Journal of
Data Mining in Genomics & Proteomics

Jakupciak et al., J Data Mining Genomics Proteomics 2013, 4:4
DOI: 10.4172/2153-0602.1000138

Research Article Open Access

Population Analysis of Bacterial Samples for Individual Identification in Forensics Application

John P Jakupciak[1]*, Jeffrey M Wells[1], Jeffrey S Lin[2] and Andrew B Feldman[2]

1Cipher Systems, 2661 Riva Road, Annapolis, MD 21401, USA
[2]The Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Rd., Laurel, MD 20723, USA

Abstract

Biodefense preparedness begins with the ability to detect and respond to bio-threats, based on accurate interpretation of genetic information with sophisticated, yet easy-to-use bioinformatics tools. Microbial forensics further enables attribution of microbial pathogen samples back to a suspected source. Sample characterization and traceability back to source are dependent on genome identification of specific targets within samples, comprehensive analysis of mixtures of populations' present, and detection of major/minor variations in the identified genomes and comparison of sample genetic profile against other samples. Commercial Next Generation Sequencing (NGS) platforms offer the promise of dramatically higher detection sensitivity and resolution of forensic DNA samples than is possible with methods in current use. Before applying these technologies for forensic analyses of bacterial samples, however, it is critical to fully elucidate the benefits, caveats and pitfalls of NGS for hypothesis testing in comparative analyses, as ultimately this will be required for NGS use both as an investigative tool and tool for attribution in courts of law.

Methods: We developed and evaluated novel probabilistic algorithms to process metagenomic sequence data from direct sample sequencing to identify genomes present in mixtures.

Results: We present a pipeline for reference-free sample-to-sample comparisons to improve target characterization beyond one microorganism to characterization of comprehensive sample content. Our tools strengthen statistical confidence to trace the ancestry of samples and attribute samples to source with probabilistic certainties on many targets instead of a single genome.

Conclusion: This study developed a novel reference free, bioinformatics strategy to account for and identify genetic diversity in samples. Sequence variants must be non-arbitrarily confirmed in both forward and reverse reads at a rate above the background noise level of sequencer machine error. A similarity distance metric compares genomes within a range of near relationships. Using sequence data from bio-threat agents, we successfully attributed known related strains together, and excluded near relation of known unrelated strains. The major strengths of this forensic method are the non-arbitrary determinations of data validation and relatedness metrics, as well as the ability to compare microbial genomes with or without a reference database of related genomes.

Keywords: Population-Sequencing; Bioinformatics; Mapping pipeline; Biothreats; Forensics; Metagenomics; Probability; Calibrant; Hypothesis testing

Introduction

Our technical approaches were driven by anticipated requirements for forensic analysis of DNA samples [1]. First, to, mitigate effects of protocol choices and data processing pipeline parameter values [2-4]. We minimize the nu MBer of initial assumptions and developed algorithms with a minimum nu MBer of parameters. Second, it is important to address run-to-run variability in measured DNA sequences due to chemical-reagent-lot variability, sequencing artifacts and sequencing errors, and experimental conditions. To address this second need, we recommend and have developed approaches for internally calibrating sequencing runs. Two types of standards are needed: 1) internal standards to evaluate system errors and 2) bioinformatics calibrant to correct for sequencing artifacts [5]. Finally, we address confidence in conclusions because reference genome databases (DB) contain only limited sampling of real-world biological diversity.

Forensics characterization of bacterial constituents

Metagenomics is an emerging discipline for microbial population(s) analysis based on sequence information obtained directly from samples without culture purification enabled by a growing nu MBer of bioinformatics tools being developed to address analysis of mixtures [6]. Genome mixture analysis for forensic characterization of constituent organisms is driven by several considerations.

First, reference DB such as Genbank will always be inherently limited due to biases introduced by selection of microorganisms to sequence and the vast dynamic genetic diversity of microorganisms. The degree to which sequences found to be "discriminatory" among reference DB genomes are actually unique is knowable to a degree of probability. This class of analysis has complementary value, however, in that certain sequence motifs, e.g., those that confer pathogenicity, etc., are informative, for other purposes.

In addition, reference DB is rapidly growing due to reductions in sequencing costs [7]. Forensic tools must therefore be computationally scalable. Using off-the-shelf tools, e.g. BLAST, to compare sequencing

***Corresponding author:** John P Jakupciak, Cipher Systems, 2661 Riva Rd, Annapolis, MD 21401, USA, Tel: (410) 412-3326; Fax (410) 897-1066; E-mail: j.jakupciak@cipher-sys.com

Received May 16, 2013; **Accepted** August 12, 2013; **Published** August 19, 2013

Citation: Jakupciak JP, Wells JM, Lin JS, Feldman AB (2013) Population Analysis of Bacterial Samples for Individual Identification in Forensics Application. J Data Mining Genomics Proteomics 4: 138. doi:10.4172/2153-0602.1000138

reads against entire DB content is computationally costly and cu MBersome [8,9]. Further, as the cost of sequencing decreases, the cost of data analysis of volumes of sequence information increases [10]. There is a critical need for efficient bioinformatics tools [11].

Unlike reference DB entries, single sequences do not capture the diversity of genomes in a population. Population structure-the distributions of genetic variability-within the organism constituents in forensic samples (cultured and non-culture) add uncertainty to genomic compositional description of samples. With respect to forensic applications, sample sequences inherently contain unknown genetic variation with respect to known reference DB genomes on account of multiple subtle and stressful environmental selection pressures.

Lastly, for sample constituents present in very low proportion (minor sample content), sampling statistics and sequencing process errors place limits on their detection. Therefore, quantitative metrics for characterizing uncertainty in detection relative to relevant noise measures are needed [12-16]. With appropriate accommodation of these considerations, accurate identification of genomes in mixtures can be successfully addressed. We evaluated a three-step approach to characterizing genome constituents in samples: 1) Rapid filtering of sequencer reads using non-alignment methods. 2) Characterization of reads for consistency with DB genomes using a quantitative hypothesis test based on characterization of reference genome coverage breadth as a function of coverage depth. 3) Application of hypothesis testing with conservative p-value thresholds to identify candidate reference genomes for detailed, alignment-based mapping of total reads to genetic variation with respect to closest known organisms in DB. The overall approach is depicted in Figure 1.

Our read filtering scheme is based on representation of individual reads and database genomes as a collection of *n-mer* words determined by sequentially passing a window *n*-base in width across the read or genome sequence one base at a time and recording *n-mer* words found in the sequence. By requiring a certain fraction of *n-mers* per read to match those found in reference genomes, we can rapidly filter reads according to their *n-mer* content with respect to references.

For example, for an *n-mer* length of 25 and a read length of 100,

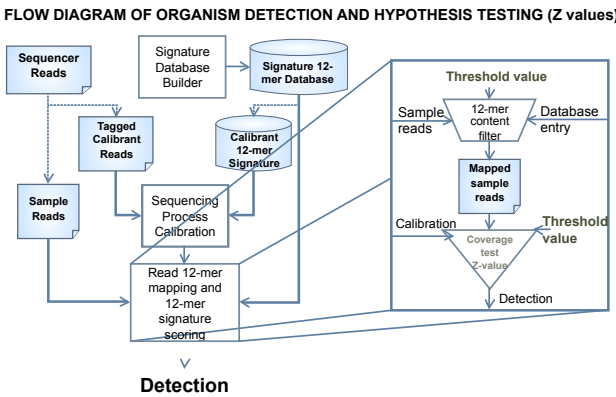

Figure 1: Process flow for detecting constituent organisms in a test sample using a "noise" calibration, *n*-mer read filtering to deconvolve a mixture, and a hypothesis test (Z-value). The hypothesis test measures deviation of the observed matching rate of *n*-mers data against a reference genome's *n*-mers, and compares it to the mean deviation of a known organism (calibrant) against its reference when sequenced in same run. The Z value is the deviation expressed in terms of standard deviations of measurements for the calibrant.

there are 76 *n-mer* words, 25 bases long. A matching fraction of ~0.2 requires matching 15 *n-mers* from the read against those found in the reference DB. While such a low fraction comes at the expense of specificity (i.e., the shorter n-*mers* capture diversity with respect to population variation (for example a single base change at one site or small insertion) and sequencing noise, which could produce. For *n-mer* length 25, a single based change (noise or single nucleotide polymorphism–SNP) will modify up to 49 of the 76 *n-mers* per read per 100 base long read. Thus, use of thresholds below ~0.3 enable rapid *n-mer* filtering of reads and provides balances assessment of noise and population variability. Although alignment-based methods perform well against such genetic differences between a reference and a read, the matching process is much more computationally onerous, particularly for large reference DB and matching millions of reads [17,18]. Further, sequence artifacts need to be removed, which will decrease accuracy of bioinformatics tools. The threshold is a floating value for read filtering dependent on each sequence run. It is selected by the user to establish how much data to include; alternatively, it can be viewed as the cut-off of noise. The decision to include/exclude reads is a balance of long *vs* short *n-mers* and a balance of accurate identification versus tolerating noise, artifacts, and error. The example in the text explains the value, but specifically, the value of 0.3 translates into processing *n-mers* with matching 22 bases from the read to the reference. The current text points out that a value of 0.2 translates into matching *n-mers* of approximately 15 base pairs. 15 bp *n-mers* are common across genomes and this effects accuracy. A new reference has been added that illustrates the size of *n-mers* and their "uniqueness" value [19]. Thresholds greater than 0.3 would reduce the nu MBer of *n-mers* used in the hypothesis testing to reads specific to single genomes, but they would not include in the analysis reads that may contain point mutations, important to understanding the population.

The choice of *n-mer* length comes with, trade-off of resolution of genetic differences from reference against the specificity advantage of using longer *n-mers* (~20-25 long) which are more unique among bacterial genomes and provide higher specificity in the filtering step. The choice also represents a balance of computing storage, memory, and rapid access requirements. To address memory, storage, and rapid access requirements, we developed a unique hashing approach to storing and retrieving long *n-mers* from reference database genomes. For a given n-*mer*, the total possible nu MBer of words is 4n; thus the counts of a specific *n-mer* in a genome can be recorded in an array, dimension N=4n, entities (8-bit character, 32-bit integer, etc…). For example, for n=12, we have N=16,772,216, which uses a modest amount of computer memory, N*4 bytes or ~48 MB, for 32-bit integer storage. As n increases, we begin to hit the limit of available storage in computer random access memory (RAM). For n=16, we require ~17.2 GB of RAM for integer storage or 4.3 GB for 8-bit storage. For a given *n-mer*, each sequence can be hashed to a unique integer array index by mapping specific bits in the integer to A,T,C,G, and their positions within an *n-mer* word. This allows a rapid array lookup. We benchmarked the software on a MacBook Pro and Dell Server with 8 processors. Scoring 2,500,000, 100 base reads from a *Y. pestis* C090 sample against the entire Genbank database took ~3 hours on the MacBook Pro and 80 minutes on the Dell system.

To populate a *n-mer* database and rapidly retrieve *n-mer* counts for a specific ~25-*mer*, we break up the *n-mer* into two separate *n-mers,* one that is used as rapid indexing into an array and the second that is used in an ordered binary tree, with the root of this tree associated with the array index determined by the first n-*mer*. This algorithm provides a tunable compromise between memory storage and speed

of lookup. For 25-*mers*, using a 12-*mer* first index and 13-*mer* second index into the tree, typical memory usage for a complete genome is ~2 GB. For purposes of filtering, each read is tested against all DB genomes and reads satisfying the threshold criterion are retained and scored collectively against the particular reference genome using hypothesis test. A set of notional results is shown in Table 1.

Our hypothesis test for the presence of specific reference genome's presence/absence in samples is based on quantitative metrics reflecting coverage breadth dependence on coverage depth of each sequencing run. The metrics were derived using a rapid non-alignment approach and is estimated by measuring the specific relationship for an internal calibrant organism that is an exact consensus sequence match to its DB reference genome. To measure coverage depth and breadth, the reference genome is indexed into its constituent *n-mers*. This collection of *n-mers* is the signature for the organism. For a genome of ~5 MBases, the total signature is ~8 MBases, including forward/reverse reading and allowing for considerable non-uniqueness of *n-mers* for n=12. Total possible 12-*mers* are 412 or about 16 million; thus for any given bacterial reference genome, the probability, p, of presence/absence of an *n-mer* is ~ 0.5. For larger *n-mers* than 12, the DB will be more sparsely populated (more unique *n-mers*) and for smaller *n-mers* more densely populated (less unique *n-mers*).

The relationship between signature coverage depth and signature coverage breadth when NGS reads *n-mers* are accumulated and then compared for presence/absence to those of a reference genome is a function of several factors: uniformity of representation of reference genome in the reads; fraction of reads derived from population variants in the sample; accumulation of sequence artifacts and sequencing error rate during the run. Assuming uniform coverage during DNA extraction, shearing, and other sample processing steps prior to the NGS run and a single population variant exactly matching the reference genome, the signature *n-mer* detection rate as will increase monotonically with depth of signature *n-mer* coverage. We define coverage as the count of signature *n-mers* accumulated divided by the total nu MBer of signature *n-mers*.

Calibration analysis using expected signature coverage breadth as a function of signature depth segregates genomes in a mixture from near neighbors, which share genome content. This metric of the totality and uniformity of genome coverage provides a critical fundamental parameter for definitively detecting presence of a specific organism in samples. It can be used to establish a profile of the genomes of the populations encompassed within the unique genetic boundaries of the sample.

Results

Quantitative distinguishing genome content

Data from specifically designed proof-of-concept experiments were collected using samples prepared by CUBRC and NGS sequenced by the US Army Edgewood Chemical and Biological Center (ECBC) at Aberdeen Proving Ground in MD. To demonstrate the concept, we analyzed NGS data for select agent pathogens. A single colony was passaged into 12 different plates. These twelve bacterial cultures were maintained separately over the course of seven more passages. Each culture passage was started with a single clonal colony streaked out on a petri dish. This created a single genome bottleneck at each passage step. Mutational variations differentiating each lineage were thus a result of initial variation in the source clonally derived culture plus mutations accumulated during the course of the eight growth and passage steps. Five clones of the same lineage after passage 8 were sequenced and compared. Since the start of each passage begins with a single cell, any differences between different clones in this lineage must be due to mutations that arose during this single growth stage, laboratory error that passaged more than one cell to the next flask, or it must be due to sequencer machine error.

A summary of the raw single ended read data for the five genome samples is given below.

Nu MBer of runs: 28

Average reads/run: 3.0M

Calculated nu MBer of reads needed for 15X average coverage depth=~1M

Two runs had below 15X average coverage depth. Strains: *Yersinia pestis* CO92 (BEI#NR-61),

Burkholderia pseudomallei MSHR668 (BEI#NR-9922), *Burkholderia mallei* China 7 (BEI#NR-23).

Each organism was cultured on alternating media consisting of standard broth followed by selective media agar plates for a total of seven (7) passages, followed by DNA isolation from the eighth passage liquid culture. *Burkholderia mallei* and *B. pseudomallei* were alternately cultured in tryptic soy broth and PC agar. *Yersinia pestis* was alternately cultured in BHI broth and CIN agar".

Two sequencing runs were performed using colony isolates of *Y. pestis, B. mallei, B. pseudomallei, and B. globigii (B.g.)/ B. atropheus*. Sequencing was performed on an Illumina HiSeq 2000 system and up to 100 M, 100 bases were acquired in a single Illumina flow cell lane. Figure 2 shows the calibration curves (probability of detecting a signature *n-mer* as a function of mean signature *n-mer* coverage) obtained in Run 1 and Run 2. The error bars reflect the upper value of the 95% confidence interval for standard deviation of the measurement obtained over 3 sets of statistically independent reads in the single flow cell data. The plot also indicates values obtained for *B. globigii* sequencing during the same runs but in a separate flow cell lane using unfiltered reads. The data clearly indicate that the calibration captures run specific conditions, and that these conditions are not identical between runs. Run 1 and Run 2 represent isolate runs. They are not mixtures. This is an important point and not readily discussed in papers involving sequence analysis. The variance observed in Run 1 and Run 2 is significant because the isolates analyzed are from the identical source, but they are different sequencing runs. The data illustrate that the "same" material does not have the "same" sequence information.

Calculate % of 12-mers per read that match each database entry

Bacteria Genomes

Read #	A	B	C	D
1	**95**	5	0	23
2	**98**	0	3	4
3	**92**	3	4	0
4	5	0	**95**	2
5	3	2	**100**	3
6	2	4	**98**	14

Genome A assigned reads: 1, 2, 3, …
Genome C assigned reads: 1, 2, 3, …

Table 1: Notional figure showing filtered reads (#1-5) and assignment to genomes for hypothesis testing based percent of matching *n*-mers per read to *n*-mers in reference genomes (A, B, C, D). The *n*-mers matched in the genome need to be contiguous, as they are in the read.

Hence run to run variation is greater than the 1% "error" referenced in consensus sequencing. This is a critical difference, because consensus sequencing is much different approach than using sequencing to characterize populations in samples.

The probability, P (detect) at given mean signature coverage in Run 2 is significantly higher than in Run 1, indicating a higher noise rate in Run 1. This can also be seen in Figure 3, which shows mean signature coverage as a function processed reads and indicates a higher base calling error rate in Run 1 than in Run 2.

Hypothesis testing: we calculate P (detect) for observed mean signature coverage from our reads and compute a Z-value (statistic) by subtracting actual measured P (detect) from the interpolated value and represent this difference in units of standard deviations. Tail probabilities (probability of observing a particular value greater than or equal to Z by chance) associated with these normalized Z-values can be calculated using error function integrals. Z values <~2.0, given probability values <~0.01. To illustrate distribution of Z values obtained for unfiltered Bg reads at mean signature coverage up to 150x, we plot values measured using within run calibration and external calibration.

The value represents the expected value from true match to a reference. The measured value (the hypothesis test) is the measurement on actual run data. To achieve accurate genome identification from populations of reads against reference database genomes using the calibration, we compute mean signature coverage for the reads against the reference, then interpolate from the calibration to estimate the predicted P (detect) for an observed mean signature coverage for a true match. We then assume a Gaussian distribution with mean given by the interpolated value and standard deviation given by the maximum standard deviation in the interpolated portion of the calibration curve. We compute a Z-value by subtracting the actual measured P (detect) from the interpolated value and represent this difference in units of standard deviations. Tail probabilities (probability of observing a particular value greater than or equal to Z by chance) associated with these normalized Z-values can be calculated using error function integrals or looked up in published tables. |Z| values <~2.0 given probability values <~0.05. The interpolated value could be considered a pre-determined value, because it is based on a reference genome.

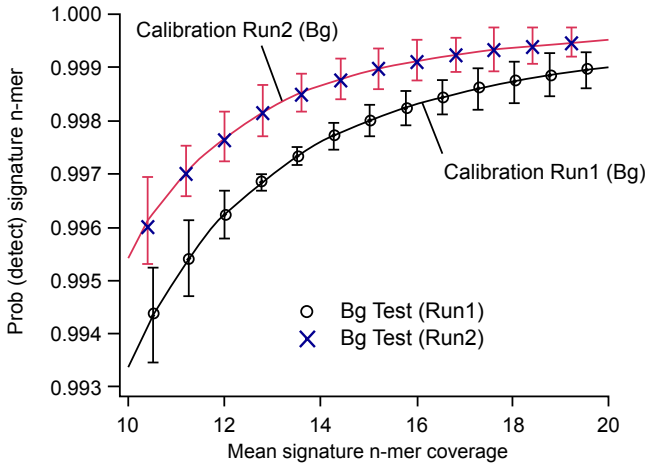

Figure 2: Calibration curves plotting the mean signature *n*-mer coverage of the reference genome (Bg strain) versus probability of detecting *n*-mers for two different Illumina runs. The errors represent the upper 95% confidence interval of the standard deviation for the triplicate measurements for each data point. Test measurement of a separate preparation of the Bg calibrant run in a separate flow cell lane of each run is indicated by the circle (Run 1) and X (Run 2). Note the calibration captures the noise conditions in each run with high precision, and that the curves are not identical between runs.

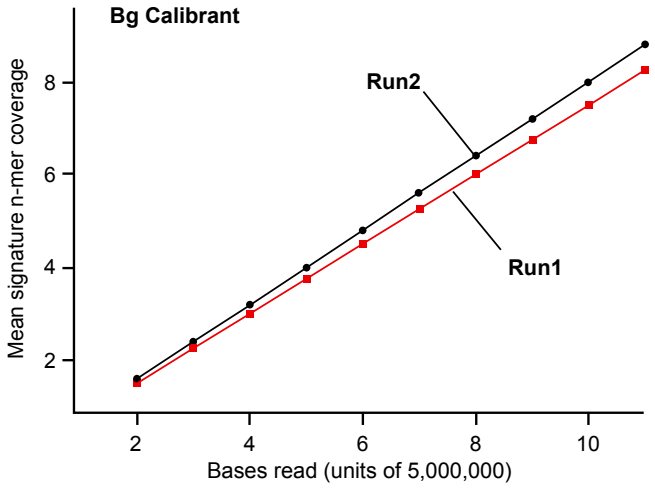

Figure 3: Example data used to compute the calibration curves in Figure 2 showing the mean signature *n*-mer coverage as a function of reads for the Run 1 and Run 2 calibrant. Note *n*-mer matching is higher at fixed read count in Run 2 than Run 1 indicating a reduced base-calling error rate for the conditions in Run 2.

In each case, a reference sequence with the same strain name as that sequenced organism was in the Genbank database: *Yersinia pestis* strain C090, *Burkholderia mallei* strain China 7 (ATCC 23744), and *Burkholderia pseudomallei* strain 668. In the general case, a single colony isolate of a cultured bacterium will correspond to a single constituent of the parent population and will accrue new random (and perhaps fitness-selected) mutations during propagation in media resulting in differences in sequence content with respect to the reference genome. In the case of novel isolate of an organism, its genetics differences with respect to reference strains could be significantly high. Hence, we introduced contextual Z-value, where genetic differences with respect to genome DB signature content is compared to typical within species differences. The computed contextual Z-values enable inference of whether a novel organism is consistent with its nearest neighbor in the reference DB within a resolution comparable to select genus/species/strain contexts. In our biothreat agent experiments, all Genbank genome strains of each organism were used to compute P (detect) for pair-wise comparisons. The mean and standard deviation of these values are then used to provide context: for the measured signature *n-mer* coverage on its nearest neighbor genome, we compute the calibration curve value of P (detect) as scale the pair-wise context values by this weighting to determine the context Z values.

The hypothesis testing results for *Y. pestis* C090 are shown in panel A of Table 2. A Z-value of 0.12 is obtained. This result is consistent with the exact database genome entry being present in the sample within the sequencing error and population structure variation as measured in calibrant sequences. The analysis was performed for 10,000,000 reads. The Z-value against strain KIM is borderline acceptable detection and it would be the nearest neighbor if strain C090 were not in the Genbank database. In this case, the *Y. pestis* context Z-value is below 2.0 and indicates that while the sample is not likely an exact match for species KIM, Nepal, etc., it would most certainly be consistent with a *Y. pestis*. Panel B in Table 2 shows the results from *Y. pestis* C092 and *B. globigii*

A

	Mean n-mer coverage	P_d n-mer	Z_{CAL}	Context Z_{Yp}
Y.pestis CO90	64.83	0.99992	0.12	-1.96
Y.pestis KIM	64.23	0.99966	2.87	-1.93
Y.pestis Nepal	64.95	0.99939	5.53	-1.91
Y.pestis Angola	65.28	0.99937	5.78	-1.91
*Y.pestis*Pestoides	64.11	0.99888	10.91	-1.85

B

			Contextual		
	P_d n-mer	Z_{CAL}	Z_{CAL}	Z_{CAL}	Z_{CAL}
B.globigiiDugway	**0.9843**	0.02	-0.78	-1.96	-2.76
B.subtilus	0.6810	68.07	615.42	30.14	2.17
Y.pestis CO90	**0.9790**	1.22	9.99	-1.40	-2.67
Y.pestis KIM	0.9780	1.45	12.02	-1.29	-2.66
*Y.pestis*Pestoides	0.9725	2.7	23.20	-0.71	-2.57
Y.pestis Nepal	0.9760	1.90	16.08	-1.08	-2.62
Y.pestis Angola	0.9755	2.02	17.01	-1.03	-2.62
Y.pseudotuberculosis	0.9080	17.30	154.24	6.12	-1.52
Y.enterocolitica	0.9080	17.30	154.24	6.12	-1.52

Mean signature n-mer coverage = 7.5

Table 2: A: Signature matching results for a pure *Y. pestis* C090 run showing reference genome database Z values, probability of detecting signature *n*-mers (P_d) Z values < ~2.0 are considered consistent with the reference organism being present in the sample within the limits imposed by base calling noise and population structure ("biological noise"). The contextual Z values represent the deviations of the observed P_d from that expected based on the calibration at the same coverage in units of standard deviation of P_d among all other known Yp reference sequences. If *Y. pestis* C090 was not in the database, KIM strain would be the best scoring, but with Z value > 2.0. However, here the contextual Z value would tell us that the distance from Yp KIM is still well within the distances of Yp among themselves, so treating the sample as an unknown we could state it is very likely a *Y. pestis*. **B**) Z value and contextual Z values for Yp and Bg and much lower coverage of the signature *n*-mers. Here the data for Yp are consistent with 4 known Yp strains. Note the high Z value for *B. subtilis,* the nearest known neighbor of Bg. It has an ~80% signature homology with Bg, showing the high sensitivity of this method to small genomic change.

sequence analysis at much lower signature coverage of signature *n-mers*. There is greater uncertainty in the scores against the various *Y. pestis*. All 5 strains are consistent with being the sample sequenced, yet other *Yersinia entercolitica* and *pseudotuberculosis*) are clearly excluded even at lower coverage. We have included *B. globigii* in the table to illustrate the context Z value concept further. Here *B. globigii* is scored against the *B. globigii* reference genome and its nearest known neighbor in Genbank, *B. subtilus*, which shares 80% genetic identity with against the *B. globigii* reference genome and its nearest known neighbor in Genbank, *B. subtilus*, which shares 80% genetic identity with *B. globigii*.

The results for Bm and Bp sequences (10,000,000 reads) are shown Tables 3 and 4, respectively. In each case the lowest Z value is associated with the correct DB strain. The low Bm contextual Z values for the Bm even when scored against *B. pseudomallei* reference genomes would indicate that the sequenced samples was consistent with a *Burkholderia* even if *B. mallei* were not in the reference database. Note that Z=3.06 is a bit high compared to the Bg and Yp cases. The same is true for the *B. pseudomallei* sequence in Table 4.

Forensic comparison of consensus sequences of bacteria

Forensic analysis of bacterial samples requires characterization of specific genetic variations of candidate constituent organisms that were identified through the triaging process. Reads were aligned using various algorithms, such as NCBI BLAST, SOAP, BFAST, etc., to reference genomes for elucidation speed and specificity.

For example, we compared *B. anthracis* Illumina sequence data from two independent sequence alignment pipelines with two independent investigators choosing parameter values. Pipeline A used the SOAP alignment tool to map the reads. SOAP is capable of modeling small contiguous insertions and deletions as well as mismatches (1-2 bases) and has a read length limitation of 60 bases. SOAP has difficulty accurately mapping reads that have more than two non-contiguous mismatches in a single stretch of bases. Pipeline B used an integrated set of public domain tools and a custom SNP calling method that uses minimal assumptions. Pipeline B is shown in Figure 4. The BFAST algorithm is used to find a candidate alignment position for each read. The mapping depends on a set of index masks to determine which locations in a read require matching as part of the scoring process. Following selection of the best scoring alignment for each read, the reads are annealed to the local reference using Smith-Waterman algorithms. The aligned reads are converted to the public domain SAM format, sorted according to position along the reference and then submitted for mPileUP analysis via the Samtools software suite. The mPileUP tool provides a useful output for quantifying local genome coverage, base calls at each reference genome position, indels, and whether base calls came from a forward or reverse direction.

Declaration of a SNP with respect to the reference genome depends on a threshold count for the non-reference matching base calls at each genome position. The subtlety of the problem is illustrated in Figure 5, which plots distribution of the fraction of non-reference calls per

c	Mean n-mer coverage	P_d n-mer	Z_{CAL}	ContextualZ_{Bm}
*B.mallei*ATCC 23744	**116.50**	0.99985	3.06	-0.73
*B.mallei*SAVP1	121.36	0.99800	49.93	-0.66
*B.mallei*NCTC-_10229	116.75	0.99660	79.13	-0.60
*B.mallei*NCTC_10247	116.13	0.99650	79.13	0.61
*B.pseudomallei*1106a	97.88	0.94000	1015.70	1.63
*B.pseudomallei*668	97.65	0.93720	1064.30	1.74
*B.pseudomallei*K96243	94.59	0.92850	1191.05	2.09
B.pseudomallei1710b	93.50	0.92400	1201.05	2.25
Burkholderia 383	70.86	0.82600	1770.90	6.12
Burkholderiathailendensis	88.56	0.88030	1839.60	3.99
Burkholderiamultivorans	80.85	0.83960	2055.80	5.59

Table 3: Z values and contextual Z values *for Burkholderia mallei* strain China 7, which is most closely matched to *B.mallei* ATCC 23744. Scores against other *Burkholderia* are shown for context. While the Z value of 3.06 is high, we show that this is due to far greater population structure ("biological noise") in *B. mallei* samples as compared to the Bgcalibrant sample, as well as consensus SNPs.

	Mean n-mer coverage	P_d n-mer	Z_{CAL}	Contextual Z_{Bp}
B.pseudomallei 668	**98.86**	0.99982	2.87	-1.65
*B.mallei*SAVP1	113.50	0.98680	278.50	-1.54
*B.mallei*ATCC_23744	108.55	0.98550	282.00	-1.52
*B.mallei*NCTC_10247	108.50	0.98530	287.91	-1.53
*B.mallei*NCTC_10229	109.04	0.98520	293.79	-1.53
*B.pseudomallei*1106a	95.55	0.96830	528.05	-1.38
*B.pseudomallei*K96243	92.37	0.95630	682.25	-1.28
B.pseudomallei 1710b	91.42	0.95310	755.40	-1.24
Burkholderia 383	70.86	0.82600	1770.90	6.12
Burkholderiathailendensis	88.56	0.88030	1839.60	3.99
Burkholderiamultivorans	80.85	0.83960	2055.80	5.59

Table 4: Z values and contextual Z values for *Burkholderia pseudomallei* strain 668. As for the *B. mallei*, the elevated Z value against the reference is attributable to greater *n*-mer diversity (population structure) as well as additional consensus SNPs compared to the Bg calibration sample.

BFAST Algorithm
- Index reference genome using mask and hash for speed
- Match reads to reference (candidate alignments)
- Perform local alignment with gaps (Smith Waterman)
- Filter for best alignments (randomly assign tied scores)
- Convert output to SAM format

SAMTOOLS
- Sort mapped reads according to genome positions
- mPileUP (accumulate base calls at each reference site)

SNP Calling
- Determine fraction *f* of identical non-reference base calls
- SNP*: f > 0.5* and confirmed in forward and reverse reads

Figure 4: Description of the Pipeline A alignment algorithm used to SNP calling and population diversity measures. BFAST and SAMTOOLS are available in the public doman. The SNP calling required a non-reference matching base fraction at a genome position exceeding 0.5 (i.e., a new consensus) and confirmation of the base call at that position in 1 or more forward reads and reverse reads.

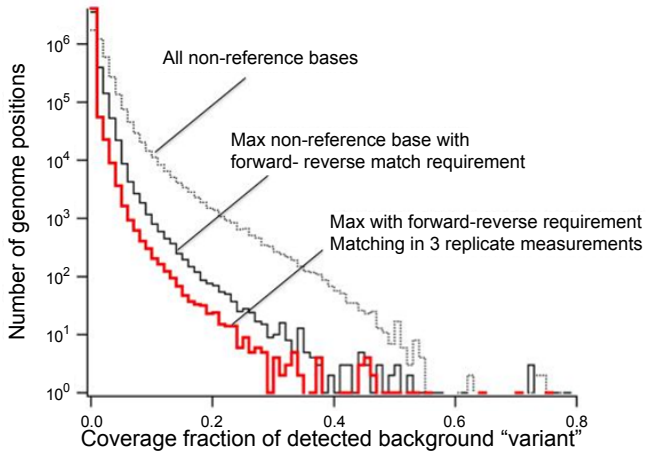

Figure 5: Distribution of non-reference matching bases per genome position after reads alignment by Pipeline A. The dotted line represents the fraction of local coverage that is any non-reference base, the solid black line is the maximum of the identical non-reference bases that have at least one read in both the forward and reverse directions, and the red line represents those variants meeting the criteria of the black curve, but further confirmed in 3 replicate, statistical identical, read sets. The continuum of the fractions is indicative of population structure and criteria of SNP calling threshold, as those variants exceeding a fraction of 0.5 (the consensus).

genome position and is shown for our Bg calibration run. The dotted line is the total of the non-reference-matching fraction, while the black line is the maximum fraction of all identical non-reference base calls, with the requirement of confirmation in at least one read in forward and in the reverse directions. The red line is the fraction of black line calls that are confirmed in replicate statistically independent sets of reads. This true biological diversity in the bacterial population sequenced and/or systematic errors of reads inducing phantom diversity [20].

We compared SNP calls through Pipeline B to those of Pipeline A and as expected, there was concordance as well as discordance. The most typical discrepancies are shown in Figure 6, which shows the mPileUP output around four distinct discrepant sites. In the figure, a period indicates a match to the reference in the forward read direction and a comma indicates a match in the reverse direction. Non-matching base calls in the forward reading and reverse reading directions are indicated by upper case and lower case letters, respectively. The left most column is genome position, the second column the reference base, and the third column depth of coverage (nu MBer of reads). The upper panel shows two SNPs called by Pipeline A, but not by Pipeline B. In both cases, Pipeline B did not call the SNP because it was only observed in one reading direction. While this is an algorithmic choice, it seems a prudent one in the setting of forensics where caution and conservatism should be the general rule for a quantitative analysis.

Table 5 summarizes specific alignment results from three threat agents and our calibrant organism Bg using the Pipeline A analysis for comparison to our non-alignment triaging (z-value) analysis. Two *Burkholderia* species had statistically high z-values outside the expected range (of z-value <2.0), as compared to Bg and *Y. pestis*. These differences can be explained by data in the third column in the table, which shows total fraction of non-reference base calls for reads

Organism	Z-score Non-mapping	Non-Reference Base Fraction (mapped reads)	Mapped Read Fraction	Unmapped Reference Bases	SNP calls
*B.globigii**	0.02	0.018	0.99	0	0
Y.pesis	0.12	0.019	0.93	134,750	47
B.mallei	3.06	0.025	0.94	166,439	431
B.pseudomallei	2.87	0.027	0.98	859	365

Representative Mapping Statistics: 10,000,000 Illumina reads, 100bp
*Calibrant

Table 5: Mapping statistics for three threat agents and our Bgcalibrant. Non-alignment Z values are included for comparison. The higher Z values for Bukholderia is reflective of the higher fraction of non-reference-matching base calls in these samples and is indicative of a greater population diversity compared to Bg. *Bukholderia* had much higher SNP calls (Pipeline B) compared to the *Y. pestis*. The larger unmapped base counts along the reference genomes (column 5) for Yp and Bm are due to insertion elements that are highly mobile within these genomes and promote re-arrangements. The BFAST default parameters for assigning candidate locations to reads when there is a high multiplicity of candidate alignments across the reference genome resulted in these gaps.

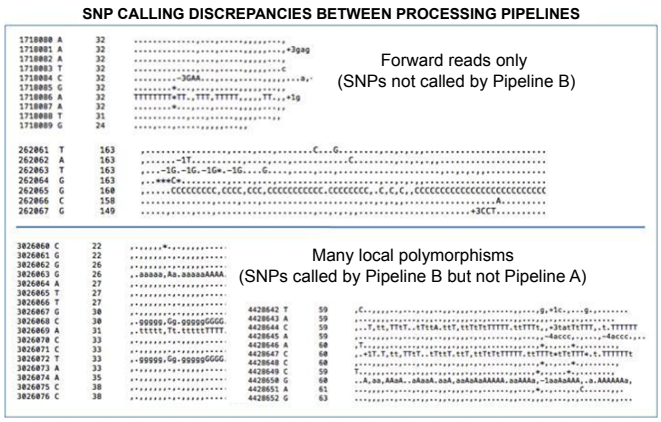

Figure 6: Analysis of SNP calling discrepancies between processing Pipelines A and B. The graphic shows mPileUP output, which is described in the text. Discrepancies can be attributed to an algorithmic limitation (Pipeline A) and stringency criterion (Pipeline B). These data suggest use of the most conservative criteria for SNP calling should be used for comparisons of two bacterial samples for forensic applications. Pipeline B can be improved to further reduce any spurious SNP calls by requiring all base calls exceed a specific quality score.

successfully aligned to the reference genome. For the calibrant Bg, the value of 0.018 reflects both the co MBination of base calling and indel error rate of sequencing process, as well as the population structure. The value is slight higher for *Y. pestis*, which had a slightly high z-value than Bg. For the *Burkholderia*, the non-matching fraction is ~30% higher than Bg and this additional biological "noise" is reflected in the higher z-values. While our hypothesis test is sensitive to this population's structure, it indicates something very useful about the analyzed sample and the importance of calibrants to account for the wide diversity of potential population structure, accumulation of sequence artifacts and sequence-dependence of sequencer error processes.

Another interesting aspect of the alignment/mapping pipeline analysis is displayed in columns four and five. While the *Burkholderia* had greater genetic diversity than *Y. pestis* and Bg samples, *Y. pestis and B. mallei* samples had lower fractions of successfully mapped reads, and a high fraction ~3% of unmapped based positions (gaps) in the reference genome. Despite this high fraction of the reference genome being unmapped, the *Y. pestis n-mer* signature coverage was >99.99%. This can be explained by the fact the both the *Y. pestis and B. mallei* genomes are unusually rich in identical insertion sequences which are mobile within the genome and promote genomic re-arrangements. A deeper analysis of the gaps in the reference revealed they exist due to algorithmic choices in the BFAST tool for aligning reads that have a large nu MBer of candidate locations.

Forensic detection of bacterial population variants

The next level of analysis in forensic comparison of bacterial samples is detection and comparative analysis of non-consensus population variants within samples. This is akin to analysis performed in the Amerithrax case, where population variants with a specific phenotype were targeted in field analysis for their forensic value. However, we expand on the SNP analysis above to include targeting of specific variants in populations that have local fractions below consensus (<0.5 fraction of the total coverage depth). As shown in Figure 5, the non-reference matching bases, that also match one another in forward and reverse directions, form a continuum. Detecting specific variants from a targeted set must account for background of population variants in a sample that could match targets by chance. A framework for this analysis is constructs of Receiver Operator Characteristic (ROC) curves, which plot probability of detection a target against the probability of falsely detecting the target from the background "noise". The ROC curve is a form of parametric curve: for each threshold values (in our case non-consensus fraction) that a variant must exceed to be "detected", we compute the probability of detecting the target variant when it is actually in the sample, and the probability of accidentally detecting it when it is not introduced explicitly into the sample (accidental detection). As the detection threshold is swept across the range (0.0 ->1.0), the probabilities trace out the ROC curve.

The potential that a forensically targeted variant could randomly appear during growth, coupled with the high sensitivity of NGS to detecting minor variants, affirms the need for conservatism and caution in the use of NGS for comparative population analysis from culture.

Forensic comparison of total genomic content–population sequence analysis

Comparative analysis of two samples for identity of total genomic content is of interest, such as bioinformatics analysis of sequence information of two sample vials to determine extent of matching compositions (direct theft). Two considerations drive the algorithmic approach: multiple, unknown contaminating organism constituents can exist in the samples, and the impacts of sequencing noise need to be minimized. Further, from the perspective of communicating an algorithmic result, a further constraint of a threshold for detection of difference is always implicit for a particular chosen depth of sequencing for the analysis. Our analysis includes of matching *n-mer* content and, but also characterization of non-matching content, the extent to which this content is or is not consistent with the expected sequencing noise. We define a metric, the scalar product distance, s, as,

$$s = \frac{v_1.v_2}{\|v_1\|\|v_2\|} v_i = \{1 m \geq t; 0 \text{ otherwise};$$

where v_1 and v_2 are vectors of dimension 4n formed by the complete set of all *n-mers* of length n, for read set 1 and read set 2 in the comparison analysis, respectively; m is the multiplicity of the *n-mer*, i.e., the nu MBer of times the *n-mer* is observed in the data set, and t is a threshold count that determines whether a particular *n-mer* has been detected. The distance s is 0.0 when there is no overlap in genomic content and 1.0, when *n-mer* content vectors are identical. The introduction of the detection threshold, t, enables us to optimally reduce the impact of noise on our comparison of total genomic content based on *n-mer* composition of the two samples.

The importance of the threshold concept is illustrated in Figure 7, which shows read *n-mer* count distributions for Bg reads for those matching reference genome signature *n-mers* (signal) and those that do not match (sequencing noise and population variants). The data in Panel A is the multiplicity, m, per *n-mer*, for the particular sequencing depth. As seen in Panel B, as the threshold for detecting an *n-mer* is increased from 0, noise is reduced at some expense of signal, but a minimum is attained for s for v_1 derived from read data and v_2 derived from the reference genome. This is the threshold value, t, we use to perform reference-free comparisons between sequencer read sets for the chosen depth of sequencing. The threshold t is related to the

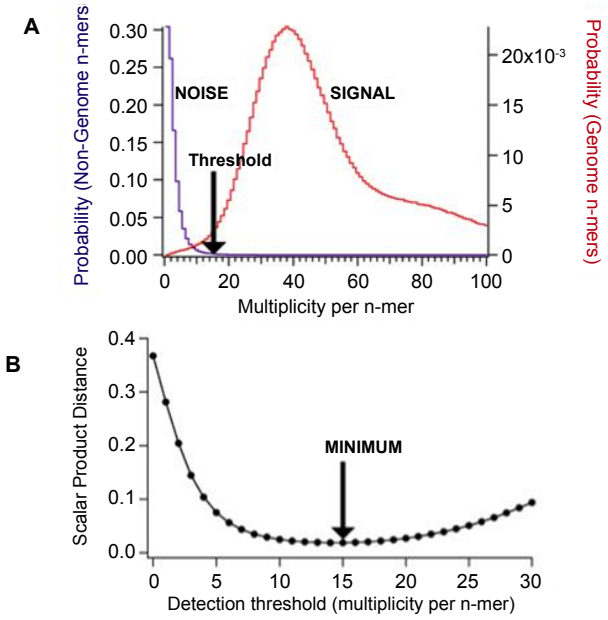

Figure 7: A) Distributions showing the probability of observing a specific multiplicity per signature *n*-mer across a reference genome and probability of observing a non-signature *n*-mer when processing NGS reads. **B)** Curve showing that the scalar product measure (defined in the text) has a minimum when a specific genomic content detection threshold is used.

"threshold-based value." The nu MBer of occurrences of an *n-mer* (the multiplicity), which is dependent on the nu MBer of genomes in the sample and the depth of sequencing. Panel B, Figure 7 illustrates the differences in multiplicity as compared to length of *n-mer*. The ratio of matching reads versus amount of non-matching reads approaches a minimum. User selection of *n-mers* to the right of the minimum will result incorporation of noise, artifacts and errors into the matching processes and result in misleading acceptance of noise as important sequence data for sample characterization. Selection of *n-mers* to the right of the minimum will increase accuracy, but at the same time reduce ability to characterize populations. This plot justifies the rationale to use a threshold-based value of 0.2 – 0.3 because this range incorporates the lowest amount of noise while capturing the greatest amount of unique reads important for accurate genome identification."

Our read-set to read-set comparison approach asks the question: are differences in the metric observed between read sets comparable to those observed when a single read set is compared to multiple statistically independent realizations of itself? Here, fluctuations in a metric reflect statistical sampling coupled with end-to-end sample processing effects for the conditions of the particular run. To measure these fluctuations, we calculate the distribution of metrics for selfversus-self comparisons. This was performed by dividing an ultra deep Bg read set into 20 individual data sets, 4,500,000 reads long (~100× genome coverage) and then computing s, and the numerator of s, which is the probability of self-matching an *n-mer* for all 190 possible unique pair-wise co MBinations of read sets. The same was performed for a second Bg sample executed within the same run. These point pairs are shown as the red dots in the scatterplot in Figure 8. We define an ellipse encompassing the self-versus-self data points and use it as a model to represent the data. The ellipse is defined such that the self-versus-self points have a 5% probability, p, of falling outside the ellipse. We can now compute pair-wise comparisons between Bg and another read set from another sample and count the nu MBer of data points, n, that fall outside of the ellipse; then compute the probability Pr (p, n) that the observed nu MBer or more could have occurred from a self-versus-self comparative analysis by chance.

$$\Pr(p,n) = \sum_{k=n}^{K} \binom{K}{k} p^k (1-p)^{K-k}$$

Here p=0.05, and K is the nu MBer of trials (pair-wise comparisons). Since the Bg read sets were identical in composition after removal of sequence artifacts, computing cross-wise paired comparisons between the two read sets from the run should yield a Pr value that is not statistically significant. The results of the comparisons are shown as the blue dots in Figure 8. Here n=13, K=190, and Pr (p, n)=0.07, which is not significant. To test the sensitivity of the approach to contamination, we introduced 20 copies of a segment of extraneous DNA 5000 bases long (equivalent of a very small plasmid) in the Bg run data, which constitutes as sample with ~0.02% contamination. We performed 10 comparative read set analyses against the uncontaminated Bg and these results are shown in Figure 9. In this case, 50% of points fell outside the model, giving a statistically significant Pr value of 0.001.

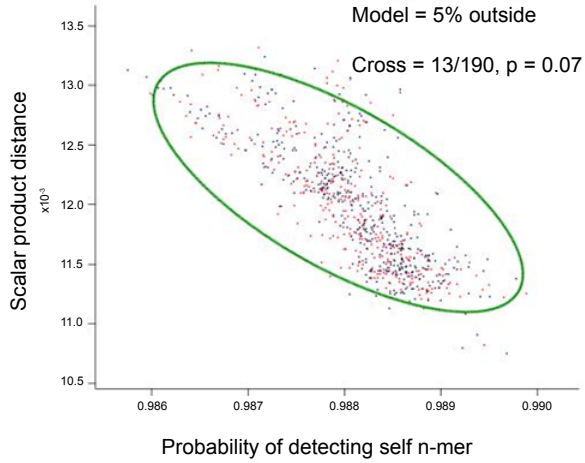

Figure 8: Plot showing scatter distribution of the scalar product metrics versus its numerator (probability of detecting self *n*-mer) for genome-reference-free comparisons of identical (statistically) read sets from the same sample (red) and between two identical samples of the same organism prepared separately, but sequenced in the same run (blue). The model in green is derived from scatter of self-versus-self points and is used to compute a p-value for observing measured fraction of test points outside the ellipse. In this case, the separately prepared, but identical organism sample gave an insignificant p-value, indicated that the samples are consistent with being identical in genomic content.

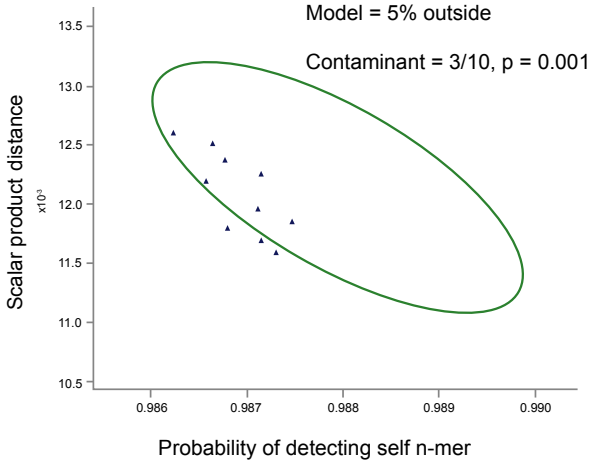

Figure 9: Plot showing the model from Figure 8 and data points from a artificial sample of Bg that is contaminated with 20 copies of an extraneous genome fragment 5000 bases long, corresponding to a 0.0002 fraction contamination. The p-value for detecting 5 out 10 trials outside the ellipse is ~ 0.001, which means the observation was unlikely to have occurred by chance, and thus the genomic content of the two samples is NOT identical.

Summary

The main cause why direct population sequencing analysis has not been widely adopted is the paucity of accurate, effective, easy-to-use bioinformatics tools that guide interpretation of major and minor genome content [21]. We examined the application of NGS bioinformatics tools for forensic analyses of bacterial samples. We evaluated them against specially prepared samples and used these results to elucidate the benefits, caveats, and potential pitfalls of direct-sequence-analysis technologies. We emphasized the importance of quantitative approaches and statistical analysis tools for hypothesis testing as the basis for forensic analysis of NGS bacterial genomic data. As with the case of forensic analysis of human DNA for identification [22,23], distinguishing sequence observations from noise and random chance due to frequencies in a population is the first critical step in the use of such data in courts of law. This task is particularly daunting in the forensic application of direct-NGS for metagenomic samples, as sequencer output depends not only on reagent lots and sequence-dependent biases in the sample processing, generation of artifacts and

base calling output, as well as "biological noise", which is the unknown population structure in a given sample, and its detailed dependence on the culturing conditions [24,25]. These factors all highlight the need for high levels of standardization in bacterial sample preparation and end-to-end sample preparation in the NGS analysis of forensic samples.

We evaluated a calibrant as a means to capture the noise impacts on detection results. Our approach was successfully applied in the context of triaging for constituent organisms in a sample. The calibration procedure provided a basis for a hypothesis test for detection, and the approach enables selection of only a small subset of reference genomes to be use in a subsequent step for detailed, computationally intensive alignment-based analysis. This triaging was incorporated into a software tool called statMap, which triages on a time scale ~2 hours per 0.5 GB of sequencing data analyzed against the entire Genbank database of reference genomes. Our work highlights the value of calibration (different noise characteristics were observed in separate NGS runs on identical samples), but also effects of bacterial population structure on such analyses, and the potential need for multiple calibrant organisms in an analysis. Our methodology also addressed the fact that organisms in forensic samples will never be exact matches to reference database genomes. This is especially important for novel organisms, where the use of a contextual analysis, when genomic distances are placed in the context of typical distances among known bacterial species, such as all *Y. pestis* strains, can be useful in the characterization of a novel organism constituent.

For alignment-based comparisons of organisms, we illustrated potential impacts of population structure and algorithm parameters on SNP calling with respect to a reference genome. We compared NGS processing pipelines based on off-the-shelf tools that make different speed-versus-accuracy tradeoffs to reveal discrepancies in results and elucidate their origins. Our observation, in our Pipeline B analysis, of detected variants from a single colony extract, strongly suggests uniquely defining SNPs as the consensus variants, i.e., the variant that exceeds 50% of the base calls at a given genome position following alignment. The potential that this continuum of variation and the nu MBer of variants exceeding this threshold will depend on specific growth conditions, suggesting a need to develop standards for growth prior to forensic analysis, even at the coarse SNP-calling level of analysis of a bacterial population [26]. Clearly, avoiding the growth requirement altogether is the most desirable scenario.

We also investigated an approach to minor variant detection in the presence of the background "noise" of population structure in a sample. We applied the concept of the ROC curve to encode the relationship between sensitivity of detecting variants against the noise background and the associated false alarm rate (or probability of observing the nu MBer of detected variants or more by chance). To illustrate the concept, we diluted one sample of Bg into a different Bg strain at ratios of 1:3 and 1:10. In both cases, we could detect all the variants at approximately the fractions expected; however, one variant was detected in both samples. As none of the other variants were detected in this control, we conclude that contamination was unlikely, and that the local variant observed is either a minor contributor to the Bg stock parent population structure, or, was generated by mutations during colony growth. In either case, cautious analysis of control samples in the use of NGS analysis for minor variant detection is required and application of hypothesis tests (i.e., ROC curve analysis) is recommended when quoting detection results using direct NGS analysis. Alternatively, estimating population constituent genomes through hyper-dilution and consensus sequencing of each colony reduces the noise, but is prohibitively labor intensive and costly.

Finally, we developed a novel approach for analysis for identity of total genomic content by direct comparative analysis of NGS reads for two samples. Critical to such analyses is the mitigation of base-calling noise and the fine population structure on the measurement. We devised a metric that mitigates noise contribution with the expense of a copy nu MBer requirement for detecting a sample constituent, whereby applying a threshold for detecting all *n-mer* words in samples. The use of this metric enables development of a model based on self-versus-self comparisons of NGS read sets for each sample. The analysis provides a hypothesis test p-value that can be used to exclude two samples as coming from the same source, and does so without use of a reference genome (the test is based purely on *n-mer* composition) and is valid for total unknowns. We believe such a test has the potential for field use in investigations, following future demonstration of its efficacy in larger data sets and in blind panel analyses.

Herein, we both developed and critiqued candidate algorithmic approaches to illuminate benefits, caveats and pitfalls of NGS use in four specific bacterial forensics contexts:

- Forensic characterization of bacterial constituents
- Forensic comparison of consensus sequences of bacteria
- Forensic detection of bacterial population variants
- Forensic comparison of total genomic content

Calibrant-based analysis demonstrates not only dependence of alignment and SNP-calling results on algorithmic parameters, but the continuum of population variants. The population structure within a single colony may arise from multiple factors, such as growth time and conditions. While there are a nu MBer of factors on how culture impacts population diversity in a sample, it is conceivable that the nu MBer of SNPs called based on a fraction exceeding 0.5 could indeed depend on such factors, making standardization of protocols a high priority for any comparative analysis requiring growth in culture. It would be most advantageous to avoid a culturing step and apply direct analysis by NGS, use our calibrant tool, characterize sample genome profile and assess attribution.

As a final note, our research illustrates utility robustness of this potentially near term, field-able approach for genome-reference-free forensic comparison. We have demonstrated that mixtures can be directly interrogated and with appropriate sequence information processing result in accurate identity of genomic content of samples. Hence, warranting our methods as an NGS-based rapid exclusion tool to support criminal investigation.

Acknowledgements

This study was funded by Department of Homeland Security contract Whole Genome Approach to Microbial Forensics (WGAMF) HSHQDC-10-C-00140. Samples were prepared and handled at the CUBRC BSL facility and cultured under standard procedures. DNA from serial passages of biothreat agents was extracted and sequenced. Sequencing was conducted at ECBC as per Illumina recommended protocols.

References

1. Budowle B, Schutzer SE, Breeze RG, Keim PS, Morse SA (2010) Microbial Forensics, Elsevier science.
2. Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. Nat Biotechnol 27: 455-457.
3. Nagasaki H, Mochizuki T, Kodama Y, Saruhashi S, Morizaki S, et al. (2013) DDBJ Read Annotation Pipeline: A Cloud Computing-Based Pipeline for High-Throughput Analysis of Next-Generation Sequencing Data. DNA Res 20: 383-390.

4. Camerlengo T, Ozer HG, Onti-Srinivasan R, Yan P, Huang T, et al. (2012) From sequencer to supercomputer: an automatic pipeline for managing and processing next generation sequencing data. AMIA Summits Transl Sci Proc 2012:1-10.
5. Jakupciak JP (2013) Population-Sequencing as a Biomarker for Sample Characterization, J. Biomarkers. In press.
6. Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaya I, et al. (2013) MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol 14: R2.
7. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet11: 31-46.
8. Altshul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
9. Jakupciak JP, Colwell RR (2009) Biological agent detection technologies. Mol Ecol Resour s1: 51-57.
10. Fritz MHY, Leinonen R, Cochrane G, Birney E (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Res. 21: 734- 740.
11. Bertelli C, Greub G (2013) Rapid bacterial genome sequencing: methods and applications in clinical microbiology. Clin Microbiol Infect.
12. Allhoff M, Schönhuth A, Martin M, Costa IG, Rahmann S, et al. (2013) Discovering motifs that induce sequencing errors. BMC Bioinformatics S1.
13. Fraser CM, Read TD, Nelson KE (2004) Microbial Genomes. Humana Press.
14. Kircher M, Stenzel U, Kelso J (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol 10: R83.
15. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: characterizing errors in Ion Torrent PGM data. PLoS Comput Biol: e1003031.
16. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12: 38.
17. Liu YC, Schmidt B (2012) Long read alignment based on maximal exact match seeds. Bioinformatics 28: i318-i324.
18. Blom J, Jakobi T, Doppmeier D, Jaenicke S, Kalinowski J, et al. (2011) Exact and complete short-read alignment to microbial genomes using Graphics Processing Unit programming. Bioinformatics. 27: 1351-1358.
19. Whiteford N, Haslam N, Weber G, Prügel-Bennett A, et al. An analysis of the feasibility of short read sequencing. Nucleic Acids Res 33: e171.
20. Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13: 36-46.
21. Sherry NL, Porter JL, Seemann T, Watkins A, Stinear TP, et al. (2013) Outbreak Investigation Using High-Throughput Genome Sequencing within a Diagnostic Microbiology Laboratory. J Clin Microbiol 51: 1396-1401.
22. Keating B, Bansal AT, Walsh S, Millman J, Newman J, et al. (2013) First all-in-one diagnostic tool for DNA intelligence: genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip. Int J Legal Med 127: 559-572.
23. Vallone PM, Jakupciak JP, Coble MD (2007) Forensic application of the Affymetrix human mitochondrial resequencing array. Forensic Sci Int Genet1: 196-198.
24. Etienne KA, Gillece J, Hilsabeck R, Schupp JM, Colman R, et al. (2012) Whole genome sequence typing to investigate the Apophysomyces outbreak following a tornado in Joplin, Missouri, 2011. PLoS One 7: e49989.
25. Huson DH, Richter DC, Mitra S, Auch AF, Schuster SC (2009) Methods for comparative metagenomics. BMC Bioinformatics 10: S12.
26. Cummings CA, Bormann Chung CA, Fang R, Barker M, Brzoska P, et al. (2010) Accurate, rapid and high-throughput detection of strain-specific polymorphisms in Bacillus anthracis and Yersinia pestis by next-generation sequencing. Investig Genet 1: 5.

This article was originally published in a special issue, **Bioinformatics for High-throughput Sequencing** handled by Editor: Dr. Heinz Ulli Weier, Lawrence Berkeley National Laboratory, USA