representing nucleotide bases, capable of discriminating targets within a group of interest from a user-defined background group. These signatures are sufficiently unique to a set of targets and sufficiently dissimilar from any sequence within the related user-defined background. We define the group of interest as the "inclusion group," the background as the "exclusion group," and a reference sequence as any inclusion target from which to extract signatures. Targets will typically comprise of fully-assembled or draft genomes. Signature discovery aims to locate unique and conserved regions within the inclusion group that are absent within the exclusion group. Neptune signatures will be specific and precisely identified for the user-defined groups and data; however, sensitivity and specificity should be verified for application to broader data sets not used in the discovery process.

A naive approach to signature discovery involves exhaustively comparing all sequences using alignments to locate signature regions. However, such approaches do not scale effectively. An approximation to exhaustive comparisons is sequence clustering; yet clustering without optimization may remain too slow. An effective signature discovery algorithm needs to be both sensitive and specific, while remaining computationally tractable. There are two common approaches to achieve sensitivity, which trade speed for sensitivity. The first approach requires inclusion sequence to match exactly [20]. This approach is extremely fast, but will miss divergent alleles and be confounded by regions that are not highly conserved. The second approach involves grouping similar sequences together using multiple sequence alignments [23], seeding techniques [21], or leveraging clustering information [2]. While these latter approaches are more sensitive, they are necessarily slower than exact-matching techniques. TOFI [22] avoids this problem by only locating signatures for a single target and not a group. Specificity of signatures is gener-

*Corresponding author:
gary.vandomselaar@phac-aspc.gc.ca

ally verified using computationally expensive alignments of signature candidates [21–23] against the background, which typically involves using BLAST [1] alignments. However, verification is performed after significant data reduction, making this operation more feasible. KPATH [23] performs verification by comparing a consensus sequence produced from inclusion targets to a large non-target database. KPATH achieves acceptable speeds by leveraging suffix trees to find matches.

To efficiently perform signature discovery, especially under time constraints, a significant data reduction is essential [21–23]. The data set is generally reduced by identifying and removing sequences that are "definitely not unique" [23] in a computationally inexpensive manner. Insignia [20], TOFI [22], and TOPSI [21] use MUMmer [12] to precompute exact matches within inclusion targets and also in an exclusion background. However, depending on the size of the background database, this may remain computationally expensive. CaSSiS [2] approaches the problem of signature discovery more thoroughly than other signature discovery pipelines. It produces signatures simultaneously for all locations in a hierarchically clustered data set, such as a phylogenetic tree, thereby producing candidate signatures for all possible subgroups. However, this process requires the input data to be provided in a hierarchically clustered format, which will additionally increase processing time.

Neptune leverages existing strategies for signature detection by using an exact-matching k-mer strategy for speed, while making allowances for inexact matches to enhance sensitivity. However, unlike other existing exact-matching approaches [20], Neptune performs signature discovery without precomputation or restriction on targets, i.e., on-the-fly. Furthermore, Neptune locates signatures with inexact matches and hence, are not perfectly conserved. Lee and Sheu [13] remark that existing signature discovery approaches are not readily parallelizable. Neptune also is designed to operate on a high performance computing cluster. Neptune extracts signatures from one or more targets, in a highly parallelizable manner, and is independent of multiple sequence alignments. Finally, Neptune's signature discovery pipeline is guided with probabilistic models, rather than heuristics, and therefore makes signature identification decisions with a measure of certainty.

2 Methods

Neptune uses the distinct k-mers found in each inclusion and exclusion target to identify sequences that are conserved within the inclusion group and absent from the exclusion group. Neptune evaluates all sequence and may therefore produce signatures that correspond to intergenic regions or contain entire operons. The k-mer generation step produces distinct k-mers from all targets and aggregates this information, reporting the number of inclusion and exclusion targets that contain each k-mer. The signature extraction step identifies candidate signatures from one or more references, which are assumed to be drawn from inclusion targets. Candidate signatures are filtered by performing an analysis of signature specificity using pairwise sequence alignments. The remaining signatures are ranked by their Neptune-defined sensitivity and specificity scores, representing a measure of signature confidence.

We provide descriptions of the different stages of signature discovery below and an overview of the signature discovery process is found in Figure 1. The majority of parameters within Neptune are automatically calculated for every reference. However, the user may specify any of these parameters. A full description of the mathematics used in the software is available in supplementary materials. In our probabilistic model, we assume that the probability of observing any single nucleotide base in a sequence is independent of all other positions and the probability of all single nucleotide variant (SNV) events (e.g. mutations, sequencing errors) occurring is independent of all other SNV events.

2.1 *k*-mer Generation

Neptune produces the distinct set of k-mers for every inclusion and exclusion target and aggregates these k-mers together before further processing. The software is concerned only with the existence of a k-mer within each target and not with the number of times a k-mer is repeated within a target. Neptune converts all k-mers to the lexicographically smaller of either the forward k-mer or its reverse complement. This avoids maintaining both the forward and reverse complement sequence [17]. The number of possible k-mers is bound by the total length of all targets. The k-mers of each target are determined independently and, when possible, in parallel. In order to facilitate parallelizable k-mer aggregation, the k-mers for each target may be organized into several output files. The k-mers in each file are unique to one target (e.g., isolate genome or sequence) and all share the same initial sequence index. This degree of organization may be specified by the user.

The k-mer length is automatically calculated unless provided by the user. A summary of recommended k-mer sizes for various genomes can be found in supplementary material. We suggest a size of k such that we do not expect to see two arbitrary k-mers within the same target match exactly. This recommendation is motivated by wanting to

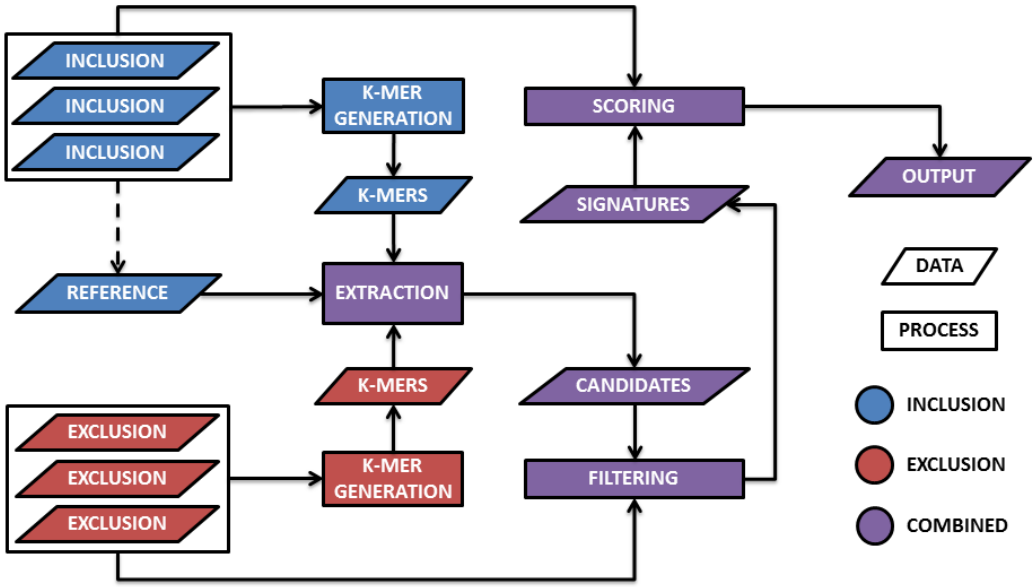

Figure 1: An overview of Neptune's signature discovery process for a single target. The first step involves generating k-mers from all inclusion and exclusion targets. These k-mers are aggregated and provided as input to signature extraction. Signature extraction produces candidate signatures, which are filtered and then sorted by their sensitivity and specificity scores.

generate distinct k-mer information, thereby having matching k-mers most often be a consequence of nucleotide homology. Let λ be the most extreme GC-content of all targets and ω be the size of the largest target in bases. The probability of any two arbitrary k-mers, k_X and k_Y, matching exactly, $P(k_X = k_Y)_A$, where $x \neq y$, is defined as follows:

$$P(k_X = k_Y)_A = \left(2\left(\frac{1-\lambda}{2}\right)^2 + 2\left(\frac{\lambda}{2}\right)^2\right)^k \quad (1)$$

We use the probability of arbitrary k-mers matching, $P(k_X = k_Y)_A$, to approximate the probability of k-mers matching within a target, $P(k_X = k_Y)$. This is an approximation because the probability of $P(k_{X+1} = k_{Y+1})$ is known to not be independent of $P(k_X = k_Y)$. However, this approximation approaches equality as $P(k_X = k_Y)_A$ decreases, which is accomplished by selecting a sufficiently large k, such that we do not expect to see any arbitrary k-mer matches. We suggest using a large enough k such that the expected number of intra-target k-mer matches is as follows:

$$\sum_{x<y} P(k_X = k_Y) \approx \binom{\omega - k + 1}{2} \cdot P(k_X = k_Y)_A < 0.05 \quad (2)$$

The distinct sets of k-mers from all targets are aggregated into a single file, which is used to inform signature extraction. This process may be performed in parallel by aggregating k-mers sharing the same initial sequence index and concatenating the aggregated files. Aggregation produces a list of k-mers and two values (the number of inclusion and exclusion targets containing the k-mer, respectively). This information is used in the signature extraction step to categorize some k-mers as inclusion or exclusion k-mers.

2.2 Extraction

Signatures are extracted from one or more references, which are drawn from all inclusion targets, unless specified otherwise. However, our probabilistic model assumes all references are included as inclusion targets. In order to identify candidate signatures, Neptune reduces the effective search space of signatures by leveraging the spatial sequencing information inherent within the references. Neptune evaluates all k-mers in each reference, which may be classified as inclusion or exclusion k-mers. An inclusion k-mer is observed in a sufficient number of inclusion targets and not observed in a sufficient number of exclusion targets. The sufficiency requirement is described below. Inclusion and exclusion k-mers are used to infer inclusion and exclusion sequence, with signatures containing primarily inclusion sequence. An inclusion k-mer may contain both inclusion and exclusion sequence because, while they may contain exclusion sequence, k-mers that overlap inclusion and exclusion sequence will often be unique to the inclusion group. An exclusion k-mer is, by default, any k-mer that has been

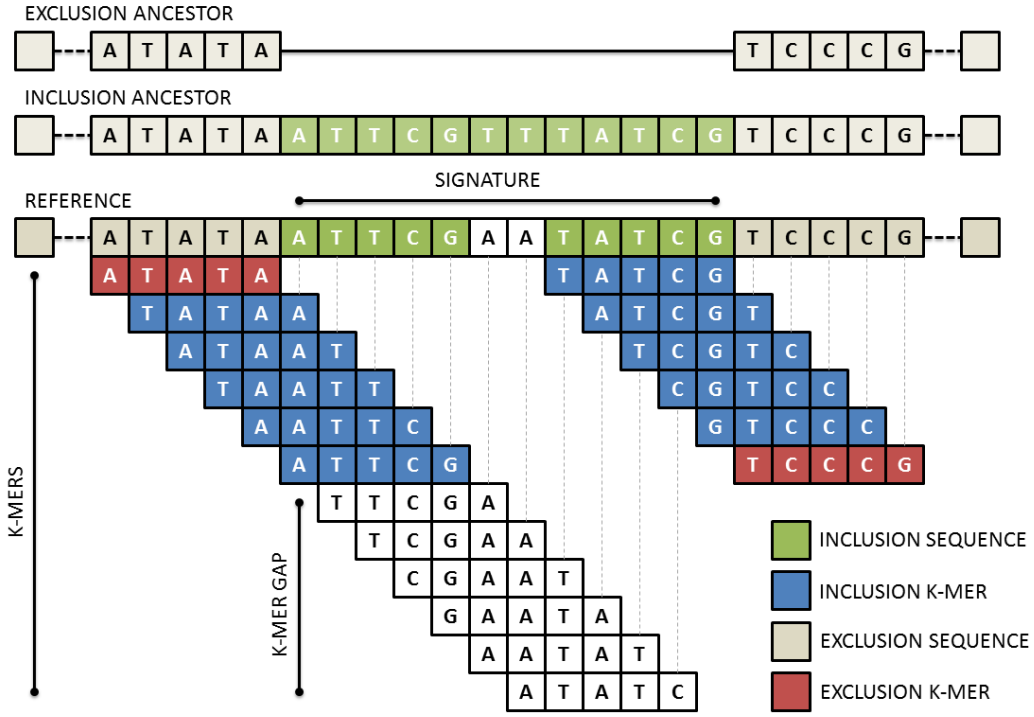

Figure 2: An overview of Neptune's signature extraction process. The reference is decomposed into its composite k-mers. These k-mers may be classified as either inclusion or exclusion and are used to infer inclusion and exclusion sequence in the reference. A signature is constructed from inclusion k-mers containing sufficiently small k-mer gaps and no exclusion k-mers.

observed at least once in any exclusion target. However, in some applications it may be desirable to relax this stringency. For example, leniency may be appropriate when the inclusion and exclusion groups are not fully understood. This may be the case when meta data is incomplete or unreliable. An exclusion k-mer should, by design, not contain any inclusion sequence. Neptune outputs several "candidate signatures", which begin with the last base position of the first inclusion k-mer, contain an allowable number of k-mer gaps and no exclusion k-mers, and end with the first base position of the last inclusion k-mer (Figure 2). This process is conceptually similar to taking the intersection of inclusion k-mers and allowable k-mer gaps. Furthermore, it avoids generating a candidate containing exclusion sequence found in inclusion k-mers that overlap inclusion and exclusion sequence regions.

An inclusion k-mer is considered sufficiently represented when it is observed in a number of targets exceeding a minimum threshold. We assume that if there is a signature present in all inclusion targets, then the signature will correspond to homologous sequences in all these targets and these sequences will produce exact matching k-mers with some probability. We start with the probability that two of these homologous bases, X and Y, match is:

$$P(X = Y)_H = (1-\varepsilon)^2 + (\varepsilon)^2 \cdot P(X_M = Y_M)_H \quad (3)$$

where ε is the probability that two homologous bases do not match exactly, and $P(X_M = Y_M)_H$ is the probability that two homologous bases both mutate to the same base. The default probability of ε is 0.01. We assume that when the homologous bases do not match, the observed base is dependent on the GC-content of the environment. Let λ be the GC-content of the environment. The probability of $P(X_M = Y_M)_H$ is defined as follows:

$$P(X_M = Y_M)_H = \left(2\left(\frac{\lambda}{\lambda+1}\right)^2 + \left(\frac{1-\lambda}{\lambda+1}\right)^2\right)(1-\lambda) + \left(2\left(\frac{1-\lambda}{2-\lambda}\right)^2 + \left(\frac{\lambda}{2-\lambda}\right)^2\right)(\lambda) \quad (4)$$

This probability depends significantly on GC-content of the environment. We assume that the probability of each base matching is independent. Therefore, the probability that two homologous k-mers, k_X and k_Y, match:

$$P(k_X = k_Y)_H = (Pr(X = Y)_H)^k \quad (5)$$

We model the process of homologous k-mer matches with a binomial distribution. If we are observing a true signature region in a reference, we expect that corresponding homologous k-mers exist in all inclusion targets and infer this homology from aggregated k-mer information. An observed reference k-mer will exactly match a corresponding homologous k-mer in another inclusion target with a probability of $p = P(k_X = k_Y)_H$ and not match with a probability of $q = 1 - p$. The expected number of exact k-mer matches with a reference k-mer will be $\mu = (n-1) \cdot p$ and the variance will be $\sigma^2 = (n-1) \cdot p \cdot q$, where n is the number of inclusion targets. We require $n - 1$ because the reference is an inclusion target and its k-mers will exactly match themselves. However, we compensate for this match in our expectation calculation. We assume the probability of each k-mer match is independent and that k-mer matches are a consequence of homology. When the number of inclusion targets and the probability of homologous k-mers exactly matching are together sufficiently large, the binomial distribution is approximately normal. Let α be our statistical confidence and $\Phi^{-1}(\alpha)$ be the probit function. The minimum number of inclusion targets containing a k-mer, $\wedge_{in}$, required for a reference k-mer to be considered an inclusion k-mer is defined as follows:

$$\wedge_{in} = 1 + \mu - \Phi^{-1}(\alpha)\sigma \tag{6}$$

The $\wedge_{in}$ parameter is automatically calculated unless provided by the user and will inform candidate signature extraction. However, there may be mismatches in the reference, which exclude it from the largest homologous k-mer matching group. We accommodate for this possibility by allowing k-mer gaps in our extraction process. We model the problem of maximum k-mer gap size between exact matching inclusion k-mers as recurrence times of success runs in Bernoulli trials. The mean and variance of the distribution of the recurrence times of k successes in Bernoulli trials is described in Feller 1960 [9]:

$$\mu = \frac{1 - p^k}{q \cdot p^k} \tag{7}$$

$$\sigma^2 = \frac{1}{(q \cdot p^k)^2} - \frac{2k+1}{q \cdot p^k} - \frac{p}{q^2} \tag{8}$$

This distribution captures how many bases we expect to observe before we see another homologous k-mer match. The probability of a success is defined at the base level as $p = P(X = Y)_H$ and the probability of failure as $q = (1 - p)$. This distribution may not be normal for a small number of observations. However, we can use Chebyshev's Inequality to make lower-bound claims about the distribution:

$$P(|X - \mu| \geq \delta\sigma) \leq \frac{1}{\delta^2} \tag{9}$$

where δ is the number of standard deviations, σ, from the mean, μ. Let $P(|X - \mu| \geq \delta\sigma)$ be our stastical confidence, α. The maximum allowable k-mer gap size, $\vee_{gap}$, is calculated as follows:

$$\vee_{gap} = \mu + \sqrt{\frac{1}{1 - \alpha}} \cdot \sigma \tag{10}$$

The $\vee_{gap}$ parameter is automatically calculated unless specified. Candidate signatures are terminated when either no additional inclusion k-mers are located within the maximum gap size, $\vee_{gap}$, or an exclusion k-mer is identified. In both cases, the candidate signature ends with the last inclusion k-mer match. The consequence of terminating a signature early is that a large, contiguous signature may be reported as multiple smaller signatures. We require the minimum signature size, by default, to be four times the size of k. However, for some applications, such as designing assay targets, it may be desirable to use a smaller or larger minimum signature size. Signatures cannot be shorter than k bases. We found that smaller signatures were more sensitive to the seed size used in filtering alignments. There is no maximum signature size. As a consequence of Neptune's signature extraction process, signatures extracted from the same target may never overlap each other.

2.3 Filtering

The candidate signatures produced will be relatively sensitive, but not necessarily specific, because signature extraction is done using exact k-mer matches. The candidate signatures are guaranteed to contain no more exact matches with any exclusion k-mer than was specified in advanced by the user. However, there may exist inexact matches within exclusion targets. Neptune uses BLAST [1] to locate signatures that align with any exclusion target and, by default, removes any signature that shares 50% identity with any exclusion target aligning to at least 50% of the signature, anywhere along the signature. This process is done to avoid investigating signatures that are not discriminatory. The remaining signatures are considered filtered signatures and are believed to be sensitive and specific, within the context of the relative uniqueness of the input inclusion and exclusion groups, and the parameters supplied for target identification.

2.4 Scoring

Signatures are assigned an overall score corresponding to their highest-scoring BLAST [1] alignments

with all inclusion and exclusion targets. This score is the sum of a positive inclusion component and a negative exclusion component, which are analogous to sensitivity and specificity, respectively, with respect to the input data. Let $|A(S, I_i)|$ be the length of the highest-scoring aligned region between a signature, S, and an inclusion target, I_i. Let $|S|$ be the length of signature S, $PI(S, I_i)$ the percent identity (identities divided by the alignment length) between the aligned region of S and I_i, and $|I|$ be the number inclusion targets. The negative exclusion component is similarly defined. The signature score, $score(S)$, is calculated as follows:

$$score(S) = \sum_{i=0}^{|I|} \frac{|A(S, I_i)| \cdot PI(S, I_i)}{|S||I|} - \sum_{i=0}^{|E|} \frac{|A(S, E_i)| \cdot PI(S, E_i)}{|S||E|} \quad (11)$$

This score is maximized when all inclusion targets contain a region exactly matching the entire signature and there exists no exclusion targets that match the signature. Signatures are sorted based on their scores with highest-ranking signatures appearing first in the output.

2.5 Output

Neptune produces a list of candidate, filtered, and sorted signatures for all references. The candidate signatures are guaranteed to contain, by default, no exact matches with any exclusion k-mer. However, there may still remain potential inexact matches within exclusion targets. The filtered signatures contain no signatures with significant sequence similarity to any exclusion target. Sorted signatures are filtered signatures appearing in descending order of their signature scores. A consolidated signature file is additionally provided as part of Neptune's output. This file contains a consolidated list of the top-scoring signatures produced from all reference targets, such that homologous signatures are reported only once. However, because this file is constructed in a greedy manner, it is possible for signatures within this file to overlap each other. To identify redundancy across the reference targets, we recommend evaluating the signatures identified from each individual reference target in combination with this consolidated file when evaluating signatures.

3 Results

We employed Neptune to identify signatures for several distinct bacterial genomes of differing phyla. In order to validate our method and to highlight mathematical considerations, we applied Neptune to locate signatures within an artificial *Bacillus anthracis* data set. We then applied Neptune to identify signatures within a clinically-relevant *Listeria monocytogenes* data set to demonstrate Neptune's behaviour when operated on clonal (i.e. not highly divergent) isolate populations. Lastly, we employed a clinically-relevant *Escherichia coli* data set to demonstrate Neptune's capacity to locate signatures for a more diverse data set.

ID	Length (bp)	Summary
1	23338	O-antigen transport
2	50038	toxin pilus
3	12259	phage replication
4	9652	phage integrase
5	4282	N-acetylneuraminate lyase
6	10155	neuraminidase

Table 1: Genomic islands naturally found within *Vibrio cholerae* (NC_012578.1) chromosome I. These islands were used as *in silico* signatures and artificially inserted within a *Bacillus anthracis* genome. These islands were identified with IslandViewer 3 [8].

3.1 Artificial *Bacillus anthracis*

In order to show that Neptune identifies signatures as expected, the software was run with an artificially created data set. We created an initial inclusion genome by inserting non-overlapping, virulence- and pathogen-associated genes from *Vibrio cholerae* (NC_012578.1) into a *Bacillus anthracis* genome (NC_007530) (Table 1). We selected 6 signature regions varying from 4 to 50-kb in size and spaced these signatures evenly throughout the *B. anthracis* genome with each signature represented only once. The initial exclusion genome represented a copy of the original (NC_007530) *B. anthracis* genome lacking modification. Lastly, we broadened both the inclusion and exclusion groups to 20 genomes each, by generating copies of the corresponding original inclusion or exclusion genome and incorporating a 1% random nucleotide mutation rate, with all mutations being equally probable.

Neptune was used to identify inserted pathogenic and virulence regions in our artificial *B. anthracis* data set. We specified a k-mer size of 27 and used Neptune's default SNV rate of 1%. The k-mer size was derived from Equation 2, given a genome size of 5337 kb and a GC-content of 0.36. Neptune produced signatures from all 20 inclusion targets (supplementary material) and these signatures were consolidated into a single signature file. We aligned these signatures to the initial inclusion genome and used GView Server [19] to visualize the identified signatures from all references (Figure 3). Neptune

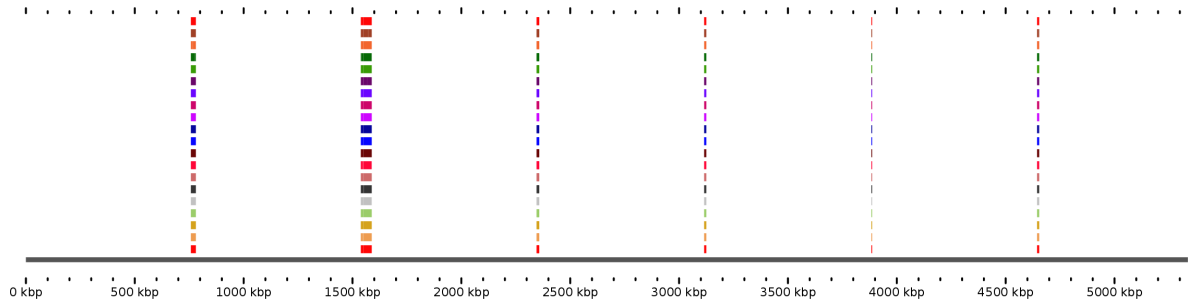

Figure 3: An array of *V. cholerae*-based *in silico* signatures produced using Neptune. All of the artificially inserted *V. cholerae* pathogenic regions were extracted consistently from several, nonidentical artificial *B. anthracis*-*V. cholerae* inclusion group targets against an endogenous *B. anthracis* exclusion group.

identified 7 consolidated signatures, corresponding to the 6 expected *V. cholerae* regions, with the largest signature region (50 kb) misreported as two adjacent, but incomplete (i.e., gapped), signatures. However, by Equation 9, we expect to see erroneous signature breaks with a frequency inversely proportional to our confidence level (95%) when extending signatures over k-mer gaps. This is not a serious issue because these events are relatively rare and these broken signatures are several thousand nucleotide bases in length. Importantly, we observed that all Neptune-identified identified signatures corresponded to the artificially inserted *V. cholerae* regions and were consistently detected for all references. Neptune reported all of the *in silico* signatures and reported no false positives. Hence, we conclude that Neptune is able to locate all *in silico* signatures; although some regions identified are reported as two adjacent signatures.

3.2 *Listeria monocytogenes*

Neptune was next used to locate signature regions within two distinct serotypes of *Listeria monocytogenes*. *L. monocytogenes* is an opportunistic environmental pathogen that causes listeriosis, a serious and life-threatening bacterial disease in humans and animals [18]. Consumption of listeria-contaminated food products have caused several recent nation-wide outbreaks in the United States and Canada and are a significant concern to the food industry and to public health [7, 15, 16]. *L. monocytogenes* is a clonal organism and recent *L. monocytogenes* evolution has been characterized by deletion events of horizontally acquired bacteriophage and genomic islands. We therefore expected to find signatures corresponding to these events.

We employed a draft genome data set produced by and analyzed for the Canadian Listeria Detection and Surveillance using Next-Generation Genomics (LiDS-NG) project (SRA PRJNA301341). Listeria isolates were serotyped using standard laboratory serotyping procedures [10]. Serotypes 1/2a and 4b were selected for evaluation as they represent distinct bacterial lineages and are clinically relevant [18]. Of the 13 *L. monocytogenes* serotypes, serotype 1/2b and 4b (lineage I) and serotype 1/2a (lineage II) are most commonly associated with human illness globally [18]. *L. monocytogenes* lineage I is characterized by low diversity and low recombination and strains from this lineage are overrepresented among human isolates, as compared to lineage II strains, which have exhibited higher levels of genomic diversity, owing to recombination and horizontal gene transfer and have an overrepresentation among food, food-related and natural environments [18]. In total, 112 serotype 1/2a and 40 serotype 4b targets were available to be used as inclusion and exclusion groups. These were independently assessed to identify 1/2a signatures as well as the reciprocal 4b signatures, by reversing the inclusion and exclusion groupings. These groups were randomly subdivided into a training set and a validation set.

Neptune was executed on the *L. monocytogenes* training data in order to produce both 1/2a and 4b signatures for validation. We specified a k-mer size of 25, derived given a genome size of 3048 kb, the length of the largest isolate in nucleotides, and a GC-content of 0.38, the most extreme GC-content of all our isolates (Equation 2). Neptune produced 101 1/2a consolidated signatures and 65 4b consolidated signatures from their respective inclusion targets. We further evaluated the top-scoring (≥ 0.95) 1/2a and 4b consolidated signatures. The signatures produced from 1/2a targets (1/2a inclusion, 4b exclusion) were aligned against *L. monocytogenes* 1/2a strains EGD-e (NC_003210) and 08-5578 (NC_013766), whereas the signatures generated from 4b targets (4b inclusion, 1/2a exclusion) were mapped to *L. monocytogenes* strain 4b F2365 (NC_2973)

The top-scoring signatures (≥ 0.95) identified for *L. monocytogenes* serotype 1/2a are listed in Table 2. These signatures included phosphoenolpyruvate (PEP)-dependent phosphotransferase systems

Rank	Score	Length (bp)	Locus Information	*L. monocytogenes* serotype 1/2a str. EGD-e coordinates
1	0.99	5455	PTS system, glucose–glucoside (Glc) family	664364 - 770791
2	0.99	4059	*bvrABC* locus, beta-glucoside-specific sensory system	2872894 - 2876953
3	0.99	5336	PTS system, L-ascorbate (L-Asc) family	2042111 - 2047448
4	0.99	4830	peptidoglycan-bound protein colossin A	2653184 - 2658014
5	0.98	4468	two-component response regulator and ABC transport systems	1086579 - 1091047
6	0.98	1943	hypothetical	776414 - 778356
7	0.97	2567	lineage II specific heat-shock system	441513 - 444080
8	0.97	1673	glycosyl-transferase	532558 - 534231
9	0.96	968	hypothetical	2717382 - 2718349

Table 2: A summary of *L. monocytogenes* serotype 1/2a signatures generated by Neptune relative to a serotype 4b background. These genomic signatures are ordered by their signature score, which is comprised on a positive inclusion component and a negative exclusion component. We show all signatures with a score ≥ 0.95. As some signatures contain multiple genes, the "Locus Information" column contains a highlight of the region.

(PTS) belonging to L-ascorbate (PTSAsc) and glucose–glucoside (PTSGlc) families [24], and a 4468 bp locus containing a two-component response regulation system and an ABC transport system [6]. In keeping with a predilection for human clinical disease, the presence of a variety of PTS systems and transport systems provides *L. monocytogenes* serotype 1/2a with a competitive advantage to survive under different environmental conditions due to its ability to utilize a variety of carbon sources. A bvrABC locus was found among these 1/2a signatures which is known to be involved in regulating virulence genes in response to environmental cues [4]. Also found was a surface-exposed internalin protein gene, which is known to be a critical factor for human pathogenesis [3]. Furthermore, a lineage II-specific heat-shock system [25] constituting an operon with 3 genes, encoding RNA polymerase factor sigma C, LstR thermal regulator, and a cell division related protein were present among those high scoring signatures. Other 1/2a signatures included sequences coding for a peptidoglycan-bound protein, glycosyl-transferase, and hypothetical proteins.

The top-ranking signatures (≥ 0.95) identified for *L. monocytogenes* serotype 4b (Table 3) included the following: a *gltA-gltB* operon [14]; a hypothetical protein gene, a 4 kb region involved in N-acetylmuramic acid metabolism spanning an RpiR family transcriptional regulator and a downstream operon consisting of *murQ* gene and a PTS system sucrose-specific transporter subunit IIBC, cell wall anchor protein gene; a 5.9 kb region with 7 genes including a two-gene operon coding for hypothetical proteins, a cell-wall surface anchor protein gene, a GntR family regulator and an ABC transport system containing 3 genes in an operon potentially coding for multidrug efflux system; a very large gene coding for RHS-repeat-family protein [11]; a pyruvyl transferase; a serine protease gene; a teichioc acid biosynthetic protein gene.

These training-generated signatures were then compared against the web-lab verified validation data sets to evaluate their *in silico* sensitivity and specificity. We used BLAST [1] to independently align the top-scoring signatures against our validation data sets. The complete alignment output can be found in supplementary material. With a percent identity threshold of 95% and a minimum alignment length of 95% the size of the signature length, 502 out of 504 (99.6%) 1/2a signature alignments against the 1/2a validation targets met our sensitivity criteria. Likewise, 179 out of 180 (99.4%) 4b signature alignments against 4b validation targets met this strictness. Similarly, with a percent identity threshold of 50% and a minimum alignment length of 50% the size of the signature length, we found no 1/2a hits against 4b validation targets and no 4b hits against 1/2a validation targets, indicating that the signatures were specific to the inclusion group. These results suggest that our top-scoring Neptune-identified *L. monocytogenes* serotype 1/2a and 4b signatures are highly sensitive and specific to their respective serotypes against the other serotype background.

3.3 *Escherichia coli*

In an attempt to model a real application of signature discovery, we employ Neptune to locate signatures corresponding to Shiga-toxin producing *Escherichia coli* (STEC). Specifically, we investigate *E. coli* that produce the Stx1 toxin. This toxin requires both the *stx1a* and *stx1b* subunits to be functional. Therefore, we expected to locate genes for these subunits using Neptune. As *E. coli* exhibits significantly increased genomic di-

Rank	Score	Length (bp)	Locus Information	*L. monocytogenes* serotype 4b str. F2365 coordinates
1	0.99	3081	gltA-gltB operon	2787943 - 2791024
2	0.99	223	hypothetical	478246 - 478469
3	0.99	4004	N-acetylmuramic acid metabolism	1685737 - 1689739
4	0.98	1709	cell wall anchor	2684246 - 2685955
5	0.97	5917	multiple, including: hypothetical, cell surface membrane anchor, multidrug efflux transporter like	228382 - 434299
6	0.97	7064	RHS repeat protein	466603 - 473668
7	0.97	1785	pyruvyl-transferase	117970 - 119755
8	0.95	1741	serine protease	1924193 - 1925934
9	0.95	1654	teichoic acid biosynthesis	2190231 - 2191884

Table 3: A summary of *L. monocytogenes* serotype 4b signatures generated by Neptune relative to a serotype 1/2a background.

versity over *L. monocytogenes*, we expect it makes identifying related signatures a more computationally challenging problem.

The inclusion and exclusion data sets were comprised of 6 STEC (Stx1) and 11 non-STEC draft assemblies, respectively. Neptune was run with a k-mer size of 25 (Equation 2) and produced 371 consolidated signatures. The top-scoring signature had nearly 100% sensitivity and 100% specificity with respect to the inclusion and exclusion groups. We further investigated the top-scoring (≥ 0.95) consolidated signatures (Table 4) by aligning these signatures against an *E. coli* O157:H7 str. Sakai reference (NC_002695.1) to infer sequence annotations. This alignment included the chromosome and both plasmids. The Sakai reference was selected because it contains a copy of the Stx1 toxin and is well characterized.

As expected, Neptune identified the Stx1 region as the highest scoring signature (Table 4). Other salient Neptune-identified signatures included several virulence regions such as the urease gene cluster containing *ureA*, *B*, *C*, *D*, *E*, *F*, *G* genes, various phage-related genes, intimin transcription regulator (*perC*) sequences, hemolysin gene cluster and type 3 secretion system (T3SS)-related regions (Table 4). In the plasmid alignments, the hemolysin-predicted signature was the only top-scoring signature located on the pO157 plasmid. Furthermore, using BLAST [1], we found that many of the Neptune top-scoring signatures aligned to known *E. coli* O157:H7 O-Islands (a set of mobile genetic islands known to carry virulence factors). This included signatures 1-3, 5, 7-15; notably Shiga toxin I (as predicted), a urease gene cluster, and several phage elements. We conclude that Neptune is effective at locating known pathogenic regions and horizontally acquired regions within STEC with high sensitivity and high specificity.

4 Discussion

4.1 Parameters

While many of Neptune's parameters are automatically calculated, there are a few parameters that deserve special mention. We recommend odd-sized k-mers to avoid the possibility of a k-mer being the reverse complement of itself. The minimum number of inclusion hits and maximum gap size are sensitive to the SNV rate and the size of k. When estimating these parameters, a slightly higher than expected SNV rate is recommended. This conservative approach will avoid false negatives at the expense of false positives. However, many of these false positives will be removed during the filtering stage at the expense of increased computational time.

4.2 Memory and Computation Time

Neptune is highly parallelizable and performs well on high-performance computing clusters. When identifying signatures within a data set of 112 *L. monocytogenes* inclusion genomes of approximately 3,000 kb in length and 40 related *L. monocytogenes* exclusion genomes, Neptune required 27 minutes on a 40-node computing cluster. The memory requirements of all individual processes never exceeded more than 10G. Neptune benefits significantly from parallelization and will run much slower in a single-CPU environment.

4.3 Limitations

Neptune's signature extraction step avoids false negatives at the expense of false positives. The software attempts to locate signatures that may not contain an abundance of exact matches. This approach produces some false positives. However, false positives are removed during signature filtering and requires increased computational time. As signatures are extracted from a reference, repeated

Rank	Score	Length (bp)	Locus Information	*E. coli* O157:H7 Sakai coordinates
1	1.00	1375	Shiga toxin	2924383 - 2925757
2	0.99	5433	urease gene cluster	1390114 - 1395545
3	0.98	3291	bacteriophage related, integrase and other	2593022 - 2596313
4	0.98	438	*perC*, transcriptional activator of EaeA/BfpA, partial	1183201 - 1183639
5	0.98	1223	phage tail length tape measure protein, partial	2170250 - 2171473
6	0.97	7697	hemolysin gene cluster: *hylC*, *hylA*, *hylB*, *hylD*	15716 - 23412 (pO157)
7	0.96	1260	colonization factor	1769157 - 1767898
8	0.96	962	hypothetical	2200204 - 2201165
9	0.96	495	hypothetical	2186614 - 2186120
10	0.96	796	phage origin, serine/threonine protein phosphatase	3488405 - 3489201
11	0.96	1364	hypothetical, colicin-like & small toxic polypeptide	1397029 - 1398393
12	0.96	987	hypothetical, putative membrane protein	3486570 - 3487557
13	0.95	916	putative serine acetlyltransferase of prophage	2605160 - 2606076
14	0.95	300	hypothetical, potential T3SS effector	2209466 - 2209765
15	0.95	1136	T3SS effector protein NleH	1804974 - 1806122

Table 4: A summary of Stx1-containing *E. coli* signatures generated by Neptune relative to a background of non-toxigenic *E. coli*.

regions do not confound signature discovery. However, if a repeated region is a true signature, then Neptune will report each region as a separate signature. In this circumstance, user curation may be required.

Neptune cannot locate isolated SNVs and small mutations. Any region with a high degree of similarity to the exclusion group will either not produce candidate signatures or be removed during filtering. Neptune is designed to locate general-purpose signatures of arbitrary size and does not consider application-specific physical and chemical properties of signatures. Furthermore, Neptune is not capable of selecting the best substring within a signature region (e.g., an assay-compatible primer). This operation would have the effect of optimizing signature efficacy for applications where smaller signature lengths are desirable. While Neptune is capable of producing signatures as small as the k-mer size, we observed that very short signatures (< 100 bases) may not contain any seed matches with filtering targets during the alignment process, thereby preventing the signature from being evaluated correctly. We recommend either using smaller seed sizes during pairwise alignments, at the expense of significantly increased computation time, or discretion when evaluating very short signatures.

Finally, Neptune makes assumptions about the probabilistic independence of bases and SNV events; while these events do not occur independently in nature, they allow for significant mathematical simplification. Nonetheless, Neptune is capable of producing highly sensitive and specific signatures using these assumptions.

4.4 Biological Implications

This study demonstrates that Neptune can be a very useful tool for the rapid characterization and classification of pathogenic bacteria of public health significance, as it can efficiently discover differential genomic signatures. Although both *L. monocytogenes* 4b and 1/2a serotypes, belonging to lineages I and II respectively, are associated with human illness, lineage I strains are overrepresented among human cases whereas lineage II isolates are widespread in food-related, natural and farm environments. Among LiDS-NG project isolates used in our study, 43% of 4b and 17% of 1/2a serotype isolates had a clinical human host origin. Among the signatures for serotype 1/2a, multiple PTS systems and ABC transport systems were found (Table 2 and supplementary material) which may be correlated to the fact that the presence of a variety of PTS and transport systems provides *L. monocytogenes* serotype 1/2a with a competitive advantage to survive under broad environmental conditions. Among the *L. monocytogenes* 4b serotype signatures genes coding for cell-wall anchor proteins, RHS protein known to be associated with mediating intercellular competition and immunity [11], and cell wall polysaccharides and teichoic acid decoration enzymes were found (Table 3 and supplementary data). Such cell-surface components play a role in bacterial-host interactions [5]. The potential involvement of these genes in the virulence and pathogenesis of serotype 4b should be an interesting area of future inquiry.

Interestingly, two very large, but divergent signature sequences corresponding to the 4b (rank 14; score 0.93; length 12685 nt) and 1/2a (rank 14; score 0.92; length 17698) inclusion groups were found by Neptune (supplementary data). These

contained non-homologous teichoic acid biosynthesis and transport system genes in a genomically equivalent location, indicating the serotype specificity of these signatures and the discriminatory power of Neptune for their identification. The signature information retrieved from Neptune analysis may be very useful for further investigating the association of identified regions with serotypes, virulence and niche-specificity of bacteria under study. Additionally, the signature sequences may be used to develop rapid diagnostic assays, such as oligonucleotide primer sets for high throughput PCR-based screening assays for the identification and characterization of bacterial isolates and are not limited to coding sequences (may include intergenic regions).

5 Conclusion

We demonstrate that Neptune is capable of locating signatures in an artificial data set. While some signatures are reported as two smaller, adjacent signatures, Neptune reports all the expected signature regions. We apply Neptune to a *L. monocytogenes* data set and show that top-scoring Neptune-identified signatures have high *in silico* sensitivity and specificity to a wet-lab verified validation data set. Finally, we employ Neptune to locate pathogen-associated signatures related to Shiga-toxigenic *E. coli* STEC, notably Stx1-encoding strains. Neptune locates many expected signature regions with high confidence. As expected, no top-scoring signatures corresponded to rDNA or housekeeping genes. The signatures found in groups of pathogenic bacteria can also provide an array of gene candidates to further investigate their possible role in pathogenesis. We conclude that Neptune is a powerful and flexible tool for locating signature regions with minimal prior knowledge.

6 Availability

The data used in the manuscript is stored under the PRJNA301341 NCBI Short Read Archive accession. Neptune is developed in Python using DRMAA, NumPy, SciPy, and Biopython libraries. The software requires a standard 64-bit Linux environment. The software is available at: *http://github.com/phac-nml/neptune*

7 Acknowledgements

The authors would like to thank Franklin Bristow and Eric Enns of the NML-PHAC for their feedback on various aspects of the software design and implementation.

References

[1] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[2] Kai Christian Bader, Christian Grothoff, and Harald Meier. Comprehensive and relaxed search for oligonucleotide signatures in hierarchically clustered sequence datasets. *Bioinformatics*, 27(11):1546–1554, 2011.

[3] H Bierne, C Sabet, N Personnic, and P Cossart. Internalins: a complex family of leucine-rich repeat-containing proteins in listeria monocytogenes. *Microbes and Infection*, 9(10):1156–1166, 2007.

[4] Klaus Brehm, María-Teresa Ripio, Jürgen Kreft, and José-Antonio Vázquez-Boland. The bvr locus of listeria monocytogenes mediates virulence gene repression by β-glucosides. *Journal of bacteriology*, 181(16):5024–5032, 1999.

[5] Filipe Carvalho, Sandra Sousa, and Didier Cabanes. How listeria monocytogenes organizes its surface for virulence. *Frontiers in cellular and infection microbiology*, 4, 2014.

[6] Ron Caspi, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A Fulcher, Timothy A Holland, Ingrid M Keseler, Anamika Kothari, Aya Kubo, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 42(D1):D459–D471, 2014.

[7] Andrea Currie, Jeffrey M Farber, Céline Nadon, Davendra Sharma, Yvonne Whitfield, Colette Gaulin, Eleni Galanis, Sadjia Bekal, James Flint, Lorelee Tschetter, et al. Multiprovince listeriosis outbreak linked to contaminated deli meat consumed primarily in institutional settings, canada, 2008. *Foodborne pathogens and disease*, 12(8):645–652, 2015.

[8] Bhavjinder K Dhillon, Matthew R Laird, Julie A Shay, Geoffrey L Winsor, Raymond Lo, Fazmin Nizam, Sheldon K Pereira, Nicholas Waglechner, Andrew G McArthur, Morgan GI Langille, et al. Islandviewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic acids research*, page gkv401, 2015.

[9] Vilim Feller. *An Introduction to Probability Theory and Its Applications: Volume 1*. J. Wiley & sons, 1960.

Contributor	Credit	Contributions
EM	Lead	manuscript; math; software; experiments
RZ	Co	manuscript; biological analysis
CB	Co	listeria experiment design, data; *E. coli* data
KW	Co	manuscript; experiment design
MD	Co	math
PM	Co	background work
NK	Co	background work
AR	Co	listeria experiment design
MG	Co	manuscript; exp, software design; resources
LC	Co	provided listeria data
GVD*	Anchor	manuscript; exp, software design; resources
FB	Ack	software discussions
EE	Ack	software discussions

Table 5: Author contributions. * = Corresponding

[10] Matthew W Gilmour, Morag Graham, Gary Van Domselaar, Shaun Tyler, Heather Kent, Keri M Trout-Yakel, Oscar Larios, Vanessa Allen, Barbara Lee, and Celine Nadon. High-throughput genome sequencing of two listeria monocytogenes clinical isolates during a large foodborne outbreak. *BMC genomics*, 11(1):120, 2010.

[11] Sanna Koskiniemi, James G Lamoureux, Kiel C Nikolakakis, Claire t'Kint de Roodenbeke, Michael D Kaplan, David A Low, and Christopher S Hayes. Rhs proteins from diverse bacteria mediate intercellular competition. *Proceedings of the National Academy of Sciences*, 110(17):7032–7037, 2013.

[12] Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12, 2004.

[13] Hsiao Ping Lee and Tzu-Fang Sheu. An algorithm of discovering signatures from dna databases on a computer cluster. *BMC bioinformatics*, 15(1):339, 2014.

[14] Xiang-He Lei, Franz Fiedler, Zheng Lan, and Sophia Kathariou. A novel serotype-specific gene cassette (glta-gltb) is required for expression of teichoic acid-associated surface antigens in listeria monocytogenes of serotype 4b. *Journal of bacteriology*, 183(4):1133–1139, 2001.

[15] Michael J Linnan, Laurene Mascola, Xiao Dong Lou, Veronique Goulet, Susana May, Carol Salminen, David W Hird, M Lynn Yonekura, Peggy Hayes, Robert Weaver, et al. Epidemic listeriosis associated with mexican-style cheese. *New England Journal of Medicine*, 319(13):823–828, 1988.

[16] Jeffrey T McCollum, Alicia B Cronquist, Benjamin J Silk, Kelly A Jackson, Katherine A O'Connor, Shaun Cosgrove, Joe P Gossack, Susan S Parachini, Neena S Jain, Paul Ettestad, et al. Multistate outbreak of listeriosis associated with cantaloupe. *New England Journal of Medicine*, 369(10):944–953, 2013.

[17] Pall Melsted and Jonathan K Pritchard. Efficient counting of k-mers in dna sequences using a bloom filter. *BMC bioinformatics*, 12(1):333, 2011.

[18] Renato H Orsi, Henk C den Bakker, and Martin Wiedmann. Listeria monocytogenes lineages: Genomics, evolution, ecology, and phenotypic characteristics. *International Journal of Medical Microbiology*, 301(2):79–96, 2011.

[19] Aaron Petkau, Matthew Stuart-Edwards, Paul Stothard, and Gary Van Domselaar. Interactive microbial genome visualization with gview. *Bioinformatics*, 26(24):3125–3126, 2010.

[20] Adam M Phillippy, Kunmi Ayanbule, Nathan J Edwards, and Steven L Salzberg. Insignia: a dna signature search web server for diagnostic assay development. *Nucleic acids research*, page gkp286, 2009.

[21] Ravi Vijaya Satya, Kamal Kumar, Nela Zavaljevski, and Jaques Reifman. A high-throughput pipeline for the design of real-time pcr signatures. *BMC bioinformatics*, 11(1):340, 2010.

[22] Ravi Vijaya Satya, Nela Zavaljevski, Kamal Kumar, and Jaques Reifman. A high-throughput pipeline for designing microarray-

based pathogen diagnostic assays. *Bmc Bioinformatics*, 9(1):185, 2008.

[23] Tom Slezak, Tom Kuczmarski, Linda Ott, Clinton Torres, Dan Medeiros, Jason Smith, Brian Truitt, Nisha Mulakken, Marisa Lam, Elizabeth Vitalis, et al. Comparative genomics tools applied to bioterrorism defence. *Briefings in Bioinformatics*, 4(2):133–149, 2003.

[24] Regina Stoll and Werner Goebel. The major pep-phosphotransferase systems (ptss) for glucose, mannose and cellobiose of listeria monocytogenes, and their significance for extra-and intracellular growth. *Microbiology*, 156(4):1069–1083, 2010.

[25] Chaomei Zhang, Joe Nietfeldt, Min Zhang, and Andrew K Benson. Functional consequences of genome evolution in listeria monocytogenes: the lmo0423 and lmo0422 genes encode σc and lstr, a lineage ii-specific heat shock system. *Journal of bacteriology*, 187(21):7243–7253, 2005.