Furstenau *et al. BMC Bioinformatics* (2018) 19:222
https://doi.org/10.1186/s12859-018-2225-z

BMC Bioinformatics

SOFTWARE Open Access

Variant site strain typer (VaST): efficient strain typing using a minimal number of variant genomic sites

Tara N. Furstenau[1], Jill H. Cocking[1,2], Jason W. Sahl[2] and Viacheslav Y. Fofanov[1,2*]

Abstract

Background: Targeted PCR amplicon sequencing (TAS) techniques provide a sensitive, scalable, and cost-effective way to query and identify closely related bacterial species and strains. Typically, this is accomplished by targeting housekeeping genes that provide resolution down to the family, genera, and sometimes species level. Unfortunately, this level of resolution is not sufficient in many applications where strain-level identification of bacteria is required (biodefense, forensics, clinical diagnostics, and outbreak investigations). Adding more genomic targets will increase the resolution, but the challenge is identifying the appropriate targets. VaST was developed to address this challenge by finding the minimum number of targets that, in combination, achieve maximum strain-level resolution for any strain complex. The final combination of target regions identified by the algorithm produce a unique haplotype for each strain which can be used as a fingerprint for identifying unknown samples in a TAS assay. VaST ensures that the targets have conserved primer regions so that the targets can be amplified in all of the known strains and it also favors the inclusion of targets with basal variants which makes the set more robust when identifying previously unseen strains.

Results: We analyzed VaST's performance using a number of different pathogenic species that are relevant to human disease outbreaks and biodefense. The number of targets required to achieve full resolution ranged from 20 to 88% fewer sites than what would be required in the worst case and most of the resolution is achieved within the first 20 targets. We computationally and experimentally validated one of the VaST panels and found that the targets led to accurate phylogenetic placement of strains, even when the strains were not a part of the original panel design.

Conclusions: VaST is an open source software that, when provided a set of variant sites, can find the minimum number of sites that will provide maximum resolution of a strain complex, and it has many different run-time options that can accommodate a wide range of applications. VaST can be an effective tool in the design of strain identification panels that, when combined with TAS technologies, offer an efficient and inexpensive strain typing protocol.

Keywords: Targeted PCR Amplicon sequencing, Bacterial strain typing, Single nucleotide polymorphisms

Background

High-resolution strain identification is vital in applications ranging from tracking of disease outbreaks and surveillance of virulent or antimicrobial resistant pathogens [1–3] to the investigation of bioterrorism and other crimes [4–6]. One of the most promising methods for molecular-based strain identification is targeted multiplex PCR amplicon sequencing (TAS) using high throughput sequencing (HTS) platforms [7]. From an unknown isolate, targets are amplified together in a multiplexed PCR reaction and sequenced, the sequences are then analyzed and compared to sequences of known isolates for identification. PCR enrichment of target sequences allows TAS to be more cost effective than whole genome sequencing and tolerant to low amounts of starting material [8]. Combining this with HTS technology allows scaled processing of hundreds to thousands of samples on

*Correspondence: Viacheslav.Fofanov@nau.edu
[1]The School of Informatics, Computing, and Cyber Systems, Northern Arizona University, 1295 S Knoles Dr., Flagstaff, Arizona 86001, USA
[2]Pathogen and Microbiome Institute, Northern Arizona University, 1395 S Knoles Dr., Flagstaff, Arizona 86001, USA

a single machine. The challenge is then deciding which targets to choose to achieve the desired outcome.

The targeted sequences have often been either a single housekeeping gene (e.g. the 16S rRNA gene [9]) or in the case of multi-locus sequence typing (MLST), a collection of a few housekeeping or well-conserved genes [10]. The variation within these genes is used to define a well curated set of different sequence types (ST) that distinguish bacterial species or strains. Depending on the amount of diversity, MLST can provide decent resolution and, as HTS techniques are increasingly applied, it is becoming more scaleable and cost-effective [11]. For some applications, however, the resolution from only a few genes can be insufficient, especially for differentiating between closely related or highly clonal variants [12]. When identifying genetic variation that distinguishes specific strains there is not always enough variation found among the established targets.

VaST was designed to find a minimal set of target loci that provide a desired level of resolution across a given strain complex. It can add resolution to an existing MLST assay or it can generate a complete set of targets from scratch when MLST loci have not been established. Either way, the goal of VaST is to provide flexibility and control to the design of specialized strain-typing assays for a number of different applications that can be customized for specific sequencing technologies. This begins with the user defining the level of strain resolution that they desire from the panel. If resolution among a specific group of strains is particularly important, this can be defined and VaST will focus on maximizing resolution for those strains. Next, established targets of variation (such as loci from a MLST assay [10, 13–19] or canonical SNPs [20–33]) can be added as a starting point which will override the VaST optimization function to guarantee their inclusion in the final set. Other targets, such as those associated with virulence or antimicrobial resistance can also be included. VaST will search for additional targets, considering many different types of genetic variation including: single nucleotide polymorphisms (SNPs), microsatellites, variable number tandem repeat (VNTRs), and small insertion/deletions (indels). These targets will be contained within a user-specified amplicon size that is appropriate for the desired sequencing technology. Because the selected targets must be amplifiable across all the strain variants, VaST will pre-filter any target that does not have sufficiently well conserved flanking primer sequences. VaST will identify and add new targets until either maximum resolution is reached, a predetermined resolution level is reached, or a specified number of targets have been identified.

Finding the minimal number of targets to achieve the desired resolution is important because it keeps costs low and it limits the potential for adverse primer interactions during multiplex PCR. Given a set of variable genomic sites to choose from, this task is, in essence, a minimum spanning set problem — the minimum set of genomic features that is capable of uniquely identifying each strain. Naively, one would hope to find a single polymorphic site per strain that uniquely distinguishes it from all other strains. In practice, finding a signature polymorphism for each strain is unlikely and the significance of such a signature may erode when additional strains are considered. Instead, our approach seeks to identify a "haplotype" or a collection of polymorphisms which in concert, provide a composite signature that is unique for any given strain. The resulting set of targets needs to be robust enough to proactively handle the rapid expansion of sequences for new strains that come with the genomic age. For this reason, we believe that the best set of targets should include basal genomic features that are stable across entire clades of strains and allow accurate placement of strains that have not been seen before. Our minimum spanning set algorithm selects each new target site based on its ability to evenly split up groups of unresolved strains. An important aspect of evenly splitting the strain complex at each step is that the early additions to the minimum spanning set tend to be more phylogenetically basal. Due to an abundance of "deep" phylogenetic markers, our approach, as we demonstrate, is very robust to characterizing previously unseen strains.

Several groups have developed approaches for identifying a minimum set of target markers for various purposes. Pan-PCR [34] and the Loci Selector Module of PanSeq [35] are the most *thematically* similar approaches as they both focus on strain typing; however, there are other methods which focus on different problems like finding a minimum set of haplotype tagging Single Nucleotide Polymorphisms (htSNPs) for identifying haplotype blocks [36–40]. The Pan-PCR algorithm uses whole genome sequence data from closely related strains to find a minimum number of gene targets whose presence or absence in a PCR product can be used to distinguish a set of input strains. Primers are designed specifically for each target to ensure that they produce different sized PCR products and the amplified targets are separated in a gel, producing a unique banding pattern that acts as a fingerprint for each of the strains of interest. In contrast, VaST's minimum spanning set algorithm is able to take advantage of variation that exist in both coding and non-coding regions of the genome which provides a larger pool of options for strain differentiation. This is critical when expanding this approach to viral organisms. VaST is also intended to be used in a sequencing-based approach which will maximize the information content of polymorphic sites, making it possible to detect presence of previously unseen strains and to place them within existing phylogenies. The Loci Selector (LS) module of the PanSeq program is

another algorithm which attempts to find loci that offer maximum discriminatory power between certain strains. Like, VaST, the LS module is agnostic with respect to the type of sequence variation that is provided as input. Unlike VaST however, the goal of the LS module is not to find a minimum set of sites that together provide maximum resolution, but rather to find a set (of a provided size) of the most discriminatory loci that have the least amount of overlap. In this case, loci that are "deeper" in the phylogeny are not prioritized because they resolve clades rather than individual strains. The resulting set of targets provides strain resolution but are less robust to correctly placing "new" strains – those not part of the original panel.

In this paper we present the VaST algorithm which computes a minimum set of targets for the purpose of bacterial strain differentiation. We provide benchmarks, computational and experimental validation, and resolution comparisons to the LS module of PanSeq and MLST assays to demonstrate how VaST can help streamline the development of fast, efficient, and cost-effective strain identification assays.

Implementation

VaST is written in Python and is designed to convert a set of genomic features from different strains into a minimum spanning set of targets which will achieve a maximum (or user-defined) level of strain differentiation. The set of genomic features can be identified using a number of available software packages that detect variant sites across a collection of genomes (we utilized NASP, a single nucleotide polymorphism (SNP) detection pipeline [41]). VaST accepts a variant site matrix where each row represents a genomic site that varies across the columns of strains; the values in the matrix characterize the state of each strain at the variable sites (See example in Table 1).

Table 1 SNP Matrix example

LocusID	Strain A	Strain B	Strain C	Strain D	Strain E
genome123::115::115	A	T	A	A	T
genome123::120::120	G	C	G	G	C
genome123::121::121	T	C	C	C	C
genome123::130::130	C	G	G	G	C
genome123::209::209	A	C	C	N	G
genome123::405::405	-	-	X	C	C
genome123::511::541	10	8	8	10	8
⋮	⋮	⋮	⋮	⋮	⋮

The first column of a variant site matrix contains a genome identifier, a start position, and an end position, each separated by two colons. The start and end position should be the same for SNPs. Each additional column represents a strain and the calls made at each variant site for that strain. The first five rows contain SNPs, the sixth row contains an indel with missing data for Strain C, and the last row contains the lengths of VNTRs (the stopping position is based on the longest repeat of 3 in this case)

Many different types of genomic variation can be included in this matrix (SNPs, indels, VNTRs, etc.) provided that the variable region is short enough to be captured in a single amplicon.

VaST is able to correctly interpret variant site matrices that contain missing data and ambiguous base calls; although, such sites can slow down the processing of the matrix. To speed up the preprocessing, VaST can be run in a strict mode which will ignore any site with ambiguous or missing data. By default, missing data is represented by an "X", and deletions are represented by a "-", and VNTRs can be represented by the number of repeats. The only other permissible character states in the matrix are DNA bases and IUPAC ambiguous base codes [42].

To run the Amplicon Filter Module (Fig. 1a), VaST requires information about the regions upstream and downstream of each of the variant sites. Therefore, a full genome matrix must be provided which should include a call for each position in the genome for all of the strains. This matrix can be generated through the alignment of genome assemblies to a reference genome or from Variant Call Format (VCF) files [43] that contain calls for each position in the genome.

Finding candidate amplicons from target sites

It is assumed that the target sites identified by VaST will ultimately be amplified using PCR and sequenced. Therefore, we included an Amplicon Filter Module which treats each variant site as a potential amplicon, combining adjacent sites as necessary, and filters out any amplicons that may be difficult to amplify in all strains.

When multiple variant sites are clustered together, it is more efficient to consider them together as a single amplicon which can be amplified with one pair of primers. The combination of sites in such an amplicon may sometimes provide more strain resolution than any one of the sites individually, and these more efficient amplicons will naturally be favored during the VaST Pattern Selection Module (Fig. 1a). The maximum distance between adjacent variant sites is defined by a window size parameter. The window starts at the position of the first variant site, and the algorithm checks to see if any of the next variant sites are captured within the window. If the window contains only the original site, this single target amplicon will be sent to the filtering step. If the window contains multiple variant sites, as shown in Fig. 1b, then the amplicon containing all of the sites will be sent to the filter. If this multi-target amplicon fails the filter, the last target site in the window will be removed and this modified amplicon will be sent to the filter. This will be repeated until either an amplicon passes the filter or there are no more target sites in the amplicon. Once the options at the first position are exhausted, the window shifts down to the next variant site. It is possible for the same region to be captured in multiple

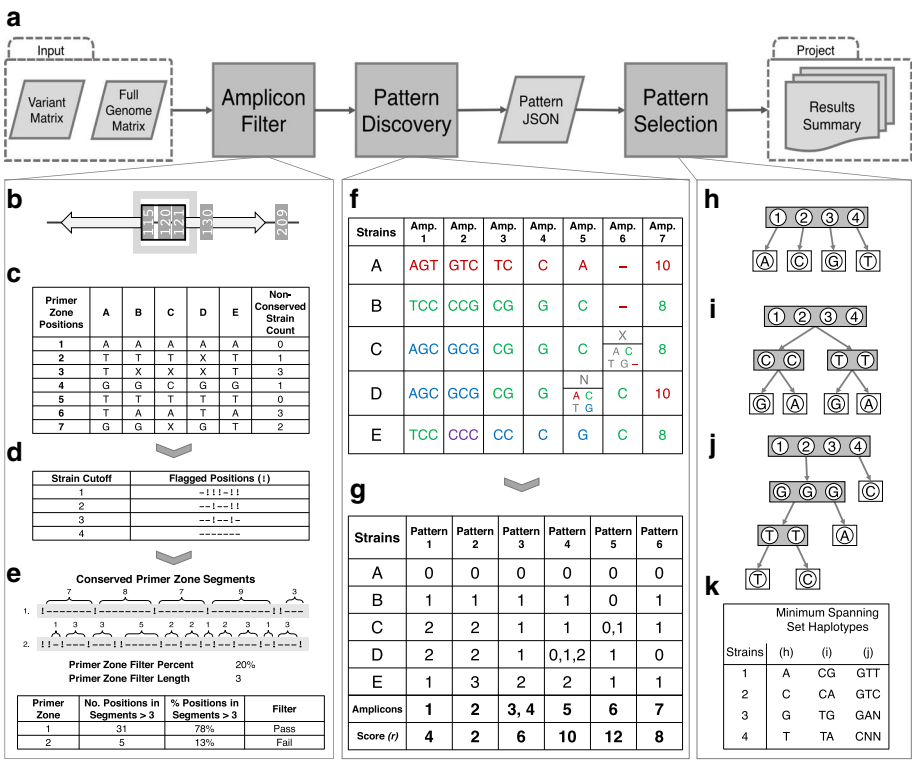

Fig. 1 VaST Pipeline Schematic. **a** Overview of the VaST pipeline. **b** The window (gray box) starts at the first site (115) and captures two additional sites (120 and 121). The amplicon (black box) extends from the first to the last variant site in the window and the primer zones (arrows) extend in opposite directions. **c** The primer zone region is extracted from the full genome matrix and the number of strains that are missing data (X) or have a base call that differs from the reference are counted for each position. **d** A position in the primer zone is flagged (!) when the number of poorly conserved strains is greater than or equal to the strain cutoff value. **e** To pass the filter in this example, 20% of the primer zone positions must be a member of a conserved segment that is longer than three positions. **f** The table shows the variant sequence features of the amplicons **g** The resolution pattern of each amplicon is determined and the amplicons that contain redundant information are combined (e.g. Amplicon 3 & 4 into Pattern 3). For ambiguous (N) or missing calls (X), all of the possibilities are enumerated and the strain simultaneously belongs to all of the feature categories that overlap with those of the other strains. The bottom row is the resolution score, *r*, for each pattern. The minimum spanning set algorithm favors patterns that evenly split up groups of strains. Using SNPs as an example, **h** is the best case scenario where *N* strains can be resolved with $\log_4(N)$ SNPs; however, **i** $\log_2(N)$ is more likely with bi-allelic SNPs. **j** In the worst case, highly unbalanced splitting can occur which can require at most $N-1$ SNPs to resolve *N* strains. **k** The associated haplotypes for each of the minimum spanning sets in (**h**-**j**)

amplicons so VaST will avoid choosing overlapping amplicons in the final solution. Customizing window lengths allows VaST to be optimized for a wide range of sequencing platforms, which vary widely in the lengths of genomic sequences that can be produced.

To amplify the target sites in a PCR, primers must be designed to anneal in the regions upstream and downstream of the target. If a single set of primers is to be designed that will amplify the target across all of the strains, the primer region must be well conserved. While VaST does not attempt to design the primers themselves, it does consider the conservation of the upstream and downstream primer regions and filters out targets that contain too much variation. During the filtering step, the proposed upstream and downstream PCR primer zones are analyzed and if they contain too much variation between the known strains (based on the number of strains with an alternative allele), or if there are too many strains with missing data, the amplicon is removed from consideration. This ensures that any remaining target sites will have highly conserved primer zones, and thus, have many options for primer design. The cutoffs for acceptable amounts of variation and number of missing strains are user-defined.

More specifically, amplicon filtering is determined by a number of user-provided parameters: the size of the primer zone, a strain cutoff, a primer zone filter percent, and a primer zone filter length. For each amplicon, the base calls for the upstream and downstream primer zone are retrieved from the full genome matrix (Fig. 1c). For each position in the primer zone, the number of strains with a variation or with missing data are counted and, if the count is greater than or equal to the strain cutoff, the position is flagged (Fig. 1d). The segments of the primer

zone that are not interrupted by flagged positions are highly conserved and are appropriate for primer design (Fig. 1e). However, in order to pass the filter, a certain percent (primer zone filter percent) of the primer zone positions must be present in segments that are longer than the primer zone filter length. This ensures that the conserved sections of the primer zone are long and contiguous. The primer zone filter is applied separately to the upstream and downstream primer zones, and both zones must pass the filter in order for the amplicon to remain. Table 2 provides a summary of the parameters required for the Amplicon Filter Module.

Characterizing the discriminatory power of candidate amplicons

A resolution pattern is calculated for each amplicon after it passes the amplicon filter. The resolution pattern describes which strains share the same features for a given amplicon (Fig. 1f). The Pattern Discovery Module maps the vector of strain features, **q**, for each amplicon to a pattern vector, **p**, which contains sets denoting the membership of each strain in a unique feature category (Eq. 1 and Fig. 1g). Strains will typically belong to a single feature category but they may belong to multiple categories when they have ambiguous or missing base calls at the target sites within the amplicon (Fig. 1g, Pattern 4, Strain D). When operating under strict mode, the algorithm can assume that there are no missing or ambiguous calls and Eq. 1 simplifies to Eq. 2.

$$\begin{aligned} \mathbf{q} &= [s_1, s_2, \ldots, s_n]\,; \text{ where } s \text{ is the set of feature states for each of the } n \text{ strains} \\ \mathbf{p} &= \left[f(s_1), f(s_2), \ldots, f(s_n)\right] \\ f\left(s; a = \{\mathbf{q} : |s_i| = 1\}\right) &= \begin{cases} g(s;a) & \text{if } g(s;a) \neq \emptyset \\ f(s; a \cup s) & \text{otherwise} \end{cases} \\ g\left(s; a\right) &= \{i\colon a_i \cap s \neq \emptyset\} \end{aligned} \tag{1}$$

Assuming there are no missing or ambiguous calls, Eq. 1 simplifies to:

$$f_s\left(s; a = \{\mathbf{q}\}\right) \mapsto \{i\colon a_i \in s\} \tag{2}$$

Despite differences in the specific sequence information of each amplicon, many amplicons will contain redundant strain differentiating information (e.g. Fig. 1f, Amplicon 3 & Amplicon 4). Therefore, instead of storing all of the amplicons individually, they are grouped together based on their strain resolution pattern (Fig. 1g, Pattern 3). Each of these patterns along with the start and stop positions of their associated amplicons are saved in a JSON file that can be passed repeatedly to the Pattern Selection Module without rerunning the preprocessing steps.

Table 2 Amplicon Filter Module parameter descriptions and considerations

Parameter	Description	Notes
Strict mode	VaST ignores missing or ambiguous data in input matrix	Speeds up preprocessing but some sites are lost
Window size	Maximum distance between adjacent sites that can be combined into a single amplicon	The desired amplicon length should be considered when setting the window size. A larger window may increase the number of variant sites that are included in the amplicons making them more efficient
Primer zone size	Size of the region upstream and downstream of the target to evaluate in the amplicon filter	The primer zones begin immediately before the first and immediately after the last target site in the window, so the maximum amplicon size is 2 × primer zone size + window size. A smaller primer zone may limit the number of primer options.
Strain Cutoff	The number of strains at a primer zone site that can have a non-conserved call before the site is flagged.	A strain cutoff greater than one will not guarantee that the primer zone sequences are conserved across all of the strains but it may be appropriate in cases where one or a few strains have low sequence coverage
Primer zone filter percent	The percent of primer zone positions that must be present in un-flagged segments of the primer zone that are longer than the primer zone filter length.	A higher primer zone filter percent will increase the total number of primer options in amplicons that pass the filter
Primer zone filter length	The length of un-flagged primer zone segments that count toward the primer zone filter percent	The primer zone filter length should be at least as long as the minimum acceptable primer length to ensure that conserved primers can be found within the primer zone

Constructing the minimal set of targets

The primary goal of the Pattern Selection Module is to find a minimum spanning set, which we define as the minimum number of patterns that are required to achieve maximum strain resolution. A naive brute-force approach to solving for the minimum spanning set requires an exhaustive search of all possible subsets of variant sites, starting from size 1 to N where N is the size of the minimum spanning set. In the worst case, this approach has exponential complexity ($\mathcal{O}(2^n)$), which quickly becomes an intractable problem even for relatively small sets of variant sites. For example, given a set V of 1,000 variant sites, the size of the search space, $|S|$, that is required to find a minimum spanning set of size 50 is on the order of 10^{85} combinations — more than the estimated number of atoms in the universe. For reference, a typical SNP matrix for a well-studied bacterial strain complex contains 10-30 thousand SNPs.

$$|S| = \sum_{k=1}^{N} \frac{|V|!}{k!\,(|V|-k)\,!}; \text{ where } V \text{ is the set of variant sites and } N \text{ is the size of the first minimum spanning set.} \tag{3}$$

Because a brute-force approach is intractable, we take a greedy approach which does not guarantee that the absolute minimum spanning set will be found but it will find a locally-optimal, minimized spanning set in a reasonable amount of time. The minimum spanning set algorithm implemented in VaST takes advantage of the exponential increase in discriminatory power with each additional pattern that is added to the set. For example, a single SNP can differentiate at most three strains because there are 4 DNA bases and at least one of the variants must be repeated for any group of more than four strains. When two SNPs are combined into a haplotype the number of possible combinations increases to 16, and a maximum of 15 strains may be uniquely identified. The discriminatory power increases exponentially at $4^n - 1$ where n is the number of SNPs in the haplotype. In contrast, binary variant (presence/absence or wild-type/mutant) approaches (c.p. [34]) can achieve a maximum discriminatory power of only $2^n - 1$.

For SNPs, the theoretical minimum spanning set requires $\log_4(N)$ SNPs to resolve N strains (Fig. 1h). To achieve this minimum, each SNP must contain all four allelic variants and the variants must evenly split up each group of unresolved strains. In practice, many SNPs are only bi- or tri-allelic so a more realistic minimum would be $\log_2(N)$ which may still be difficult to achieve when working with a limited set of available patterns (Fig. 1i). In the worst case, each SNP is only able to differentiate a single strain which causes highly uneven splitting and can require up to $N - 1$ SNPs (Fig. 1j).

In order to get as close as possible to the minimum number of variant sites, VaST favors the addition of sites that do the best job of evenly splitting up the most remaining groups of unresolved strains. In practice, this predisposes VaST to prefer at least some phylogenetically basal variants in its solutions (stable variants that occurred sufficiently far in the organism's past to be established in multiple clades' lineages). This confers significant advantages when encountering previously unobserved strains. More specifically, the algorithm iteratively incorporates patterns into the set by choosing the pattern that provides the greatest reduction in the set resolution score, r, (Eq. 4, Fig. 1g, bottom row). Before any sites are added, each value in the minimum spanning set pattern vector is zero because all of the strains are members of the same null haplotype category. The resolution score is also set to the maximum value of $N\,(N-1)$ where N is the number of strains. At the beginning, a resolution score is also calculated for each of the amplicon pattern vectors and they are sorted from lowest (best) to highest (worst). Due to the nature of greedy algorithms, it is likely that pattern choices that are locked in the early stages can lead to a suboptimal solution. Therefore, a number of the top patterns from the sorted list can be selected to seed several distinct, independently-built sets and the best solution will be returned at the end.

When the first pattern is added, the minimum spanning set pattern vector is updated (Eqs. 5 or 6 in strict mode), the resolution score is recalculated and the selected pattern is removed from further consideration. The remaining pattern vectors are then updated so they reflect their resolution combined with the resolution of the current minimum spanning set (Eqs. 5 or 6) and their scores are recalculated (Eq. 4). The pattern with the best score is then added to the minimum spanning set. Patterns are continually added in this manner until (1) full resolution is reached at which point each strain will have a unique haplotype and the set resolution score is zero; (2) when none of the remaining patterns are able to improve the current resolution of the set; (3) when some predefined number of sites or resolution threshold is reached; (4) no more patterns remain.

$$r = \sum_{i=0}^{\max(\mathbf{p})} s_i^2 - s_i; \tag{4}$$

where $\mathbf{p}$ is a pattern vector and s_i is the number of strains in the i^{th} feature category.

$$\mathbf{p}_{\text{update}} = \left[f(p_{t1} \times p_{s1}), f(p_{t2} \times p_{s2}), \ldots, f(p_{tn} \times p_{sn})\right];$$

where $p_{ti} \times p_{si}$ is the cartesian product between sets in a pattern vector, $\mathbf{p_t}$, and the current minimum spanning set pattern vector, $\mathbf{p_s}$.

$$a = \left\{p_{ti} \times p_{si} \forall i \in \{1, 2, \ldots, n\} \colon |p_{ti} \times p_{si}| = 1\right\}$$

$$f(p_t \times p_s; a) = \begin{cases} g(p_t \times p_s; a) & \text{if } g(p_t \times p_s; a) \neq \emptyset \\ f(p_t \times p_s; a \cup (p_t \times p_s)) & \text{otherwise} \end{cases} \quad (5)$$

Assuming there are no missing or ambiguous calls, Eq. 5 simplifies to:

$$f_s(p_t \times p_s; a = \{p_{ti} \times p_{si} \forall i \in \{1, 2, \ldots, n\}\}) \mapsto \{i \colon a_i \in (p_t \times p_s)\} \quad (6)$$

If multiple patterns tie for the best score, the one that is further up in the original sorted list is chosen because it will provide the greatest redundancy in the final set of patterns. This is due to the fact that higher ranking patterns offer more diversity, and therefore are more likely to complement other patterns in the set and partially compensate for them if they are missing. This added tolerance is beneficial because some of the targets might not be successfully amplified and sequenced.

As patterns are added, their associated amplicons are checked for overlap with the amplicons that are already included in the set. If a conflict cannot be resolved by removing one of the amplicons, then the new pattern is skipped and the pattern with the next best score is added and checked.

Customizing the VaST workflow

Several user-defined parameters change the way the Pattern Selection Module handles the input data. Certain strains that are included in the preprocessing step can be marked for removal and will therefore not be considered in determining the final resolution. Lists of variant sites can be flagged either for removal or for mandatory inclusion in the final set. By default, VaST attempts to achieve maximum strain resolution; however, there are settings which will force VaST to stop once a certain number of amplicons have been added or when a resolution threshold has been met. Finally, an additional input array may be supplied which defines an alternative resolution objective. By default, VaST will not prioritize the resolution of any particular strains. If an alternative resolution objective is provided, VaST will favor patterns that help attain the alternative resolution before attempting full resolution. Alternative resolution objectives are useful when it is more critical to resolve certain strains over others. To summarize, VaST can be run using any of the following workflow options: the full workflow which provides full strain resolution using any of the amplicon candidates, the abridged workflow which stops once a user-specified number of amplicons are added or a resolution threshold is met, the weighted workflow which prioritizes the resolution of certain groups of strains using an alternative resolution objective, and the set extension workflow which appends to an existing set of targets.

Results

Benchmarking

We benchmarked VaST's performance using 6 bacterial strain complexes: 537 strains of *Escherichia coli* using 189,570 SNPs, 373 strains of *Burkholderia pseudomallei* using 94,647 SNPs, 269 strains of *Yersinia pestis* using 11,249 SNPs, 186 strains of *Bacillus anthracis* using 11,989 SNPs, 64 strains of *Francisella tularensis* using 16,720 SNPs, and 122 strains of *Staphylococcus aureus* using 169,382 SNPs. These pathogens were chosen based on their relevance to human disease outbreaks and their potential for use as biothreat agents. The strains we used were drawn from previously published and well-established strain complexes [44–47]. We generated minimum spanning sets for each strain complex to demonstrate how well VaST performs in a number of genomic contexts. The *E. coli* minimum spanning set was the most efficient by resolving all 537 strains with only 69 amplicons which is 88% fewer than the number required in the worst case (dotted gray line in Fig. 2). For the other species, the number of required sites was relatively higher, providing only a 66%, 52%, 32%, 22%, and 17% reduction in the number of required sites over the worst case for *B. pseudomallei*, *Staphylococcus aureus*, *Y. pestis*, *B. anthracis*, and *F. tularensis*, respectively. The resolution index — the difference between the number of strains and the average unresolved group size — increases dramatically within the first few sites which suggests that most of the resolution is achieved early on, generally within the first 20 sites for the species we tested. The remaining sites typically resolve only a couple of strains each.

The haplotype-based approach to building a minimum spanning set (as opposed to using a single unique marker to identify each strain) adds a large amount of redundancy. For example, no matter how early in the set a strain is resolved, its haplotype will still consist of all the target sites (e.g. Fig. 1j, strain 4). Similarly, if two strains are not resolved until the last site, all of the previous sites are redundant and do not provide any useful information for resolving the two strains (e.g. Fig. 1j, strains 1 & 2). All of this redundancy is useful because it makes the set more robust to missing targets. This is evident in Fig. 3 which shows how tolerant the *Y. pestis* minimum spanning set is to an increasing number of missing sites. Even when different combinations of 20 sites are missing, the median resolution index is 267.9 which is only slightly lower than the maximum resolution index of 269.

The entire VaST pipeline can be run on a laptop computer. The preprocessing modules (Amplicon Filter and Pattern Discovery) require the most computing resources,

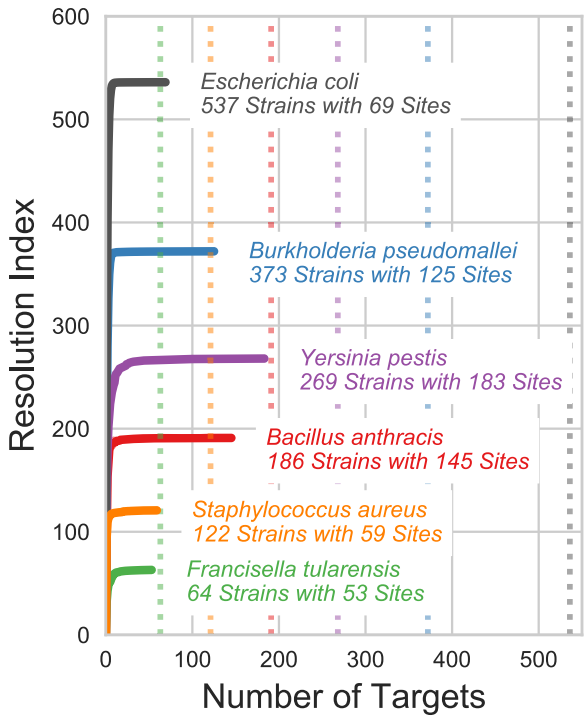

Fig. 2 Most of the resolution is achieved within the first few targets. Minimum spanning sets were generated for strains of *Bacillus anthracis*, *Burkholderia pseudomallei*, *Escherichia coli*, *Francisella tularensis*, *Staphylococcus aureus*, and *Yersinia pestis*. The plot shows how the resolution index ($N_{strains}$ − average group size ± *SD*) increases with each additional site.The number of differentiable strains included in the panel design and the size of the minimum spanning set is indicated next to each plot. The dashed vertical lines indicate the number of sites expected in the worst-case ($N - 1$ sites)

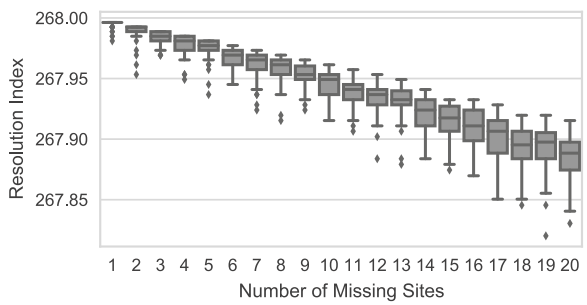

Fig. 3 The redundancy built into the minimum spanning set design makes it tolerant to missing sites. The plot shows how well the *Yersinia pestis* minimum spanning set tolerates missing sites. The x-axis is the number of missing sites and the y-axis is the expected resolution index. Each box-plot shows the distribution of resolution values for different panels ($N = 50$) with 1 to 20 sites randomly removed. The resolution index of the full panel is 269 and the median resolution when 20 sites are missing is 267.9 — a difference of only 1.1

but the amount of time and memory required is highly dependent on the size of the initial variant site matrix and whether or not strict mode is activated. As an example, using a single core of a laptop with a 2.4 GHz Intel Core i5 processor and 8GB of RAM, the preprocessing for the *Y. pestis* data set took approximately 4 hours. If more computing resources are available, VaST can use multiprocessing to speed up the preprocessing steps. The Pattern Selection module runs relatively quickly, and took under an hour for the *Y. pestis* data.

Computational validation

We tested the performance of the full *Y. pestis* minimum spanning set using publicly available HTS data from NCBI's Sequence Read Archive. We aligned reads generated from five different strains (Harbin35 (SRR1283952) [48], Pestoides B (SRR2177700) [49], Angola (SRR2153449) [50], Antiqua (SRR2176134) [51] from [52], and KIM10 (SRR2084698) [53] from [54]) to a reference genome (NC_003143.1 [55]) using bowtie2 [56] and analyzed the calls at each of the target locations. In all five cases the haplotype collected from the sequencing data matched the expected strain.

Sometimes samples will contain strains that were not a part of the original target panel design. To see how well the panel can perform when identifying such samples, we redesigned the *Y. pestis* panel after removing 5 of the original strains. The new panel required 176 sites to achieve full resolution and the removed strains were treated as if they were samples of new strains. Using the calls at the 176 target sites, we identified the strains that were most closely related to the sample strains based on how many of the calls matched. In each case, the strain that was the best match was also very closely related in the phylogentic tree (based on patristic distance) and the size of the clade that included both strains was small (Table 3).

Comparison to other methods

We compared the resolution achieved using VaST to the Loci Selector module of Panseq [35] to demonstrate how our approach is different. Using a matrix of 96 SNPs identified from *E. coli* O157:H7 [57], the LS module identified a collection of 20 SNPs that each individually offered the best discrimination for unique sets of strains. Combined, these 20 SNPs completely resolved 12 of the 19 strains, leaving a group of 7 unresolved strains. However, only 7 of the identified sites increased the resolution and the remaining 13 provided only redundant information. Because VaST prioritizes targets that evenly split up groups of strains rather than finding the most discriminatory targets at each step, it was able to completely resolve 13 strains (with a group of 6 remaining) using 6 sites. As the number of strains considered increases, we would expect an even larger improvement in performance.

Table 3 New strains that were not used to build the minimum spanning set are identified as closely related strains

Assembly accession	Strain name	Patristic distance	In same clade	Clade size
GCA_000255875.1 [61]	Biovar Orientalis AS200901434	2	Yes	3
GCA_000186725.1 [62]	Biovar Medievalis Harbin 35	7	Yes	2
GCA_000182545.1 [63]	Pestoides A	1	Yes	3
GCA_000006645.1 [64]	KIM10	9	Yes	2
GCA_000013805.1 [65]	Nepal516	10	Yes	3

The *Y. pestis* minimum spanning set was regenerated with 5 of the original strains removed. These strains were then treated as samples and identified using the new minimum spanning set. In each case, the strain that most closely matched the sample strain's haplotype was closely related. The table shows the assembly accession and name of each of the strains that were removed. The patristic distance between the sample strain and the strain it was identified as was calculated using the full tree. The clade size is the size of the clade that included both strains

We also compared the strain resolution achieved with VaST to that of a traditional MLST assay using a total of 159 *S. aureus* whole genome sequences from the NCBI RefSeq database. Using these sequences, we generated a SNP matrix using NASP [41] and identified the ST from 7 housekeeping genes (*arcC, aroE, glpF, gmk, pta, tpi,* and *yqiL*) using an open-source MLST program (https://github.com/tseemann/mlst). A total of 41 different groups were resolved using MLST genes, with group sizes ranging from a single strain ($n = 20$) to 44 strains and a mean size of 4.0. Using a total of 59 amplicons, VaST resolved 138 groups, with group sizes ranging from a single strain ($n = 122$) to 8 strains and a mean size of 1.2. Figure 4 compares the resolution and it is clear that the VaST targets can resolve strains within very closely related groups.

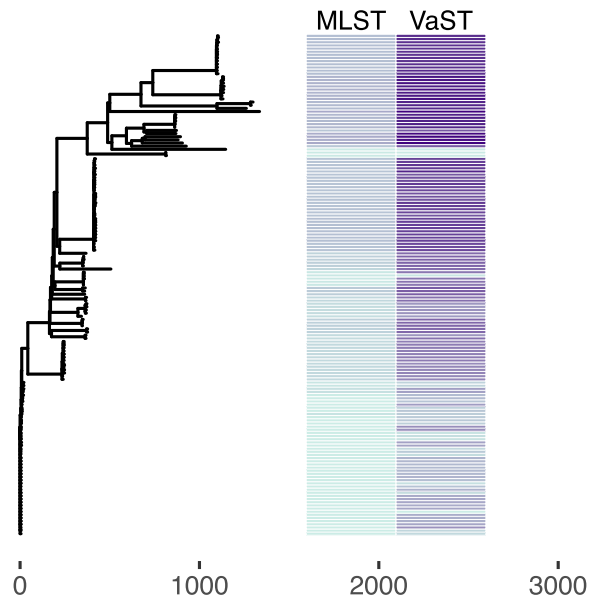

Fig. 4 VaST identifies more targets than a traditional MLST and provides greater strain resolution. The neighbor joining tree was built using 5,000 SNPs from 159 strains of *Staphylococcus aureus*. The colors in the heatmap represent different strain groups ranging from 1-138. The MLST loci only resolved 41 groups as indicated by the smaller range of colors compared to VaST which resolved 138 groups

Experimental validation

We experimentally validated the *Y. pestis* minimum spanning set that VaST produced by performing a TAS assay. Due to the challenges associated with optimizing a multiplex PCR reaction for a large number of targets, we opted to use a truncated version of the panel which included only the first 42 amplicons. This truncated panel had a slightly lower resolution index (266.1 compared to 269 for the full panel) but it was able to resolve most of the major clades. Table 4 shows the number of unresolved groups of different sizes which were used to calculate the resolution index for the truncated panel. Using only 42 of the 183 sites, 38 strains can be uniquely identified (group size 1). The largest unresolved group consisted of 20 very similar biovar Orientalis strains that were all isolated from

Table 4 Resolution of truncated Yersenia pestis minimum spanning set

Group size	Count
1	38
2	18
3	9
4	6
5	5
6	2
7	6
8	1
11	1
12	1
15	1
20	1

The table shows the expected resolution using only the first 42 of the 183-site *Y. pestis* minimum spanning set. The group size indicates a number of strains that could not be differentiated from one another and the count is how many groups of each size exist. A total of 28 strains were fully resolved and the largest group contained 20 unresolved strains

rodents in Peru. The median group size is 5 so at least half of the strains are in groups of 5 or smaller.

The targets of the truncated minimum spanning set were amplified in sample DNA from six different *Y. pestis* strains (Pestoides A, Pestoides F, KIM10, Harbin35, Nepal515, and Antiqua) and the amplicons were sequenced. The calls made at each of the target sites placed every sample strain within the correct clade (Fig. 5). In each case, the maximum resolution expected for the minimum spanning set was achieved.

Discussion

We have developed, benchmarked, and tested a desktop-compatible pipeline which identifies a minimum set of targets that are appropriate for bacterial strain identification. We anticipate that this software will aid in the design of customized, high-resolution typing assays that will be useful for forensic and epidemiological applications, or even for identifying and maintaining laboratory stocks of bacterial isolates. The minimum spanning algorithm implemented in VaST optimizes a combinatorially complex problem in a minimal amount of time even on a desktop computer. The haplotypes produced by VaST provide built-in redundancy which allows the panel to tolerate the likely failure of some amplicons without sacrificing much resolution. The many different run-time options available in VaST provide flexibility to accommodate many different situations. When some strains have particularly low coverage (lots of missing or ambiguous sites), turning off strict mode will open up many more target options for better results. On the other hand, when there is fairly even coverage across the strains, enabling strict mode will speed up the preprocessing steps. The set extension workflow can easily extend existing panels when additional strains or clades are identified or sequenced.

Compared to other strain typing methods, VaST offers a several advantages. Unlike the Pan-PCR method [34], VaST is able to take advantage of variation that exists in both coding and non-coding regions of the genome which provides a larger pool of options for strain differentiation. This is critical when expanding this approach to viral organisms. As a sequencing based approach, opposed to presence/absence detection, VaST is also able to maximize the information content of polymorphic sites, which makes it possible to detect the presence of previously unseen strains and place them within existing

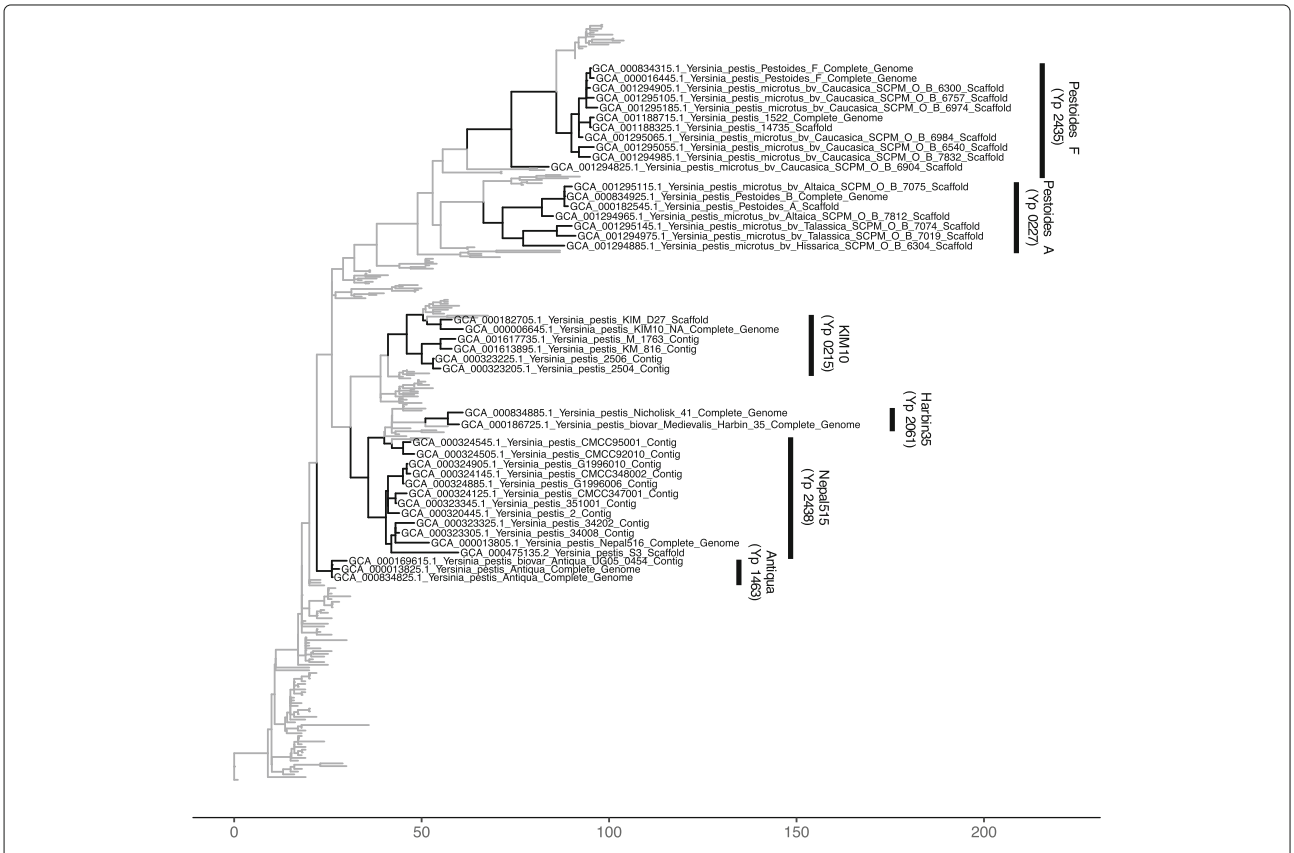

Fig. 5 The *Y. pestis* samples were correctly identified using the target sites identified by VaST. The placement and resolution of the sample strains on a neighbor joining tree produced using the full SNP matrix (11,249 SNPs). The group of strains indicated for each sample represent the strains that were most similar to the sample strain at each of the targets analyzed in the truncated panel. The branch lengths indicate the number of SNP differences

phylogenies. A failed target amplification in the Pan-PCR assay can easily corrupt the expected presence/absence signal and lead to a complete mis-characterization of a strain sample. In a VaST panel, the failure of certain targets will reduce resolution but will not result in a mis-identified strain.

The LS Module of PanSeq focuses on finding the variant sites that offer the most discriminatory power and thus it does not prioritize the addition of variants that are deeper in the phylogeny, as they resolve clades rather than individual strains. The resulting set of targets will therefore be less robust when new strains are introduced that were not a part of the panel design process. In contrast, VaST prioritizes sites that evenly split strain complexes at each step so that the early additions to the minimum spanning set tend to be more phylogenetically basal — stable variation that occurred earlier in the evolution of the organism. In essence, this approach seeks to resolve the full phylogeny, rather than just the leafs of the species tree. As a result, an important feature of VaST is its ability to characterize previously unseen strains, due to abundance of "deep" phylogenetic variants. This was demonstrated in our computation simulations which consistently place strains that were not included in the design of the panel into the correct clade with their most closely related neighbors.

Finally, over the last 20 years, a number of well validated variant markers and MLST profiles have been proposed for the purpose of identifying bacterial clades, particularly for identifying strains that are important in the biodefense sector and clinically relevant strains [30, 45, 58–60]. Using the information from previously established markers, VaST can add targets that are specifically designed to improve resolution, in a user-defined way, starting from the resolution provided by these markers. This allows for backwards compatibility and consistency with previous work thus avoiding the need to repeat the validation of well-established markers.

Conclusions

Fine-scale resolution of bacterial strains is vital when narrowing down potential sources of a pathogen in forensic investigations, providing an accurate prognosis when diagnosing an infection, and establishing the transmission pattern of an infectious strain outbreak. As more and more strains are being identified and sequenced, it is important to be able to rapidly design, implement, and update strain identification panels. Strain typing using TAS technology can provide high resolution (hundreds or thousands of targets can be run simultaneously), scalability (many samples can be processed in a single sequencing run), and sensitivity (PCR amplification allows samples to be identified using small amounts of DNA). Using the ever-growing collection of variant sites identified through whole genome sequencing, VaST provides a tool which will automate the task of finding efficient strain typing markers for use in TAS panels.

Availability and requirements

Project Name: VaST
Project Home Page:
https://github.com/FofanovLab/VaST.git
Operating system(s): Platform independent
Programming language: Python
Other requirements: Anaconda (to use virtual environment)
License: MIT License

Abbreviations

HTS: High-Throughput Sequencing; Indel: Insertion or deletion; PCR: Polymerase Chain Reaction; SNP: Single Nucleotide Polymorphism; TAS: Targeted PCR amplicon sequencing; VNTR: Variable Number Tandem Repeat

Funding

This work was funded by the Department of Homeland Security, Homeland Security Advanced Research Projects Agency, Chemical Biological Division under contract number HSHQDC-16-C-B0031.

Availability of data and materials

The SNP matrices and identified targets are available in the FigShare repository at https://doi.org/10.6084/m9.figshare.5536744.v1 (DOI: 10.6084/m9.figshare.5536744.v1). The sequencing reads used for the computational validation are publicly available in the National Center for Biotechnology Information (NCBI) Sequence Read Archive at https://www.ncbi.nlm.nih.gov/sra. [48–51]. The *S. aureus* whole genome assembly sequences are publicly available in the NCBI RefSeq database and the GCF accession IDs are listed in the FigShare repository. The primers and sequencing reads used in the experimental validation of the *Y. pestis* panel are available from the corresponding author on reasonable request. The VaST source code is available at https://github.com/FofanovLab/VaST.git and has been archived at https://doi.org/10.5281/zenodo.1036007 (DOI: 10.5281/zenodo.1036007).

Authors' contributions

TNF wrote the program, and performed the computational validations and analysis. JHC performed the amplicon sequencing assay. JWS provided the SNP matrices for the strain complexes and the *Y. pestis* DNA. TNF and VYF conceived and designed the algorithm and experiments and wrote the manuscript. JHC and JWS critically reviewed the manuscript. All authors read and approved the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

TNF, JWS, and VYF declare that they have applied for a patent for the truncated *Y. pestis* primer panel.

Publisher's Note

Received: 6 November 2017 Accepted: 30 May 2018
Published online: 11 June 2018

References

1. Brzuszkiewicz E, Thürmer A, Schuldes J, Leimbach A, Liesegang H, Meyer F, et al. Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). Arch Microbiol. 2011;193(12):883–91. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3219860/.

2. Deng X, den Bakker HC, Hendriksen RS. Genomic Epidemiology: Whole-Genome-Sequencing Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. Annu Rev Food Sci Technol. 2016;7(1):353–74. PMID: 26772415 Available from: https://doi.org/10.1146/annurev-food-041715-033259.
3. Pires dos Santos T, Damborg P, Moodley A, Guardabassi L. Systematic Review on Global Epidemiology of Methicillin-Resistant *Staphylococcus pseudintermedius*: Inference of Population Structure from Multilocus Sequence Typing Data. Front Microbiol. 2016;7:1599. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2016.01599.
4. Rasko D, Worsham P, Abshire T, Stanley S, Bannan J, Wilson M, et al. *Bacillus anthracis* comparative genome analysis in support of the Amerithrax investigation. Proc Natl Acad Sci U S A. 2011;108:5027–32.
5. Schmedes SE, Sajantila A, Budowle B. Expansion of Microbial Forensics. J Clin Microbiol. 2016;54(8):1964–74. Available from: http://jcm.asm.org/content/54/8/1964.abstract.
6. Yang R, Keim P. Microbial forensics: A powerful tool for pursuing bioterrorism perpetrators and the need for an international database. J Bioterr Biodef. 2012;S3:007. https://doi.org/10.4172/2157-2526.S3-007.
7. Bybee SM, Bracken-Grissom H, Haynes BD, Hermansen RA, Byers RL, Clement MJ, et al. Targeted Amplicon Sequencing (TAS): A Scalable Next-Gen Approach to Multilocus, Multitaxa Phylogenetics. Genome Biol Evol. 2011;01(3):1312–23. Available from: https://doi.org/10.1093/gbe/evr106.
8. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;01(7):111–8. Available from: https://doi.org/10.1038/nmeth.1419.
9. Weisburg WG, Barns SM, Pelletier DA, Lane DJ. 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol. 1991;173:697–703. Available from: http://jb.asm.org/content/173/2/697.abstract.
10. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci. 1998;95(6):3140–5. Available from: http://www.pnas.org/content/95/6/3140.
11. Boers SA, van der Reijden WA, Jansen R. High-Throughput Multilocus Sequence Typing: Bringing Molecular Typing to the Next Level. PLoS ONE. 2012;7(7):1–8. Available from: https://doi.org/10.1371/journal.pone.0039630.
12. Fournier P, Dubourg G, Raoult D. Clinical detection and characterization of bacterial pathogens in the genomics era. Genome Med. 2014;6(11):114. Available from: https://doi.org/10.1186/s13073-014-0114-2.
13. Bartual SG, Seifert H, Hippler C, Luzon MAD, Wisplinghoff H, Rodríguez-Valera F. Development of a Multilocus Sequence Typing Scheme for Characterization of Clinical Isolates of *Acinetobacter baumannii*. J Clin Microbiol. 2005;43(9):4382–90. Available from: http://jcm.asm.org/content/43/9/4382.abstract.
14. Blanchard AM, Jolley KA, Maiden MCJ, Coffey TJ, Maboni G, Staley CE, et al. The Applied Development of a Tiered Multilocus Sequence Typing (MLST) Scheme for *Dichelobacter nodosus*. Front Microbiol. 2018;9:551. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2018.00551.
15. Boonsilp S, Thaipadungpanit J, Amornchai P, Wuthiekanun V, Bailey MS, Holden MTG, et al. A Single Multilocus Sequence Typing (MLST) Scheme for Seven Pathogenic *Leptospira* Species. PLoS Negl Trop Dis. 2013;7(1):1–10. Available from: https://doi.org/10.1371/journal.pntd.0001954.
16. Curran B, Jonas D, Grundmann H, Pitt T, Dowson CG. Development of a Multilocus Sequence Typing Scheme for the Opportunistic Pathogen *Pseudomonas aeruginosa*. J Clin Microbiol. 2004;42(12):5644–9. Available from: http://jcm.asm.org/content/42/12/5644.abstract.
17. King SJ, Leigh JA, Heath PJ, Luque I, Tarradas C, Dowson CG, et al. Development of a Multilocus Sequence Typing Scheme for the Pig Pathogen *Streptococcus suis*: Identification of Virulent Clones and Potential Capsular Serotype Exchange. J Clin Microbiol. 2002;40(10):3671–80. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC130843/.
18. Shibata Y, Tien LHT, Nomoto R, Osawa R. Development of a multilocus sequence typing scheme for *Streptococcus gallolyticus*. Microbiology. 2014;160(1):113–22. Available from: http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0%.071605-0.
19. Woo PC, Teng JL, Tsang AK, Tse H, Tsang VY, Chan KM, et al. Development of a multi-locus sequence typing scheme for *Laribacter hongkongensis*, a novel bacterium associated with freshwater fish-borne gastroenteritis and traveler's diarrhea. BMC Microbiol. 2009;9(1):21. Available from: https://doi.org/10.1186/1471-2180-9-21.
20. Chanturia G, Birdsell DN, Kekelidze M, Zhgenti E, Babuadze G, Tsertsvadze N, et al. Phylogeography of *Francisella tularensis* subspecies *holarctica* from the country of Georgia. BMC Microbiol. 2011;11(1):139. Available from: https://doi.org/10.1186/1471-2180-11-139.
21. Griffing SM, MacCannell DR, Schmidtke AJ, Freeman MM, Hyytiä-Trees E, Gerner-Smidt P, et al. Canonical Single Nucleotide Polymorphisms (SNPs) for High-Resolution Subtyping of Shiga-Toxin Producing *Escherichia coli* (STEC) O157:H7. PLoS ONE. 2015;10(7):1–13. Available from: https://doi.org/10.1371/journal.pone.0131967.
22. Gyuranecz M, Birdsell DN, Splettstoesser W, Seibold E, Beckstrom-Sternberg SM, László M, et al. Phylogeography of *Francisella tularensis* subsp. *holarctica*, Europe. Emerg Infect Dis. 2012;18(2):290. Available from http://wwwnc.cdc.gov/eid/article/18/2/11-1305.
23. Hornstra HM, Priestley RA, Georgia SM, Kachur S, Birdsell DN, Hilsabeck R, et al. Rapid Typing of *Coxiella burnetii*. PLoS ONE. 2011;6(11):1–8. Available from: https://doi.org/10.1371/journal.pone.0026201.
24. Karlsson E, Svensson K, Lindgren P, Byström M, Sjödin A, Forsman M, et al. The phylogeographic pattern of *Francisella tularensis* in Sweden indicates a Scandinavian origin of Eurosiberian tularaemia. Environ Microbiol. 2013;15(2):634–45. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/1462-2920.12052.
25. Karlsson E, Macellaro A, Byström M, Forsman M, Frangoulidis D, Janse I, et al. Eight New Genomes and Synthetic Controls Increase the Accessibility of Rapid Melt-MAMA SNP Typing of *Coxiella burnetii*. PLoS ONE. 2014;9(1):1–12. Available from https://doi.org/10.1371/journal.pone.0085417.
26. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, et al. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet. 2010;42(10):1140–3. Available from: http://dx.doi.org/10.1038/ng.705.
27. Okinaka RT, Henrie M, Hill KK, Lowery K, Van Ert M, Pearson T, et al. Single Nucleotide Polymorphism Typing of *Bacillus anthracis* from Sverdlovsk Tissue. Emerg Infect Dis. 2008;14(4):653–6. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2570946/.
28. Simonson TS, Okinaka RT, Wang B, Easterday WR, Huynh L, U'Ren JM, et al. Bacillus anthracis in China and its relationship to worldwide lineages. BMC Microbiol. 2009;9(1):71. Available from: https://doi.org/10.1186/1471-2180-9-71.
29. Svensson K, Granberg M, Karlsson L, Neubauerova V, Forsman M, Johansson A. A Real-Time PCR Array for Hierarchical Identification of rancisella Isolates. PLoS ONE. 2009;4(12):1–14. Available from: https://doi.org/10.1371/journal.pone.0008360.
30. Van Ert MN, Easterday WR, Simonson TS, U'Ren JM, Pearson T, Kenefic LJ, et al. Strain-Specific Single-Nucleotide Polymorphism Assays for the Bacillus anthracis Ames Strain. J Clin Microbiol. 2007;45:47–53. Available from: http://jcm.asm.org/content/45/1/47.abstract.
31. Van Ert MN, Easterday WR, Huynh LY, Okinaka RT, Hugh-Jones ME, Ravel J, et al. Global Genetic Population Structure of *Bacillus anthracis*. PLoS ONE. 2007;2(5):1–10. Available from: https://doi.org/10.1371/journal.pone.0000461.
32. Vogler AJ, Birdsell D, Price LB, Bowers JR, Beckstrom-Sternberg SM, Auerbach RK, et al. Phylogeography of *Francisella tularensis*: Global Expansion of a Highly Fit Clone. J Bacteriol. 2009;191(8):2474–84. Available from: http://jb.asm.org/content/191/8/2474.abstract.
33. Vogler AJ, Chan F, Wagner DM, Roumagnac P, Lee J, Nera R, et al. Phylogeography and Molecular Epidemiology of *Yersinia pestis* in Madagascar. PLoS Negl Trop Dis. 2011;5(9):1–11. Available from: https://doi.org/10.1371/journal.pntd.0001319.
34. Yang JY, Brooks S, Meyer JA, Blakesley RR, Zelazny AM, Segre JA, et al. Pan-PCR, a Computational Method for Designing Bacterium-Typing Assays Based on Whole-Genome Sequence Data. J Clin Microbiol. 2013;51:752–8. Available from http://jcm.asm.org/content/51/3/752.abstract.
35. Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, et al. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics. 2010;11:461. Available from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949892/.

36. Ding K, Zhang J, Zhou K, Shen Y, Zhang X. htSNPer1.0: software for haplotype block partition and htSNPs selection. BMC Bioinformatics. 2005;6:38. Available from: https://doi.org/10.1186/1471-2105-6-38.
37. Frei UK, Wollenweber B, Lübberstedt T. "PolyMin": software for identification of the minimum number of polymorphisms required for haplotype and genotype differentiation. BMC Bioinformatics. 2009;10(1): 176. Available from: https://doi.org/10.1186/1471-2105-10-176.
38. Hao K, Liu S, Niu T. A Sparse Marker Extension Tree Algorithm for Selecting the Best Set of Haplotype Tagging Single Nucleotide Polymorphisms. Genet Epidemiol. 2005;29:336–52. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2712933/.
39. Ke X, Cardon LR. Efficient selective screening of haplotype tag SNPs. Bioinformatics. 2003;19:287–8. Available from: http://dx.doi.org/10.1093/bioinformatics/19.2.287.
40. Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS, Ramoni MF. Minimal haplotype tagging. Proc Natl Acad Sci. 2003;100:9900–5. Available from: http://www.pnas.org/content/100/17/9900.abstract.
41. Sahl JW, Lemmer D, Travis J, Schupp JM, Gillece JD, Aziz M, et al. NASP: an accurate, rapid method for the identification of SNPs in WGS datasets that supports flexible input and output formats. Microb Genom. 2016;2: e000074. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5320593/.
42. Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984. Nucleic Acids Res. 1985;13(9):3021–30. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC341218/.
43. Danecek P, Auton A, Abecasis G, Al bers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/.
44. Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, Kusecek B, et al. Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc Natl Acad Sci U S A. 2004;101:17837–42. Available from http://www.pnas.org/content/101/51/17837.abstract.
45. Johansson A, Farlow J, Larsson P, Dukerich M, Chambers E, Byström M, et al. Worldwide Genetic Relationships among *Francisella tularensis* Isolates Determined by Multiple-Locus Variable-Number Tandem Repeat Analysis. J Bacteriol. 2004;186:5808–18. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC516809/.
46. Sahl JW, Schupp JM, Rasko DA, Colman RE, Foster JT, Keim P. Phylogenetically typing bacterial strains from partial SNP genotypes observed from direct sequencing of clinical specimen metagenomic data. Genome Med. 2015;7:52. Available from https://doi.org/10.1186/s13073-015-0176-9.
47. Sahl JW, Pearson T, Okinaka R, Schupp JM, Gillece JD, Heaton H, et al. A *Bacillus anthracis* Genome Sequence from the Sverdlovsk 1979 Autopsy Specimens. mBio. 20167. Available from: http://mbio.asm.org/content/7/5/e01501-16.abstract.
48. Johnson SL, Daligault HE, Davenport KW, Jaissle J, Frey KG, Ladner JT, et al. *Yersinia pestis* strain:Harbin35 Genome sequencing. SRR1283952 [Sequence Read Archive]. 2015. Available from https://www.ncbi.nlm.nih.gov/sra. Accessed 27 Sept 2017.
49. Johnson SL, Daligault HE, Davenport KW, Jaissle J, Frey KG, Ladner JT, et al. Whole Genome Sequencing of *Yersinia pestis* str. Pestoides B. SRR2177700 [Sequence Read Archive]. 2015. Available from: https://www.ncbi.nlm.nih.gov/sra. Accessed 27 Sept 2017.
50. Johnson SL, Daligault HE, Davenport KW, Jaissle J, Frey KG, Ladner JT, et al. *Yersinia pestis* Angola Genome sequencing. SRR2153449 [Sequence Read Archive]. 2015. Available from: https://www.ncbi.nlm.nih.gov/sra. Accessed 27 Sept 2017.
51. Johnson SL, Daligault HE, Davenport KW, Jaissle J, Frey KG, Ladner JT, et al. Yersinia pestis Antiqua Genome sequencing. SRR2176134 [Sequence Read Archive]. 2015. Available from: https://www.ncbi.nlm.nih.gov/sra. Accessed 27 Sept 2017.
52. Johnson SL, Daligault HE, Davenport KW, Jaissle J, Frey KG, Ladner JT, et al. Thirty-Two Complete Genome Assemblies of Nine *Yersinia* Species, Including *Y. pestis*, *Y. pseudotuberculosis*, and *Y. enterocolitica*. Genome Announc. 2015;3:e00148–15. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4417686/.
53. Johnson SL, Minogue TD, Daligault HE, Wolcott MJ, Teshima H, Coyne SR, et al. *Yersinia pestis* Antiqua Genome sequencing. SRR2084698 [Sequence Read Archive]. 2015. Available from: https://www.ncbi.nlm.nih.gov/sra. Accessed 27 Sept 2017.
54. Johnson SL, Minogue TD, Daligault HE, Wolcott MJ, Teshima H, Coyne SR, et al. Finished Genome Assembly of *Yersinia pestis* EV76D and KIM 10v. Genome Announc. 2015;3:e01024–15. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4574367/.
55. Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, et al. Yersinia pestis CO92 chromosome, complete genome. NC_003143.1 [NCBI Reference Sequence]. 2015. Available from: https://www.ncbi.nlm.nih.gov/refseq/. Accessed 27 Sept 2017.
56. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
57. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, et al. Variation in virulence among clades of Escherichia coli O157:H7 associated with disease outbreaks. Proc Natl Acad Sci. 2008;105(12): 4868–73. Available from: http://www.pnas.org/content/105/12/4868.
58. Birdsell DN, Johansson A, Öhrman C, Kaufman E, Molins C, Pearson T, et al. *Francisella tularensis* subsp. *tularensis* Group A.I, United States. Emerg Infect Dis. 2014;20:861–5. Available from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4012810/.
59. Li Y, Cui Y, Cui B, Yan Y, Yang X, Wang H, et al. Features of Variable Number of Tandem Repeats in Yersinia pestis and the Development of a Hierarchical Genotyping Scheme. PLoS ONE. 2013;8:e66567.
60. Vogler AJ, Driebe EM, Lee J, Auerbach RK, Allender CJ, Stanley M, et al. Assays for the rapid and specific identification of North American *Yersinia pestis* and the common laboratory strain CO92. Biotechniques. 2008;44: 201–7. Available from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3836605/.
61. Gibbons HS, Krepps MD, Ouellette G, Karavis M, Onischuk L, Leonard P, et al. Comparative Genomics of 2009 Seasonal Plague (*Yersinia pestis*) in New Mexico. PLoS ONE. 2012;7:1–11. Available from: https://doi.org/10.1371/journal.pone.0031604.
62. Plunkett GI, Anderson BD, Baumler DJ, Burland V, Cabot EL, Glasner JD, et al. *Yersenia pestis* biovar Medievalis str. Harbin 35 (enterobacteria). GCA_000186725.1 [GenBank Assembly]. 2011. Available from: https://www.ncbi.nlm.nih.gov/genbank/. Accessed 27 Sept 2017.
63. Anisimov AP, Dentovskaya SV, Svetoch TE, Panfertsev EA. Variability of the Protein Sequences of LcrV Between Epidemic and Atypical Rhamnose-Positive Strains of *Yersinia pestis*. In: The Genus *Yersinia*: From Genomics to Function. New York: Springer New York; 2007. p. 23–7.
64. Deng W, Burland V, Plunkett III G, Boutin A, Mayhew GF, Liss P, et al. Genome Sequence of *Yersinia pestis* KIM. J Bacteriol. 2002;184:4601–11. Available from: http://jb.asm.org/content/184/16/4601.abstract.
65. Chain PSG, Hu P, Malfatti SA, Radnedge L, Larimer F, Vergez LM, et al. Complete Genome Sequence of *Yersinia pestis* Strains Antiqua and Nepal516: Evidence of Gene Reduction in an Emerging Pathogen. J Bacteriol. 2006;188:4453–63. Available from: http://jb.asm.org/content/188/12/4453.abstract.