FOOD BIOLOGICAL CONTAMINANTS

Baseline Practices for the Application of Genomic Data Supporting Regulatory Food Safety

DOMINIC LAMBERT
Canadian Food Inspection Agency, Ottawa Laboratory (Carling), Ottawa, ON, Canada
ARTHUR PIGHTLING
U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition, College Park, MD
EMMA GRIFFITHS
Simon Fraser University, Department of Molecular Biology and Biochemistry, Burnaby, BC, Canada
GARY VAN DOMSELAAR
Public Health Agency of Canada, National Microbiology Laboratory, Winnipeg, MB, Canada
PETER EVANS
U.S. Department of Agriculture, Food Safety and Inspection Service, Washington, DC
SHARON BERTHELET
Canadian Food Inspection Agency, Ottawa Laboratory (Carling), Ottawa, ON, Canada
DUNCAN CRAIG
Food Standards Australia New Zealand, Canberra, Australia
P. SCOTT CHANDRY
Commonwealth Scientific and Industrial Research Organisation, Melbourne, Australia
ROBERT STONES
Food and Environment Research Agency, Sand Hutton, York, United Kingdom
FIONA BRINKMAN
Simon Fraser University, Department of Molecular Biology and Biochemistry, Burnaby, BC, Canada
ALEXANDRE ANGERS-LOUSTAU and JOACHIM KREYSA
European Commission, Joint Research Center, Ispra, Italy
WEIDA TONG
U.S. Food and Drug Administration, National Center for Toxicological Research, Little Rock, AR
BURTON BLAIS[1]
Canadian Food Inspection Agency, Ottawa Laboratory (Carling), Bldg 22, 960 Carling Ave, Central Experimental Farm, Ottawa, ON, Canada K1A 0Y9

The application of new data streams generated from next-generation sequencing (NGS) has been demonstrated for food microbiology, pathogen identification, and illness outbreak detection. The establishment of best practices for data integrity, reproducibility, and traceability will ensure reliable, auditable, and transparent processes underlying food microbiology risk management decisions. We outline general principles to guide the use of NGS data in support of microbiological food safety. Regulatory authorities across intra- and international jurisdictions can leverage this effort to promote the reliability, consistency, and transparency of processes used in the derivation of genomic information for regulatory food safety purposes, and to facilitate interactions and the transfer of information in the interest of public health.

Received August 17, 2016. Accepted by AH November 22, 2016.

The opinions expressed by the authors do not necessarily reflect the opinions or policies of their respective institutions. Any statements in this article should not be considered present or future policy of any regulatory agency.

The authors declare no conflicts of interest.

[1] Corresponding author's e-mail: burton.blais@inspection.gc.ca; burton.blais@canada.ca

DOI: 10.5740/jaoacint.16-0269

The role of regulatory food safety agencies worldwide is to pursue a scientifically informed approach in protecting consumers from preventable illnesses. The application of leading-edge analytical technologies for the detection and characterization of foodborne pathogens is one of the underpinnings of an effective risk-based regulatory food safety system. Approaches capable of maximizing the amount of information obtained in the course of conducting laboratory testing of samples will foster the most appropriate regulatory responses, e.g., by informing the health risk assessment process undertaken to categorize the degree of risk attending a contamination incident.

In the present era of globalization and the introduction of new food manufacturing, distribution, and consumption practices, food microbiology testing programs require high-throughput analytical technologies that provide actionable results within a short time frame. The classic approach has relied on the recovery of microorganisms (particularly bacterial pathogens) from food samples by enrichment, their identification on the basis of phenotypic characteristics elucidated by biochemical and serological techniques, and their typing by molecular methods such as pulsed-field gel electrophoresis analysis of genomic DNA banding patterns (1, 2). Although effective under many circumstances, there are shortcomings to this limited, low-resolution approach, such as difficulty identifying certain classes of pathogens with irregular or unconventional features

[e.g., virulent Shiga-toxigenic *Escherichia coli* (STEC) strains)] and attributing contamination sources.

Next-generation sequencing (NGS) technologies offer new possibilities for comprehensive analyses of microbial isolates recovered from inspection samples. For example, whole-genome sequencing (WGS) can now render a bacterial genome much faster and at a significantly lower cost than previously possible, making it feasible to sequence foodborne isolates in near real time (e.g., during foodborne illness outbreak investigations). Currently available bioinformatics tools are sufficiently advanced to enable the rapid processing of raw sequence data into a usable form for many purposes. Sequencing pathogenic bacteria, whether in the context of food safety investigations or information-gathering in the course of research, can bring an unprecedented quality of information regarding the presence of virulence and other marker genes of relevance to pathogen identification and risk characterization (3, 4).

The role of genomics technology in food microbiology inspection programs is still being defined (5), and there are several possible avenues for its integration in food safety regulation. The following examples illustrate how genomic technologies are used:

(*1*) *Source attribution for food safety incidents*.—Foodborne pathogens may be characterized (typed) to a high degree of resolution using WGS data, enabling both the determination of the degree of relatedness among clinical, food, and environmental isolates and the attribution of contamination sources.

A global foodborne illness investigation focusing on *Salmonella* Bareilly isolates recovered from clinical and implicated food samples featured high-resolution typing analysis of WGS data, enabling identification of the geographic origin of the initial contamination event with pinpoint accuracy (6).

WGS data analyses have also been used in the typing of foodborne pathogens during epidemiological and attribution studies that sought to describe related clusters of bacteria arising from common sources (7–9).

(*2*) *Identification, characterization, and genotyping of pathogenic bacteria on the basis of genomic markers*.—Pathogenic bacteria recovered from foods can be identified on the basis of their definitive genetic characteristics, enabling delivery of test results days sooner than traditional biochemical techniques. Genomic approaches for the identification of foodborne bacterial isolates have the potential to support more timely regulatory interventions, as well as enhance the evidence base for food safety risk analysis (e.g., policy/standards development).

In one documented approach, colonies of priority STEC were subjected to WGS with raw data mapping and analysis during the early stages of the sequencing process, enabling completion of the procedure within a single working day (10). This approach may be regarded as an identification system that offers ultimate multiplexing capacity in terms of the number of different genomic markers that can theoretically be investigated at once, providing inherent flexibility that enables the determination of any relevant genomic marker on an ad hoc basis.

Analyses of WGS data are also used for the identification of genetic markers and during functional genomic studies for the purpose of making phenotypic predictions regarding an organism's potential for virulence, pathogenicity, and antimicrobial resistance (11, 12).

Although many WGS and bioinformatics operations may be common to all bioanalytical scenarios, food safety applications, which for present purposes are confined to the analysis of bacterial pathogens, may have particular requirements that warrant the development of specific guidelines. Indeed, quality requirements for sequence calling stringency, depth, and breadth of coverage and extent of genome assembly will vary depending on the intended purpose of the analysis. Moreover, the nature of biological data used to assess sequence data, or the utilization of reference genomes selected from large (though possibly incomplete) databases to test bioinformatics processes, will need to reflect the properties of the types of bacteria implicated as food safety concerns. Processes that may be amenable to the analysis of eukaryotic sequence data may not be suited for bacterial genomes. Likewise, the organization of a certain bacterial species genome (i.e., synteny) may lend itself to certain types of analytical procedures (e.g., reference-based assembly) that would not be biologically appropriate for another species.

Therefore, the implementation of best practices should be flexible in the determination of both the appropriate WGS data quality attributes and the technical requirements necessary for the development and adaptation of bioinformatics tools that address local end-user applications, i.e., be based on the principle of fitness-for-purpose. Their adoption on an international scale would not only enhance consumer protection but also facilitate information transfer, scientific collaboration, and potentially, trade of agri-food commodities among partners in the world economy.

The intent of the present guidelines is to focus on WGS applications for microbiological food safety, although the general principles expounded herein can be extended to other types of genomics technologies. Even within WGS, the intent of the present guidelines is not to prescribe any particular sequencing chemistry or analytical approach but, rather, to establish best practices that assure that essential details for bioinformatics processes are adequately benchmarked and the outputs are captured in sufficient detail to ensure retrospective scrutiny. Establishing best practices for both the utilization of bioinformatics analysis and reporting of WGS data among food safety partners in the international community will promote the systematic application of quality criteria. This will ensure that procedures and processes used to support regulatory decisions are reliable, transparent, reproducible, and auditable (13) and that provisions are made for the long-term storage of data so that analyses can be repeated under conditions as close as possible to the original.

Metadata Generation and Standardization

The purpose of metadata is to help identify, organize, and summarize a data set, to facilitate the discovery of relevant information about a data set, and to convey that information to an informed user community. During food safety investigations, experimental results (whether based on traditional microbiology or WGS) and contextual information (e.g., epidemiology questionnaires) should be shared among different government agencies, stakeholders, and end users to assess and manage risks to human health. Therefore, it is crucial that not only a sufficiency of metadata describing an isolate itself, its source (e.g., clinical, food, or environmental), and associated

laboratory results is captured, but also that stakeholders agree on a defined set of standardized parameters allowing the correct interpretation of the data used to support a regulatory action.

The requirement for standardized, high-quality metadata has been recognized by the international regulatory food safety community, and a number of metadata standardization initiatives have sought to harmonize approaches among different jurisdictions and across all segments of the end user community. The development of robust, internationally relevant food vocabularies has been an area of particular interest. The intrinsic attributes of foods, such as variability in pH, water activity, and antimicrobial properties, have a considerable impact on the resident microbiome and survival characteristics of pathogens. Given the immense complexity of the modern global food supply (e.g., the large variation in the composition and properties of meats and their derived products), the establishment of a precise, universally understood vocabulary for the identification of food sources is paramount. Precision in specifying the food of origin using a universal consensus language is essential to enable investigators, present and future, to make rational use of databases and have a sufficiency of information to support any inferences that the data may suggest. A key consideration is the implementation of a metadata approach that utilizes a controlled vocabulary to ensure consistency among contributions to the database.

LanguaL (Langua aLimentaria, or "language of food"; 14), first developed in the late 1970s by the U.S. Food and Drug Administration Center for Food Safety and Applied Nutrition, is a standardized language for describing food. LanguaL contains around 35 000 foods and is internationalized with equivalent terms in Czech, Danish, English, French, German, Hungarian, Italian, Portuguese, and Spanish. The equivalent European initiative, FoodEx, is a food dictionary constructed by the European Food Safety Authority. The latest version (Food Ex2) provides a comprehensive classification of terms and is designed to facilitate food exposure assessment (15). These food commodity databases are further outlined in Table 1.

Although the aforementioned databases allow detailed descriptions of food matrices, they were not primarily developed to integrate metadata related to WGS-based microbial identification and characterization in support of regulatory food safety. As a result, the structure and organization of these hierarchies do not favor the genomic context, and the generation and standardization of metadata describing genomic data sets needs further development. The principles of best practices implementation for metadata have been broadly described (16), and may be summarized as follows:

Table 1. Examples of food commodity databases

Food commodity databases	Comments
LanguaL	Standardized language for describing food based on facets of food composition, preservation, and labeling (http://www.langual.org/)
FoodEx2	Standardized food classification and description system developed by the European Food Safety Authority, with descriptions of a large number of individual food items by food groups and broader food categories in a hierarchical relationship (http://www.efsa.europa.eu/en/data/data-standardisation) (http://www.efsa.europa.eu/sites/default/files/assets/804e.pdf)

(*1*) Identify the scope of information to be captured in the metadata standard.

(*2*) Develop a simple structural framework to store and organize metadata [e.g., minimum information (MI) checklist].

(*3*) Develop minimum requirements to make the metadata accessible, exchangeable, and minable (e.g., metadata syntax and semantics).

(*4*) Build consensus and encourage uptake by the community. Achieving a consensus requires the creation of a consortium of subject experts, stakeholders, and end users that will generate and use the metadata content, syntax, and semantics.

(*5*) Monitor overlap between metadata standards. Related standards should be tracked and linked to avoid duplication. Metadata standards require calibration and maintenance over time. Standards with broad support and funding that can maintain relevancy over the long term are most likely to enjoy wide acceptability in the community.

Scope of Information

The requirements of end users who must interpret the results of genomic analysis to inform their decisions are an important consideration in defining the elements to be captured. For present purposes, it may be generally understood that a key goal of genomic analysis is to provide evidence supporting the implementation of regulatory strategies for the management of public health risks in the food supply chain. Therefore, there should be a sufficiency of information to enable risk assessors and risk managers to understand fully the context pertaining to a data set and the tools used in its generation. For regulatory food safety applications, the complement of metadata should include information describing (*1*) salient features of the pathogenic isolate itself (genus, species, etc.); (*2*) the circumstances of its isolation (source, location, time, etc.); and (*3*) the methods and analyses performed (*see Good Laboratory Practices and Prevention of Procedural Errors*) and the laboratory records, in order to ensure traceability from the sampling event to the provision of results.

MI Checklists

MI checklists aid in organizing as well as ensuring the consistency and completeness of the metadata content (rather than data format). The standardization of MI checklists ensures that the data are easily verified, accessible, interoperable, and readily interpretable by the regulatory food safety community. One of the most widely recognized MI lists is the MI about Any Sequence (MIxS) checklist (17), developed by the Genomic Standards Consortium (GSC). It consists of three core standards, i.e., MI about a Genome Sequence, MI about a Metagenome Sequence, and MI about a Marker Gene Sequence. Similarly, the Genome Sequencing Centers for Infectious Diseases and the Bioinformatics Resource Centers developed the Project and Sample Application Standard (18) to define metadata types that should be attached to human pathogen and vector genomic sequences. Required information includes characteristics of the organism or environmental source of the specimen, spatial-temporal information about the specimen isolation event, phenotypic characteristics of the pathogen/vector isolated, project leadership, and support. Another MI checklist relevant to the use of genomics in support of foodborne illness outbreak investigation and regulatory food safety is the MI about a Phylogenetic

Table 2. MI checklists relevant to the implementation of bioinformatics analyses in regulatory food safety

MI checklist resource	Comments
MI about a Neuroscience Investigation (MINI)/MI for Biological and Biomedical Investigations (MIBBI)	Common portal to a group of nearly 40 MI checklists for various biological disciplines (MIxS core standards can be found via this portal); the MIBBI Foundry is developing a cross-analysis of these guidelines to create an intercompatible, extensible community of standards (https://biosharing.org/standards/?selected_facets=isMIBBI:true)
MI about a Phylogenetic Analysis (MIAPA)	Application to formalize annotation of phylogenetic data (http://www.ontobee.org/ontology/MIAPA)

Analysis (19), which details the information necessary to evaluate phylogeny results (e.g., topology, alignment, and tree inference methods). These checklists have been aggregated by the MI for Biological and Biomedical Investigations project (20), along with many other MI standards developed for biomedical applications. An extended listing of relevant checklists and minimal metadata standards is presented in Table 2.

Minimum Requirements for Metadata Syntax and Semantics

To make metadata accessible, exchangeable, and minable, minimum requirements for metadata format standard (syntax) and meaning (semantics) must be developed. The format should facilitate metadata communication. Typically, an "object model" is created and then translated into a metadata exchange format (e.g., XML, YAML, and JSON). The meaning of the metadata is communicated with the use of "descriptors" and is best achieved via the creation of ontology (21, 22). Ontologies are open-source, well-defined standardized terms interconnected by logical relationships. These logical interconnections provide a layer of intelligence with which to query engines, making ontologies much more powerful than flat lists of information. Ontologies describe entities (universals and instances), classes (concepts), attributes, and relations, and present this information in a hierarchy that provides a searchable framework for integrating and sharing diverse types of information (23, 24).

Best practices for creating ontologies have previously been described (24, 25). These include (*1*) the designation of an existing ontology to reduce the creation of redundant standards and to simplify data integration; (*2*) the careful ordering of terms in a hierarchy, from the root to the highest node, to ensure logical coherence for reasoning; and (*3*) the formalization of the hierarchy in a computer-usable language that can be implemented as a computable framework. Principles of good practice in ontology development are now being put into practice within the framework of the Open Biomedical Ontologies (OBO) consortium through its OBO Foundry initiative (26). OBO principles include open-source software, well-referenced and defined terms, orthogonality, plurality of users, clearly bounded subject matter, good syntax, and community evaluation and feedback (27). The OBO Foundry family of ontologies can be searched through the Ontobee portal (28), whereas other useful ontologies can be found through the National Center for Biomedical Ontology BioPortal (29).

A number of existing ontologies can be leveraged by food safety authorities to facilitate the implementation of international metadata standards for the use of genomics to support regulatory actions. For instance, the Environmental Ontology (30) includes a description of food products and could be further developed to attain the level of detail found in LanguaL or FoodEx2. Likewise, the Infectious Disease Ontologies (31), a set of interoperable ontologies, each describing a specific organism, could be expanded to include common foodborne pathogens. Other relevant ontologies include the Ontology for Biomedical Investigations (32), which describes the protocols, instruments, materials, analyses, and results used during investigations, and the Genomic Epidemiology Ontology (33), which is being developed by the Integrated Rapid Infectious Disease Analysis consortium to describe the vocabulary necessary to identify, document, and research foodborne pathogens and associated outbreaks. Extended listings of relevant ontologies for regulatory food safety and metadata management resources for public health are presented in Tables 3 and 4, respectively.

Table 3. Ontologies relevant to the implementation of bioinformatics analyses to regulatory food safety

Ontology resource	Comments
Ontology for Biomedical Investigations (OBI)	Integrated description of biological and clinical investigations (http://www.ontobee.org/ontology/OBI)
Systematized Nomenclature of Medicine (SNOMED)	Collection of medical terms, in human and veterinary medicine, to provide codes, terms, synonyms, and definitions that cover anatomy, diseases, findings, procedures, microorganisms, substances, etc. (https://www.nlm.nih.gov/healthit/snomedct/index.html)
Sequence Ontology (SO)	Controlled vocabulary of sequence types and features for sequence annotation, exchange of annotation data, and description of sequence objects in databases (http://www.ontobee.org/ontology/SO)
Infectious Disease Ontology (IDO)	Description of entities generally relevant to both the biomedical and clinical aspects of infectious diseases (http://www.bioontology.org/wiki/index.php/Infectious_Disease_Ontology)
Environment Ontology (ENVO)	Specification of a wide range of environments and habitats relevant to multiple life science disciplines (http://www.ontobee.org/ontology/ENVO)
Human Disease Ontology (DOID)	Classification of human diseases organized by etiology, including some common foodborne pathogens (http://www.ontobee.org/ontology/DOID)
EMBRACE Data and Methods (EDAM)	Common bioinformatics operations, topics, types of data (including identifiers), and formats (http://www.ontobee.org/ontology/EDAM)
Genomic Epidemiology Ontology (GenEpiO)	Developed as part of the Integrated Rapid Infectious Disease Analysis Platform (https://github.com/Public-Health-Bioinformatics/IRIDA_ontology)

Table 4. Other metadata management resources for public health applications

Resource	Comments
Adverse Event Reporting (AERO)	Ontology aimed at supporting clinicians at the time of data entry, increasing quality and accuracy of reported adverse events (http://www.ontobee.org/ontology/AERO)
BRENDA Tissue/Enzyme Source (BTO)	A structured controlled vocabulary for the source of an enzyme, including tissues, cell lines, cell types, and cell cultures (http://www.ontobee.org/ontology/BTO)
Common Anatomy Reference Ontology (CARO)	An upper-level ontology to facilitate interoperability between existing anatomy ontologies for different species (http://www.ontobee.org/ontology/CARO)
Antimicrobial Resistance Ontology (ARO)	Classification of antibiotic resistance gene data

Good Laboratory Practices and Prevention of Procedural Errors

Because many regulatory laboratories generating and processing genomic data will likely subscribe to recognized QA standards (e.g., ISO/IEC 17025:2005; 13), any proposed new guidelines should be compatible with the established requirements of accrediting bodies. Whereas the present guidelines focus on the integration of WGS data and the relevant bioinformatics analyses in the regulatory food safety arena, best practices currently used by analytical laboratories certified to International Organization for Standardization (ISO) standards may simply be extended to capture relevant information. Many of the steps used to process food samples can indeed be reiterated to process bacterial isolates for WGS to ensure that laboratory records are traceable to the sampling event. Ideally, isolates sent for sequencing are linked to the food samples from which they are derived using a Laboratory Information Management System or an alternative system comprising management strategies for the food samples and bacterial isolates, the sequencing workflow, and the data being processed and stored. For instance, the documentation process begins with the reception of samples at the sequencing laboratory. Isolate details such as the reception date, sender and receiver identity, number of isolates shipped, and unique identifiers (derived from food sample identifiers; *see Metadata Generation and Standardization*) are typically captured on a controlled worksheet. This ensures that sample information is acquired in a consistent fashion and that there is no confusion between information that was missing versus that which was not properly captured at the time of reception. Moreover, the physical location of bacterial isolates within a culture collection should be documented. Likewise, information pertaining to the sequencing workflow, i.e., methods, instruments, and spreadsheets involved in genomic DNA extraction, quantification, and normalization, as well as details related to sequencing library preparation (lot numbers and expiration dates of reagents, kits, etc.), the identity of the analyst(s) who performed the procedures, and the instrument used, should be recorded on a controlled worksheet, along with the date on which the results were obtained. Standard operating procedures should be established to ensure consistency among different analysts. Deviations from methods, sometimes necessary to ensure business continuity, should also be documented using a controlled form, and data supporting the appropriateness of deviations should be included. Permanent modifications to methods should be recorded as new versions to ensure that previous iterations remain accessible for traceability purposes. Analysts should keep familiarization records related to training on the procedures and the instrument(s) used. The latter should be maintained and operated according to manufacturer specifications (e.g., environmental conditions and calibrations), and a log of maintenance is required.

Most errors reported after analyses of sequencing results appear to be due to mishandling of samples. Countermeasures to prevent such events, such as the use of barcoding, easy-to-read printed labels, or verification at critical control points, are highly recommended. Errors should be documented using a controlled form, recording the specific nature of the nonconformance along with details of the appropriate corrective action. Simple practices such as keeping the isolate, DNA extracts, and leftover libraries until analysis is complete and approved should be encouraged in case problems arise during the sequencing run. In addition, running duplicates of test samples during high-priority investigations, although more costly, can avoid delays in the event of poor sequence quality.

Information Technology Solutions

Laboratories accredited by internationally recognized laboratory QA standards (e.g., ISO 17025:2005) must demonstrate uninterrupted competence to produce "technically valid results" (13). In the case of WGS data analyses, this relies entirely on the use and maintenance of information technology (IT) to support data collection, analysis, transfer, and storage. The validation of IT applications confirms their performance according to pre-established specifications. Laboratories are required to develop and implement procedures to formally document validation of IT solutions in support of operations. A risk analysis may be used to determine the extent of validation for a given application. Here again, there is considerable experience to be gained from the implementation of IT solutions in analytical environments that subscribe to quality management system principles to inform the development of guidelines for pipeline validation and adequate data management. Specific QA requirements for these two elements are addressed in the following sections.

Software Validation

Genome analyses may require multiple procedures, or "steps," to be performed with different pieces of software. These steps can be carried out manually or, more often, using automated bioinformatics workflows (pipelines) that execute multiple applications with minimal user input (34, 35). Pipelines may also include scripts written in-house or by outside developers to perform a variety of tasks, such as refining the output of single-nucleotide polymorphism (SNP) callers (36, 37). Establishing that bioinformatics pipelines are fit-for-purpose and suitably validated entails adherence to guidelines

and standards developed by accreditation and international standard–setting bodies [e.g., the ISO and the Institute of Electrical and Electronics Engineers (IEEE) Standards Association] and documentation of the "software life cycle," including the following steps (modified from Gogates, 38):

(*1*) Develop software requirements through end user consultations.

(*2*) Document software design and follow programming best practices.

(*3*) Perform rigorous software testing that ensures that requirements of end users are met.

(*4*) Ensure robustness of software installation on user platform(s).

Laboratories should use these validation and verification activities to determine whether bioinformatics workflows conform to the requirements of a given activity and whether the software satisfies its intended use and user needs. The extent of documentation required for each of these activities depends on whether the software to be used is commercial or custom (38). For instance, validation of proprietary bioinformatics applications that are used without changes to source code (39) is accomplished by testing the products to establish that end user requirements are met. The process should be repeated for any updates to the software. Validation of custom software products that are downloaded, installed, and used freely (i.e., open-source software or freeware), as well as software written in-house, requires more extensive documentation before implementation. In any case, software validation should be well documented through user manuals, software descriptions, and testing methodologies, with results published in peer-reviewed journals. Ideally, data sets, source code, parameters, protocols, and analytical workflows (including any manual operations) should be made publicly available and be sufficient in detail to permit successful reproduction of analyses based on the original test data sets so that the software validation and benchmarking efforts may be repeated and appropriately vetted by the research community (40–42). Software archives such as GitHub (43) are especially well suited for software development and providing public access to software projects.

User Requirements

It is important to document end user requirements by determining precisely what users need the system to do. This document need not be technical in nature, and anyone generally familiar with the activity should be able to understand it. For example, in the case of trace-back investigations of foodborne pathogens using reference-guided alignment of WGS data to identify nucleotide differences, a simple set of end user requirements could be

- System accepts multiple FASTQ data sets and assesses their qualities.
- System produces sequence quality report.
- System accepts reference sequence and maps the reads contained in FASTQ data sets to the reference.
- System produces reports of nucleotide differences between provided genome sequence data sets, relative to the reference.

Variations or additions to these steps by individual laboratories will almost certainly be necessary; e.g., laboratories may require read quality filtering and trimming of reads prior to read mapping. Additional requirements may also be necessary when technicians require a graphical user interface to interact with the system, as opposed to using the command line. It is critical that technicians carefully consider this step and implement a rigorous process, including interviews with end users (such as risk assessment and food recall specialists), because it will facilitate proper system design and testing.

Software Design Documentation and Programming Best Practices

After the end user requirements are developed, software developers can translate them into their technical components, providing a workflow that will aid in documenting the software design. Software developers can document the design graphically, along with descriptions that demonstrate how the software will satisfy the requirements. Typical software design descriptions outline individual components or entities (e.g., module and subprogram) and the interfaces between them. Recommended practices for software design can be found in IEEE Recommended Practice for Software Design Descriptions (IEEE Standard 1016-2009). Good principles for software construction, such as the use of version-controlled software components in bioinformatics pipelines, have been previously described (44, 45) and can expedite software development and subsequent validation work.

Testing Procedures

Testing is required, whether laboratories use proprietary applications, freeware, or products written in-house, to ensure that the software performs the tasks prescribed by the user requirements in a reliable, reproducible, and robust manner. This activity usually includes an assessment of the reproducibility of the results, identification of the types and frequencies of errors, systematic tendencies (bugs) that generate incorrect or unexplained results, and factors that can affect whether the software works properly (e.g., dependencies). Moreover, appropriate benchmarking requires that laboratories develop (*1*) relevant data sets comprising genome sequences obtained from foodborne bacteria, (*2*) representative databases of genome sequences, and (*3*) application-specific quality thresholds and performance criteria. It should be noted that the selection of error thresholds needs to be validated and justified and match criteria for inclusion or exclusion of samples (46). Note, also, that benchmarking needs to be performed in order to test the protocols put in place by individual laboratories, including the evaluation of any conditions and parameters that have been selected, especially when laboratories deviate from software default or recommended settings. Verifications using standardized *in silico* data sets, such as those under development through the Global Microbial Identifier initiative (47), may be very useful for this purpose. Technicians should carefully consider these elements in order to demonstrate that the bioinformatics approaches that laboratories use during scientific investigations of foodborne pathogens are highly accurate and precise and that the entire workflow is as transparent as possible.

Laboratories should test software by using genome sequence data sets that are representative of the types and ranges of samples that will normally be analyzed (48). Technicians should also select data sets that challenge the software with very closely and somewhat distantly related genomes, as well

as data generated by a variety of sequencing platforms, if applicable. Genome sequence data with known relationships and differences [i.e., SNPs, insertions and deletions (indels), and structural variants] are extremely useful (49). Microbial genomic DNA reference material (e.g., DNA sequence data) made available by the National Institute of Standards and Technology can be accessed at the National Center for Biotechnology Information (NCBI) Sequence Read Archive (BioProject identification No. PRJNA252728). In addition, a set of *Listeria monocytogenes* short-read sequence data generated by the Listeriosis Reference Centre for Canada from sequencing runs of varying qualities is available at NCBI (BioProject identification No. PRJNA251727; 50). It is also possible to add artificial mutations to a known reference sequence to simulate nucleotide substitutions (50, 51).

It may be necessary for laboratories to develop their own benchmarking materials if suitable data sets are not readily available, as well as to generate simulated sequence data. The use of simulated sequence data provides the advantage that it can be generated within individual laboratories (52) and provides users with a great amount of control over the quality of the sequence data. Users can simulate sequence data with known numbers and locations of nucleotide differences, providing the ability to compare results to a ground truth. However, studies have shown that simulated reads do not accurately represent real sequence data and may, therefore, yield unrealistic results (50, 53). Thus, it is usually desirable that data sets comprising both real and simulated sequence data be used for testing purposes. All test data sets should be provided to external users for evaluation under specific running parameters with the software being tested.

It is also important that WGS databases used in trace-back investigations and genome-wide association studies be fit-for-purpose (i.e., databases should reflect the full breadth of biological variability of organisms that laboratories are studying), because the availability of genome sequence data can significantly influence the results of bioinformatics analyses (50, 54). Furthermore, it may be useful for laboratories to report the numbers of genome sequences in databases that are analyzed and state their criteria (e.g., minimum sizes of databases) for achieving the level of statistical power appropriate for their needs.

After laboratories select test data sets and ascertain the comprehensiveness of testing databases, it is necessary to develop criteria for measuring the performance of bioinformatics software products. In addition, to determine sequence application-specific quality thresholds, laboratories will need to conduct benchmarking experiments using WGS data of varying quality (e.g., depth and breadth of coverage and raw reads versus assemblies). For example, bacterial genotyping pipelines that use single-nucleotide differences (i.e., SNP; 55, 56) should be tested using data sets comprising subject and reference genomes for which the differences between them are known. The performance of these pipelines is commonly gauged by counting the numbers of four types of calls (i.e., true positive, in which the software correctly identifies differences; false positive, in which the software indicates a difference when none actually exists; true negative, in which the software makes no call and there are is no difference between the subject and reference; and false negative, in which the software makes no call and there is actually a difference; 57), by calculating performance metrics (49, 58), and lastly, by reporting results against performance criteria (e.g., minimum percentage sensitivity and selectivity) deemed acceptable by the regulating agency. Pipelines using a defined set of genomic features (e.g., multilocus sequence typing; 59) should be tested using a comprehensive allelic database and sequence data sets for which sequence types have been confirmed in order to assess the ability of the software to correctly identify a predetermined number of markers. In addition, it may be informative to evaluate whether the complete genome features were delineated completely or truncated (60). These approaches also apply to pipelines designed for the identification of genomic features that are strongly associated with phenotypes of interest (e.g., antimicrobial resistance gene or serotype), which should be tested using a set of genomes derived from strains thoroughly characterized biochemically and serologically. However, differences between phenotypic results and genotypic predictions should be further investigated to assess whether they are due to natural genetic events (e.g., SNPs and indels) before being accounted for as erroneous results (61). In any case, replicate experiments should be run during the testing phase with unique sets of genome sequences.

Installation

Pipelines may be run from local workstations (either stand-alone or as part of a sequencing platform), local servers, or off-site servers (e.g., Web-based applications). If pipelines are run on several different workstations, then validation of the software should be performed for each station (38). In addition, if changes in servers or Web-based applications occur, such as updates to operating systems or changes in hardware configurations, then additional testing and validation will be necessary. Note that for auditability purposes, the computing facility should keep records of operating system updates and hardware maintenance (date, time, and analyst), as well as documentation that demonstrates that the infrastructure is secure.

Data Management

Accredited laboratories must use data transfer and storage measures that ensure the integrity, control, accessibility, confidentiality, and security of data, documents, and records. The rationale supporting the choice of data transfer and storage methods should be documented (34). Although some sequencing platforms offer integral analytical tools to process the sequencing reads, many laboratories rely on alternative commercial or in-house programs to perform their analyses. This requires that the FASTQ files be transferred to separate computers, an additional step that must be accounted for. Sensitive or confidential information should be encrypted if it is to be transmitted to a separate site via a nonsecure network, and a controlled document (as defined in quality system guidelines such as ISO 17025:2005; 13) should be used to ensure that the identities of the sender and the receiver, as well as the date and time of transfer, are documented. A sequencing facility may wish to develop a procedure in which the receiver must acknowledge that the information was received in full and unadulterated (e.g., use of checksum), and that this confirmation is logged on the sender's controlled form. For auditability, the transfer strategy should be aligned to the regulatory body's policy on the use of external IT providers.

Versions of draft genomes, associated metadata, related controlled worksheets, and reports used to support a regulatory action, as well as all the raw FASTQ files and the version of the pipeline and all its components, should be archived securely and indexed to be easily accessible. Common approaches in testing laboratories for ensuring the integrity of the electronic data and records include controlling access, use of passwords, use of read-only storage media, backup strategies that allow restoration to the most-recent condition, and documented procedures that call for tracking information pertaining to the amendment of electronic records (i.e., a standard operating procedure defining who may access or modify data and who can approve the modification). To ensure continued accessibility, data and records must also be stored at a secure off-site location, and consideration must be given to maintenance of the attendant hardware and software, along with an analysis of the potential longevity of file format(s) and storage medium. A standard operating procedure defining who may access the information and make (and approve) different types of amendments should be developed. The procedure should also describe the frequency and conditions for conducting verifications of data integrity and format conversions. Records for all the above operations should be maintained.

Reporting and Interpreting Results

Match Criteria

To achieve the high levels of confidence required of conclusions that arise from epidemiological and attribution studies, it is necessary that the software used reliably and robustly identifies phylogenetic clusters or clades of pathogens, indicating common sources of contamination within the food supply. In the case of food processing facilities, pathogens could be introduced to products by the incorporation of tainted ingredients, through contact with contaminated surfaces within the production line, or both. The enhanced resolving power afforded by robust analysis of WGS data may enable analysts to distinguish between the two types of contamination events. Therefore, it is important that laboratories validate the conditions (usually levels of genome sequence identity calculated from SNP data) that indicate matches between bacterial genomes. Commonly, phylogenetic relationships between organisms (illustrated with trees or cladograms) and numbers of nucleotide differences between DNA sequences are used to determine inclusion or exclusion of subject samples within a larger group of genomes. Guidelines for the interpretation of these results should be provided, and deviations from those guidelines should be documented and justified. Similarly, for studies that seek to predict phenotypes of pathogens, the criteria used for determining correlations between genotypes and phenotypes should be explicit. Such studies usually require the use of mathematical models and analyses of population structures to make predictions about the relationships between genome features and characteristics that are ultimately tested in the laboratory. It should be kept in mind that databases for genome sequences obtained from foodborne bacteria are currently expanding at a rapid rate. It is likely they will continue to do so as campaigns to generate WGS data for pathogens continue to expand, and the characteristics of genome sequence data are likely to change in the future (e.g., read length and sequence quality) due to advancements in genome sequence technologies and chemistries. Therefore, it may be more appropriate to establish a process for setting thresholds and clearly defining where they can be used than to set specific cutoffs that are based on data that represent only a single point in time.

Factors Contributing to the Uncertainty of Test Results

Errors in WGS data may arise during data generation, including sample processing and DNA sequencing (62). Subsequent bioinformatics analyses of raw sequencing data, such as read mapping, assembling, and base calling, may also introduce errors (63). Although bioinformaticians have not fully evaluated the influence of these errors for all the comparative methods presented here, researchers have documented increased rates of false-positive and false-negative calls due to sequence data quality and selection of software during SNP analyses with reference sequences (50). Therefore, bioinformaticians should fully evaluate each application they use with regard to these common sources of error (in addition to any unique tendencies) and their ability to process data that contain common systematic errors. For example, errors may be introduced due to PCR amplification bias that arises from DNA library construction methodologies and to GC bias due to sequencing chemistries (64). It is also known that the selection of evolutionarily distant reference genome sequences results in both false-positive and false-negative calls during SNP detection, which influences subsequent phylogenetic analyses (50, 54). Therefore, analysts need to be aware of how the use of closely and distantly related genome sequences for references may influence the results of analyses, and have written procedures in place to guide the proper selection of reference genome sequences. Similarly, computational steps within the bioinformatics workflow (e.g., quality filtering and trimming of reads before analysis, removal of duplicate reads, and realignment of reads around indels) are known to influence the results of analyses (58, 65–67). The rationale for using each of these steps and their effects on results of analyses and interpretation of the data should be well documented. In regulatory food safety communities, the assessment and documentation of all aspects of bioinformatics pipelines and their impact on the interpretation of results are of utmost importance.

Data Quality Assessment Parameters and Traceability Information

A test record should capture all salient sample information, such as the sample metadata and the information that allows quality assessment of the sequencing run itself (e.g., read lengths, number of reads, number of clusters, indices used, and sequencer-generated statistics). When a QC sample is processed alongside an investigative isolate, documentation regarding its production, quantification, and storage should be kept (*see Good Laboratory Practices and Prevention of Procedural Errors*). A record of genomic analyses performed should include the date, time, and identity of the analyst(s); the analytical pipeline used and its version (allowing back-tracing of all software components and versions; *see Data Management*); and the quality assessment parameters for the bioinformatics analysis. The latter may vary according to the functions performed by the pipeline. For instance, the report from an assembly pipeline

could include well-known assembly metrics and information facilitating quality evaluation, e.g., depth of coverage, total length (number of nucleotides assembled in "contigs"), expected genome size, and the reference genome used (if applicable).

Legal Admissibility

NGS data and its interpretation may be challenged through a judicial system. In the United States, the standard for legal admissibility of scientific evidence arose from the Daubert decision (Daubert et al. v. Merrell Dow Pharmaceuticals, Inc., 1993; 68), which emphasized that the focus of judicial review must be predicated on the principles and methodology of the approach, not the conclusions generated. In this particular ruling, the U.S. Supreme Court provided guidance for trial judges considering the admissibility of expert scientific testimony. These include testability, error rate, peer review, standards, and widespread acceptance. Significance or weight of the forensic association is another consideration, distinct from admissibility (46). Therefore, the development of standards to encourage adoption and compliance with best practices for NGS data can play an important role in supporting the admissibility of testimony of findings based on this type of data.

Conclusions

The power of genomics resides in the tremendous amount of information it provides that allows different questions relevant to the analysis of pathogens to be addressed. As such, the best traditions of scientifically informed, risk-based approaches for the management of microbial hazards in the food supply are well supported. Genomics provides an unprecedented opportunity to determine salient features of foodborne bacterial isolates such as identity (e.g., species and serotype), risk assessment attributes (e.g., virulence profiles), molecular type (high-resolution SNP and multilocus sequence typing analyses) and "value-added" markers (e.g., antibiotic resistance profile), thus rendering highly informative test results to support risk assessment and management decisions. Although some analyses may follow a predefined protocol, other analyses may be more ad hoc in nature to suit a particular investigative scenario. For example, during the course of a food safety investigation, new questions may arise requiring the determination of unanticipated features to shed light on a particular aspect of the food safety concern at hand, such as the occurrence of a new virulence factor or other trait of public health significance (e.g., novel antimicrobial resistance gene). Genomics provides for routine, high-resolution characterization of isolates and for exploratory, case-based investigations, both of which provide valuable public health information over the short- and long-term timescales.

A key strength of WGS applied to bacterial isolates is its amenability to providing open data and analysis transparency, thus enabling retrospective scrutiny so that all parties affected by regulatory decisions can ascertain their basis and even make their own independent verifications. Genomic data constitute an important record for legacy purposes, supporting future investigations into new scientific problems. There is great benefit to be derived from long-term retention of genomic data for research and development (e.g., development of diagnostic reagents and therapeutic targets), trend analysis (e.g., strain matching through comparisons of historical and contemporary data), and the determination of specific attributes to gain a better understanding of public health events (e.g., prevalence of genes conferring resistance to important antibiotics). The availability of reliable genomic data lends itself to examining the properties of a given isolate informatically irrespective of remoteness from the original point of isolation, obviating the need to transfer the actual strain itself, thus avoiding potential traceability, biosecurity, and cost burdens. Therefore, it is imperative that appropriate measures be implemented in the food microbiology genomics laboratory to ensure transparency and reproducibility of analytical processes and to safeguard the integrity of the data to ensure its future usability. This would entail (*1*) the implementation of provisions at the originating laboratory for clear documentation of the key steps that underpin decisions, (*2*) the assurance of accessibility of data and the tools used in data generation and analysis, (*3*) the protection of data integrity and confidential information, and (*4*) the retention of important contextual information such as the metadata associated with the original sample.

Perhaps the full potential of this powerful technology can best be achieved by avoiding prescriptive approaches in which users are committed to the application of predetermined analytical routines, which are inherently limiting. Instead, a performance model could be adopted, when practicable, in which analysts could select the tools of their choice according to site-specific circumstances while demonstrating proficiency in meeting a common performance standard. It is our hope that the development of a consensus on best practices for the application of bioinformatics in support of regulatory food safety will contribute in a significant manner to the full realization of all that genomics technology has to offer in serving the needs of regulators seeking to protect consumers from preventable illnesses.

Acknowledgments

This work was undertaken under the auspices of the Bioinformatics Working Group of the Global Coalition for Regulatory Science Research. The authors are grateful to Martine Gauthier (Canadian Food Inspection Agency, Ottawa, ON) for helping to prepare this manuscript.

References

(1) Herschleb, J., Ananiev, G., & Schwartz, D.C. (2007) *Nat. Protoc.* **2**, 677–684. doi:10.1038/nprot.2007.94

(2) Peters, T. (2009) *Methods Mol. Biol.* **551**, 59–70. doi:10.1007/978-1-60327-999-4_6

(3) Elkins, C.A., Kotewicz, M.L., Jackson, S.A., Lacher, D.W., Abu-Ali, G.S., & Patel, I.R. (2013) *Food Addit. Contam. Part A* **30**, 1422–1436

(4) EFSA Panel on EFSA Biological Hazards (2014) *EFSA J.* **12**, 3784. doi:10.2903/j.efsa.2014.3784

(5) Tong, W., Ostroff, S., Blais, B., Silva, P., Dubuc, M., Healy, M., & Slikker, W. (2015) *Regul. Toxicol. Pharmacol.* **72**, 102–106. doi:10.1016/j.yrtph.2015.03.008

(6) Hoffmann, M., Luo, Y., Monday, S.R., Gonzales-Escalona, N., Ottesen, A.R., Muruvanda, T., Wang, C., Kastanis, G., Keys, C., Janies, D., Senturk, I.F., Catalyurek, U.V., Wang, H., Hammack, T.S., Wolfgang, W.J., Schoonmaker-Bopp, D., Chu, A., Myers, R., Haendiges, J., Evans, P.S., Meng, J., Strain, E.A., Allard, M.W., & Brown, E.W. (2015) *J. Infect. Dis.* **213**, 502–508. doi:10.1093/infdis/jiv297

(7) Gilmour, M.W., Graham, M., Van Domselaar, G., Tyler, S., Kent, H., Trout-Yakel, K.M., Larios, O., Allen, V., Lee, B., & Nadon, C. (2010) *BMC Genomics* **11**, 120. doi:10.1186/1471-2164-11-120

(8) Holch, A., Webb, K., Lukiancenko, O., Ussery, D., Rosenthal, B.M., & Gram, L. (2013) *Appl. Environ. Microbiol.* **79**, 2944–2951. doi:10.1128/AEM.03715-12

(9) Stasiewicz, M.J., Oliver, H.F., Wiedmann, M., & den Bakker, H.C. (2015) *Appl. Environ. Microbiol.* **81**, 6024–6037. doi:10.1128/AEM.01049-15

(10) Lambert, D., Carrillo, C.D., Koziol, A.G., Manninger, P., & Blais, B.W. (2015) *PLoS One* **10**, e0122928. doi:10.1371/journal.pone.0122928

(11) Amaral, G.R., Dias, G.M., Wellington-Oguri, M., Chimett, L., Campeão, M.E., & Thompson, C.C. (2014) *Int. J. Syst. Evol. Microbiol.* **64**, 357–365. doi:10.1099/ijs.0.057927-0

(12) Gordon, N.C., Price, J.R., Cole, K., Everitt, R., Morgan, M., Finney, J., Kearns, A.M., Pichon, B., Young, B., Wilson, D.J., Llewelyn, M.J., Paul, J., Peto, T.E., Crook, D.W., Walker, A.S., & Golubchik, T. (2014) *J. Clin. Microbiol.* **52**, 1182–1191. doi:10.1128/JCM.03117-13

(13) International Organization for Standardization/International Electrotechnical Commission (2005) ISO/IEC 17025:2005, General Requirements for the Competence of Testing and Calibration Laboratories, http://www.iso.org/iso/catalogue_detail.htm?csnumber=39883

(14) Ireland, J.D., & Møller, A. (2010) *Eur. J. Clin. Nutr.* **64**, S44–S48. doi:10.1038/ejcn.2010.209

(15) European Food Safety Authority (2015) EFSA Supporting Publications Technical Report. doi:10.2903/sp.efsa.2015.EN-804

(16) Field, D., & Sansone, S.A. (2006) *OMICS* **10**, 84–93. doi:10.1089/omi.2006.10.84

(17) Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J.R., Amaral-Zettler, L., Gilbert, J.A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., Birren, B.W., Blaser, M.J, Bonazzi, V., Booth, T., Bork, P., Bushman, F.D., Buttigieg, P.L., Chain, P.S., Charlson, E., Costello, E.K., Huot-Creasy, H., Dawyndt, P., DeSantis, T., Fierer N., Fuhrman, J.A., Gallery, R.E., Gevers, D., Gibbs, R.A., San Gil, I., Gonzalez, A., Gordon, J.I., Guralnick, R., Hankeln, W., Highlander, S., Hugenholtz, P., Jansson, J., Kau, A.L., Kelley, S.T., Kennedy, J., Knights, D., Koren, O., Kuczynski, J., Kyrpides, N., Larsen, R., Lauber, C.L., Legg, T., Ley, R.E., Lozupone, C.A., Ludwig, W., Lyons, D., Maguire, E., Methé, B.A., Meyer, F., Muegge, B., Nakielny, S., Nelson, K.E., Nemergut, D., Neufeld, J.D., Newbold, L.K., Oliver, A.E., Pace, N.R., Palanisamy, G., Peplies, J., Petrosino, J., Proctor, L., Pruesse, E., Quast, C., Raes, J., Ratnasingham, S., Ravel, J., Relman, D.A., Assunta-Sansone, S., Schloss, P.D., Schriml, L., Sinha, R., Smith, M.I., Sodergren, E., Spor, A., Stombaugh, J., Tiedje, J.M., Ward, D.V., Weinstock, G.M., Wendel, D., White, O., Whiteley, A., Wilke, A., Wortman, J.R., Yatsunenko, T., & Glöckner, F.O. (2011) *Nat. Biotechnol.* **29**, 415–420. doi:10.1038/nbt.1823

(18) Dugan, V.G., Emrich, S.J., Giraldo-Calderón, G.I., Harb, O.S., Newman, R.M., Pickett, B.E., Schriml, L.M., Stockwell, T.B., Stoeckert, C.J. Jr., Sullivan, D.E., Singh, I., Ward, D.V., Yao, A., Zheng, J., Barrett, T., Birren, B., Brinkac, L., Bruno, V.M., Caler, E., Chapman, S., Collins, F.H., Cuomo, C.A., Di Francesco, V., Durkin, S., Eppinger, M., Feldgarden, M., Fraser, C., Fricke, W.F., Giovanni, M., Henn, M.R., Hine, E., Hotopp, J.D., Karsch-Mizrachi, I., Kissinger, J.C., Lee, E.M., Mathur, P., Mongodin, E.F., Murphy, C.I., Myers, G., Neafsey, D.E., Nelson, K.E., Nierman, W.C., Puzak, J., Rasko, D., Roos, D.S., Sadzewicz, L., Silva, J.C., Sobral, B., Squires, R.B., Stevens, R.L., Tallon, L., Tettelin, H., Wentworth, D., White, O., Will, R., Wortman, J., Zhang, Y., & Scheuermann, R.H. (2014) *PLoS One* **9**, e99979. doi:10.1371/journal.pone.0099979

(19) Leebens-Mack, J., Vision, T., Brenner, E., Bowers, J.E., Cannon, S., Clement, M.J., Cunningham, C.W., dePamphilis, C., deSalle, R., Doyle, J.J., Eisen, J.A., Gu, X., Harshman, J., Jansen, R.K., Kellogg, E.A., Koonin, E.V., Mishler, B.D., Philippe, H., Pires, J.C., Qiu, Y.L., Rhee, S.Y., Sjölander, K., Soltis, D.E., Soltis, P.S., Stevenson, D.W., Wall, K., Warnow, T., & Zmasek, C. (2006) *OMICS* **10**, 231–237. doi:10.1089/omi.2006.10.231

(20) Taylor, C.F., Field, D., Sansone, S.A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C.A., Binz, P.A., Bogue, M., Booth, T., Brazma, A., Brinkman, R.R., Michael Clark, A., Deutsch, E.W., Fiehn, O., Fostel, J., Ghazal, P., Gibson, F., Gray, T., Grimes, G., Hancock, J.M., Hardy, N.W., Hermjakob, H., Julian, R.K. Jr., Kane, M., Kettner, C., Kinsinger, C., Kolker, E., Kuiper, M., Le Novère, N., Leebens-Mack, J., Lewis, S.E., Lord, P., Mallon, A.M., Marthandan, N., Masuya, H., McNally, R., Mehrle, A., Morrison, N., Orchard, S., Quackenbush, J., Reecy, J.M., Robertson, D.G., Rocca-Serra, P., Rodriguez, H., Rosenfelder, H., Santoyo-Lopez, J., Scheuermann, R.H., Schober, D., Smith, B., Snape, J., Stoeckert, C.J. Jr., Tipton, K., Sterk, P., Untergasser, A., Vandesompele, J., & Wiemann, S. (2008) *Nat. Biotechnol.* **26**, 889–896. doi:10.1038/nbt.1411

(21) Gruber, T.R. (1993) *Knowl. Acquis.* **5**, 199–220. doi:10.1006/knac.1993.1008

(22) Erfianto, B., Mahmood, A.K., & Rahman, A.S.A. (2007) 5th Student Conference on Research and Development, Selangor, Malaysia. doi:10.1109/SCORED.2007.4451440

(23) Ferreira, J.D., Paolotti, D., Couto, F.M., & Silva, M.J. (2013) *J. Epidemiol. Community Health* **67**, 385–388. doi:10.1136/jech-2012-201142

(24) Arp, R., Smith, B., & Spear, A.D. (2015) *Building Ontologies with Basic Formal Ontology*, The MIT Press Scholarship Online, Cambridge, MA, 43–84. doi:10.7551/mitpress/9780262527811.003.0003

(25) Paslaru, E., Simperl, B., & Tempich, C. (2006) in *On the Move to Meaningful Internet Systems 2006: COOPIS, DOA, GADA, and ODBASE*, Lecture Notes in Computer Science Vol. 4275, R. Meersman & Z. Tari (Eds), Springer-Verlag, Berlin, Germany, pp 836–854. doi:10.1007/11914853_51

(26) Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Consortium, O.B.I., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., & Lewis, S. (2007) *Nat. Biotechnol.* **25**, 1251–1255. doi:10.1038/nbt1346

(27) OBO Foundry, http://obofoundry.org/wiki/index.php/OBO_Foundry_Principles

(28) Ontobee, http://www.ontobee.org

(29) National Center for Biomedical Ontology BioPortal, http://bioportal.bioontology.org/

(30) Buttigieg, P.L., Morrison, N., Smith, B., Mungall, C.J., Lewis, S.E. & the ENVO Consortium (2013) *J. Biomed. Semantics* **4**, 43. doi:10.1186/2041-1480-4-43

(31) Cowell, L.G., & Smith, B. (2010) in *Infectious Disease Informatics*, V. Sintchenko (Ed.), Springer, New York, NY, pp 373–395

(32) Brinkman, R.R., Courtot, M., Derom, D., Fostel, J.M., He, Y., Lord, P., Malone, J., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Soldatova, L.N., Stoeckert, C.J. Jr., Turner, J.A., Zheng, J., & the OBI consortium (2010) *J. Biomed. Semantics* **1**, S7. doi:10.1186/2041-1480-1-S1-S7

(33) Genomic Epidemiology Ontology, GenEpiO, https://github.com/GenEpiO/genepio/

(34) Gargis, A.S., Kalman, L., Berry, M.W., Bick, D.P., Dimmock, D.P., Hambuch, T., Lu, F., Lyon, E., Voelkerding, K.V., Zehnbauer, B.A., Agarwala, R., Bennett, S.F., Chen, B., Chin, E.L., Compton, J.G., Das, S., Farkas, D.H., Ferber, M.J., Funke, B.H., Furtado, M.R., Ganova-Raeva, L.M., Geigenmüller, U., Gunselman, S.J., Hegde, M.R., Johnson, P.L., Kasarskis, A., Kulkarni, S., Lenk, T., Liu, C.S., Manion, M., Manolio, T.A., Mardis, E.R., Merker, J.D., Rajeevan, M.S., Reese, M.G., Rehm, H.L., Simen, B.B., Yeakley, J.M., Zook, J.M., & Lubin, I.M. (2012) *Nat. Biotechnol.* **30**, 1033–1036. doi:10.1038/nbt.2403

(35) Leipzig, J. *Brief. Bioinform.* 2016, 1–7. doi:10.1093/bib/bbw020

(36) Davis, S., Pettengill, J.B., Luo, Y., Payne, J., Shpuntoff, A., Rand, H., & Strain, E. (2015) *PeerJ Comp. Sci*. **1**, e20. https://doi.org/10.7717/peerj-cs.20

(37) Umarji, M., Seaman, C., Koru, A.G., & Liu, H. (2009) "Software Engineering Education for Bioinformatics," 22nd Conference on Software Engineering Education and Training, February 17–20, 2009, Hyderabad, Andhra Pradesh, India, pp 216–223. doi:10.1109/CSEET.2009.44

(38) Gogates, G.D. (2012) Software Validation in Accredited Laboratories: A Practical Guide, http://www.a2la.org/guidance/adequate_for_use.pdf

(39) Smith, D.R. (2014) *Brief. Bioinform.* **16**, 700–709. doi:10.1093/bib/bbu030

(40) Mesirov, J.P. (2010) *Science* **327**, 415–416. doi:10.1126/science.1179653

(41) Peng, R.D. (2011) *Science* **334**, 1226–1227. doi:10.1126/science.1213847

(42) Sandve, G.K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013) *PLOS Comput. Biol.* **9**, e1003285. doi:10.1371/journal.pcbi.1003285

(43) GitHub, http://www.github.com

(44) Leprevost F.d.V., Barbosa, V.C., Francisco, E.L., Perez-Riverol, Y., & Carvalho, P.C. (2014) *Front. Genet.* **5**, 199, 1–3. doi:10.3389/fgene.2014.00199

(45) Wilson, G., Aruliah, D.A., Brown, C.T., Chue Hong, N.P., Davis, M., Guy, R.T., Haddock, S.H., Huff, K.D., Mitchell, I.M., Plumbley, M.D., Waugh, B., White, E.P., & Wilson, P. (2014) *PLoS Biol.* **12**, e1001745. doi:10.1371/journal.pbio.1001745

(46) Wilson, M.R., Allard, M.W., & Brown, E.W. (2013) *Cladistics* **29**, 449–461. doi:10.1111/cla.12012

(47) Global Microbial Identifier, http://www.globalmicrobialidentifier.org/

(48) EURACHEM/CITAC Measurement Uncertainty Working Group (2012) Quantifying Uncertainty in Analytical Measurement, 3rd Ed., S.L.R.Ellison & A. Williams (Eds), http://www.eurachem.org/index.php/publicaitons/guides/quam

(49) Olson, N.D., Lund, S.P., Coman, R.E., Foster, J.T., Sahl, J.W., Schupp, J.M., Keim, P., Morrow, J.B., Salit, M.L., & Zook, J.M. (2015) *Front. Genet.* **6**, 235. doi:10.3389/fgene.2015.00235

(50) Pightling, A.W., Petronella, N., & Pagotto, F. (2014) *PLoS One* **9**, e104579. doi:10.1371/journal.pone.0104579

(51) McTavish, E.J., & Timme, R. (2015) TreeToReads, https://github.com/snacktavish/TreeToReads

(52) Huang, W., Li, L., Myers, J.R., & Marth, G.T. (2012) *Bioinformatics* **28**, 593–594. doi:10.1093/bioinformatics/btr708

(53) Caboche, S., Audebert, C., Lemoine, Y., & Hot, D. (2014) *BMC Genomics* **15**, 264. doi:10.1186/1471-2164-15-264

(54) Bertels, F., Silander, O.K., Pachkov, M., Rainey, P.B., & van Nimwegen, E. (2014) *Mol. Biol. Evol.* **31**, 1077–1088. doi:10.1093/molbev/msu088

(55) Allard, M.W., Luo, Y., Strain, E., Li, C., Keys, C.E., Son, I., Stones, R., Musser, S.M., & Brown, E.W. (2012) *BMC Genomics* **13**, 32. doi:10.1186/1471-2164-13-32

(56) Lienau, E.K., Strain, E., Wang, C., Zheng, J., Ottesen, A.R., Keys, C.E., Hammack, T.S., Musser, S.M., Brown, E.W., Allard, M.W., Cao, G., Meng, J., & Stones, R. (2011) *N. Engl. J. Med.* **364**, 981–982. doi:10.1056/NEJMc1100443

(57) Farrer, R.A., Henk, D.A., MacLean, D., Studholme, D.J., & Fisher, M.C. (2013) *Sci. Rep.* **3**, 1512. doi:10.1038/srep01512

(58) Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., & Nielsen, H. (2000) *Bioinformatics* **16**, 412–424. doi:10.1093/bioinformatics/16.5.412

(59) Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., & Spratt, B.G. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 3140–3145. doi:10.1073/pnas.95.6.3140

(60) Carrillo, C.D., Kruczkiewicz, P., Mutschall, S., Tudor, A., Clark, C., & Taboada, E.N. (2012) *Front. Cell. Infect. Microbiol.* **2**, 57. doi:10.3389/fcimb.2012.00057

(61) Read, T.D., & Massey, R.C. (2014) *Genome Med.* **6**, 109. doi:10.1186/s13073-014-0109-z

(62) Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., & Gu, Y. (2012) *BMC Genomics* **13**, 341. doi:10.1186/1471-2164-13-341

(63) Nielsen, R., Paul, J.S., Albrechtsen, A., & Song, Y.S. (2011) *Nat. Rev. Genet.* **12**, 443–451. doi:10.1038/nrg2986

(64) Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., & Jaffe, D.B. (2013) *Genome Biol.* **14**, R51. doi:10.1186/gb-2013-14-5-r51

(65) Dettman, J.R., Rodrigue, N., Melnyk, A.H., Wong, A., Bailey, S.F., & Kassen, R. (2012) *Mol. Ecol.* **21**, 2058–2077. doi:10.1111/j.1365-294X.2012.05484.x

(66) Liu, Q., Guo, Y., Li, J., Long, J., Zhang, B., & Styr, Y. (2012) *BMC Genomics* **13**, S12. doi:10.1186/1471-2164-13-S8-S8

(67) Van der Auwera, G.A., Carneiro, M.O., Harti, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., & DePristo, M.A. (2013) *Curr. Protoc. Bioinformatics* **43**, 1–33. doi:10.1002/0471250953.bi1110s43

(68) Daubert et al. v. Merrell Dow Pharmaceuticals, Inc., 1993, http://www.exify.com/daubertOne.htm