Investigative Genetics

REVIEW Open Access

Validation of high throughput sequencing and microbial forensics applications

Bruce Budowle[1,2*], Nancy D Connell[3], Anna Bielecka-Oder[4], Rita R Colwell[5,6,7,8], Cindi R Corbett[9,10], Jacqueline Fletcher[11], Mats Forsman[12], Dana R Kadavy[13], Alemka Markotic[14], Stephen A Morse[15], Randall S Murch[16], Antti Sajantila[1,17], Sarah E Schmedes[1], Krista L Ternus[13], Stephen D Turner[18] and Samuel Minot[13]

Abstract

High throughput sequencing (HTS) generates large amounts of high quality sequence data for microbial genomics. The value of HTS for microbial forensics is the speed at which evidence can be collected and the power to characterize microbial-related evidence to solve biocrimes and bioterrorist events. As HTS technologies continue to improve, they provide increasingly powerful sets of tools to support the entire field of microbial forensics. Accurate, credible results allow analysis and interpretation, significantly influencing the course and/or focus of an investigation, and can impact the response of the government to an attack having individual, political, economic or military consequences. Interpretation of the results of microbial forensic analyses relies on understanding the performance and limitations of HTS methods, including analytical processes, assays and data interpretation. The utility of HTS must be defined carefully within established operating conditions and tolerances. Validation is essential in the development and implementation of microbial forensics methods used for formulating investigative leads attribution. HTS strategies vary, requiring guiding principles for HTS system validation. Three initial aspects of HTS, irrespective of chemistry, instrumentation or software are: 1) sample preparation, 2) sequencing, and 3) data analysis. Criteria that should be considered for HTS validation for microbial forensics are presented here. Validation should be defined in terms of specific application and the criteria described here comprise a foundation for investigators to establish, validate and implement HTS as a tool in microbial forensics, enhancing public safety and national security.

Keywords: Microbial forensics, Validation, High throughput sequencing, Sample preparation, Library preparation, Bioinformatics

Background

Microbial forensics involves analysis of microbe-related materials found at a crime scene, suspected laboratory, and so on, for forensic attribution and, thus, can be pivotal for developing investigative leads. Attribution (assigning to a source) can be defined as the characterization of a sample with the greatest specificity, which in the case of a microorganism would be at the species or strain level and ideally at the level of the isolate or even the culture vessel (for example, flask) from which the sample originated. High throughput sequencing (HTS) vastly improves the possibility that the forensic and scientific communities will be able to assign features to bio-forensic evidence, such as specific identity for unknown or emerging pathogens, sample or microbe origin, antibiotic sensitivity, evidence of genetic engineering and virulence profile. Now that a number of laboratories can afford HTS systems, community-accepted validation guidelines or standards are needed. As with any analytical tool(s) for forensic application, the utility of HTS operating conditions and tolerances and interpretation guidelines must be carefully defined. Guiding principles must be established to validate HTS systems. Here we define the criteria and offer a process for validation of HTS systems in microbial forensics. If methods are validated within the framework outlined here, microbial forensics will achieve an ever

* Correspondence: bruce.budowle@unthsc.edu
[1]Department of Molecular and Medical Genetics, Institute of Applied Genetics, University of North Texas Health Science Center, Fort Worth, Texas, USA
[2]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia
Full list of author information is available at the end of the article

higher level of power and analytical value and, ultimately, greater protection for the public and the nation's safety and security.

Introduction

More than a decade ago the United States experienced a simple but effective biological attack in which *Bacillus anthracis* endospores were placed in envelopes and delivered by the US postal service to intended victims [1-4]. The Federal Bureau of Investigation initiated the Hazardous Material Response Unit in 1996 to undertake a forensic investigation of bioterrorism events. Despite this effort, in 2001 the forensic infrastructure was inadequately prepared to analyze and interpret the available microbiological evidence to assist in determining who did and did not have the capacity to perpetrate such an attack. In fact, much of the needed forensic science applications had not yet been developed or validated. As part of an immediate national response to investigate such crimes, the field of microbial forensics was born [5-7] and its emergence was accelerated by the urgent requirement to investigate the anthrax mailing attacks.

The foundations of the field of microbial forensics lie in public health epidemiology and its practices [6-10] as well as agriculture practices [11-13]. Microbial forensics involves analysis of microbe-related materials found at a crime scene, suspected laboratory, and so on for forensic attribution (assigning to a source) and, thus, can be pivotal for developing investigative leads. Attribution in the case of microbial forensics can be further defined as the characterization of microorganisms within a sample to the species or strain level and ideally to the specific isolate or culture vessel from which the sample originated. Indeed, metagenomic approaches to assess microbial composition of samples also may provide strong microbial forensics evidence (either phylogenetically by identifying a specific target organism in the complex sample or by abundance spectrum profile) to attribute sample(s) to source. Scientific attribution also eliminates as many other candidate isolates or sources as possible and supports both investigation and legal proceedings.

The standards and requirements for microbial forensic practices are less well defined than those within human identification and other established forensic disciplines. However, establishing the validity of microbial forensic methods and their use and interpretation contributes to acceptance, admissibility, confidence, value and weight of physical evidence in the jurisprudence process [14] as well as within the military, intelligence and homeland security sectors that have the responsibility to act upon data and reports associated with suspected bioterror activities. Within two years following the anthrax letter attacks, the FBI's Scientific Working Group for Microbial Genetics and Forensics (SWGMGF) formalized and published Quality Assurance (QA) guidelines [7]. The motivation for establishing a QA system was to put quality practices in place to ensure that microbial forensic evidence was analyzed using the best practices possible and that the interpretation of results was based on extant data and sound principles.

The SWGMGF QA guidelines were a good first step in establishing a QA system for microbial forensics and for increasing confidence in the data generated. However, as technologies advance and application of microbial forensics expands beyond the traditional law enforcement communities, it becomes increasingly important to continue to build upon the SWGMGF guidance and previously published microbial validation methods [7] to reflect the current state-of-the practice and foster greater community wide acceptance. Significant drivers to expand validation guidance include the substantial developments and applications of next-generation or HTS. For perspective, the first bacterial genomes that were sequenced in 1995 by the Institute of Genome Research (Rockville, MD, USA) [15,16] took more than three months to complete. Although HTS technology was initially developed, in part, for characterizing human genomes [17-19], these instruments have increasingly been used successfully to characterize unknown microbes in samples of varying complexity [20-42]. Within the field of microbial forensics [7,43,44], HTS combined with powerful bioinformatics capabilities offers a powerful tool to characterize forensic bio-evidence, including unknown microorganisms, genetically-engineered microorganisms and low-abundance (or trace) microorganisms present in complex mixed samples with extremely high sensitivity [45]. HTS technologies have features that make them more desirable and accessible for microbial forensic analyses than Sanger sequencing [46], including high throughput, reduced cost (on a per nucleotide or per genome basis) [47] and large-scale automation capability. Millions of sequencing reactions can be performed in a massively parallel fashion in a single instrument run [48-53]. With many copies sequenced at any desired portion of the genome (known as coverage), consensus sequence accuracy can be increased far beyond the per-read accuracy rate. As the throughput and accuracy of HTS continues to increase, more samples can be multiplexed in a single run without sacrificing depth of coverage or more complex samples may be analyzed at a greater depth of coverage.

Several HTS platforms are available and currently used for microbial sequencing, usually based on massively parallel sequence by synthesis strategies with high accuracy in a reduced footprint compared with Sanger sequencing. The primary HTS platforms include the HiSeq and MiSeq from Illumina (San Diego, CA, USA), the Ion PGM and Ion Proton Sequencers from ThermoFisher (South San Francisco, CA, USA) and the 454 systems from Roche

(Pleasanton, CA, USA). The Illumina NextSeq 500 system is the latest platform on the market in this desktop category with 150 Gigabase throughput and 'push-button simplicity'.

Another type of sequencing chemistry, developed by Pacific Biosciences (PacBio, Menlo Park, CA, USA), is the first to utilize single molecule real time (SMRT) sequencing, in which each base is detected in real time as a polymerase adds fluorescently tagged nucleotides along single DNA template molecules. SMRT sequencing is distinct from the other HTS technologies in providing very long read lengths. The average read length with the PacBio RS instrument is approximately 3,000 bp and can reach up to 20,000 bp [54]. Furthermore, examining the polymerase kinetics of SMRT sequencing allows for direct detection of methylated DNA bases [55]. This intrinsic capability of the SMRT sequencing workflow does not affect primary sequence determination, while yielding yet another forensic signature that is not captured with standard protocols on other HTS instruments.

A new and potentially revolutionary sequencing platform in development by Oxford Nanopore (Oxford, United Kingdom) will allow a single DNA molecule to pass through a protein nanopore set within an electrically resistant membrane bilayer. The resulting cross-membrane current fluctuations are used to identify the targeted nucleotide [56]. The company projects sequencing rates initially will be 20 bases per second, increasing to 1,000 bases per second in the future, and providing read lengths up to tens of thousands of bases [57]. While these individual reads will contain a larger number of errors than the other mentioned HTS instruments, the PacBio (and potentially the Oxford Nanopore) errors are random. With redundant interrogation of the same base of a circular template with SMRT sequencing and with sufficient depth of coverage, highly accurate consensus calls can be obtained [54].

HTS vastly improves the possibility that the forensic and scientific communities will be able to assign features (for example, strain identity, virulence profile, and so on) and, ultimately, attribution to bio-forensic evidence. However, these improvements cannot be realized or known with any level of statistical confidence without effective and validated bioinformatics tools to process, analyze and interpret the large amounts of HTS data generated. Most application-oriented laboratories are unlikely to have in-house bioinformaticians, and even for laboratories with such resources, a comprehensive data analysis pipeline must be defined and validated to establish that the software algorithm(s) reliably analyze sequence data and produce accurate final results. Many bioinformatic tools are available within commercial, academic and other open sources. However, the specific tools employed or developed are highly dependent on the need and intended use of that laboratory and may not have been rigorously tested. An appropriate data analysis pipeline must be implemented and fully validated, including understanding the uncertainty and error associated with each step of the process, as well as the collective uncertainty. The appropriate interpretation and weight of the evidence must be employed successfully and effectively communicated.

Now that laboratories are implementing HTS systems, community-accepted validation guidelines or standards are needed. Development of HTS technologies and associated bioinformatics tools will continue to progress rapidly, and, no doubt, increasingly powerful tools will be available to support microbial forensics. HTS applications for microbial forensics include assembly of draft and finished single genomes of microorganisms, targeted site sequencing, metagenomics (both amplicon sequencing of conserved genes for microbial community structure and shotgun sequencing for profiling the content of a sample), and source attribution, including profiling, sample comparison, sample engineering, and other microbial evolution or epidemiology applications. As with any analytical tool(s) for forensic application, the utility of HTS operating conditions and tolerances must be carefully defined. Regardless of the variation in technologies and software, guiding principles, such as the criteria listed in Table 1, must be established to validate HTS systems. Here we define the criteria and offer a process for validation of HTS systems in microbial forensics. Rather than delineating a set of protocols for a particular set of tools and reagents that apply to a limited set of instances, which may quickly become obsolete, those tools and reagents universally needed for protocol validation are described. By addressing each area described below, an investigator will be able to establish, validate and implement HTS as a tool for microbial forensics.

Application and validation of HTS for microbial forensics

Microbial forensic applications of HTS include single isolate sequencing with *de novo* assembly, read mapping, targeted sequencing of specified genes or other regions of interest (which generally include diagnostic markers, for example, SNPs, indels, and so on) [63,64], and metagenomics. Metagenomics analyzes by sequencing DNA (or RNA) samples to identify or describe microbial community composition of environmental samples such as soil [65], plants [41,42], sea water [66,67], hospital environments [68] and human-associated habitats [69,70]. HTS makes metagenomics readily feasible since culturing is not required for sample enrichment. HTS and associated bioinformatic technologies make it possible to detect microorganisms of interest when they are present in low abundance and differentiate them from near neighbors by using diagnostic genomic signatures.

Table 1 Validation criteria for analytical performance metrics

Criteria	Definitions
Analytical sensitivity	Likelihood that the assay will detect a target (for example, organism variant, sequence region, functional element, and so on) in a sample (that is, target), if present; can include target attribution when defined as strain- or isolate-level detection. Also known as the true positive rate. Calculated by dividing number of true positives by the sum of true positive and false negatives (TP/(TP + FN)).
Analytical specificity	Likelihood that the assay will not detect a target, if not in the sample; can include false target attribution. Also known as the true negative rate. Calculated by dividing true negatives by the sum of true negatives plus false positives (TN/(TN + FP)). May be impractical to calculate for methods designed to detect the known universe of organisms.
Precision	The degree that individual measurements of the same sample are similar with regard to the presence and absence of target. Determined by the distribution of random errors and not the true or underlying value.
Accuracy	Degree that the material measured is similar to its true value. Calculated by (TP + TN)/(TP + FP + FN + TN).
Reproducibility	The degree to which the same result(s) is obtained for a sample when the assay is repeated between/among different operators and/or detection instruments.
Repeatability	The degree to which the same result(s) is obtained for a sample when the assay is repeated by the same operator and/or detection instrument.
Limit of detection	Minimum level of input material for a target as a proportion of the total at which all replicates are consistently positive for that target.
Reportable range	The region(s) of genome(s) that are sequenced and from which information is drawn for comparison or attribution.
False positive rate	The rate at which a target is incorrectly called as present. Also known as Type I error. Calculated as 1 – specificity
False negative rate	The rate at which a target organism is incorrectly called as absent. Also known as Type II error. Calculated as 1 – sensitivity.
Assay robustness	Stability of analytical performance under variable conditions, that is, likelihood of assay success.
Reference materials[a]	Materials/samples used to test the performance of the assay (for example, reference panels of the target and mock or non-probative materials) relevant to the intended application of the assay.
Databases[a]	Collection of data and reference genomes, genes and genomic elements to be used for interpretation of results.
Interpretation criteria for results[a]	Analysis (quantitative or qualitative) used and confidence level of a result (match, association, most recent common ancestor, and so on).

[a]These last three items – Reference materials, Databases, and Interpretation criteria – typically have not been considered validation criteria. However, they have been included here primarily because interpretation of results is an essential part of generating reliable and appropriate results, which should be described within a standard operating protocol (SOP). The data used to test a system are reliant on reference materials and, depending on the situation, databases. See [58-62].

Customers, stakeholders, the judicial system and the public expect forensic methods to be validated, when feasible, prior to use on forensic evidence. A validation process is essential in the development of methods for microbial forensics, and such methods must be reliable, defensible and fit for purpose.

Validation has been described as the process that:

1. Assesses the ability of procedures to obtain reliable results under defined conditions.
2. Rigorously defines the conditions that are required to obtain the results
3. Determines the limitations of the procedures.
4. Identifies aspects of the analysis that must be monitored and controlled.
5. Forms the basis for the development of interpretation guidelines to convey the significance of the findings [58].

While these general principles apply to HTS technologies and guidelines specifically for HTS used in metagenomic profiling already exist [71], there are challenges that arise when validating HTS for microbial forensics that require further consideration. Here we describe the specific guidelines for validating HTS technologies so that the microbial forensics community (and others) will have a common protocol and lexicon to leverage the exciting potential of HTS while maintaining high quality and confidence under rigorous scrutiny when this technology is used to support investigations of bioterrorism or biocrimes.

General considerations for validation

The requirements for validation will vary according to the process in question and should be defined in terms of the specific application. While full developmental and internal validation is ideal [7,58], this requirement may not be practical for all situations, such as an attack involving a novel agent not in previously validated systems. Indeed, the use of multilocus variable number of tandem repeat (VNTR) analysis [72] to determine that the strain of *B. anthracis* in the 2001 letter attack was Ames was not a fully validated procedure in casework analysis. Yet, it was sufficiently developed for investigative lead value [73].

Because of the vast and incompletely described biological diversity of microbes and the potential of having to deal with a large number of samples in a microbial forensic case, it is not possible to validate every scenario. Moreover, HTS and bioinformatics technologies are changing rapidly and will continue to be improved in the immediate and long-range future. Lastly, exigent circumstances may require immediate response, and microbial forensics should be able to lend support using all available tools. For such unforeseen circumstances preliminary validation may be 'carried out to acquire limited test data to enable the evaluation of a method for its investigative-lead value, with the intent of identifying key parameters and operating conditions and of establishing a degree of confidence in the methods of collection, extraction, and analysis' [74]. However, once general validation is accomplished for instrumentation, bioinformatics data analysis, and Standard Operating Protocols (SOPs), only novel aspects of validation for new targets may be needed to generate informative leads and to make public health decisions with associated levels of confidence. Therefore, it is extremely important to establish comprehensive criteria for validation of HTS technologies with all aspects of the validation study documented. The fact that a validation study is preliminary should be stated clearly, with the limitations of the assay and validation study clearly described. However, validation of finalized SOPs is essential for reliable and defensible use of HTS technologies in microbial forensics. Sample collection and storage have been addressed elsewhere [75] and will not be described here. Validation of the HTS process addressed here relies, in part, on reports available in the literature [59-61,76] that have defined validation requirements for HTS applied to human clinical genetic analyses. The validation guidelines for the three major technical components of HTS (sample preparation, sequencing and data interpretation) as related to the field of microbial forensics, are presented in the following sections.

Sample preparation

Nucleic acid extraction – quantity and purity

Validation should include anticipated sample types and matrices of those sample types. A range of routinely anticipated types of samples incorporating an array of quality and quantity of nucleic acids, environmental matrices, inhibitors of downstream analytical processes and biological contaminants expected to impact reliability, specificity and obtaining results, should be included.

Template DNA (or RNA, even though DNA is referenced here) must be of sufficient quantity and quality for library preparation and sequencing. The amount of DNA available will influence the library preparation method used. At the time of preparation of this manuscript, for example, the TruSeq (Illumina, Inc.) sequencing preparation method requires approximately 100 ng to 1 μg [77], Haloplex (Agilent, Santa Clara, CA, USA) 225 ng [78], Nextera XT (Illumina) 1 ng [79], and polymerase chain reaction (PCR)-based methods, though variable, may require less than 1 ng. Minimum and maximum DNA requirements for analysis should be established using a laboratory's work flow. A set of guidelines is needed to establish what levels of prepared DNA may be insufficient or compromised and how to proceed under such circumstances (for example, analyze anyway, stop, or select an alternate assay). Metrics based on precise quantitative pre-analytical sample characterization are needed to assess the fraction of template molecules that meet the requirements for downstream analyses, which is important for amplicon sequencing and shotgun sequencing. It is likely that samples from which the DNA is insufficient, damaged and/or inaccessible will be encountered, especially when collected from the environment. This information will be helpful to assess and compare potential downstream partial and/or complete loss of target data. The DNA extraction method used should be tested for yield and sufficient purity for downstream analytical processes. Additional extraction processes may include separating a particular genome from a metagenomic sample or selective filtration to separate specific types of microbes, such as virus particles in a metagenomic sample [71,80] or methylated DNA from non-methylated DNA [81]. Since host DNA or background genome(s) may comprise a major component(s) of a given metagenomic sample, the ability to sequence minor components of complex samples may be affected. Purification procedures used to maximize the yield of targets of interest should be evaluated the same as the nucleic acid purification process. Lastly, proper positive and negative controls should be included to assess process performance and laboratory background contamination, respectively.

Enrichment and library preparation

DNA samples, single source or metagenomic, may be enriched for specific target regions of genomes using a capture approach or PCR. For many enrichment processes the desired genomic regions should be known and defined in order to design the protocol. However, whole genome amplification methods such as non-specific or degenerate PCR primers, [82,83] including multiple displacement amplification [84], can be used. The methods used for genome amplification can impact the results by introducing contaminating chimera formation and sequence bias [71], and should be considered, depending on the method or assay during validation.

Capture- and PCR-based methods have both advantages and limitations. PCR-based methods provide greater sensitivity of detection, but are likely to produce greater error

from mis-incorporation by the polymerase than would be generated with a capture approach. PCR-based methods, in which a multiplex panel of markers may be considered, will require development of primer sets that amplify the targeted sites in a balanced fashion (or at least describe any significant imbalance) and do not cross-hybridize to unspecified targets. In contrast, capture methods will require more template DNA and would not provide the limit of detection necessary for microbial forensic analyses of trace materials. Regardless of the methods listed here or new ones subsequently introduced, it is incumbent upon the analyst to define validation criteria that address the advantages and limitations of enrichment.

Whether or not a sample is enriched, the next step in sample preparation is library preparation where the DNA sample is modified for sequencing. DNA is typically fragmented into shorter pieces by mechanical shearing (for example, sonication) or enzymatic fragmentation (for example, tagmentation [79,85]). Adapters are added to each fragment of DNA to facilitate clonal amplification prior to sequencing of the cloned fragments. Adapters can be incorporated into existing amplicon fragments during PCR. With long PCR amplicons, fragmentation may be required. DNA fragments and/or PCR amplicons then are size-selected for the range appropriate for down-stream sequencing and quality assessment. This process generates a library of millions of cloned fragments that are ready for sequencing. Quality must be assured by testing reproducibility of library preparations and robustness of indexing (described below) to identify (or misidentify) labeled fragments. Internal controls to monitor enrichment and library quality should be considered.

Multiplexing

Multiplexing with HTS can be achieved by barcoding (or indexing) [86,87]. Short unique sequence tags are added to every fragment of a sample during library preparation to 'tag' the fragments unique to a sample. Thereby, samples can be pooled (or multiplexed) and data separated (that is, demultiplexed) after sequencing, based on the unique tagged sequences. With the high throughput capacity afforded by HTS, many different samples may be sequenced simultaneously. For example, the MiSeq and Reagent Kit V2 (Illumina) is capable of generating more than 7.5 to 8.5 Gbp using a 2 × 250 paired-end run (about 39 hours sequencing run time). With 8 Gbp of sequence data, 16 samples can be multiplexed on a single run assuming desired 100× coverage of a 5 Mb bacterial genome ($5e^6$ bp genome × 100× coverage × 16 samples = $8e^9$ bp MiSeq output). This calculation is just an example and will change as throughput and read lengths increase, which is likely to occur relatively quickly and often. As the throughput of HTS continues to increase, more samples could be multiplexed in a single run without sacrificing depth of coverage or more complex samples may be analyzed at a greater depth of coverage. In theory, hundreds to thousands of barcodes could be synthesized, but currently 12 to 384 different reference samples can be pooled in a single reaction [86,87]). The Earth Microbiome Project provides >2,000 barcodes that could be combined, theoretically enabling multiplexing of >4 million samples in a single sequencing run [88]. Depending on the target, for example, single source samples, the number of samples that can be barcoded and sequenced in a single run should be predictable. The performance of barcoding to identify specifically tagged samples should be evaluated and documented. Furthermore, when feasible, use of different indexes in sequential sequencing runs on an instrument can indicate if carry-over contamination has occurred, which offers another quality control for monitoring potential impact of contamination on sequencing results.

Sequencing

Each HTS system employs a unique chemistry for sequence determination and each will have to be validated in general and then specifically according to applicable features of the system [51,52,89-93]. For example, chemistries employed by 454 and Ion Torrent systems tend to be less accurate than Illumina-based chemistry for sequencing homopolymers. The type of sequencing, either single-end (fragments sequenced at one end only) or paired-end (both ends are sequenced) can impact coverage and stringency in different ways. Sanger sequencing, still considered the gold standard, allows for some concordance testing (that is, comparative analysis of the same target sequence with different analytical systems). However, there is no guarantee that the gold standard always provides the correct result. For example, Harismendy *et al.* [94] sequenced 266 kb of portions of six ion channel-related genes using Sanger sequencing, three HTS platforms, and one microarray platform and compared the results. The portion of false negative (FN) and false positive (FP) single nucleotide polymorphisms (SNPs) attributed to Sanger sequencing were 0.9% and 3.1%, respectively. Moreover, the lower throughput and coverage of Sanger sequencing makes it impractical for concordance testing with HTS generated data. The data generated by HTS are so much greater per run than those generated by Sanger sequencing that only limited sampling and very short regions can be reasonably compared. Instead concordance testing may be better achieved by testing orthogonal HTS systems with templates of 'known' genome sequence. Potential errors and biases inherent in each HTS system may be determined and documented better in this manner. For each sample type and platform, the error rate (and error profile) of sequencing can be determined only by empirical testing. The data can be used to define limitations of the current system that should be part of an

interpretation SOP. In addition, orthogonal testing allows for identifying weaknesses and enables assay improvements before implementation. Where possible, orthogonal analyses should be employed for validating HTS methods.

Data analysis and interpretation

The final major components of HTS validation are data analysis and interpretation of results. Bioinformatics is essential and critical because of the massive amount of data, the requirement to answer forensic and investigative questions using the data, and the questions that may be addressed with trace samples, complex samples, potential genetic engineering, and background endemicity (that is, microorganisms that generally are found at a location). Comprehensive data analysis pipeline(s) should be defined and validated. It is important to establish that the software algorithms reliably analyze sequence data to produce accurate final results. The flow of data generally progresses through base calling, quality control and, finally, downstream taxonomic, functional and/or comparative analysis (which is generally either alignment- or fragment-based, if assembly is not performed) (Figure 1).

Quality metrics, generated during the analytical process, include: quality scores for base calling, read-level quality control (QC) (to include trimming of low quality bases on fragment ends), alignment, GC content, depth of coverage, strand bias and variant calling. Base calling, the identification of the specific nucleotide present at each position in a single read, should be part of instrument software. A quality threshold of base scoring is typically set with a Q score. A threshold of Q20 sets the minimum base call accuracy at 99% allowing for an incorrect base call per read at 1 in 100, while a Q30 score sets the accuracy at 99.9% and incorrect base call per read at 1 in 1,000 [95]. A Q score threshold should be set for both validation studies and subsequent implementation. However, there are no guidelines that suggest that, for example, a Q20 score is a requirement. A less than Q20 score may not impact accuracy as subsequent coverage and annotation may be adequate. Under defined conditions and for investigative leads or exigent circumstances the quality score may be relaxed; however, the justification or reliability of a lower score must be documented. Each HTS platform will have specific sequencing limitations and errors: signal-intensity decay over the read, erroneous insertions and deletions, strand bias, and so on. These limitations should be described and defined.

Accuracy of identifying sequence variants (for example, SNPs, indels, chromosomal rearrangements, copy number variants) is dependent on a number of factors that include base calling and alignment as well as choice of reference genome, depth of sequence coverage (as well as average coverage), and sequence chemistry/platform. Because alignment involves arranging a read with a reference sequence (targeted or whole genome), different alignment strategies can and do produce different results (Figure 2). Differences in alignment will vary with software, so rules for alignment should be defined for consistency and traceability.

Choice of a reference genome, if used for alignment is important. Because the reference will vary by species and circumstance, specific criteria for selection are not provided here. However, it is incumbent upon the analyst to develop criteria for the reference genome that is selected. For example, from a microbiological perspective, a reference may be one that is accessible, is relevant as the Type strain, is relevant from a public health perspective, or is well-defined microbiologically; and from a computational perspective, it may be one, several or a collection of genomes, or the optimal computational sequence, and should be curated, such as a finished genome. Validation should define the purpose of the reference genome and describe the criteria for selection.

Minimum criteria should be established by the laboratory for the output of the selected analytical method, such as depth and uniformity of coverage. Defining output thresholds for metagenomic samples may be difficult given the immense quantity of data and microbial diversity; therefore, single source samples and defined mixtures can be used as a guide. These limitations may be necessary in defining FNs and FPs. Clearly, there will be ambiguous calls due to sequencing noise and novel genome composition. The specific parameters and settings used to establish thresholds, FP and FN rates should be detailed thoroughly to enable sound interpretation and accurate comparison to alternative methods and protocols.

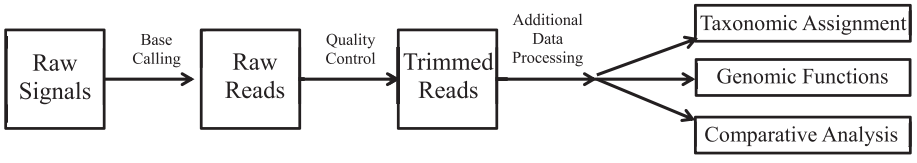

Figure 1 Basic schematic of data flow through an analysis process. The first step of base calling generally is completed by the instrument software, and each downstream step must be included in the validated analytical pipeline. Additional data processing after generating sequence reads is required, for example with contig building and/or alignment, and will depend on the application.

```
Reference genome:  GGCCGCATCTCTTGAAGGCC
Aligned read 1:    ....GCA--TCTTGAA....
Aligned read 2:    ....GCATCT--TGAA....
```

Figure 2 Alternate alignments of identical sequences. Reads 1 and 2 are aligned in equally optimal ways that indicate different locations for a 2 bp deletion relative to the reference. Differences in alignment can be problematic when an evidence sample's consensus alignment is based on a different approach than that of the reference sample or entries in a database.

Software may be adequate or somewhat limited with respect to the accuracy of variant calling. The limitations should be described and quantified, and algorithms may need to be modified to address specific limitations. The method(s) of identification and annotation should be described. Different formats are available for exporting variants and annotations [59]. The file format should include 'a definition of the file structure and the organization of the data, specification of the coordinate system being used, e.g., the reference genome to which the coordinates correspond, whether numbering is 0-based or 1-based, and the method of numbering coordinates for different classes of variants, and the ability to interconvert to other variant formats and software' [59].

The FP and FN rate often are determined for most analytical assays. However, there are additional considerations with HTS and microbial identification. Similar to homologous regions and pseudogenes for human genetic variation testing [60], the homologous regions of a near neighbor (for example, genes shared across the bacterial kingdom) become important for target identification (target being the species or strain of interest). Generating a metagenomic profile at the resolution of the phylum level, or even the genus level, may indicate a general environment from which a sample originates but often cannot identify the microorganism of interest at the species or strain level. However, newer approaches have started to achieve strain level identification [96-99] by exploiting higher throughput and novel algorithms. The results can be accurate and reliable and can translate into identification of the target agent in an unknown complex sample. Many reliable reads of the sequence of any particular species will share the same sequence, particularly so with near neighbors. For example, while *Bacillus* species may be sufficiently diverse to discriminate in a particular assay, strains of *B. anthracis* are nearly indistinguishable from one another [100]. FPs must be defined by specificity and the ability to phylogenetically differentiate a species (or strain) from near neighbors, such as *Bacillus anthracis* and *Bacillus cereus*. Testing that a known single source sample fits in a phylogenetic schema is not the same as identifying a particular species in a simple or complex sample. Methods for identification of targets should be validated based on intended use. FN rate may be difficult to determine for metagenomic sample analyses as stochastic effects and sampling variance may impact detection of the target(s). Known data sets can be helpful to define the FN rate.

Once assay conditions and pipeline configurations have been established, the entire method should be tested prior to use. Although individual components may have been validated, it is imperative to demonstrate that valid and reliable results are obtained when the components are combined. The standard microbial forensics validation criteria [7,58] apply to HTS methods as well. Special attention should be given to accuracy, precision, analytical sensitivity and specificity, reproducibility, limits of detection, robustness, reportable range, reference range, either FN/FP or confidence, statements of findings and databases used (Table 1). The laboratory must select and be able to clearly and defensibly state the parameters and thresholds necessary to determine whether the overall sequencing run is of sufficient quality to be considered successful. Criteria should include error rate, percentage of target captured, percentage of reads aligned, average and range of coverage depth, and so on.

Reference materials

Reference materials (RMs) should be used during test validation. Well-characterized reference samples should be included to establish baseline data to which future test modifications also can be compared [60]. Many different types of samples can serve as RMs for HTS, including characterized DNA derived from specimens prepared from microbial cultures, samples collected from several different endemic regions with high incidence of microorganisms of interest, samples from several non-endemic regions discovered accidentally and described as isolated outbreaks or findings, synthetic DNA (sets of sequences of known isolates), or electronic data (that is, generated *in silico*). A gold-standard reference genome would contain a single gap-less sequence for each chromosome or plasmid in the source organism, with no ambiguous bases. RMs are homogeneous for single source samples, stable and defined. Because complex mixtures are likely to be unstable and subject to stochastic effects, simple mixtures should be used. *In silico* complex samples, which can be considered stable, are suitable for testing the bioinformatics pipeline. The sequences

used and parameters employed for testing should be documented. The same rationale can be applied to positive controls, which must be defined. Negative controls may include no-template controls, blank controls for different phases of the analytical process or DNA samples void of the target.

Bioinformatics software management

The bioinformatics community has not yet defined uniform guidelines or protocols for benchmarking software. Thus, users must fully validate and document their bioinformatics pipeline. Software may be open source, purchased from commercial entities, developed in-house, or come from a combination of sources. The software programs should perform general quality metrics assessment, but the software likely will differ in performance and potentially yield different results. Therefore, accurate versioning of the state of the software is essential [76], not just for validation but also for data analyses. The software and modifications must be tracked. Settings that can be modified by the user should be documented. Documentation also should include the specific version(s) of each component of the pipeline, the hardware, dates of use and changes to software. Each software upgrade requires revalidation of the steps downstream of HTS. Virtual Machines [101], which are software simulation(s) of a machine, encompass the entire computational environment used for analysis and can help accomplish comprehensive version control on this complete system. By maintaining informative curated reference datasets, validation of updates or changes to software pipelines may be facilitated without any additional HTS or with only minimal effort.

Analysis by computer software is an essential component of using HTS data. Two general criteria addressing software performance are verification and validation. According to the Institute of Electrical and Electronics Engineers (IEEE) Std 610.12-1990 [102], verification is 'the process of evaluating a system or component to determine whether the products of a given development phase satisfy the conditions imposed at the start of that phase,' and validation is 'the process of evaluating a system or component during or at the end of the development process to determine whether it satisfies specified requirements.' Since verification applies to whether the software system was built correctly and validation is whether the intended use was met, most users will only have tools at their disposal to perform a validation of software. To be considered validated, software must be tested using input data that replicate challenging real-world situations. Software can be integrated within the HTS process (for example, instrumentation) for data generation and external to the HTS system for interpretation (for example, phylogenetics, attribution, and so on). The software specifications should conform to the intended use [103]. Performance characteristics must define the entire process addressed by the software and individual process steps, as appropriate. Much of the above discussion on data generation addressed the criteria that relate to performance of data generation and subsequent interpretation and, thus, serves as a guide for software performance characteristics. Users who create software for intended applications should refer to the standards by the IEEE [102]. However, the majority of users in the application-oriented laboratory will obtain commercially-available software, and so validation likely will be at the 'black box' level. Even without access to the source code, users still are responsible for ensuring that the software performs the intended functions correctly. Regardless, since software requirements often are derived from overall system requirements for the HTS analytical tool, a systems approach is recommended for validation. The user's intended use and needs [103] drive the criteria for validation. When possible, the user can rely on the manufacturer's data for some of the validation, if the data are made available, or on the scientific literature. Nevertheless an internal validation demonstrating that performance criteria are met is required. Software is different than instrumentation in that it does not wear out [103], and likely will be modified for better performance over the lifespan of its use (however, the computer operating system can change, rendering the software incompatible with the newer system). Since software will be modified and updated, a validation analysis should be conducted for the specific change(s) following the same principles of validation. New problems may arise with the intended update and, therefore, any impact that modification may have on software beyond the change should be tested using a systems approach [103].

Data storage

Permanent storage of all raw HTS data is not practical as the raw data are exceedingly large. After base-calling, this information is routinely discarded. Sequence data should be in conventional, widely used format(s), for example, .fastq files for sequence reads, or be readily convertible to a standard format [59]. Many bioinformatics pipelines create output data structures that may be compressed and stored as an alternative to the .fastq read files. Other compression systems have been proposed for HTS data [104,105], but there may be utility in widely used methods like gzip compression that will likely remain in use for years to come. A best practice should be to create a process so that results can be re-analyzed as necessary when updates are made to the downstream bioinformatics systems. Other files to archive include: SAM/BAM (mapping data) and vcf (variants). These or

similar formats could be used to store alignments and data about known SNPs of special diagnostic power, such as canonical SNPs. Where possible, DNA samples should be stored for re-testing [76]. Because of limitations with large amounts of data, it is necessary that reference datasets are documented and maintained in order to perform validation of future software updates. Lastly, conversion of data from one format to another could create unforeseen transcription errors; therefore, defined data sets should be tested before and after data conversion for accuracy.

Interpretation and reporting

Interpretation of results for attribution should be defined clearly and documented. Equally important, the level of resolution possible with a particular system should be stated. Also, the database(s) used for validation (and for casework analysis) is likely to be expanded and improved with HTS technologies on a relatively rapid basis; so the records of the database(s) used for individual analyses must be maintained. The target areas that define a species or strain and resolve it from near neighbors are critical [100,106]. One or more sites may be required depending on phylogenetic resolution. A minimum number of targets and degree of confidence with the number of targets should be established [107]. The means by which a software pipeline determines attribution may not be accessible to the user, in which case all relevant output data and associated thresholds should be documented and stored in a standard way according to the SOP. The minimum number of reads is essential for limits of detection, stochastic effects, and FNs and FPs and should be defined empirically for obtaining a reliable result(s). An interpretation statement(s) and degree of confidence (qualitative or quantitative) should be developed regarding attribution of the sample, and that confidence, when feasible, should be based in a rigorous statistical framework.

Resequencing assembly can be effective if the reference dataset contains sequences of closely related reference genomes [71]. *De novo* sequencing is computationally more demanding. Thus, the software and, just as importantly, reference data sets are critical to result quality. There are a number of assembly algorithms that can take millions of short reads generated by HTS and translate them into a portion or complete genome sequence [108-112]. Each approach has benefits and limitations affecting quality and efficiency. Therefore, specific software used, standard metrics (for example, N50, coverage, contig sizes) assumptions and criteria applied should be documented [113,114].

While there may be some situations in which assembly is required, it is less likely to be used or even necessary for the foreseeable future in microbial forensics analyses and especially with mixtures or metagenomic samples where near neighbors, strains and coverage constraints reduce the practicality of assembly. Alignment strategies or fragment counting strategies are preferable to assembly. Sequence alignment compares DNA sequences (although it can apply to any kind of sequence data) and seeks to identify regions of homology. More often a reference sequence will be maintained, targeted short reads will be aligned with that reference, and differences with respect to the reference will be listed (as 'variants') [115-117]. In addition to the same documentation requirements for assembly strategies, the reference sequence should be fully documented. While we believe that alignment strategies will be favored over assembly strategies for metagenomic microbial forensic applications, if capabilities improve that enable effective assembly, then it is incumbent upon the user to validate the process.

Taxonomic assignment

Methods for read-based taxonomic classification of metagenomics data fall into two broad categories: composition-based and alignment-based. Composition-based approaches rely on comparing signatures of short motifs from a query fragment to a reference genome – for instance, a particular GC content, gene and protein family content, or k-mer frequency and distribution [71]. Composition based approaches include Phylopythia [118], PhylopythiaS [119], Phymm [120], the Naive Bayes Classifier [121], Sequedex [122], the Livermore Metagenomic Analysis Toolkit (LMAT) [97], GENIUS [96] and Kraken [99]. Alignment-based approaches compare reads to a set of labeled reference genomes using a basic local alignment search tool (BLAST)-based approach. Alignment based approaches include MEGAN, Bowtie, MetaPhlAn, MetaPhyler, CARMA, WebCARMA, IMG/M, MG-RAST, and others [98,116,123-132]. Additionally, methods for direct taxonomic classification of sequencing reads use a combination of both composition and sequence similarity approaches, such as MetaCluster [133], Rapid Identification of Taxonomic Assignments [134], and PhymmBL [127,128,135]. A more comprehensive review of sequence classification methodology and software is presented elsewhere [136].

Many programs use a phylogenetic approach to classify sequences and summarize results by taxonomic group. A sequence(s) can be assigned at any level from the phylum down to the species and strain. The output of the program may potentially assign a sequence(s) to any taxonomic level. Most commonly, a program will summarize the overall abundance of each taxonomic level it detects. If a species is detected and no other higher resolving sequence data are available, then strains within that species cannot be resolved based on that sequence data. Many programs may achieve assignment to the genus level, but not to species level attribution. Some programs conduct

classification down to either genus or species, while other programs will assign to a variety of levels depending on the level of specificity of the input data. Programs designed to make assignment at the strain level for bacteria will need to be validated for that level of specificity as well as congruency with genus and species level summaries. Viral strain assignment poses additional challenges, as some viruses (for example, RNA viruses) can have high rates of mutation and form quasi-species for which no clear reference genomes are available [107]. Bacterial and virus level assignments are likely to improve as the number of sequenced microbial genomes continues to increase. Since phylogenetic assignments are based on extant data, the databases and software (and version) used to perform the phylogenetic analyses should be documented.

Software typically is run with thresholds for assignment likelihood that can be set at either the initiation of analysis or at the time of interpretation of output. The thresholds used for analysis should be defined and documented thoroughly. Documentation should include the step(s) at which thresholds are specified, either by user input, within configuration files, in output interpretation, or at any other step in the analytical process. Thresholds should not be assumed to be equivalent between programs or within different versions of the same program, as every step of the analysis can impact the odds or strength of assignment. While many thresholds for taxonomic assignment are set automatically, the user has a responsibility to design experiments that test the impact of thresholds on the output of known samples on taxonomic assignment and set those thresholds accordingly.

Abundance levels

The most basic measure of the abundance of an organism in a sample is binary abundance (that is, presence/absence). Proportional abundance provides information on a continuous scale, but usually does not accurately convey relative level of abundance. The current state-of-the-art is generally composed of abundance measures with no associated confidence values [97,122,127]. Because of the complex nature of metagenomic samples a stochastic threshold (or minimum abundance threshold) for detection should be implemented. The abundance threshold can be set empirically to where anything above that value is present and anything lower (below the limit of detection) is either inconclusive, not detected, or absent, but then should be used consistently to measure corresponding error rates. The degree of accuracy is tied to the threshold of detection that is set. Internal standards are useful. Most studies to date have collected metagenomic data in a relative framework, in which abundance of genes or messages is calculated as percent or proportion of the sample content. However, the abundance level can be more accurate if internal genomic DNA is added at the sample processing stage. If these control molecules are mixed into and processed alongside the sample-derived nucleic acids, more effective quantification and inter-sample comparisons may be performed. Internal controls also may provide information on the extent or directionality of changes in any particular gene or organisms present. For example, in tracking a particular source of a contamination, measuring a gradient pointing towards the source may be useful. When drawing a conclusion that the presence of a microorganism is, for example, inconclusive or absent, it should be stated as being below the limit of detection that is determined both by the amount of sequence data and the parameters at which the analysis program was benchmarked.

Organism classification

Taxonomic classification of bacteria can sometimes create the misconception that microbial species are discrete and unique entities. Rather, some species are extremely closely related to each other and may form a continuum that is not readily resolved, while others are extremely distant from other microorganisms and can be categorized effectively [106]. Unfortunately, some separately named strains have almost identical genomes, while others are more dissimilar than some pairs of species. Therefore, when evaluating the power with which genomics can be used to distinguish between and among microorganisms and, thereby, define attribution under the circumstance of the analysis (for species to strain level identification or for determining similarity between two or more samples), it is important to understand the level of genomic similarity that they share (with known diversity of extant isolates). Also, the diversity in sequence within a species should be appreciated.

When constructing a test dataset for benchmarking, a decision first must be made regarding the level of genomic discrimination required. The level of genomic discrimination will likely be based on a list of known microorganisms of interest compared to their near neighbors. Whether that value is 5%, 1%, 0.1% or less, the microorganisms used for thresholding must have degrees of similarity consistent with that threshold. When calculating the similarity of two genomes, there are at least two methods that could be used: 1) calculating the similarity of regions and genes that are shared, or 2) normalizing that alignment value to the proportion of each genome that can be aligned. The second approach may account for plasmids or horizontally-transferred elements that may distinguish two strains of the same species. However, those strain-specific genes or regions may not provide any added discriminatory power to an algorithm depending on how it is constructed. One approach may be the percent identity of common (shared)

genes or regions to characterize the similarity of different genomes, so that the relationship of strains with a high degree of similarity within the core genome is not confounded by the presence of mobile elements. The performance of an algorithm should be presented only in the context of the degree of similarity between the organisms used for validation, with probability estimate, if possible.

Another strategy for selecting microorganisms for benchmarking is to use specific microorganisms that are of particular interest. For example, discriminating between a threat agent (such as *B. anthracis*) and a close relative (such as *B. cereus*) may be a higher priority than discriminating between all known species that are differentiated by at least 1%. It is important to note that such a specific target approach cannot be applied to benchmarking studies of other microorganisms as they may not, and likely will not, have a comparable level of genomic dissimilarity. The documented goal(s) of the user will determine whether the validation is designed to assess global similarity measures or the similarity of specific target organisms to their near neighbors.

Community structure

In addition to containing many different microorganisms, whether the same ones or very different ones, metagenomic samples will differ dramatically according to the relative abundances of microorganisms comprising the sample. Abundances of each microorganism (or taxonomic level of resolution) will vary widely, so that performance will be judged across orders of magnitude. It is difficult to predict how the presence of one microorganism may modulate the detection of another (due to similar elements in those genomes and power of discrimination of the bioinformatic method). The relative abundances of each organism can be varied across a number of replicates if the method lacks discriminatory power. This evaluation is performed best *in silico*.

The output data from a series of validation tests should consist of a set of records containing:

1. Microorganism (or taxonomic level resolved).
2. Known abundance, for example, controls.
3. Measured abundance (either proportional or binary).
4. If possible, a confidence measure (or qualitative/quantitative statement).

Sets of independent tests and repetitive tests will allow for summary statistics to be applied for assessing attribution capabilities, as well as the performance of the analytical system as a whole. Since empirical data generation is demanding and costly, the use of simulation data is strongly recommended. Power testing also can be defined, based on the number of samples to be analyzed. Comparisons of abundance values of microbes in two or more samples may be used for potentially indicating association [137]. Relevant to such analyses may be population genetic quantities, such as alpha and beta diversities [138]. The appropriate criteria for abundance distributions and comparisons should be established during validation. Current software may perform such data analyses to a degree and it is anticipated that novel programs will become available.

Rates of FPs and FNs are important measures and correspond to the sensitivity and specificity of the assay. If a proportional abundance measure is given, an abundance threshold should be set to render an interpretation of presence/inconclusive/absence. If a confidence measure is given, a more stringent threshold can be used along that dimension as well. Threshold values are dependent on the parameters of the sequencing run, as well as the program used and reference database. A validation process that establishes confidence values for a particular set of output data will only be applicable to other samples that are processed on the same platform, using the same settings (read length, and so on), filtered and processed with the same Q-score cutoffs, and then analyzed with the same taxonomic assignment program run with identical settings. This process is extremely important because the results of the validation process cannot be extended directly to an analysis in which any of those parameters have been changed or do not match.

The accuracy of proportional abundance can be measured with a correlation coefficient, either parametric (for example, Pearson) or nonparametric (for example, Spearman). Pearson's test could indicate how closely the absolute values generated resemble the known composition, while Spearman's test could indicate how closely the generated rank-order of each organism resembles the known composition. The utility of a program in determining the proportional abundance of individual microorganisms within a sample depends on the value of the correlation coefficient with data for controls included in the analysis. However, for many forensic applications the relative abundance of an organism is far less important than the presence or absence of that organism, along with designation to the strain level of identification. Nevertheless, for applications in which relative abundance is to be reported with confidence, thorough validation must satisfy all requirements of a binary presence analysis, with the added dimension of the correlation coefficient.

Standard operating protocols or procedures

All validated assays require SOPs, which must be based on the results of validation that encompass all appropriate aspects of the assay process, including but not limited to: sample collection and storage, nucleic acid extraction, enrichment, library preparation, sequencing, data analysis and interpretation of results. SOPs for implementation of

HTS as a diagnostic tool include: (1) standardization; (2) optimization; (3) validation; and (4) automation [139]. These concepts, while initially developed for HTS-based microbial clinical diagnostics, apply equally to developing HTS SOPs for microbial forensics. Standardization, in this context, requires selecting a set of methods, software and workflows, along with setting thresholds for making a forensic interpretation based on features present in the HTS data set. SOPs themselves must be validated, ideally with blinded prospective studies using static data analysis workflows. Finally, data analysis and interpretation SOPs ideally should be fully automated, if possible, to reduce user-configurable parameters to a minimum [139].

Conclusions

Conveying confidence in a test or process is essential in microbial forensics because the consequences are serious and the conclusions must be based on data and resultant interpretations of evidence in the case of a biothreat event. Therefore, the limitations of methods used in microbial forensics to generate results must be reliable and defensible and the process(es) of validation will contribute substantially in defining confidence associated with an assay, method, or system. HTS is an invaluable tool, expanding the scope and power of microbial forensics to provide protection against and response to attacks with biological agents. The HTS process was described in some detail herein so that analysts, who are not experienced researchers, will have guidance on the features and criteria that should be addressed during a validation. An outline of the HTS validation criteria is provided in the list of elements below. The reader may consider such validation quite challenging. However, similar demands have been in place for forensic human identification and the benefits to that forensic science community outweigh the task of validation. It is difficult to lay out the highest priority or near-term goals here as these may vary with the test or application and, therefore, such decisions are left to the community of users. To accomplish a validation the investigator should develop criteria as he or she requires for each situation. However, the criteria and the results and conclusions from validation studies must be available for inspection by appropriate parties.

List of elements to consider during validation of HTS for microbial forensics[a]

I. Sample Preparation
  a. Template (DNA or RNA) quantity and quality
    i. Minimum and maximum requirements
    ii. Guidelines for action when these values fall out of range
  b. Enrichment
    i. Desired genomic regions for enrichment
    ii. Limitations of the chosen method (for example, introduces known bias, increases error) and specific circumstances for its justified use
  c. Library preparation
    i. Quality, sensitivity, reproducibility and robustness of library preparation method(s) across expected sample types
  d. Multiplexing
    i. Performance of barcoding to identify specifically tagged samples
II. Sequencing
  a. System features
    i. Platform (if feasible, multiple orthogonal platforms)
    ii. Chemistry
    iii. Quality metrics
    iv. Limitations
      1. Error
      2. Signal-intensity decay
      3. Erroneous insertions/deletions
      4. Strand bias
      5. Potential for carry over contamination
III. Data analysis
  a. Bioinformatics pipeline
    i. Functions
    ii. Quality metrics
      1. Variant/sequence identification
      2. Q score
      3. Coverage
      4. Error
      5. Allele call (SNP state, indel state, and so on)
      6. Threshold
      7. False positive and false negative rates
    iii. Reference standard
      1. Variant calling
      2. Gene or functional element assignment
    iv. Alignment- or composition-based software
      1. Functions
      2. Rules for alignment
    v. Phylogenetics software
      1. Functions
  b. Bioinformatics software management
IV. Controls
  a. Level of acceptable characterization
  b. Intended use
V. Reference materials
VI. Databases
VII. Interpretation
  a. Sample type
    i. Single source
    ii. Complex or metagenomic
      1. Abundance
    iii. FP and FN rates

b. Attribution
   i. Taxonomic assignment
   ii. Association
   iii. Reverse engineering
      1. Sample preparation
      2. Genetic engineering
   iv. FP and FN rates
c. Quantitative/Qualitative statements
   i. Confidence

VIII. SOPs
a. Sample preparation
   i. Extraction
   ii. Enrichment
   iii. Reverse transcription (if necessary)
b. Library preparation
c. Sequencing
d. Bioinformatics pipeline
   i. Data analysis
   ii. Data storage
   iii. Data transfer
   iv. Interpretation

[a]It is not possible to generate an all-inclusive element list because of the wide diversity of samples, sample types, chemistries, platforms, and bioinformatics for which HTS methods may be applied. Therefore, this outline serves as a guideline, rather than an exhaustive or prescriptive regulation. The user should evaluate these elements, select those that apply, justify why some elements were not applied, and add any elements that are method specific and not included in this outline.

The HTS validation process should, at a minimum: 1) ensure that appropriate reference and benchmarking datasets are used to establish FP and FN values within a rigorous statistical framework; and 2) require the practices, chemistries, settings, and bioinformatics programs used to generate actionable results be thoroughly documented and standardized, at least within the specific laboratory employing the method(s). It is important to remember that identification to species and strain is highly dependent on phylogenetic similarity of near neighbors used for comparison. Consequently, the validation of a process to detect a given species or strain cannot be applied indiscriminately to additional target organisms without additional validation. The ability of a process to identify to species level varies across the tree of life, and validation processes must take the phylogenetic framework into consideration.

The validation process described herein allows for translation of research tools to forensic applications so that HTS can provide the reproducibility and specificity necessary to stand up to the full weight of legal scrutiny. In addition to validation, the laboratory is urged to adopt an overall quality management and quality assurance system to provide a working foundation essential for microbial forensics, maintaining good laboratory practices and bolstering confidence in results.

As HTS technologies continue to advance, costs will continue to drop, instruments will become faster, smaller, eventually portable, and their applications continue to increase. Advanced sequencing technologies will begin to be applied to measuring DNA modifications, epigenetic factors and offer yet another layer of specificity. With longer reads, genetically engineered organisms will be detected, most likely by identification of unexpected adjacency of genomic elements. The validation criteria described in this paper may likely apply to the new analytical flourishes in the coming years and, therefore, provide a stable foundation for future implementation in microbial forensics. If methods are validated within the framework outlined here, microbial forensics will achieve an ever higher level of power and analytical value and, ultimately, greater protection for the public and the nation's safety and security.

Abbreviations

bp: base pair; FN: false negative; FP: false positive; HTS: high throughput sequencing; IEEE: Institute of Electrical and Electronics Engineers; PCR: polymerase chain reaction; QA: quality assurance; RMs: reference materials; SMRT: single molecule real time; SNPs: single nucleotide polymorphisms; SOPs: standard operating protocols; SWGMGF: Scientific Working Group for Microbial Genetics and Forensics.

Competing interests

The authors declare that they have no competing interests. Ethics: No IRB approval was required for this work.

Authors' contributions

BB and SM conceived and developed the topic and wrote the majority of the manuscript; NC, AB, RC, CC, JF, MF, DK, AM, SAM, RM, AS, SS, KT and ST provided input and review. All authors read and approved the final manuscript.

Financial disclosure

The affiliated primary institutions and agencies supported this study by providing salaries for the authors. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details

[1]Department of Molecular and Medical Genetics, Institute of Applied Genetics, University of North Texas Health Science Center, Fort Worth, Texas, USA. [2]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia. [3]Rutgers New Jersey Medical School, Center for Biodefense, Rutgers University, Newark, New Jersey, USA. [4]Department of Epidemiology, The General K. Kaczkowski Military Institute of Hygiene and Epidemiology, Warsaw, Poland. [5]CosmosID®, 387 Technology Dr, College Park, MD, USA. [6]Maryland Pathogen Research Institute, University of Maryland, College Park, MD, USA. [7]University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA. [8]Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [9]Bioforensics Assay Development and DiagnosticsSection, Science Technology and Core Services Division, National Microbiology Laboratory, Winnipeg, MB, Canada. [10]Department of Medical Microbiology, University of Manitoba, Winnipeg, Canada. [11]National Institute for Microbial Forensics & Food and Agricultural Biosecurity, Oklahoma State University, Stillwater, OK, USA. [12]Division of CBRN Defence and Security, Swedish Defence Research Agency, Umeå, Sweden. [13]Signature Science, LLC, Austin, TX, USA. [14]University Hospital for Infectious Diseases "Fran Mihaljevic" and Medical School University of Rijeka, Zagreb, Croatia. [15]Division of Foodborne,

Waterborne, and Environmental Diseases, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia. [16]Virginia Tech, National Capital Region, Arlington, VA, USA. [17]Department of Forensic Medicine, Hjelt Institute, University of Helsinki, Helsinki, Finland. [18]Public Health Sciences, Bioinformatics Core Director, University of Virginia School of Medicine, Charlottesville, VA, USA.

Received: 8 May 2014 Accepted: 9 July 2014
Published: 30 July 2014

References

1. Bush LM, Abrams BH, Beall A, Johnson CC: **Index case of fatal inhalational anthrax due to bioterrorism in the United States.** *N Engl J Med* 2001, **345:**1607–1610.
2. Traeger MS, Wiersma ST, Rosenstein NE, Malecki JM, Shepard CW, Raghunathan PL, Pillai SP, Popovic T, Quinn CP, Meyer RF, Zaki SR, Kumar S, Bruce SM, Sejvar JJ, Dull PM, Tierney BC, Jones JD, Perkins BA, Team FI: **First case of bioterrorism-related inhalational anthrax in the United States, Palm Beach County, Florida, 2001.** *Emerg Infect Dis* 2002, **8:**1029–1034.
3. Jernigan JA, Stephens DS, Ashford DA, Omenaca C, Topiel MS, Galbraith M, Tapper M, Fisk TL, Zaki S, Popovic T, Meyer RF, Quinn CP, Harper SA, Fridkin SK, Sejvar JJ, Shepard CW, McConnell M, Guarner J, Shieh WJ, Malecki JM, Gerberding JL, Hughes JM, Perkins BA: **Anthrax Bioterrorism Investigation Team: Bioterrorism-related inhalational anthrax: the first 10 cases reported in the United States.** *Emerg Infect Dis* 2001, **7:**933–944.
4. Hsu VP, Lukacs SL, Handzel T, Hayslett J, Harper S, Hales T, Semenova VA, Romero-Steiner S, Elie C, Quinn CP, Khabbaz R, Khan AS, Martin G, Eisold J, Schuchat A, Hajjeh RA: **Opening a bacillus anthracis-containing envelope, Capitol Hill, Washington, D.C.: the public health response.** *Emerg Infect Dis* 2002, **8:**1039–1043.
5. Murch RS: **Forensic perspective on bioterrorism and bioproliferation.** In *Firepower in the Laboratory. Proceedings of the Symposium on Research Needs for Laboratory Automation and Bioterrorism.* Washington DC: National Academy of Sciences Press; 2001:203–214.
6. Murch RS: **Microbial forensics: building a national capacity to investigate bioterrorism.** *Biosecur Bioterror* 2003, **1:**117–122.
7. Budowle B, Schutzer SE, Einseln A, Kelley LC, Walsh AC, Smith JA, Marrone BL, Robertson J, Campos J: **Public health.** *Building microbial forensics as a response to bioterrorism. Science* 2003, **301:**1852–1853.
8. Morse SA, Budowle B: **Microbial forensics: application to bioterrorism preparedness and response.** *Infect Dis Clin North Am* 2006, **20:**455–473.
9. Flowers LK, Mothershead JL, Blackwell TH: **Bioterrorism preparedness. II: the community and emergency medical services systems.** *Emerg Med Clin North Am* 2002, **20:**457–476.
10. Morse SA, Kellogg RB, Perry S, Meyer RF, Bray D, Nichelson D, Miller JM: **Detecting biothreat agents: the Laboratory Response Network.** *ASM News* 2003, **69:**433–437.
11. Fletcher J: **The need for forensic tools in a balanced national agricultural security program.** In *Crop Biosecurity: Assuring Our Global Food Supply.* Edited by Gullino M, Fletcher J, Gamliel A, Stacks J: Springer Science + Business Media B.V; 2008:93–101.
12. Fletcher J, Barnaby N, Burans J, Melcher U, Ochoa Corona F: **Forensic plant pathology.** In *Microbial Forensics.* Edited by Budowle B, Schutzer SE, Breeze RG, Keim PS, Morse SA: Elsevier Inc; 2010:89–105.
13. Fletcher J, Bender C, Budowle B, Cobb WT, Gold SE, Ishimaru CA, Luster D, Melcher U, Murch R, Scherm H, Seem RC, Sherwood JL, Sobral BW, Tolin SA: **Plant pathogen forensics: capabilities, needs, and recommendations.** *Microbiol Mol Biol Rev* 2006, **70:**450–471.
14. Harmon R: **Admissibility standards for scientific evidence.** In *Microbial Forensics.* Edited by Budowle B, Schutzer SE, Breeze RG: Academic Press; 2005:382–392.
15. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu LI, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, *et al*: **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.** *Science* 1995, **269:**496–512.
16. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA 3rd, Venter JC: **The minimal gene complement of Mycoplasma genitalium.** *Science* 1995, **270:**397–403.
17. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N: **The sequence of the human genome.** *Science* 2001, **291:**1304–1351.
18. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860–921.
19. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452:**872–876.
20. Brzuszkiewicz E, Thürmer A, Schuldes J, Leimbach A, Liesegang H, Meyer FD, Boelter J, Petersen H, Gottschalk G, Daniel R: **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:**883–891.
21. Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamaichi Y, Calderwood SB, Mekalanos JJ, Schadt EE, Waldor MK: **The origin of the Haitian cholera outbreak strain.** *N Engl J Med* 2011, **364:**33–42.
22. Cummings CA, Bormann Chung CA, Fang R, Barker M, Brzoska P, Williamson PC, Beaudry J, Matthews M, Schupp J, Wagner DM, Birdsell D, Vogler AJ, Furtado MR, Keim P, Budowle B: **Accurate, rapid and high-throughput detection of strain-specific polymorphisms in Bacillus anthracis and Yersinia pestis by next-generation sequencing.** *Investig Genet* 2010, **1:**5.
23. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Fitzgerald M, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng Q, Abouelleil A, Bochicchio J, Chauvin S, Desmet T, Gujja S, McCowan C, Montmayeur A, Steelman S, Frimodt-Møller J, Petersen AM, Struve C, Krogfelt KA, Bingen E, Weill FX, Lander ES, *et al*: **Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011.** *Proc Natl Acad Sci U S A* 2012, **109:**3065–3070.
24. Eisen JA: **Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes.** *PLoS Biol* 2007, **5:**e82.
25. Hasan NA, Choi SY, Eppinger M, Clark PW, Chen A, Alam M, Haley BJ, Taviani E, Hine E, Su Q, Tallon LJ, Prosper JB, Furth K, Hoq MM, Li H, Fraser-Liggett CM, Cravioto A, Huq A, Ravel J, Cebula TA, Colwell RR: **Genomic diversity of 2010 Haitian cholera outbreak strains.** *Proc Natl Acad Sci U S A* 2012, **109:**E2010–E2017.
26. Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, Bortolaia V, Pearson T, Waters AE, Upadhyay BP, Shrestha SD, Adhikari S, Shakya G, Keim PS, Aarestrup FM: **Population genetics of Vibrio cholerae from Nepal in 2010: evidence on the origin of the Haitian outbreak.** *MBio* 2011, **2:**e00157–11.
27. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, Dolecek C, Achtman M, Dougan G: **High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi.** *Nat Genet* 2008, **40:**987–993.
28. Hornstra HM, Priestley RA, Georgia SM, Kachur S, Birdsell DN, Hilsabeck R, Gates LT, Samuel JE, Heinzen RA, Kersh GJ, Keim P, Massung RF, Pearson T: **Rapid typing of Coxiella burnetii.** *PLoS One* 2011, **6:**e26201.
29. Howden BP, McEvoy CR, Allen DL, Chua K, Gao W, Harrison PF, Bell J, Coombs G, Bennett-Wood V, Porter JL, Robins-Browne R, Davies JK, Seemann T, Stinear TP: **Evolution of multidrug resistance during Staphylococcus aureus infection involves mutation of the essential two component regulator WalKR.** *PLoS Pathog* 2011, **7:**e1002359.
30. Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J, Peacock SJ: **Routine use of microbial whole genome sequencing in diagnostic and public health microbiology.** *PLoS Pathog* 2012, **8:**e1002824.
31. MacLean D, Jones JD, Studholme DJ: **Application of "next-generation" sequencing technologies to microbial genetics.** *Nat Rev Microbiol* 2009, **7:**287–296.

32. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H: **Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology.** *PLoS One* 2011, **6:**e22751.
33. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D, Klammer A, Peluso P, Lee L, Kislyuk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyert SR, Frimodt-Møller J, Struve C, Petersen AM, Krogfelt KA, Nataro JP, Schadt EE, Waldor MK: **Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany.** *N Engl J Med* 2011, **365:**709–717.
34. Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, Chen W, Pu F, Peng Y, Li J, Xi F, Li S, Li Y, Zhang Z, Yang X, Zhao M, Wang P, Guan Y, Cen Z, Zhao X, Christner M, Kobbe R, Loos S, Oh J, Yang L, Danchin A, Gao GF, Song Y, Li Y, Yang H, *et al*: **Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4.** *N Engl J Med* 2011, **365:**718–724.
35. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJ, Brinkman FS, Brunham RC, Tang P: **Whole-genome sequencing and social-network analysis of a tuberculosis outbreak.** *N Engl J Med* 2011, **364:**730–739.
36. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD: **Evolution of MRSA during hospital transmission and intercontinental spread.** *Science* 2010, **327:**469–474.
37. Harris SR, Cartwright EJ, Török ME, Holden MT, Brown NM, Ogilvy-Stuart AL, Ellington MJ, Quail MA, Bentley SD, Parkhill J, Peacock SJ: **Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study.** *Lancet Infect Dis* 2013, **13:**130–136.
38. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S: **Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study.** *PLoS Med* 2013, **10:**e1001387.
39. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE: **Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study.** *Lancet Infect Dis* 2013, **13:**137–146.
40. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ, Quick J, Weir JC, Quince C, Smith GP, Betley JR, Aepfelbacher M, Pallen MJ: **A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4.** *JAMA* 2013, **309:**1502–1510.
41. Stobbe AH, Daniels J, Espindola AS, Verma R, Melcher U, Ochoa-Corona F, Garzon C, Fletcher J, Schneider W: **E-probe Diagnostic Nucleic acid Analysis (EDNA): a theoretical approach for handling of next generation sequencing data for diagnostics.** *J Microbiol Methods* 2013, **94:**356–366.
42. Stobbe AH, Schneider W, Hoyt P, Melcher U: **Screening metagenomic data for viruses using the e-probe diagnostic nucleic acid assay (EDNA).** *Phytopathology*. in press.
43. Breeze RG, Budowle B, Schutzer SE: *(Eds): Microbial Forensics.* Amsterdam: Academic Press; 2005.
44. Budowle B, Schutzer SE, Breeze RG, Keim PS, Morse SA: *(Eds): Microbial Forensics.* 2nd edition. Amsterdam: Academic Press; 2011.
45. Budowle B, Schmedes S, Murch RS: **The microbial forensics pathway for use of massively-parallel sequencing technologies.** In *The Science and Applications of Microbial Genomics.* Washington DC: The National Academies Press; 2013:117–133.
46. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors.** *Proc Natl Acad Sci U S A* 1977, **74:**5463–5467.
47. Wetterstrand KS: *DNA sequencing costs: data from the NHGRI Large-Scale Genome Sequencing Program*; 2013. Available at: https://www.genome.gov/sequencingcosts/.
48. Brenner S, Williams SR, Vermaas EH, Storck T, Moon K, McCollum C, Mao JI, Luo S, Kirchner JJ, Eletr S, DuBridge RB, Burcham T, Albrecht G: *In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs Proc Natl Acad Sci U S A* 2000, **97:**1665–1670.
49. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, *et al*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376–380.
50. Merriman B, Rothberg JM: **Progress in ion torrent semiconductor chip based sequencing.** *Electrophoresis* 2012, **33:**3397–3417.
51. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11:**31–46.
52. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ: **A large genome center's improvements to the Illumina sequencing system.** *Nat Methods* 2008, **5:**1005–1010.
53. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nat Biotechnol* 2008, **26:**1135–1145.
54. Roberts RJ, Carneiro MO, Schatz MC: **The advantages of SMRT sequencing.** *Genome Biol* 2013, **14:**405.
55. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW: **Direct detection of DNA methylation during single-molecule, real-time sequencing.** *Nat Methods* 2010, **7:**461–465.
56. Eisenstein M: **Oxford Nanopore announcement sets sequencing sector abuzz.** *Nat Biotechnol* 2012, **30:**295–296.
57. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW: **Transforming clinical microbiology with bacterial genome sequencing.** *Nat Rev Genet* 2012, **13:**601–612.
58. Budowle B, Schutzer SE, Morse SA, Martinez KF, Chakraborty R, Marrone BL, Messenger SL, Murch RS, Jackson PJ, Williamson P, Harmon R, Velsko SP: **Criteria for validation of methods in microbial forensics.** *Appl Environ Microbiol* 2008, **74:**5599–5607.
59. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E: **Working Group of the American College of Medical Genetics and Genomics Laboratory Quality Assurance Committee: ACMG clinical laboratory standards for next-generation sequencing.** *Genet Med* 2013, **15:**733–747.
60. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA, Agarwala R, Bennett SF, Chen B, Chin EL, Compton JG, Das S, Farkas DH, Ferber MJ, Funke BH, Furtado MR, Ganova-Raeva LM, Geigenmüller U, Gunselman SJ, Hegde MR, Johnson PL, Kasarskis A, Kulkarni S, Lenk T, Liu CS, Manion M, *et al*: **Assuring the quality of next-generation sequencing in clinical laboratory practice.** *Nat Biotechnol* 2012, **30:**1033–1036.
61. Pont-Kingdon G, Gedge F, Wooderchak-Donahue W, Schrijver I, Weck KE, Kant JA, Oglesbee D, Bayrak-Toydemir P, Lyon E: **Biochemical and Molecular Genetic Resource Committee of the College of American Pathologists: Design and analytical validation of clinical DNA sequencing assays.** *Arch Pathol Lab Med* 2012, **136:**41–46.
62. SWGDAM: **Validation guidelines for DNA analysis methods.** 2012. Available at: http://swgdam.org/SWGDAM_Validation_Guidelines_APPROVED_Dec_2012.pdf.
63. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, Bittinger K, Hwang J, Chen J, Berkowsky R, Nessel L, Li H, Bushman FD: **Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags.** *BMC Microbiol* 2010, **10:**206.
64. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R: **Experimental and analytical tools for studying the human microbiome.** *Nat Rev Genet* 2012, **13:**47–58.
65. Daniel R: **The metagenomics of soil.** *Nat Rev Microbiol* 2005, **3:**470–478.
66. DeLong EF: **Microbial community genomics in the ocean.** *Nat Rev Microbiol* 2005, **3:**459–469.
67. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, *et al*: **The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific.** *PLoS Biol* 2007, **5:**e77.
68. Kembel SW, Jones E, Kline J, Northcutt D, Stenson J, Womack AM, Bohannan BJ, Brown GZ, Green JL: **Architectural design influences the diversity and structure of the built environment microbiome.** *ISME J* 2012, **6:**1469–1479.
69. Human Microbiome Project Consortium: **Structure, function and diversity of the healthy human microbiome.** *Nature* 2012, **486:**207–214.
70. Human Microbiome Project Consortium: **A framework for human microbiome research.** *Nature* 2012, **486:**215–221.

71. Thomas T, Gilbert J, Meyer F: **Metagenomics - a guide from sampling to data analysis.** *Microb Inform Exp* 2012, **2:**3.
72. Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, Jackson PJ, Hugh-Jones ME: **Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within Bacillus anthracis.** *J Bacteriol* 2000, **182:**2928–2936.
73. Hoffmaster AR, Fitzgerald CC, Ribot E, Mayer LW, Popovic T: **Molecular subtyping of Bacillus anthracis and the 2001 bioterrorism-associated anthrax outbreak, United States.** *Emerg Infect Dis* 2002, **8:**1111–1116.
74. Schutzer SE, Keim P, Czerwinski K, Budowle B: **Use of forensic methods under exigent circumstances without full validation.** *Sci Transl Med* 2009, **1:**8cm7.
75. Budowle B, Schutzer SE, Burans JP, Beecher DJ, Cebula TA, Chakraborty R, Cobb WT, Fletcher J, Hale ML, Harris RB, Heitkamp MA, Keller FP, Kuske C, Leclerc JE, Marrone BL, McKenna TS, Morse SA, Rodriguez LL, Valentine NB, Yadev J: **Quality sample collection, handling, and preservation for an effective microbial forensics program.** *Appl Environ Microbiol* 2006, **72:**6431–6438.
76. Ellard S, Charlton R, Lindsay H, Camm N, Watson C, Abb S: *Mattocks C.* Practice Guidelines for Targeted Next Generation Sequencing Analysis and Interpretation. Clinical Molecular GeneticsSociety: Taylor GR; 2012. Available at: http://cmgsweb.shared.hosting.zen.co.uk/BPGs/BPG%20for%20targeted%20next%20generation%20sequencing%20final.pdf.
77. Illumina: **PCR-free sample preparation kits for whole genome DNA sequencing.** 2013. Available at: http://www.illumina.com/products/truseq-dna-pcr-free-sample-prep-kits.ilmn.
78. Agilent: **HaloPlex target enrichment system-ILM.** 2013. Available at: http://www.chem.agilent.com/Library/usermanuals/Public/G9900-90001.pdf.
79. Illumina: **Nextera XT DNA sample preparation kit.** 2013. Available at: http://www.illumina.com/products/nextera_xt_dna_sample_prep_kit.ilmn.
80. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F: **Laboratory procedures to generate viral metagenomes.** *Nat Protoc* 2009, **4:**470–483.
81. Feehery G, Yigit E, Oyola S, Langhorst B, Schmidt V, Stewart FJ, Dimalanta ET, Amaral-Zettler LA, Davis T, Quail MA, Pradhan S: **A method for selectively enriching microbial DNA from contaminating vertebrate host DNA.** *PLoS One* 2013, **8:**e76096.
82. Telenius H, Carter NP, Bebb CE, Nordenskjöld M, Ponder BA, Tunnacliffe A: **Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer.** *Genomics* 1992, **13:**718–725.
83. Cheung VG, Nelson SF: **Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA.** *Proc Natl Acad Sci U S A* 1996, **93:**14676–14679.
84. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, Lasken RS: **Comprehensive human genome amplification using multiple displacement amplification.** *Proc Natl Acad Sci U S A* 2002, **99:**5261–5266.
85. Syed F, Gruenwald H, Caruccio N: **Next-generation sequencing library preparation: simultaneous fragmentation and tagging using in vitro transposition.** *Nat Methods* 2009. Available at: http://www.nature.com/nmeth/journal/v6/n11/full/nmeth.f.272.html.
86. Knapp M, Stiller M, Meyer M: **Generating barcoded libraries for multiplex high-throughput sequencing.** *Methods Mol Biol* 2012, **840:**155–170.
87. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R: **Error-correcting barcoded primers allow hundreds of samples to be pyrosequenced in multiplex.** *Nat Methods* 2012, **5:**235–237.
88. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R: **Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms.** *ISME J* 2012, **6:**1621–1624.
89. Berglund EC, Kiialainen A, Syvänen AC: **Next-generation sequencing technologies and applications for human genetic history and forensics.** *Investig Genet* 2011, **2:**23.
90. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, Mellmann A, Goesmann A, von Haeseler A, Stoye J, Harmsen D: **Updating benchtop sequencing performance comparison.** *Nat Biotechnol* 2013, **31:**294–296.
91. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ: **High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity.** *Nat Rev Microbiol* 2012, **10:**599–606.
92. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: **A tale of three next generation sequencing platforms: comparison of Ion Torrent.** *Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics* 2012, **13:**341.
93. Lam H, Clark M, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ, Ji HP, Snyder M: **Performance comparison of whole-genome sequencing platforms.** *Nat Biotechnol* 2011, **30:**78–82.
94. Harismendy O, Ng P, Strausberg R, Wang X, Stockwell T, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA: **Evaluation of next generation sequencing platforms for population targeted sequencing studies.** *Genome Biol* 2009, **10:**R32.
95. Ewing B, Green P: **Base-calling of automated sequencer traces using phred.** *II. Error probabilities. Genome Res* 1998, **8:**186–194.
96. Hasan NA, Young BA, Minard-Smith EM, McMillan NJ, Isom R, Abdullah AS, Bornman DM, Faith SA, Choi SY, Longmire G, Dickens ML, Cebula TA, Colwell RR: **Microbial community profiling of human saliva using shotgun metagenomic sequencing.** *PLoS One*. in press.
97. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE: **Scalable metagenomic taxonomy classification using a reference genome database.** *Bioinformatics* 2013, **29:**2253–2260.
98. Minot S, Turner SD, Ternus KL, Kadavy DR: *SIANN: Strain identification by alignment to near neighbors*; 2014. Available at: http://biorxiv.org/lookup/doi/10.1101/001727.
99. Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments.** *Genome Biol* 2014, **15:**R46.
100. Keim P, Kalif A, Schupp J, Hill K, Travis SE, Richmond K, Adair DM, Hugh-Jones M, Kuske CR, Jackson P: **Molecular evolution and diversity in Bacillus anthracis as detected by amplified fragment length polymorphism markers.** *J Bacteriol* 1997, **179:**818–824.
101. Nocq J, Celton M, Gendron P, Lemieux S, Wilhelm BT: **Harnessing virtual machines to simplify next-generation DNA sequencing analysis.** *Bioinformatics* 2013, **29:**2075–2083.
102. Board IS: **IEEE Standard 610.12 Glossary of software engineering terminology.** 1990, Available at: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?reload=true&punumber=2238.
103. Crumpler S, Cheng J, Tillman DB, Benesch B, Sawyer D, Murray J, Press H, Snipes C, Godziemski A, Bergeson D, Loreng J: *General principles of software validation; Final guidance for industry and FDA staff*; 2002. Available at: http://www.fda.gov/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm085281.htm.
104. Deorowicz S, Grabowski S: **Compression of DNA sequence reads in FASTQ format.** *Bioinformatics* 2011, **27:**860–862.
105. Bonfield J, Mahoney M: **Compression of FASTQ and SAM format sequencing data.** *PLoS One* 2013, **8:**e59190.
106. Buckley M, Roberts RJ: *Reconciling Microbial Systematics and Genomics.* Washington DC: American Academy of Microbiology; 2007. Available at: http://academy.asm.org/index.php/genetics-genomics-molecular-microbiology/454-reconciling-microbial-systematics-and-genomics.
107. Gonzalez-Candelas F, Bracho M, Wrobel B, Moya A: **Molecular evolution in court analysis of a large hepatitis C virus outbreak from an evolving source.** *BMC Biol* 2013, **11:**76.
108. Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaya I, Ondov B, Darling AE, Phillippy AM, Pop M: **MetAMOS: a modular and open source metagenomic assemblyand analysis pipeline.** *Genome Biol* 2013, **14:**R2.
109. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18:**821–829.
110. Narzisi G, Mishra B: **Comparing de novo genome assembly: the long and short of it.** *PLoS One* 2011, **6:**e19175.
111. Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J: **De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer.** *Genome Res* 2008, **18:**802–809.
112. Chaisson MJ, Pevzner PA: **Short read fragment assembly of bacterial genomes.** *Genome Res* 2008, **18:**324–330.
113. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou WC, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, *et al*: **Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species.** *Gigascience* 2013, **2:**10.
114. Gurevich A, Saveliev V, Vyahhi N, Tesler G: **QUAST: quality assessment tool for genome assemblies.** *Bioinformatics* 2013, **29:**1072–1075.

115. Hatem A, Bozdağ D, Toland AE, Çatalyürek ÜV: **Benchmarking short sequence mapping tools.** *BMC Bioinformatics* 2013, **14:**184.
116. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357–359.
117. Delcher AL, Phillippy A, Carlton J, Salzberg SL: **Fast algorithms for large-scale genome alignment and comparison.** *Nucleic Acids Res* 2002, **30:**2478–2483.
118. McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I: **Accurate phylogenetic classification of variable-length DNA fragments.** *Nat Methods* 2007, **4:**63–72.
119. Patil KR, Roune L, McHardy AC: **The PhyloPythiaS web server for taxonomic assignment of metagenome sequences.** *PLoS One* 2012, **7:**e38581.
120. Brady A, Salzberg SL: **Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models.** *Nat Methods* 2009, **6:**673–676.
121. Rosen GL, Reichenberger ER, Rosenfeld AM: **NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads.** *Bioinformatics* 2011, **27:**127–129.
122. Berendzen J, Bruno WJ, Cohn JD, Hengartner NW, Kuske R, McMahon BH, Wolinsky MA, Xie G: **Rapid phylogenetic and functional classification of short genomic fragments with signature peptides.** *BMC Res Notes* 2012, **5:**460.
123. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data.** *Genome Res* 2007, **17:**377–386.
124. Davenport CF, Neugebauer J, Beckmann N, Friedrich B, Kameri B, Kokott S, Paetow M, Siekmann B, Wieding-Drewes M, Wienhöfer M, Wolf S, Tümmler B, Ahlers V, Sprengel F: **Genometa–a fast and accurate classifier for short metagenomic shotgun reads.** *PLoS One* 2012, **7:**e41224.
125. Sharma VK, Kumar N, Prakash T, Taylor TD: **Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin.** *PLoS One* 2012, **7:**e34030.
126. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M: **Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences.** *BMC Genomics* 2011, **12**(Suppl 2)**:**S4.
127. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C: **Metagenomic microbial community profiling using unique clade-specific marker genes.** *Nat Methods* 2012, **9:**811–814.
128. Haft DH, Tovchigrechko A: **High-speed microbial community profiling.** *Nat Methods* 2012, **9:**793–794.
129. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J: **Phylogenetic classification of short environmental DNA fragments.** *Nucleic Acids Res* 2008, **36:**2230–2239.
130. Gerlach W, Jünemann S, Tille F, Goesmann A, Stoye J: **WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads.** *BMC Bioinformatics* 2009, **10:**430.
131. Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, Liolios K, Pagani I, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC: **IMG/M: the integrated metagenome data management and comparative analysis system.** *Nucleic Acids Res* 2012, **40:**D123–D129.
132. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA: **The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.** *BMC Bioinformatics* 2008, **9:**386.
133. Wang Y, Leung HC, Yiu SM, Chin FY: **MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species.** *J Comput Biol* 2012, **19:**241–249.
134. MacDonald NJ, Parks DH, Beiko RG: **Rapid identification of high-confidence taxonomic assignments for metagenomic data.** *Nucleic Acids Res* 2012, **40:**e111.
135. Brady A, Salzberg S: **PhymmBL expanded: confidence scores, custom databases, parallelization and more.** *Nat Methods* 2011, **8:**367.
136. Bazinet A, Cummings M: **A comparative evaluation of sequence classification programs.** *BMC Bioinformatics* 2012, **13:**1–13.
137. Fierer N, Lauber CL, Zhou N, Mcdonald D, Costello EK, Knight R: **Forensic identification using skin bacterial communities.** *Proc Natl Acad Sci U S A* 2010, **107:**6477–6481.
138. Whittaker RH: **Evolution and measurement of species diversity.** *Taxon* 1972, **21:**213–251.
139. Fricke WF, Rasko DA: **Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions.** *Nat Rev Genet* 2013, **15:**49–55.

doi:10.1186/2041-2223-5-9
Cite this article as: Budowle *et al.*: Validation of high throughput sequencing and microbial forensics applications. *Investigative Genetics* 2014 **5**:9.