

# Discussion Paper

<b>Title</b>	<b>High-Throughput Sequencing Technologies for Microbial Identification and Detection of Antimicrobial Resistance Markers</b>
<b>Contact</b>	<b>Heike Sichtig, Ph.D. (heike.sichtig@fda.hhs.gov)</b>
<b>Affiliation</b>	<b>CDRH/OIR/Division of Microbiology</b>
<b>Purpose</b>	<b>Discussion material for April 1<sup>st</sup>, 2014 public workshop</b>
<b>Date</b>	<b>March 24, 2014</b>

## **High-Throughput Sequencing Technologies for Microbial Identification and Detection of Antimicrobial Resistance Markers**

### **Introduction:**

FDA is issuing this discussion paper to obtain feedback on possible studies to evaluate the use of high throughput sequencing (HTS) devices as an aid in microbial diagnostics, and to gain a better understanding of the potential HTS clinical implementation strategies in microbial diagnostics. For the purposes of this discussion paper, HTS is used to describe the automated sequencing technologies used to generate large quantities of raw sequence reads in a short period of time.

In this document, we intend to discuss:

- The possible approaches to validation studies and data for the evaluation of HTS systems for potential regulatory clearance/approval,
- The use of sequence outputs from HTS devices to evaluate performance, and
- A potential new comparator paradigm that could enable the use of HTS technologies as a single comparator to determine the specific microbial composition of a clinical specimen, which in turn may aid in determining the clinical truth for multiplexed nucleic acid based assays.

This discussion paper is being released as part of the preparation for an FDA Public Workshop being held at the White Oak Campus, Maryland, April 1<sup>st</sup>, 2014. It is important to note that the information contained in this document are possible approaches for validation and are not meant to convey FDA's recommended approach, rather the information is provided to offer background and the basis for discussions at the Public Workshop.

The cornerstone for each of the concepts proposed herein is a public database that is populated with sequence entries that meet an acceptable level of quality for use in clinical and regulatory decision processes. When used as part of a comparator, the reference database would provide sequence information that could be used (1) to verify results obtained during the clinical evaluation of HTS devices and (2) to potentially establish the clinical performance of the device in support of regulatory clearance/approval. Once a HTS system is cleared/approved, a high quality public database used in combination with the HTS system is of paramount importance to ensure access to the most accurate microbial (including antimicrobial resistance markers) detection and identification to inform treatment decisions. If implemented, this comparator paradigm has the potential to provide a time-efficient, least-burdensome method to evaluate clinical performance for these investigational devices. Implementation of such a database to serve as a reference for regulatory/clinical use could enable a clear regulatory pathway for clearance/approval of HTS systems for microbial identification and potentially improve public health. Moreover, the use of cleared/approved HTS systems used as a comparator method in conjunction with a high-quality public database may have the potential to dramatically reduce the burden encountered during the clinical evaluation of new targeted multiplexed assays for microbial diagnosis.

Currently, comparator methods for targeted multiplexed assay evaluation include diverse methods or composite methods whereby multiple tests are performed to determine the presence or absence of microorganisms in a clinical specimen. The disadvantage of these approaches is

that they can still only identify a limited number of analytes in any single clinical specimen. In the past, highly validated sequencing technologies (targeted PCR followed by Sanger sequencing) have been considered a reference method when used as part of a comparator paradigm however, since an increasing number of devices detect multiple analytes in a single sample, requiring confirmation of all analytes using traditional reference methods is practically impossible and imposes an unnecessarily large regulatory burden. Thus the proposed use of HTS could significantly streamline the approaches employed by developers to determine the microbial composition in human specimens.

The primary focus of this discussion paper is a) to propose the validation concepts of a HTS assay for FDA clearance/approval as a diagnostic; b) to promote the discussion of the necessary quality criteria for sequence information to be used as a reference or as part of a comparator database; and c) to introduce the potential use of HTS as a single comparator method for targeted multiplexed devices.

This document does not address the use of HTS in a research setting.

This discussion paper provides a brief overview of the general sequencing process from nucleic acid isolation to result output (Section 1), possible approaches for the process of HTS validation for microbiology applications (Section 2), a brief description of FDA considered efforts to enhance sequence quality in the public domain (Section 3), potential clinical applications and challenges (Section 4), and possible approaches to streamline the clinical trials for molecular diagnostic device regulatory clearance using HTS (Section 5).

## **Section 1: Genome Sequencing Overview**

Sanger bidirectional sequencing, also known as dideoxynucleotide chain termination, has been often accepted as a reference technology. Because the original manual technique was labor intensive and required the use of radioactive materials, automated Sanger sequencing machines using fluorescent labeled nucleotides were developed in an effort to reduce the cost and manpower required for nucleic acid sequencing. With the introduction of the “next generation” sequencing (NGS) technologies (e.g., pyrosequencing, sequencing by synthesis, and sequencing by ligation), the cost-per-base for genome sequencing has declined even further while the throughput and resolution of sequence complexity has grown exponentially. Several types of NGS platforms have been developed and, although they differ in the technical details that underlie their unique sequencing chemistries, all NGS platforms overcome the limited scalability of Sanger sequencing by relying on miniaturized sequencing reactions performed in parallel. Specifically, all NGS platforms use massively parallel sequencing approaches to obtain large volumes of sequencing data of varying read lengths.

The HTS genome sequencing process also known as next generation or massively parallel sequencing technologies relies on a number of steps that are summarized in Figure 1 and discussed briefly in the following sections. HTS platforms can be differentiated from one another by any number of ways including chemistries, throughput, and sequencing length. This diversity results in unique performance characteristics which may define the clinical applications that each HTS technology is best suited for; however, the intention of this discussion paper is to

introduce for consideration a near universal validation approach for the systems to establish performance of HTS platforms for clinical microbiology diagnostic applications.

This section summarizes the sequencing process and highlights the areas of possible FDA regulatory oversight in microbial diagnostic applications.

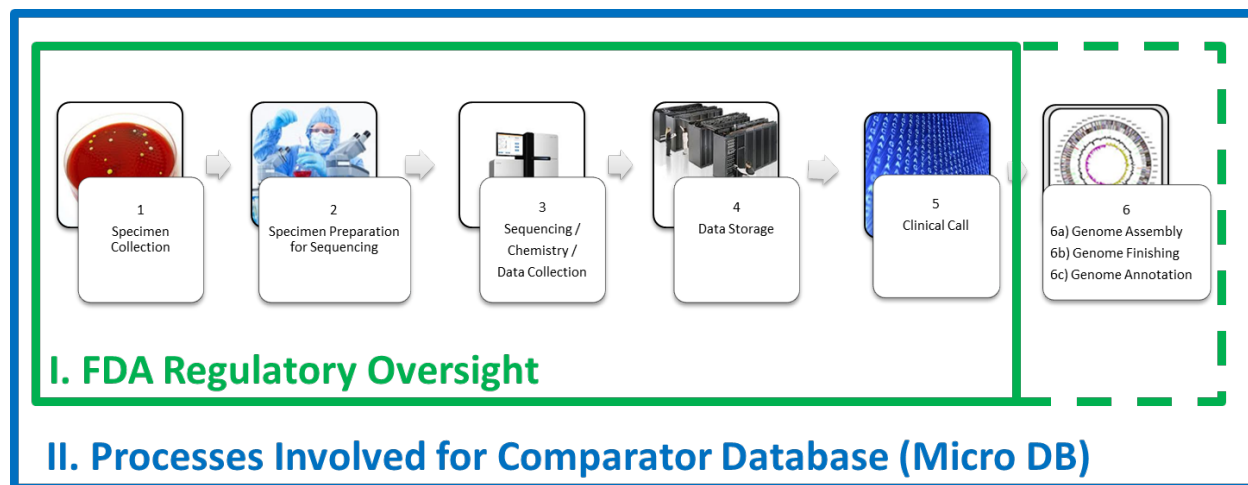


Figure 1: Sequencing Process in microbial diagnostics applications. This figure depicts the areas that would possibly come under consideration for FDA regulatory oversight (solid green box). Similar to other diagnostic devices FDA is proposing to use a more “systems” approach for evaluating HTS devices, from sample collection through the output of clinically actionable data. It is important to note that some aspects of part 6a-c may fall under regulatory purview if they are used as part of the data analysis pipeline to generate the final clinical call.

#### Specimen Collection (1):

The performance of HTS is highly dependent on the quality and quantity of the isolated nucleic acid, therefore specimen type, collection methods, and storage plays important roles in successful HTS. For the purposes of this discussion paper there are two main specimen types associated with microbial diagnostics: (1) clinical isolates, where microbes are grown as pure clonal isolates on a defined media, and (2) direct human clinical specimens, where potential pathogens may be present in a more complex environment, potentially with commensal organisms or in the presence of host cells. For the purposes of a regulatory submission and ultimately for clearance/approval, the HTS platform should be validated using any and all specimen types that are claimed in the indication for use.

#### Specimen Preparation for Sequencing (2):

Following specimen collection or clonal isolation, nucleic acid extraction and purification represent the next steps in the process. A number of methods are available for preparing purified nucleic acids and there are several commercially available kits, as well as automated systems. The integrity and purity of the extracted nucleic acids are especially important for successful HTS applications. Similarly, the presence of inhibitors and interfering substances can impact the performance of HTS devices.

The next major step in the process involves the preparation of the nucleic acids for sequencing, which is known as DNA library preparation, and this step can be performed using a number of different methods. While the exact methods employed are usually platform specific they tend to

share many similar features. For each HTS platform the methods employed during this step should be well documented and in a final “locked-down” configuration for manufacturing.

#### Sequencing / Chemistry/ Data Collection (3):

HTS platforms employ a number of sequencing mechanisms, including, but not limited to: sequencing by synthesis that is based on DNA polymerase dependent methods including cyclic reversible termination (CRT), single-nucleotide addition, and real-time sequencing; sequencing by ligation (SBL) that uses DNA ligase; and single molecule sequencing without prior amplification, for example, advanced optical detection, and nanopore technology. The majority of the HTS platforms use optical-based imaging for detection, measuring either bioluminescent or fluorescent signals generated when labeled nucleotides are sequentially incorporated into the template. In addition, there are platforms that use non optical methods for detection, such as the ion-sensitive field-effect transistor semiconductor chip. The common feature across all of these technologies is that they generate sequences of multiple DNA fragments in parallel that comprise the sequencing reaction. A number of quality metrics, discussed in greater detail later, should be applied to HTS to evaluate the performance of the instrument runs and quality of the data generated.

#### Data Storage (4):

The sequences of the multiple DNA fragments that comprise the signal outputs of the reaction must then be stored in a suitable format that allows subsequent bioinformatics analysis. There are a number of data formats applied to HTS; however, the most common are the text-based formats FASTA (stores the biological sequence - format used to search NCBI BLAST database) and FASTQ (stores both the biological sequence and its corresponding quality scores). The storage methods will be discussed in greater detail in later sections with regards to regulatory submissions.

#### Clinical Call Determination (5)

The informatics package or data analysis pipeline provided by the manufacturer for use with the sequencing platform is the final step in the process to obtain clinically actionable data. Validation of the informatics pipeline for the determination of the clinical call is discussed in later sections. It is important to note that the data analysis pipeline should be in a locked down configuration prior to device validation.

#### Additional Data Analysis (6a-6c):

All HTS platforms generate a tremendous amount of raw data that requires a significant level computational time/power to analyze the data outputs. The data analysis process can follow many paths depending on the ultimate needs of the end-user. It is important to note that whether the sequence information is to be manipulated to generate a high quality annotated draft genome or is used to generate sufficient information to guide patient management, both ultimately require a database that is composed of high-quality reference sequences. The sequences that compose this database, and ultimately their construction do not fall under the regulatory purview of FDA; however, it is critical to point out the necessity of an agreed upon level of sequence quality in the public domain so that FDA can provide the least burdensome pathway for device/assay developers and to enable a more widespread adoption of these technologies in the microbiological laboratory.

For the purposes of this discussion paper we are only concerned about the manipulation of sequence data outputs that result in actionable information to effect patient management, i.e., the clinical laboratory call (5). Steps like genome annotation, genome assembly, manual closure and genome finishing may not be needed for clinically actionable data, however, if these steps are part of a specific claim that a sponsor will make then they should be included in the validation process.

#### Genome Assembly (6a):

Genomic assembly is the process by which multiple, fragmented sequence reads generated by the HTS platform are assembled to reconstruct the original sequence. Assembly is achieved via either a reference sequence (using the existing sequence of a closely related organism) or *de novo* (contigs generated with no reference sequence). The *de novo* approach is more complex, requiring increased time and resources compared with mapping to a reference sequence. The coverage quality for *de novo* is dependent on the size and continuity of the resulting contigs. There are a number of commercial and public software packages with various alignment algorithms to aid in the alignment/assembly process.

#### Genome Finishing (6b):

Genomic finishing is the process of generating a single contiguous sequence that has no ambiguities. Genomic finishing takes the draft assembly generated in step (5) and uses gap closure and assembly validation and refinement processes to correct sequence errors/uncertainties that occur due to sequencing biases, sequencing artifacts or incorrect reconstruction. Relative to generating the draft assembly, refining an assembly to create a finished or nearly-finished genome can take significantly longer, with additional experiments often required.

#### Genome Annotation (6c):

Annotation is the process of assigning biological information/function to the final sequence. There are a number of *in silico* tools that can assist in genome annotation; however, automated annotation is often performed in conjunction with manual annotation.

## **Section 2: Device Validation**

Device validation can be a rather long and costly process for developers. As the capability to detect a greater number of organisms from a single specimen grows, so do the clinical validation studies necessary to determine “clinical truth” and performance characteristics. The effects of challenging evaluation studies are evident with the introduction of multiplexed platforms for use in clinical diagnostics. Many of these multiplex devices have the potential to detect 30 or more targets from a single specimen. However, the potential detection capabilities for high throughput sequence-based diagnostic devices can far exceed this, with the potential ability to detect virtually any organism present in a clinical specimen. This unique ability should be kept in mind when designing the analytical and clinical validation studies to substantiate performance claims in a premarket regulatory filing.

To enable a “least burdensome” approach for developers of these devices without sacrificing the necessary scientific information to substantiate performance claims, the FDA is proposing a novel comparator paradigm that uses a database populated with high-quality, highly confident genomic sequences that meets certain regulatory quality criteria and has been vetted by experts. This information, augmented with well-designed in-house analytical studies, it is thought, is expected to provide the requisite information for regulatory review and clearance of microbiological diagnostic devices based on sequencing. Each of the studies should be discussed below; however, it is important to note that the actual design and execution of these studies may require modification by the developers. To this end, the Agency suggests that all developers engage with the Agency through the pre-submission process to get feedback prior to implementation.

Once the clinical system has matured to a point where it has been “locked down” for manufacturing purposes, then the following assay validation principles could be applied. It is important to note that there are many different platforms and methodologies that can be employed for genomic sequencing for microbial detection and identification so, the following is meant to provide a broad overview of the validation studies the Agency is considering to look for when reviewing devices for a specific intended use. Briefly, the following topics should be addressed: specimen type (including collection, handling, and preparation for sequencing), limit of detection, interference (endogenous, exogenous), reproducibility/repeatability, clinical sensitivity/specificity, and data analysis pipeline validation. Furthermore, there are a number of device specific validation studies that should also be considered and will be discussed during the workshop.

All validation studies should be designed to support the intended use of the device. The intended use should specify the platform/instrument, pathogens and, if applicable, antimicrobial resistance markers that the test detects and identifies, the specimen types for which testing will be indicated, the clinical indications for which the test is to be used, and the specific population(s) for which the test is intended. The intended use should state whether detection of a pathogen or resistance marker is presumptive, and any specific conditions of use.

We are encouraging developers to communicate intended marketing of a microbial sequence-based diagnostic device with the Agency in the early phases of device development using the pre-submission process, due to the dynamic environment and the current state of the art in high throughput sequencing; however, this is not a requirement. Feedback can be provided regarding the design validation studies to substantiate all claims and demonstrate safety and effectiveness, and to suggest an appropriate regulatory strategy for marketing purposes of the device.

#### *Microbial Standard Reference Materials*

Whenever possible the use of standard microbial reference materials for the validation process is recommended. NIST, through FDA funding, is creating a set of microbial standard reference materials (SRMs) for use in the validation of genomic sequencing devices. The standard reference materials are highly characterized and have stable and homogeneous properties desirable for validation studies. At least four microbial SRMs will be available, ranging from low to high GC content and attempt to capture the physiochemical diversity of pathogens routinely encountered in clinical specimens.

### Specimen Type and Handling

All specimen type(s) for which the assay is intended to be used should be validated. Appropriate specimen types depend on a variety of factors, including the site of infection and the pathogen nucleic acid to be detected. Specifically, a clinical specimen should be collected from the appropriate anatomical site or source at the appropriate time in the clinical progression of disease. Appropriate specimen types will vary according to clinical syndrome. Many different specimen types have the potential to be used for validation studies and we suggest that sponsors consult the FDA to determine which specimens are considered appropriate for the intended test panel and if certain specimen types could be considered equivalent and thus could be combined.

The quality and quantity of extracted nucleic acids can be affected by multiple factors such as specimen source, collection method, and handling (e.g., transport, storage time, temperature). The acceptance criteria for all specimen stability parameters should be clearly indicated and justified and should include the following:

1. Validation of any nucleic acid extraction method to be indicated for use by the system
2. Validation that each collection method used in the system provides adequate and appropriate nucleic acid for all pathogens detected by your assay (i.e., nucleic acid from anaerobic bacteria, aerobic bacteria, acid fast bacteria, etc.)
3. The device maintains acceptable performance under the various specimen handling conditions to be claimed.

Prior to any signal generation through a sequencing-based technology, the nucleic acid needs to be prepared using any number of technology dependent methodologies. Given the significant differences that these can entail, and the effect on the overall performance that they could have, it is suggested that validation data be provided for each method used with the assay.

### Library Preparation

For devices that use library preparation steps you should address the variability on assay performance for all claimed preparation methods and reagents used. Different library preparation methods may yield nucleic acids of varying quantity and quality, and thus the extraction method can be crucial to a successful result. You should consider the steps involved in the construction and normalization of the specimen libraries, which could impact the reproducibility and reliability of the sequences generated (e.g., sample enrichment, sequencing strategy, primers, amplification efficiency, reagent lots, hybridization, etc.). Moreover, an analysis of potential inhibitors from the clinical specimen or methods employed to extract the nucleic acids should also be considered during the validation of library preparation.

### Limit of Detection (LoD)

The LoD provides a measure of the analytical sensitivity of an assay for a particular target and is defined as the lowest concentration of target that can be sequenced reliably and distinguishable from negative specimens and is consistently detected in  $\geq 95\%$  of the specimen replicates. Proper determination of the LoD is critical since bacterial pathogens may be present in a patient specimen at very low levels. Depending on the sequencing format, ranging from a targeted-PCR multiplex sequencing approach to a high throughput sequencing approach, there may be different



considerations for how the LoD is established and validated. Determination of the LoD for each target included in the assay menu and each specimen type may not be feasible given the large number of pathogens that can be detected when using a high throughput sequencing approach, however the ability to potentially detect any pathogen present could negatively impact the analytical sensitivity of the device. If a targeted multiplexed sequencing panel approach is used then validation concepts similar to those of other multiplexed devices could be applied.

#### Interference (endogenous, exogenous)

An evaluation of interfering substances found in the clinical specimen that could interfere with signal generation and sequencing should be considered. Potential sources of interfering substances from the clinical specimens include exogenous substances (i.e., prescription/non-prescription drugs, anticoagulants, etc.) and endogenous substances (i.e., proteins, lipids, hemoglobin, bilirubin, etc.). The selection of interferents for inclusion in the device validation studies would be determined by the indicated clinical specimen type. Additionally, a thorough evaluation of potentially interfering substances that could be introduced by the sequencer should be considered during the validation process and could include residual chemicals from previous treatments or wash cycles.

#### Carryover

Evaluation of the affects from carryover contamination should be considered. This would include evaluation of the entire device, including sample preparation and library preparation, where known positive samples (at a high target concentration) and negative samples are alternated. The carryover rate from previous runs should be calculated and reported. This information would be included in the device labeling to caution the end user. Furthermore, depending on the rate of carryover, there may need to be additional information included in the package labeling (warnings/ precautions and cleaning instructions) to direct the end user on how to reduce or eliminate this effect.

#### Inclusivity

Validation of inclusivity or analytical reactivity should be conducted based on the intended use of the device. Depending on the diagnostic claims made by the manufacturer the studies would be designed to validate the ability to specifically detect potential genetic variation among the pathogens and/or resistance markers included in the intended use. The approach to establish inclusivity could use intact cultured organisms that undergo all pre-analytical steps, or pre-extracted and defined nucleic acids. The analytes used in this evaluation should be tested at or very near the LoD of the device. The evaluation could use test panels designed to reflect the different genetic elements on which any conclusions would be based.

#### Reproducibility/Repeatability

The evaluation of reproducibility of the sequencing device would be designed to assess the variability when the same material is evaluated and multiple variables are introduced. For example, evaluation of reproducibility could be done using instruments at multiple sites with different operators running the instruments on different days. The evaluation would also determine the effect of multiple reagent lots on the variability of the performance of the device

and any impact it may have on the final results. The microbial SRM's that are under development by NIST may prove to be the ideal tool for use in this evaluation.

Similarly, an evaluation of repeatability would also be done to evaluate the precision of the device when a standard material is analyzed multiple times at a fixed condition. This evaluation would be done using a single site with as many of the non-device related variables eliminated to determine impact, if any, that the device has on precision of the sequence outputs. Similar to the evaluation of reproducibility, the evaluation of repeatability could also employ the SRM's that are currently under development by NIST.

#### Sensitivity/Specificity

The determination of clinical sensitivity and specificity of a HTS device could be done using many of the principles applied to other microbial diagnostic devices. The evaluation would still be done using multiple sites in the intended use environment, using specimens indicated for the subject device, and with the operators who are trained at the appropriate level. However, given the number of potential pathogens that these technologies may be able to detect in a single clinical specimen, the application of more traditional regulatory strategies may hinder approval/clearance of these devices by requiring extensive evaluation of every detected organism (genomic sequence) from a single specimen, or in the case of device specificity all of those that were not detected, using expensive reference methods. To promote a least burdensome regulatory approach, we are proposing a novel comparator paradigm that will rely heavily on public databases that are populated with high-quality genomic sequences that meet certain regulatory quality criteria. The genomic sequence outputs from the subject device, when compared against the high quality database with sufficient coverage, should provide adequate information to determine the specificity of the device. Clearly, there may not be adequate representation of every organism in the public domain to employ this approach in its entirety at the present time; however, there may be pathways where certain facets of this strategy can be employed until such a time as there is adequate coverage in the public domain, especially if a panel-based approach is utilized.

In addition to the reference database, the implementation of this approach is likely to rely heavily on a robust validation of the LoD for the device. Moreover, information relating the analytical sensitivity of the device to the clinically relevant range of the pathogen load in the indicated disease state is also likely needed to be provided.

#### Data Analysis Pipeline Validation – Clinical Call Determination

The breadth of the data analysis pipeline ranges from detection of a signal indicating the presence of a specific nucleotide to a final call based on the sequence. This analysis relies heavily on informatics components that are intricate parts of the analysis pipeline and would require supporting validation data. This could include information from the following specific areas of the informatics pipeline: (1) signal to base call transformation, (2) alignment to a reference sequence and (3) clinical call determination (percent identity measure).

For the signal to base call transformation component, the platform pipeline, including base caller and version, and the quality score rationale seems to be the correct parameters that should be provided.

For alignment/mapping to a reference sequence, FDA suggests to provide a protocol chronicling the steps from raw sequence data (i.e., reads) to the actionable final sequence, listing the specific alignment/mapping tool, version and parameter settings, and the reference sequence with adequate source information. Note, for *de-novo* sequencing, requiring *de-novo* assembly, the specific assembler and version is likely needed to be provided.

FDA suggest that detailed information on the algorithm and version used for the clinical call determination, including genomic coverage requirements, trimming logic and other potential factors, be provided. Also, depending on the sequencing format, ranging from a targeted PCR multiplexed sequencing approach to a high throughput sequencing approach, there may be different considerations for how a clinical call is made and validated. FDA propose that all assay specific software optimization should be addressed and properly validated.

FDA suggests to submit the following for review: (1) sequencing strategy, (2) details regarding each selected targets and reference sequences, (3) summary information and statistics on sequence runs (number of reads generated, quality of those reads, mapping statistics), and (4) quality reads metrics (i.e., quality score, total number of reads generated, percentage of total reads that pass filter, signal strength). Furthermore, quality control metrics should be provided (i.e. depth of coverage, uniformity of coverage, GC skew, Ti/Tv, base call quality scores, mapping quality, on-target coverage, removal of duplicate reads, distribution of random ends, and signal intensity decline, etc.).

### **Section 3: Reference Databases**

The Food and Drug Administration's Center for Devices and Radiological Health facilitates medical device innovation by providing predictable, consistent, transparent and efficient regulatory pathways. To advance the use of HTS in the clinical microbiology laboratory and for use by the assay development industry, a well-vetted high quality reference database of clinically relevant microorganisms is needed. Currently, the general concept for evaluating microbial diagnostic devices relies on the use of an accepted standard reference method, usually either a gold standard method normally employed by a laboratory or a composite reference method consisting of two or more complimentary approaches to determine the microbial presence in a specimen. This information is then compared with the output results from the subject device to assess clinical performance. Although this approach has worked well for many assay formats and has been streamlined over the past few years to accommodate even more complex technologies, it does not translate well to the evaluation of HTS platforms for regulatory clearance/approval. We expect sequence-based microbial diagnostic devices to seek US market clearance in the near future and acknowledge that these new devices will require an alternative validation process to demonstrate safety and effectiveness. Sequence-based devices typically allow multiplex detection in samples and, at the same time, require complex bioinformatics data analysis to arrive at the result. The ability of the FDA to streamline the regulatory evaluation process for these devices would be greatly enhanced by a publicly available high-quality reference database.

The advent of HTS brings with it the need for extensive well curated databases for organism identification. To enable premarket clearance and more widespread use in the clinical microbiological laboratory, these databases need to be populated based on agreed upon criteria to provide assurances to the end users of the quality of the deposited information. Device manufacturers may be limited to resources and genomic information at their disposal and that may result in clinical applications of limited scope that rely on proprietary databases of unknown quality. To enable the clearance and widespread adoption of HTS platforms for microbial detection in a timely manner, FDA is envisioning a well curated and public microbial reference database containing broad strain/isolate entries that are of a quality standard that is commensurate for use in regulatory and clinical decision processes.

To initiate this process, the Division of Microbiology Devices is engaged in generating an initial set of high quality, regulatory-grade microbial sequences through the FDA MicroDB project funded by the Office of Counterterrorism and Emerging Threats (OCET). These genomic sequences, their metadata, and standard operating procedures, from specimen collection to deposition in the public database, will be made available shortly (accessible from FDA via web interface or directly within NCBI databases). Our vision is a robust, high quality microbial database that contains (a) qualified regulatory-grade sequence data from public databases (NCBI/DDBJ/EMBL), (b) newly generated high quality sequence data for prioritized organisms that represent some of the information gaps in the public domain (initial set of regulatory-grade sequences from the FDA MicroDB project), and (c) future genomic sequence entries that meet the qualifications outlined here-in.

Currently, FDA envisions that the raw sequence data (in FASTA or FASTQ format) and sufficient clinical and technical metadata are required to enable the re-validation of existing sequencing-based diagnostics as other bioinformatics approaches and technologies emerge.

### ***FDA's Perspective on the Regulatory Quality Criteria for Microbial Genomic Entries***

The following sections highlight the areas of information that FDA intends to capture so that genomic sequence depositions in the public domain can be evaluated and qualified for regulatory purposes.

#### **A. Extracted Genomic DNA (gDNA)**

Extracted gDNA should be of high quality and purity, and at sufficient concentration to achieve a suitable yield to assure adequate depth and breadth of genomic coverage for the type of sequencing method employed.

#### **B. BioSample Metadata**

A minimal description of the isolate source material is necessary for traceability. We are using 14 descriptors as outlined below. (Note: Minimal metadata is modeled in part after NCBI's minimal pathogen template)

1. Unique ID	Unique Database ID for the sample
2. Organism	Scientific name of the organism (genus and species), the source of the sequenced genetic material
3. Strain/Isolate	Strain/isolate from which sequence was obtained
4. Sample Site	Anatomical sampling site (e.g., skin, wound, urine catheter)
5. Specimen Type	Specimen type (e.g., blood, stool, urine)
6. Host Disease	Name of relevant disease, e.g. sepsis
7. Collection Date	Date of sampling, in "DD-Mmm-YYYY", "Mmm-YYYY" or "YYYY" format
8. Collected By	Name of the person or lab who collected the clinical/countermeasure isolate
9. Patient Age	Age group (FDA criteria)
10. Gender	Gender (male, female)
11. Geographic Location	Geographical origin of the sample (recommended)
12. AST Method*	Antimicrobial susceptibility testing method (recommended)
13. AST Method Manufacturer*	Manufacturer of AST Method (recommended)
14. Antimicrobial Susceptibilities*	For each antibiotic (e.g., Vancomycin, Oxacillin, Tetracycline, Tobramycin) (recommended)

\*It is important to note that not every entry will have the associated antimicrobial susceptibility (AST) data, however, the lack of the AST data will not be used as a criteria for exclusion. The purpose of this information is to create a link between the phenotypic traits of particular organisms and their genomic sequence. Moreover this information is becoming increasingly critical as diagnostic technologies begin to migrate away from more traditional culture based formats.

### C. Sequencing Data

The minimum requirement for sequencing data is that the generated raw reads should be deposited in NCBI's Sequence Read Archive (SRA) and assemblies should be deposited at NCBI's Assembly division. The availability of raw reads and assemblies will provide a pathway to re-analyze the data as newer technologies emerge. Furthermore, annotation data should be deposited when available.

1. SRA	Deposit raw reads at NCBI's Sequence Read Archive (SRA) division
--------	--

- |                |   |
|----------------|---|
| 2. Assembly    | Deposit assemblies at NCBI's Assembly division                  |
| 3. Annotation* | Deposit annotations at NCBI's Annotation division (recommended) |

\* Genome annotations should be deposited at NCBI's Annotation division when available and/or should be requested to be added using NCBI Prokaryotic Genome Annotation Pipeline (PGAP)

#### D. Sequencing Metadata

A minimal description of the sequencing process is necessary for traceability. We are using 7 descriptors as outlined below including bioinformatics tool information for assembly and annotation, and genomic coverage information.

- |                    |   |
|--------------------|---|
| 1. Library         | Library manufacturer, strategy, source , selection and layout of library                                |
| 2. Platform        | Platform manufacturer and instrument model  |
| 3. Submitted by    | Name of person or sequencing center that submitted the clinical/ countermeasure isolate sequencing data |
| 4. Fold coverage   | Coverage of genome  |
| 5. Pipeline        | Processing pipeline used to generate data, sequencer platform software and version                      |
| 6. Assembler       | Assembler and version   |
| 7. Annotation Tool | Annotation tool and version (recommended when available)  |

#### E. Suggested Phenotypic Metadata

A description of the phenotypic information is suggested to create a link between the phenotypic traits of particular organisms and their genomic sequence. We are recommending 5 descriptors as outlined below (1-4 are also included in sections B and C).

- |                                   |   |
|-----------------------------------|---|
| 1. Annotation                     | Genome Annotation data  |
| 2. AST Method                     | Antimicrobial susceptibility testing method                                 |
| 3. AST Method Manufacturer        | Manufacturer of AST Method  |
| 4. Antimicrobial Susceptibilities | For each antibiotic (e.g. Vancomycin, Oxacillin, Tetracycline, Tobramycin)  |
| 5. Additional Phenotypic Data     | Information on morphology, gram stain, virulence data, metabolic data, etc. |

The FDA in collaboration with the National Center for Biotechnology Information (NCBI) has identified more than 550 clinically relevant bacterial pathogens for which validated and curated genomic sequence data is needed to advance the use of HTS as diagnostic devices and promote public health. The following is a description of the FDA MicroDB project involving the construction of an initial set of these >550 high quality microbial reference sequences. We expect to expand the database with additional sequences once regulatory quality criteria are finalized from (a) currently available sequences in the public domain and (b) newly submitted entries that meet the pre-defined criteria described above.

#### Overview of FDA MicroDB: Regulatory Microbial Database

Link to FDA MicroDB NCBI BioProject: <http://www.ncbi.nlm.nih.gov/bioproject/231221>

The FDA MicroDB project is an initiative to generate a set of high quality regulatory-grade sequences for bacterial microorganisms. It is important to note that this data set is by no means exhaustive and will require additional efforts to expand in order to encompass the breadth of microorganisms and antimicrobial resistance markers necessary to realize this effort. Funding was provided for more than 550 medical countermeasure and clinical relevant pathogens from the Office of Counterterrorism and Emerging Threats (OCET). HTS is used to generate these sequences and microorganisms are prioritized based on gaps in the public domain (diversity of genera).

We are in the process of developing minimal regulatory quality criteria that were used to inform our approach. Samples are acquired from several sources, both clinical isolates and repository entries. A high-throughput laboratory will be tasked with implementing a complementary dual platform approach to generate genomic sequence for the selected pathogens. Data analysis will include hybrid assemblies and annotation. Raw reads, assemblies and annotations will be deposited at NCBI databases. In addition, FDA will provide a web-based interface to access the sample and sequencing data from these NCBI databases.

The long term goal is to construct a microbial reference database for use as a tool to enable a regulatory and clinical pathway for implementation of sequence-based diagnostic devices in the Microbiology Laboratory. All data and procedures will be publically available to continue this effort. Traceability and repeatability are essential in this fast-paced field of emerging sequencing based technologies. The realization and adoption of such standards will improve quality of the device development process and expedite the regulatory review as it will serve the needs of “platform-agnostic” sequencing-for-diagnostics systems, such as HTS platforms, and continue to drive new developments for sequence-based systems. We are hopeful that the initial set of regulatory-grade sequences and the established regulatory criteria and procedures are a step forward towards evaluating microbial sequencing based diagnostic devices for regulatory and clinical use.

#### **Section 4: Clinical Application and Challenges:**

FDA is seeking input from academia, government, industry, clinical laboratories, and other stakeholders, at a very practical level, about current and realistic clinical applications of HTS for

microbial diagnostic use. FDA recognizes that HTS has rapidly advanced, but understands that the technology is still continuously improving and certain critical parameters and criteria listed below are influencing the use of HTS devices in the management of patients. Also, the successful adoption of HTS in the clinical microbiology laboratory still requires expertise in both molecular biology techniques and bioinformatics, which can be a challenge for minimally staffed clinical microbiology laboratories. From the FDA perspective, it is essential that any HTS assay includes sufficient information on how the generated sequence data is transformed into clinically meaningful outputs, otherwise the clearance or approval of these devices will be hindered. Some examples of clinical uses and challenges for HTS applications are discussed below. In these examples, high throughput sequencing is likely to be conducted with previous knowledge of the reference genome(s). Although *de novo* microbial sequencing (sequencing uncharacterized genomes when no reference or “gold standard” sequence information is available) has been used for several epidemiologic applications (e.g., disease surveillance, outbreak investigation, and determination of pathogen etiology), it may be difficult and unnecessary to employ this method for diagnostic purposes.

## **Clinical Applications for HTS:**

### **1.) Microorganisms Detection and Identification**

Any microbial detection and identification device used as an aid in the clinical management of microbial infection will be heavily affected by the specimen type intended to be tested. Some specimen types are established samples for HTS while others present issues that need to be seriously considered before HTS technologies are applied. Following are the possible clinical applications, with any issues they may have:

- a. Clinical samples from culture enriched clinical isolates, culture media, or pre-amplified PCR processed samples (e.g., blood cultures, blood agar, chrome agar etc.)

These samples usually represent homogeneous material, with limited background from contaminating genomes, and limited interference substances. The required nucleic acid material is usually easily and abundantly available for the purification procedures and downstream amplification reactions. The initial amplification step via culture could help circumvent issues of sensitivity for some HTS systems as discussed later. Even invalid runs or results can easily be followed up with an additional HTS reaction; consequently concerns of false negative test results with culture enriched specimens are low.

Please note, other techniques, such as mass spectroscopy technologies, have emerged for the identification of single colonies isolated from clinical specimens. Recently, MALDI-TOF diagnostic instruments were cleared that perform accurate species identification within minutes directly from a single colony at very low costs, which makes this assay a clear competitor of HTS technology. However, for certain applications, MALDI-TOF technology does not provide the necessary resolution.



- b. Clinical patient samples with a limited number of pathogenic organisms (e.g., cerebrospinal fluid, blood, saliva, urine, wound swabs/fluids)

Some clinical specimens, including certain body fluids, usually have only a limited number of pathogenic organisms in the sample. The capabilities of HTS technology to directly detect certain pathogens without any initial amplification with a short turnaround time should be robustly demonstrated. Data that support such a capability has been reported from the field of intestinal microbiota and environmental research; however, the limited amount of pathogen nucleic acid in the clinical specimen might still be challenging.

- c. Specimens with complex microbial background, high heterogeneity, or large amounts of non-target genomic information, such as human DNA, nucleic acids from commensal bacteria, and/or nucleic acid from asymptomatic shedding (e.g., feces, throat-nasal swabs, upper and lower respiratory aspirates).

The detection of a pathogen with a low copy number in the background of a complex microbial flora presents extreme challenges and approaches to enrich the nucleic acid sequence for the microbe of interest might be necessary. Single cell or individual real time DNA molecule sequencing technologies may be able to perform this identification; however, application of the appropriate random sampling/sequencing statistics might be required to avoid under sampling and false negative results. So far, concrete applications have been limited; however, possible scenarios are emerging where HTS from complex specimens is permissible, e.g., for the monitoring of at-risk patients for bloodstream invasion by vancomycin-resistant *Enterococcus faecium* where changes in the intestinal microbiota towards vancomycin-resistant *Enterococcus* domination precedes the bloodstream invasion in that patient [Ref: 1].

- d. Difficult to culture microorganisms, or uncultured organisms

The accurate detection of non-culturable or difficult-to-culture organisms, including slow growing organisms (e.g., *M. tuberculosis*), fastidious bacteria and anaerobes, and possible biothreat agents [Ref: 2], has proven to be another early application of microbial HTS. For example for drug resistant tuberculosis, it is essential to perform *Mycobacterium tuberculosis* complex identification, antimicrobial susceptibility testing, and bacterial genotyping, all of which typically takes weeks to 1-2 months due to the slow growth rate of the *Mycobacterium tuberculosis* complex. Microorganisms in which the primary diagnostic assays are serologic may be particularly difficult to diagnose in hosts with abnormal immune systems. For these microorganisms, HTS technologies could provide an alternative mechanism for microorganism identification. Additionally, in the presence of co-infection, HTS eliminates the need to isolate and culture individual microbial species as multiple pathogenic bacteria can be sequenced and identified simultaneously.

## 2.) Microbial Characterization/Antimicrobial Resistance/Marker Detection for Guided Therapy

- a. Antimicrobial resistance: examples include monitoring for genotypes related to antimicrobial resistance/susceptibility

The application of HTS in antimicrobial susceptibility testing is more limited, given that the sensitivity and robustness of phenotypic susceptibility testing is not reached by HTS technology, in part due to incomplete data linkage of genotype to phenotype. In addition, phenotypic antimicrobial susceptibility testing is inexpensive and has been highly automated and fully established in numerous clinical laboratories. However, there are certain immediate applications for the use of HTS technology for antimicrobial resistance: (i) where phenotypic testing is prohibitively slow, e.g., as mentioned previously, the genotypic antimicrobial susceptibility testing for microbes that are difficult to grow [Ref: 3], and (ii) where drug resistance traits are not caused by multiple genetic components, but instead linked to point mutations or small indels in a single gene, e.g., rifampin resistance in *M. tuberculosis*, methicillin resistance in *S. aureus* and trimethoprim-sulfamethoxazole resistance in *T. whipplei* [Ref: 4]. In addition, if the overall HTS turnaround time, including molecular and bioinformatics methods can be reduced to one day, then HTS could complement phenotypic testing to rule in resistance for certain antibiotics where known drug-resistance mutations or genes are found before phenotypic results become available [Ref: 5], and time and resources could be redirected to detect resistance encoded by novel mechanisms.

- b. Virulence factors; e.g., toxin production

Many bacteria express toxins that can cause severe disease (e.g., toxic shock syndrome caused by *Streptococcus pyogenes*). Traditionally, detection of these virulence factors was accomplished using bacterial serotyping or PCR based techniques; however, these assays can give false negative results if the toxin-encoding gene has been mutated. In this case, HTS applications could provide an alternative mechanism that would allow the sensitive detection of bacterial virulence factors even in the presence of mutations.

## 3.) Epidemiological Typing/Outbreak Investigation

Microbial genotyping has been recently carried out to support infection control teams. This can be of utmost importance for outbreaks management in hospitals and in the community. For example, the use of carbapenems for *E. cloacae* infections has been further limited by the emergence of acquired carbapenemase-producing strains of *Enterobacteriaceae*. Here, HTS can deliver additional information by identifying the precise subtype of the resistance gene in question and providing a list of resistance determinants beyond the dominant determinants implicated in carbapenem resistance. The genomic data can also be used for epidemiological tracking to investigate onward transmission. In these cases, extended antimicrobial susceptibility patterns can provide timely evidence of the

introduction and transmission of a new strain if it has a different susceptibility pattern to those seen in the preceding weeks or months [Ref:5]. For example, the genotypic natural variation observed between *Salmonella* strains has shown to be both stable and sufficient to allow high resolution traceback of clinical isolates and foods. This so called next generation sequencing strain “nanotyping” holds the potential to revolutionize the manner in which responses to outbreaks are managed [Ref: 6]. Sequencing of the entire genome provides the ultimate resolution for epidemiological studies, as demonstrated by several recent studies, including outbreaks of cholera in Haiti and *E. coli* O104:H4 in Germany. Routinely, epidemiological typing is used to detect laboratory cross-contamination, to define transmission pathways of pathogens [Ref: 7], and for organisms whose rate of genomic change is sufficiently high, the resolution obtained may make it possible to reconstruct transmission pathways between healthcare centers, hospital wards, or even patients on the same ward. The use of HTS in this setting would provide a mechanism for monitoring outbreaks in real-time and highlight daily opportunities for infection control [Ref: 5].

#### 4.) Novel Pathogens or Specialty Cases

HTS could also be used in cases where standard diagnostic tests consistently fail to identify the causative pathogen, either due to it being completely novel or due to it being a variant of a known pathogen that leads to false negative results. [Ref: 8, 9].

### **Clinical Challenges for HTS:**

#### 1.) Turnaround Time

The turnaround time for microbial HTS must be sufficiently short to support individual patient treatment or contemporary infection control decisions. The technological advances have been so steady that the latest generation of benchtop DNA sequencing platforms now provide an accurate high throughput sequence for a broad range of bacteria in less than a day. Preferentially, this one day turnaround time includes clone isolation, nucleic acid extraction, sample preparation, and bioinformatics assembly and analyses, all of which reduce the need for highly trained and qualified technical HTS specialists.

#### 2.) LoD/Sensitivity Issue

HTS must become sensitive enough to sequence DNA from specimens without the need for sub-culturing or a DNA pre-amplification step. In the best case scenario, parallel execution of phenotypic antimicrobial susceptibility testing and DNA sequencing from an isolate, done overnight, should deliver phenotypic susceptibility and high throughput epidemiological typing. The analysis of low abundant species is reported to be possible with single cell sequencing approaches [Ref: 10].

#### 3.) Data Processing and Results Interpretation

Massive amounts of data are generated by HTS devices and the originally obtained raw data from the instruments requires complex bioinformatics based data processing to generate a final result to be used for patient management. For clinical and diagnostic laboratories to effectively use these HTS devices, it is of utmost importance to have robust and streamlined data processing pipelines and almost fully automated sequence interpretation software and tools available that deliver clinically relevant information in a format that is understood and can be acted on by healthcare providers with no specialist knowledge of genome sequencing. In addition, the time needed for these complex bioinformatics based computations should not significantly extend the overall process. Data processing software that is platform and organism independent and could perform different tasks ranging from epidemiological tracking to susceptibility testing of any organism would be the ideal approach. It is important to note that the utility of the interpretation software will depend on a continuously updated database containing sequence variants or genes that account for drug resistance. Samples that have previously unknown mutations in genes that confer drug resistance of other clinically relevant bacterial factors could be flagged for phenotypic follow-up.

## **Section 5: Streamlining Clinical Evaluation for Regulatory Submissions**

The cornerstone of most, if not all, regulatory submissions to obtain clearance/approval to legally market a diagnostic device for microbial identification is the clinical evaluation. In this evaluation the subject device is evaluated against an already cleared device or a standard laboratory reference method which is used to establish the presence/absence of a pathogen. As diagnostic devices become more complex, with higher orders of multiplexing, the methods employed to establish sensitivity and specificity of the subject device become more diverse and can become burdensome on the manufacturer. Moreover, many of the comparator methods are not standardized and results can vary between various laboratories and manufacturers. The capabilities of HTS technologies may enable a streamlined, cost-effective, and harmonized approach to establish comparative performance for a variety of diagnostic devices.

Many infectious organisms that are detected using molecular methods cannot be cultured, or are extremely fastidious, and therefore, traditional reference methods cannot readily be used for comparative analysis to establish performance. In many cases, when these pathogens are detected by the subject device there is an array of sub-optimal comparators that must be utilized to establish performance. The implementation of HTS as a comparator method could help to provide a more uniform approach to demonstrating performance of devices that detect these types of organisms. However, this approach is not without issues that should be considered when comparing HTS results with those obtained from other device types for microbial identification.

First are the potential differences in the limit of detection (LoD) between the HTS platforms and the subject devices. It is expected that the LoD for HTS sequencing instruments may be higher than that of the more traditional molecular approaches, as the trade off in performance capabilities to detect any pathogen present may be at the expense of assay sensitivity. Consequently, when comparing device performance to HTS platforms this could potentially lead to a large number of false positives by the subject device due to the lower sensitivity of the HTS

comparator, in specimens that are actually positive. If this does indeed become a factor, then a more targeted sequencing approach may still need to be employed to determine the clinical truth in specimens used for the evaluation of diagnostic devices.

Second, to implement this approach significant consideration needs to be given to the protection of patient information. Although the proposed approach is for use in the determination of comparative performance for the detection and identification of microorganisms and their markers of antimicrobial resistance the capabilities of HTS may lead to unexpected findings. Furthermore, the reads that are generated using a high throughput approach from a human clinical specimen could undoubtedly contain information for markers of heritable diseases.

In this discussion paper, FDA has proposed the development of a new comparator paradigm in which a reference database would provide high throughput sequence information that could be used to verify results obtained during clinical trials for HTS platforms. Moreover, the combined use of HTS platforms and a high quality database potentially also provides device manufacturers with another tool to use to determine the microbial composition of a specimen used in the determination of sensitivity and specificity. The implementation of this novel paradigm would likely occur through an evolving process as we begin to understand the capabilities of these newly emerging technologies more fully. There will still be cases where more traditional evaluation methods may need to be employed; however, the intended use of the subject device will be the determining factor.

#### References:

- 1.) Ubeda et al., (2010) J Clin Invest 120:4332-4341
- 2.) Zaph C (2010) J Clin Invest 120: 4182-4185
- 3.) Köser et al., (2012) PLoS Pathog 8(8): e1002824.
- 4.) Bakkali et al., (2008) J, Infect Dis 2008; 198:101-108
- 5.) Köser et al., (2012) NEJM 366:2267-2275
- 6.) Allard et al., (2012) BMC Genomics 13: 32-
- 7.) Schürch AC, Siezen RJ 2010; Microb Biotechnol 3: 623-633
- 8.) Relman DA (2011) NEJM 365: 347-357
- 9.) Garcia-Alvarez, et al., (2011) Lancet Infect Dis 11:595-603
- 10.) Yilmaz & Singh, Curr Opin Biotechnol 2012 June; 23(3):437-443