Chapter 2
Global Microbial Identifier

Peter R. Wielinga, Rene S. Hendriksen, Frank M. Aarestrup, Ole Lund, Saskia L. Smits, Marion P.G. Koopmans, and Jørgen Schlundt

Introduction

Human and animal populations increasingly share a number of emerging and re-emerging infections including infections that are exchanged between these populations (i.e. zoonotic infections) either directly or indirectly through food or vectors. Recent global outbreaks, such as SARS (Severe Acute Respiratory Syndrome), avian influenza (H5N1), pandemic (swine)influenza (H1N1) and MERS (Middle East Respiratory Syndrome) have rightfully received global attention, both in relation to the disease burden, the risk of rapid spread and the additional economic cost relative to travel and trade restrictions. To complete the picture of the disease burden and economic cost of human disease related to animals a number of endemic human infections that are continuously transferred from animals (e.g. salmonellosis, brucellosis, campylobacteriosis, rabies, cysticercosis) should also be considered. It is estimated that more than six out of every ten emerging infectious diseases in humans are

P.R. Wielinga • R.S. Hendriksen • F.M. Aarestrup • O. Lund
National Food Institute, Technical University of Denmark, Lyngby, Denmark
e-mail: peter.wielinga@gmail.com; rshe@food.dtu.dk; fmaa@food.dtu.dk; lund@cbs.dtu.dk

S.L. Smits
Department of Viroscience, Erasmus Medical Centre, Rotterdam, The Netherlands
e-mail: s.smits@erasmusmc.nl

M.P.G. Koopmans
Department of Viroscience, Erasmus Medical Centre, Rotterdam, The Netherlands

Virology Division, Centre for Infectious Diseases Research, Diagnostics and Screening, National Institute for Public Health and the Environment, Bilthoven, The Netherlands
e-mail: m.koopmans@erasmusmc.nl

J. Schlundt (✉)
Nanyang Technological University, Singapore, Singapore
e-mail: jschlundt@ntu.edu.sg

X. Deng et al. (eds.), *Applied Genomics of Foodborne Pathogens*, Food Microbiology and Food Safety, DOI 10.1007/978-3-319-43751-4_2

spread from animals [1]. A number of factors, including poverty, increasing population density, disruption of wildlife habitats, increased food trade and changes in food preservation and consumption habits have resulted in increased risks of contraction of infectious diseases and subsequently their potential global spread. Globally, about 23 % of all deaths are caused by infectious diseases, with the most significant burden in developing countries [2]. Nearly all of the most important human pathogens are either zoonotic or originated as zoonoses [3–6]. Striking examples include HIV/AIDS and Spanish influenza, which started by interspecies transmission of the causative agents [7–10] and have caused millions of deaths worldwide and more recently SARS and MERS coronaviruses and H1N1 and H5N1 influenza A viruses.

Detection and surveillance form the backbone of all systems currently used to control infectious diseases worldwide. However, surveillance is still typically targeted at a relatively limited number of specified diseases, and, maybe more importantly, there is a very significant global disparity in national disease detection systems and methodology. In particular, public health efforts and patient treatment are hampered by different obstacles: the use of different, specialized, expensive and difficult-to-compare detection techniques; a lack of collaboration between different microbiological fields; (inter)national politics on the disclosure of (patient) information and research data; intellectual property rights; and, a lack of sufficient diagnostic capacities particularly in developing countries. A more effective and rational approach to the prevention of microbial threats is essential at the global level. Efforts to mitigate the effects of infectious threats, focusing on improved surveillance and diagnostic capabilities, are crucial [11]. With recent technological advances and declining costs in the next generation sequencing field, these tools will play an increasingly important role in the surveillance and identification of new and previously unrecognized pathogens in both animals and humans but also for identification and characterization of traditional pathogens. Inherently an enormous increase in microbial whole genome sequences (WGS) is to be expected, providing a wealth of information to aggregate, share, mine and use to address global public health and clinical challenges. The goals of the Global Microbial Identifier (GMI) initiative in this respect will be outlined below.

Next Generation Sequencing and Whole Genome Sequencing: A New Potential for Integrated Surveillance of Infectious Diseases

Surveillance is a key component of preparedness for infectious diseases, and is done globally to monitor trends in endemic diseases (e.g. influenza, dengue, salmonellosis), to monitor eradication efforts (polio, measles, brucellosis), or to signal unusual disease activities. Molecular diagnostic tools, which rely on the recognition of short pieces of unique genome sequence (e.g. PCR and microarray (biochip) technologies) and provide sensitive and specific detection and sufficient genetic diversity for subtyping, are used routinely in clinical diagnostic and surveillance settings. Although

the partial genome information, such as epidemiological markers, often is sufficient for patient management and basic surveillance objectives, from a public health perspective the increasing capacity for more extensive sequencing most likely will increase the depth of information gathered on pathogens and disease. Recombination and reassortment of viral genomes for instance may generate future threats; influenza A viruses for example are able to undergo reassortment if a single cell is concurrently infected with more than one virus [12]. These reassortment events can dramatically change the evolution of influenza A viruses in a certain host and lead to new epidemics and pandemics. Such events may easily be missed when surveillance is relying on molecular diagnostic tools that target small microbial genome fragments.

Whole genome sequencing (WGS) is a laboratory process that determines the complete genome sequence of an organism under study providing significantly more information than routine molecular diagnostic tools. This can have important implications; for instance during the recent outbreak of MERS coronavirus in the Middle East, analysis of small genome fragments did not provide sufficient phylogenetic signal for reliable typing of virus variants [13]. Classically, whole microbial genome sequences were determined by PCR and Sanger sequencing. Nowadays next generation sequencing (NGS) techniques are used increasingly in the human medical sciences, and are now also widely used to identify and genotype microorganisms in almost any microbial setting [14–17]. There are different NGS techniques targeting single microorganisms or a complete metagenome in a sample through methods unrelated to specific sequence recognition.

A cascade of technological NGS advancements both in the analytical sequencing field (e.g. pyro- and nanopore sequencing) and in the information technological (IT) field (e.g. increasingly faster and cheaper internet, computing rates and storage capacities; and the development of NGS software tools) has decreased the cost of WGS much faster than predicted 10 years ago (Fig. 2.1). Today, the actual cost of sequencing an average bacterial genome of 5 Mb would in practise cost between USD 50–100. It is estimated that both the price and the speed of WGS analyses will decrease to a point where it can seriously compete with traditional routine diagnostic identification techniques. The enormous potential of WGS in the surveillance of infectious diseases [18,19] has been demonstrated in many studies now including the tracking and tracing of the cholera outbreak in Haiti in 2010 [20], the Enterohaemorrhagic *Escherichia coli* (EHEC) outbreak starting in Germany in 2011 [21] and others e.g., [22,23]. During the EHEC outbreak, scientists from around the globe performed NGS and shared their results for analysis. The collaboration between these researchers allowed for joint and rapid analysis of the genomic sequences, revealing important details about the involved new strain of *E. coli*, including why it demonstrated such high virulence. Similar collaborations exist globally during emerging viral infections such as MERS coronavirus. Continuing innovations, however, are required to allow NGS techniques to become standard in clinical practice. In addition, hurdles regarding ethical, legal, social and societal issues need to be overcome.

It seems certain that NGS techniques will play an increasingly important role in the identification of new and previously unrecognized pathogens and inherently a large increase in the total amount of microbial whole genome sequences is to be expected.

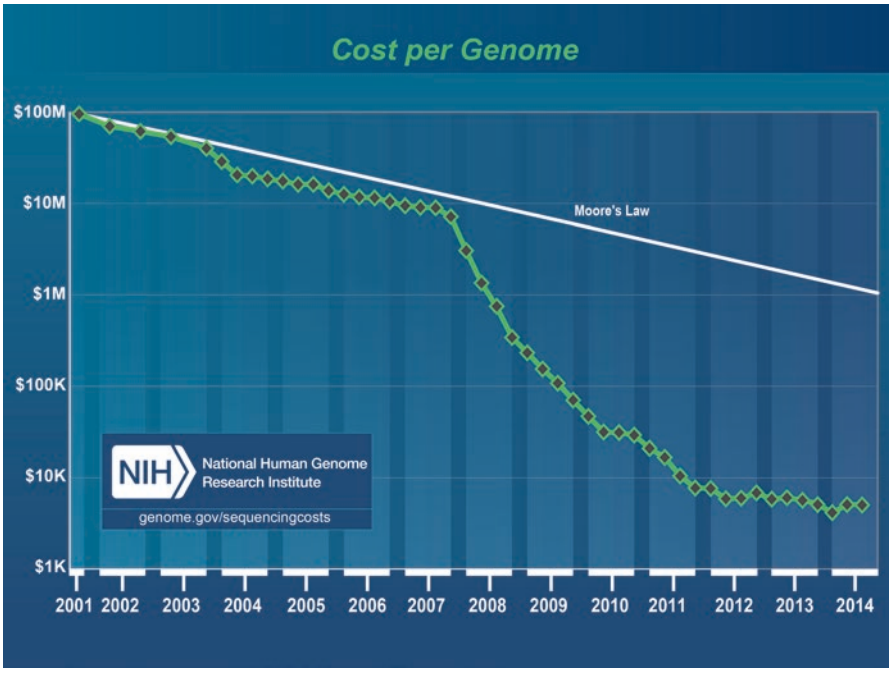

Fig. 2.1 NGS cost per raw magabase of DNA sequence. Taken from the National Human Genome Research Institute (http://www.genome.gov/sequencingcosts/)

As a consequence of the steadily decreasing costs of WGS, an increasing number of microbiological laboratories have embarked on WGS projects to characterize own stocks of infectious agents in their existing biobanks. This in turn generates huge amounts of genomic data in private databases as well as significantly increased numbers of genomes to the global DNA databases such as GenBank. This genomic information is, however, not fully interconnected and in most cases not accompanied with sufficient (national or international) metadata. The need to integrate these databases and to harmonize data collection has been generally recognized by the scientific community for some time [24]. Further integration of these databases and linking the genomic data to metadata for optimal prevention of infectious diseases, and to make it fit for other uses including routine diagnostics, is the new challenge.

Notably, while future use of WGS is likely to boom in developed countries, an even more dramatic change in developing countries creates a potential for a significant diagnostic leap-frog in these countries. While current diagnostic methods are diverse and require a lot of specialized training, NGS holds the potential of a simple one-size-fits-all tool for diagnosis of all infectious diseases, thereby dramatically improving public health in developing countries. At a systemic level, the use of NGS will enable uniform laboratory-, reporting- and surveillance-systems not only relative to human health, but reaching out to the identification of microorganisms in all other habitats, including animals and the environment: a true 'One Health' approach [25]. At the same time the development of new centralized and de-centralized diagnostic systems

will be significantly simplified with the potential of real-time characterization of microorganisms in individual, local decentralized labs with sequencers and internet link-up. Recent studies have shown that it is possible to determine the species, type as well as the antimicrobial/antiviral susceptibility of both bacterial and viral pathogens, even when using sequencing directly on clinical samples [18, 26]. This would be even more valuable for clinical laboratories in developing countries that do not currently have the same diagnostic capacities as most developed countries.

As NGS technology spreads more globally, there is an obvious potential to develop a global system of whole microbial genome databases to aggregate, share, mine and use microbiological genomic data, to address global public health and clinical challenges, and most importantly to identify and diagnose infectious diseases. Such a system should be deployed in a manner which promotes equity in access and use of the current technology worldwide, enabling cost-effective improvements in plant, animal, environmental and human health. If the system is set up in an 'open access' format it would likely enable comprehensive utility of NGS in developing countries, since the development of open databases and relevant algorithm platforms at the global level would enable immediate translation of sequence data to microbial identity and antimicrobial resistance pattern. In general, it is necessary to have a comprehensive database of all known microbial DNA sequences to make full use of locally derived DNA sequence to identify and characterize your isolate microbiologically and epidemiologically. A global system, supported by an internationally agreed format and governance system, will benefit those tackling individual problems at the frontline (clinicians, veterinarian, epidemiologists, etc.) as well as other stakeholders (i.e. policy-makers, regulators, industry, etc.). By enabling access to this global resource, a professional response on health threats will be within reach of all countries with (even relatively simple) basic laboratory infrastructure.

The Global Microbial Identifier (GMI) Initiative

The GMI initiative attempts a description of the landscape and opportunities of the global NGS/WGS field and suggests a collaborative effort to bring together different microbiological fields with the purpose of creating a global microbial identifier (GMI) tool on the basis of WGS data. To achieve this, GMI envisions a WGS database and analytic tools that are used and maintained by multidisciplinary researchers, clinical microbiologists, food scientists, (bio)informaticians, veterinarians, physicians, and other stakeholders. This database should be useful for basic research and for identification and disease diagnosis of any possible microorganism. In September 2011 the first international GMI conference was organised in Brussels[1] to discuss the possibility to use WGS as a microbiological diagnostic tool on a global scale [27]. At

[1] Perspectives of a global, real-time microbiological genomic identification system—implications for national and global detection and control of infectious diseases. Consensus report of an expert meeting 1–2 September 2011, Bruxelles, Belgium. Available at http://www.globalmicrobialidentifier.org.

this stage several preconditions for a successful initiation of an initiative of this sort seemed to have been met: (1) WGS had become mature and a potential serious alternative for other genotyping techniques, (2) the price of WGS had been falling dramatically and was now in some cases below the price of traditional methods, (3) vast amounts of IT resources and a fast internet had become available in most parts of the world, and (4) suggestions had been made that a One Health (human/animal) approach could enable improved control of infectious diseases [28].

Currently, GMI organizes annual meetings to discuss progress and future development. These meetings are organised and attended by a number of scientists and policy makers from around the world, including the World Health Organization (WHO), the UN Food and Agricultural Organization (FAO), the World Organisation for Animal Health (OIE), the United States Food and Drug Administration (US FDA), the European Commission (EC), the United States Centers for Disease Control and Prevention (CDC), the European Centre for Disease Prevention and Control (ECDC), the National Food Institute of Denmark, the European Food Safety Authority (EFSA) and several other universities, food research institutes and public health institutions. The general conclusion of the first meeting was that the spread of the WGS technology for microorganisms should be linked to the establishment of a global genomic database for microorganisms. This would entail an interactive, global, open source database supported by scientists from all regions of the world and from all fields, including human health, animal health, food safety and environmental health, and holding information on bacteria, viruses, fungi as well as parasites, together with important metadata relating to host information, environmental factors, sequencing methods, and other microbiological and epidemiological details. The structure and platform of the database(s) should be such that it could be used by different software tools (algorithms etc.) to generate meaningful results from data in the database.

Landscaping the Global Microbial WGS Field

The current steep rise in the potential of NGS has led to several developments around the globe: new fields of science have been strengthened (e.g. bioinformatics and its subfields); established scientific fields utilize NGS in novel ways; new WGS software tools are put online every week; multiple companies offer NGS and WGS equipment and services; and also at governmental level, NGS is considered in the continuous quest for public health efficiency improvements. These developments make NGS grow from a basic research tool into a mature general purpose tool; however, the constant danger is that cross-talk between these separate initiatives wanes in typical silo-fashion and that all technical development takes place in the western world (+ China), which might lead to a strong underuse of the total WGS potential.

While many researchers, clinicians as well as public and animal health professionals have made statements in support of the dramatic new potential, there currently is still no coherent description of the global (diagnostic) landscape of WGS and how it could best take over from the traditional techniques, as well as the potential benefits and costs of such development at a global scale. At the same time it

should be realized that the free sharing of genomic data will meet significant obstacles, both from the research, the public health—and the food production communities. Important examples, which can already be envisioned, are: (a) the general reluctance of researchers to share data before publication, (b) the reluctance of governments and institutions to share data when competing interests are in play (e.g. trade, tourism etc.), (c) legal and ethical issues including personal information confidentiality and intellectual property rights [29–31].

There is a need to further analyze this landscape. Such analyses should include identifying all stakeholders and their use of NGS, describing the technical and political needs, characterizing the potential future clinical and public health systems enabled by WGS, and in the process, specifically considering the need for capacity building in this area for developing countries [28]. A number of different scientific fields should be included in the analysis (e.g. public health, food safety and production, animal health, environmental health, bioinformatics, clinical science, biotechnology etc.). Likewise different societal sectors should be considered (e.g. healthcare, food and healthcare industry, agriculture, commerce, as well as developmental economics etc.). A description of existing WGS initiatives within different microbiological specializations (virology, bacteriology, parasitology etc.) will be key to understanding this field, as will be a thorough description of existing and future NGS potential in laboratory settings in developing countries.

GMI the Network

Following the inception of GMI in 2011, GMI has grown as a global network of scientists and other experts committed to improving global infectious disease and food safety prevention using WGS. A charter has been drawn up in which the network partners have agreed on its mission and vision (http://www.globalmicrobialidentifier.org/). In short, the mission is to build a global network for microbiological identification and infectious disease surveillance using an open and interactive worldwide network of databases for standardized identification, characterization and comparison of microorganisms through whole genome sequences of microorganisms. GMI's vision is a world where high quality microbiological genomic information from human, food, animal and plant domains is shared globally to improve public health, healthcare, a healthy environment and safer food for all.

The GMI network essentially is a global network of stakeholders that take part in shaping how the database and its supporting structures can best be defined, set up and used. Figure 2.2 shows a simplified impression of GMI: the GMI users, the database(s), the GMI software pipelines and other analytical tools, and the GMI organization. The users include anybody using the GMI database such as medical and veterinary labs, physicians and veterinarians, public health institutes, food science and industry etc. The GMI database is defined as all the microbiological WGS data and the linked metadata that can be accessed by GMI software. GMI software includes any software tool or software pipeline designed to interact with the GMI database to produce

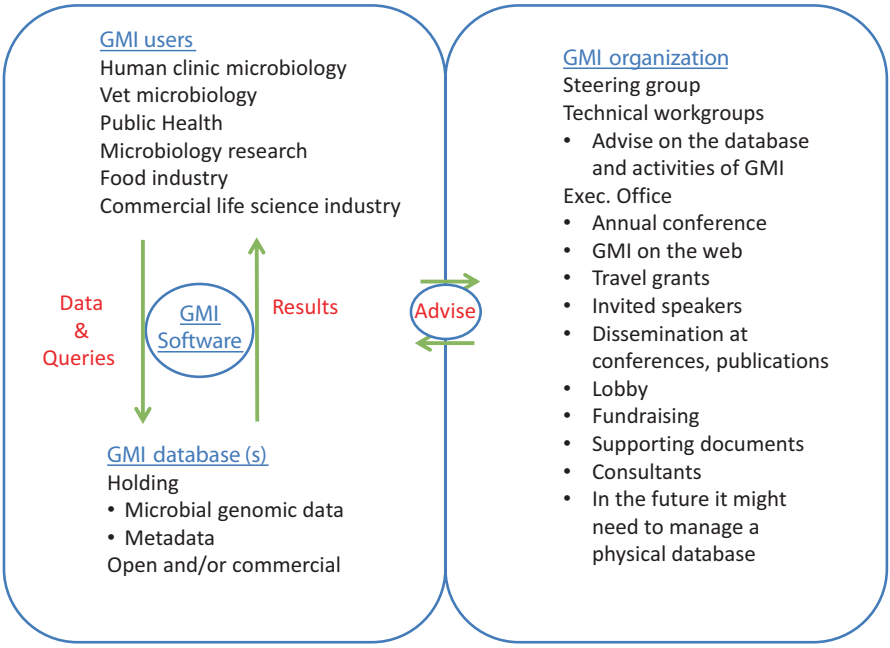

Fig. 2.2 Schematic outline of GMI

results, e.g., genome assembly, data comparison, disease diagnosis, resistance prediction, simple data extraction, data viewing etc. The GMI organization includes the people creating the database(s), people helping the development of the necessary software, and people active in the GMI working groups and steering committee.

GMI is now a global initiative with a defining Charter, annual conferences, a website and regular newsletters.[2] The eighth global GMI conference (GMI8) was held in May 2015 in Beijing, China, and GMI9 took place in FAO in Rome, Italy in May 2016. GMI is organized through a Steering Committee overseeing four working groups and supported by an executive office. The four working groups are: (1) Political challenges, outreach and building a global network; (2) Repository and storage of sequence and meta-data; (3) Analytical hard- & software and (4) Ring trials and quality assurance.

GMI the Database

The proposed GMI database will consist of all the microbiological WGS data, both annotated (including reference strains) and un-annotated, together with the relevant metadata, all to be accessed by GMI software. Questions related to the status,

[2] Homepage: http://www.globalmicrobialidentifier.org/.

separation and encryption of metadata within the database system need further consideration, including international and national political debate. While the ultimate aim might be one database or a federated database system[3] enabling fast identification through the comparison of a new isolate with many existing reference genomes worldwide, this system could be too complicated initially. Given the right tools, however, this technical complication may be neutralized and a federated system may even allow faster identification by parallel computing. The likely and preferable development in this area will depend on many other factors, including the available software and the state of global internet infrastructure.

A global reference database may be supported by additional database(s) to do the follow-up analysis after a first identification has been achieved, and these databases could potentially be located elsewhere. Considering the technical challenge of complete genome assembly, it becomes important to consider which level of (un-assembled) rough data can be input for assembly and analysis with GMI tools? This issue will potentially disappear when more powerful software is developed. Currently, however, these issues are still bottlenecks when quick turnaround of data analysis is demanded.

Compared to a federated system, centralized storage has several advantages. It will be a one stop shop and its openness may be preserved by the government(s) supporting the database. The creation of a centralized system would not prevent the future addition of regional/local databases to the structure to create a federated database system, which may potentially become necessary anyway if the future amount of data becomes too large for a single location. Such additional, federated databases may also be commercial, and this may hold both risks as well as advantages. The key will most likely be that the software adding and retrieving data can reach all relevant information. This means that either the software needs to handle multiple formats used by different databases, or the data structures of different databases should be similar. In addition, commercial databases should address how they can be accessed by GMI software and how users pay for their database use. Clearly this involves many controversial issues. Commercial involvement may on the one hand put limitations on the development of the GMI database and the speed at which it will evolve. On the other hand, in economic terms it will have a great spin-off in terms of companies that may offer services to and depend upon the GMI database, in a way somewhat similar to the functionality of the internet at present. Such spin-off activity may be beneficial for the quality and quantity of the use of the database. Taken together these are all important issues that GMI aims to discuss and solve through the work of its different work groups and through open discussion and interaction with all stakeholders in the field.

GMI the Software

In addition to the database, a proper functioning GMI system needs software. This software could be located as part of the database in a way similar to software offered by NCBI and linked to GenBank, such as BLAST, and other parties may also offer

[3] A system in which several databases appear to function as a single entity.

software that uses the data from the GMI database to generate comparisons and analyses that the GMI community asks for. An example of this could be the ResFinder software from the Center for Genomic Epidemiology at the Technical University of Denmark (see Table 2.2), which can be used to predict antibiotic resistance profiles from WGS data [18]. On the internet there is a wealth of such tools available and new bioinformatics tools are constantly released, either under an open source license or as commercial software packages or services. The current list of tools is very long and includes many different packages able to perform many different analyses. Unfortunately, there is a lack of coordination and awareness among developers and users. Some tools have been a repetition of already developed tools; some have overlapping analyses; and some are simply outdated already when they occur. On the other hand it is a welcome development that the bioinformatics community is flourishing with an abundance of tools, and GMI could take up the task of providing a portal to help users navigate among tools. Table 2.1 provides a list of several of these tools and links to webpages important for the field of WGS microbiology.

The whole field of "analytical tools" is currently developing fast and there are many different and new initiatives. Ideally, there is a need for simplicity and some of the individual tools developed are being sequentially combined into analytic pipelines. However, there is still much effort needed in this field, because not all programs are compatible with each other, some are not user friendly, are not maintained or are only available on specific platforms. GMI work groups 3 and 4 have taken initiative to investigate what is available and what would be necessary to have for a GMI database to function as a general diagnostic tool. Further advances in the software tools should aim at answering specific questions from the different fields of microbiology. Important advances in this area will be to generate more user friendly software to take the tools that now mainly are geared towards the bioinformatics and basic science communities, to the first-line users (clinicians and public health and food safety professionals) e.g. to help the clinical field with disease diagnosis or to help with complicated global tracking and tracing analyses relative to food contamination or infectious diseases. User friendliness would increase the use of the GMI database and thereby its value. It may be envisioned that this may come through the combination of apps and online software tools generated for use on (super) computers down to smart phones. In addition, bringing together different software routines that currently need to be run separately, will contribute to this. The chapter on comparative genomic epidemiology (CGE) elsewhere in the book gives an extensive overview and discussion of CGE tools for WGS microbiology.

Metadata and Depth of Analysis

Metadata is data that describes other data, and in many cases represent data that are necessary to make epidemiological sense of WGS data. Metadata relative to the sequence data of a clinical isolate in the database would for instance be patient demographics, geographical location and method of isolation etc. The more details there

Table 2.1 Short overview of some of the WGS analysis tools found on the internet

Tool	Link	Short description
Online analysis tools	http://molbiol-tools.ca/	Lists numerous bioinformatics tools
VFDB	http://www.mgc.ac.cn/VFs/	This database provides BLAST-based identification of virulence genes in 26 genera of bacterial pathogens. The database aims at being the most comprehensive database of virulence factors and hence also contains, for instance, hypothetical proteins
ResFinder	http://cge.cbs.dtu.dk/services/ResFinder/	ResFinder identifies acquired antimicrobial resistance genes in total or partial sequenced isolates of bacteria
ARDB	http://ardb.cbcb.umd.edu/	A manually curated database (ARDB) unifying most of the publicly available resistance genes and related information. Regular BLAST and RPS-BLAST tools would help the user to identify and annotate new potential resistance genes by blasting against ARDB DNA or protein sequences. Has not been maintained since 2009
BTXpred	http://www.imtech.res.in/raghava/btxpred/	The BTXpred server aims at predicting whether an amino acid sequence is a bacterial toxin or not, whether it is an endo- or exotoxin, and the function of exotoxins. It requires amino acid sequences as input
RASTA-Bacteria	http://genoweb1.irisa.fr/duals/RASTA-Bacteria/	RASTA-Bacteria is aimed at the identification of TA modules (toxins/antitoxin modules)
The comprehensive antibiotic resistance database	http://arpcard.mcmaster.ca/	The RGI provides automated annotation of your DNA sequence(s) based upon the data available in CARD, providing prediction of antibiotic resistance genes
t3db	http://www.t3db.org/	t3db is a database containing toxins and targets along with detailed information collected from various sources. It does not focus solely on bacterial virulence factors, but includes pollutant, pesticides, and drugs. Also, it is very strict with the inclusion of toxins and only includes toxins for which the structure is known
DBETH	http://www.hpppi.iicb.res.in/btox/	DBETH is a database of bacterial exotoxins for humans. As it requires amino acid sequences as input
VICMpred	http://imtech.res.in/raghava/vicmpred/	VICMpred is an SVM-based method for prediction of toxins (and other functional proteins) based on amino acid sequence

(continued)

Table 2.1 (continued)

Tool	Link	Short description
Samtools	http://samtools.sourceforge.net/	SAM Tools provides various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format
Figtree	http://tree.bio.ed.ac.uk/software/figtree/	Tree viewer
Velvet (combined with VelvetOptimiser)	http://bioinformatics.net.au/software.velvetoptimiser.shtml	*de novo* assembler
BWA	http://bio-bwa.sourceforge.net/	Sequence mapper
AdapterRemoval	https://github.com/slindgreen/AdapterRemoval	Trimming and adapter removal from raw read data

are in the metadata the more detailed the tracking and tracing of microorganism can be. However, a higher level of detail can also result in political and/or privacy/ethical complications, especially for the people publishing the data [29]. Without metadata, one would have a 'genotype' database only containing peta- to exabytes of WGS data. This would already be a giant step for mankind as we will discover many new genomes and microbial communities. However, to use genomes for infectious disease investigation and epidemiology, metadata are essential. The list and structure of metadata should be concise and include only what is defined as essential while excluding redundant or unethical information. For instance, making a distinction between men and women, children and the elderly would be very informative and may be essential for clinical data. However, it would be under discussion whether to include race in the list of metadata, even though there might be situations imaginable for which having such metadata would help solve scientific questions. Also, different fields of research or policy making may have use for different metadata. For instance, economists studying the economic cost of a certain disease will be interested in the number of outbreaks and relations between different economically important sectors e.g. specific food or food preparing sectors, while public health specialists and clinicians might be more interested in resistance phenotypes, treatment options etc.

In general, it is thus essential to generate a list of metadata that can be considered essential for each sample. In addition, per discipline this list may be extended with field specific metadata which are essential for each individual field. Also, there should be a list of metadata to be avoided by GMI. Such thinking would bring us roughly three lists of metadata: the minimal essential, the field specific list, and the list of metadata to be excluded. To help the discussion on this one may categorize each of these three lists further into essential and optional data. Table 2.2 gives a very basic and simplified example of how such lists might look like to help the discussions on what these lists should finally comprise. Generating the different field specific list in a collaborative manner as done in GMI will potentially be beneficial

Table 2.2 Example of types of metadata that may be valuable for the GMI database

General metadata	Examples work field specific metadata		Not essential for GMI
	Clinical examples	Food examples	
Essential			
Submitters contact info Submitters identifier Unique GMI identifier Name organism Name strain Alias(es) Date isolated Attribute package (if pathogenic type of pathogen) Isolation source Cultured Yes/No Geographical origin of the sample (Country and City) Lab strain Yes/No Reference strain Yes/No	Host (human/animal) Host sex Host age range Host age Name(s) disease Zoonotic Y/N Resistance profile Treatment options Confirmation tests Confirmation tests results Short (standard) case description Outbreak Y/N, plus code	Specific name source GPS location source Climate type source Location in source depth Zoonotic Y/N Resistance profile Virulence Confirmation tests Confirmation tests results Outbreak Y/N, plus code Commercial source Y/N Producer (kept secret?) Short case description	Race host HIV status host Links to hosts social network profile
Optional			
Colony color Description of the sample/source/strain Detailed geographical origin of the sample (GPS coordinates etc) Growth rate Growth on cell lines	Outbreak code Growth rate Growth on cell lines PFGE pattern codes Serovar Outcome other tests Suggestion for further testing Short (standard) case description	Host (human/animal) Host Sex Host age range Host age Name(s) disease PFGE pattern codes Serovar Growth rate Colony color Growth on cell lines Suggestions for further testing Short (standard) case description	Occupation of host Use in terror attack

These are examples of metadata that different microbiological fields would like to collect and serves to illustrate the concept. Further discussion in GMI will be needed to generate agreed lists of metadata required for each microbiology field, and as the technology progress these list may have to be updated

for the end results, since different sectors would be able to follow each other's progress and may minimize redundancy in the different lists. Furthermore, it should be decided which data should be collected but kept confidential and only accessible by the submitter and others with permission of the submitters. For instance, should it be open source information if a specific producer is linked to a specific microorganism and/or outbreak, or should such data be managed (and kept secret or open) by the relevant regulatory agencies?

There are a number of technical questions adding to the complexity of the issue. Should reference strain data have a different set of metadata than the metadata required when submitting and/or comparing one's samples to the GMI database? And, where to best store all the metadata? The different types of metadata originating from different fields might be centrally stored which will have advantages for retrieving and working with them, and for ensuring the open source character of the database. However, this is not necessary and by using the right identifiers, different metadata databases may be generated that connect to a central WGS database. Central storage of all kinds of metadata may bring the advantage of having a one stop shop for everything, and it may help to find new cross links between data from different fields. It may, however, also lead to confusion when users are overloaded with too much information that is not necessary for their purpose.

Quality Assurance and Testing

Investigating whether GMI users will be able to perform DNA extraction, library preparation, the actual sequencing, the assembly and phylogenetic analysis following different laboratory protocols, software tools, and sequence platforms will enable an evaluation of the reliability of submitted sequence data to a GMI database [32]. GMI aims to assist laboratories and partners globally to perform NGS to the highest quality level, and to prepare for this GMI in 2013 conducted a survey to identify the intended end-users, priority organisms, and quality markers for proficiency testing [33]. GMI in 2014 performed a pilot proficiency test with a limited number of laboratories to test the developed IT system and corresponding protocol. The GMI 2015 proficiency test was fully rolled out by December 15 (supported by the EU/COMPARE programme (www.compare-europe.eu) and the USFDA GenomeTrakr and Microbiologics®. This first global proficiency test in this area had a focus on *Salmonella enterica*, *Escherichia coli* and *Staphylococcus aureus*, and allowed for sign-up for each species separately (see www.globalmicrobialidentifier.org). 55 laboratories, from all continents, signed up for the test. The main objective of this proficiency test was to assess the feasibility of achieving reliable laboratory results of consistently good quality within the area of DNA preparation, sequencing, and analysis (e.g. for the use relative to phylogeny, MLST, resistance genes etc.). This will in time ensure or enable harmonization and standardization of whole genome sequencing and data analysis, with the final aim to produce comparable data for the GMI initiative, and thereby consistent data for the GMI database. A further objective is to assess and improve the data uploaded to databases such as NCBI, EBI and DDBJ. Therefore, the laboratory analysis work performed for this type of proficiency testing should be done employing the methods routinely used in the individual laboratories. The proficiency testing performed in this area has consisted of two wet-lab and one dry-lab components targeting *Salmonella*, *E. coli* and *S. aureus*. The wet-lab components assess the laboratories' ability to perform DNA preparation, sequencing procedures and analysis of epidemiological markers whereas the dry component assesses the ability to analyse a whole-genome-sequencing

dataset and distinguish between clonally related and sporadic genomes. At present (September 2016) the GMI2016 proficiency testing is ongoing. The future vision of the proficiency testing is to target lower priority bacterial pathogens as well as to develop a parallel proficiency testing regime targeting viruses (http://www.globalmicrobialidentifier.org/News-and-Events/Previous-meetings/7th-Meeting-on-GMI).

Other laboratory methods are being discussed and optimized for use in GMI, including consideration of how to include other types of microorganisms in proficiency tests and how to initiate parallel viral pilot proficiency test schemes including RNA methods.

Next to quality assurance at the laboratory level it will be important to have a reliable source of analytical tools that cover the different tasks requested by the GMI users and are of sufficient quality to be used in different (sometimes more sensitive) settings than basic research, for instance in clinical settings. GMI aims to define the functional requirements for these tools from the perspective of end-users (clinical, public health, research) in terms of applications needed (identification, outbreak detection etc.) and priority microorganisms and diseases. To do so, GMI maps the currently available analytical software tools as well as developments in the field and benchmark them against the needs of GMI end-users in order to identify implementation gaps and projects that may fill those gaps. By this mapping effort and through software testing, GMI aims to construct a central portal of tools, to indicate a quality level, and state the usefulness and the user friendliness of the different tools for the different GMI end-users. Through this effort it will be possible to provide guidance for further development of (new) analytical tools.

In addition to the development of these testing schemes, which will get a more permanent shape and place in the future GMI network, GMI plans to design *in silico* pilots using realistic scenarios based on and using data from a previous infectious disease outbreak or another event (http://www.globalmicrobialidentifier.org/News-and-Events/Previous-meetings/7th-Meeting-on-GMI). The goal of these pilots will be to help shape the process and the form that the GMI tools take, develop training skills and increase the participation level. In particular, this would be important to increase the participation level of members that currently lack the necessary laboratory capacity including members from many developing countries. These pilots will address several important issues and may help answer some important questions such as: How well does data transfer work? How well does data analysis, including species identification and outbreak clustering, work? What are the biggest challenges for coordinating an analysis that is highly dependent on metadata? What are the minimum standards required to run the system? And finally, what might be the gain in turn-around time?

Concluding Remarks

Several already existing internet-based genomic tools and databases have been presented and discussed at GMI global meetings—all of which are generally aimed at improving (inter)national detection and identification of different types

of microorganisms. For instance, the global programme PulseNet compares the PFGE 'DNA fingerprints' of bacteria from patients to find clusters of disease that might represent unrecognized outbreaks (http://www.cdc.gov/pulsenet/). MLSTnet can be used to compare various bacteria on the basis of multilocus sequence typing (MLST) analysis (http://www.mlst.net/). EuPathDB (http://eupathdb.org/eupathdb/) and ZoopNet [34] are portals for accessing genomic-scale and MLST datasets, respectively, which are associated with eukaryotic parasites. And NoroNet is a network of public health institutes and universities sharing virological, epidemiological and molecular data on norovirus and includes a tool for Norovirus identification and epidemiology on the basis of sequence comparison (http://www.rivm.nl/en/Topics/N/NoroNet). In contrast to GMI these earlier networks had to focus their effort on a single technique and often a limited group of microorganisms to make comparisons possible. With the arrival of cost-effective NGS and WGS this is no longer necessary; the different microbiological fields may now work together and different WGS analytical tools can be exchanged to maximize efficiency. Many of these earlier networks are now trying to make the move from traditional techniques to NGS. For example, PulseNet investigates how to use WGS and potentially metagenomics to replace PFGE and thus have a culture independent and faster technique (see: http://www.cdc.gov/pulsenet/next-generation.html). In such transitions it is important to implement new techniques in a way such that the old and new techniques are comparable and no data are lost.

GMI is an initiative open to anyone interested and many of the people associated with the networks summarized above have actually participated in the initiation and development of GMI. The work of GMI is to promote inter-disciplinary and international discussion of potential synergistic solutions to optimize the use of WGS globally. This process will take time and although some work may progress quickly (e.g. proficiency testing) for other issues more time is needed, as is intergovernmental debate and agreement. The roadmap for the development of the database that has been proposed with a vision of constructing an international system by 2020 is as follows:

- Development of pilot systems.
- Initiation of appropriate 'legal entity', with the formation of an international core group and governance structure
- Analysis of the present and future landscape to build the database
- Diplomacy efforts to bring the relevant groups together
- Development of a robust IT-backbone for the database
- Development of novel genome analysis algorithms and software
- Construction of a global solution, including the creation of networks and regional hubs

Initiatives with similar or overlapping goals as GMI have emerged and should be used to explore the opportunity for collaboration and synergy. Examples of such initiatives are the global alliance for genomics and health (http://genomicsandhealth.org/) and that of Google Genomics (https://developers.google.com/genomics/), both mainly focusing on human genomics, and the initiatives of CDC to use

WGS in parallel to their PFGE diagnostics and in their AMD programme (Advanced Molecular Detection) as well as the creation of the USFDA Genome Trackr Network, linking public health and university laboratories that collect and share genomic and geographic data from foodborne pathogens. GMI is presently in contact with these initiatives in order to investigate the potential for collaboration and synergy in the area of NGS/WGS use in microbiological identification and research as well as in genomic epidemiology and food microbiology.

References

1. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman P, Daszak P. Global trends in emerging infectious diseases. Nature. 2008;451:990–3.
2. Mathers CD, Boerma T, Ma Fat D. Global and regional causes of deaths. Br Med Bull. 2009;92(1):7–32. doi:10.1093/bmb/lpd028.
3. Kuiken T, Leighton FA, Fouchier RAM, et al. Pathogen surveillance in animals. Science. 2005;309(5741):1680–1.
4. Smith GJD, Vijaykrishna D, Bahl J, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 2009;459(7250):1122–5.
5. Taylor LH, Latham SM, Mark EJ. Risk factors for human disease emergence. Philos Trans R Soc Lond B Biol Sci. 2001;356(1411):983–9.
6. Woolhouse MEJ, Gowtage-Sequeria S. Host range and emerging and reemerging pathogens. Emerg Infect Dis. 2005;11(12):1842–7.
7. de Wit E, Kawaoka Y, de Jong MD, Fouchier RAM. Pathogenicity of highly pathogenic avian influenza virus in mammals. Vaccine. 2008;26 Suppl 4:D54–8.
8. Gao F, Bailes E, Robertson DL, et al. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature. 1999;397(6718):436–41.
9. Hirsch VM, Olmsted RA, Murphey-Corb M, et al. An African primate lentivirus (SIV) closely related to HIV-2. Nature. 1989;339:389–92.
10. Osterhaus A. Catastrophes after crossing species barriers. Philos Trans R Soc Lond B Biol Sci. 2001;356(1410):791–3.
11. Osterhaus ADME, Smits SL. Genomics and (Re-) emerging viral infections. In: Ginsburg GS, Willard HF, editors. Genomic and personalized medicine. 2nd ed. Amsterdam: Elsevier; 2012. doi:10.1016/B978-0-12-382227-7.00097-5.
12. Steel J, Louwen AC. Influenza A virus reassortment. Curr Top Microbiol Immunol. 2014;385:377–401.
13. Smits SL, Raj VS, Pas SD, Reusken CB, Mohran K, Farag EA, Al-Romaihi HE, AlHajri MM, Haagmans BL, Koopmans MP. Reliable typing of MERS-CoV variants with a small genome fragment. J Clin Virol. 2015;64:83–7. doi:10.1016/j.jcv.2014.12.006.
14. Liu GE. Recent applications of DNA sequencing technologies in food, nutrition and agriculture. Recent Pat Food Nutr Agric. 2011;3(3):187–95.
15. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010;11(1):31–46.
16. Rogers GB, Bruce KD. Next-generation sequencing in the analysis of human microbiota: essential considerations for clinical application. Mol Diagn Ther. 2010;14(6):343–50.
17. Smits SL, Osterhaus ADME. Virus discovery: one step beyond. Curr Opin Virol. 2013;3:1–6. doi:10.1016/j.coviro.2013.03.007.
18. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, Aarestrup FM. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. J Clin Microbiol. 2014;52(1):139–46. doi:10.1128/JCM.02452-13.

19. Reuter S, Ellington MJ, Cartwright EJ, Köser CU, Török ME, Gouliouris T, Harris SR, Brown NM, Holden MT, Quail M, Parkhill J, Smith GP, Bentley SD, Peacock SJ. Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology. JAMA Intern Med. 2013;173(15):1397–404. doi:10.1001/jamainternmed.2013.7734.
20. Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, Bortolaia V, Pearson T, Waters AE, Upadhyay BP, Shrestha SD, Adhikari S, Shakya G, Keim PS, Aarestrup FM. Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. MBio. 2011;2(4), e00157-11.
21. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. PLoS One. 2011;6(7), e22751. doi:10.1371/journal.pone.0022751.
22. Allard MW, Luo Y, Strain E, Li C, Keys CE, Son I, Stones R, Musser SM, Brown EW. High resolution clustering of *Salmonella enterica* serovar Montevideo strains using a next-generation sequencing approach. BMC Genomics. 2012;13:32. doi:10.1186/1471-2164-13-32.
23. Potron A, Kalpoe J, Poirel L, Nordmann P. European dissemination of a single OXA-48-producing *Klebsiella pneumoniae* clone. Clin Microbiol Infect. 2011;17(12):E24–6. doi:10.1111/j.1469-0691.2011.03669.x.
24. Brown EW, Detter C, Gerner-Smidt P, Gilmour MW, Harmsen D, Hendriksen RS, Hewson R, Heymann DL, Johansson K, Ijaz K, Keim PS, Koopmans M, Kroneman A, Wong DLF, Lund O, Palm D, Sawanpanyalert P, Sobel J, Schlundt J, Aarestrup FM. Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. Emerg Infect Dis. 2012;18(11), e1. doi:10.3201/eid/1811.120453.
25. Wielinga PR, Schlundt J. One health and food safety. In: Yamada A, Kahn LH, Kaplan B, Monath TP, Woodall J, editors. Confronting emerging zoonoses: the one health paradigm hardcover. New York, NY: Springer; 2014.
26. Prachayangprecha S, Schapendonk CM, Koopmans MP, Osterhaus AD, Schürch AC, Pas SD, van der Eijk AA, Poovorawan Y, Haagmans BL, Smits SL. Exploring the potential of next-generation sequencing in detection of respiratory viruses. J Clin Microbiol. 2014;52(10):3722–30. doi:10.1128/JCM.01641-14.
27. Kupferschmidt K. Epidemiology. Outbreak detectives embrace the genome era. Science. 2011;333(6051):1818–9. doi:10.1126/science.333.6051.1818.
28. Schlundt J. The time is right for a global genomic database for microorganisms. Health Dipl Monit. 2012;3(2):2–3.
29. Heger M. Next-gen sequencing shows promise for public health, but faces technical, political, socialhurdles.2011.http://www.genomeweb.com/sequencing/next-gen-sequencing-shows-promise-public-health-faces-technical-political-social.
30. Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. Bull World Health Organ. 2010;88(6):462–6. doi:10.2471/BLT.09.074393.
31. Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, Kokko H, Jennions MD, Kruuk LE. Troubleshooting public data archiving: suggestions to increase participation. PLoS Biol. 2014;12(1), e1001779. doi:10.1371/journal.pbio.1001779.
32. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA, Agarwala R, Bennett SF, Chen B, Chin EL, Compton JG, Das S, Farkas DH, Ferber MJ, Funke BH, Furtado MR, Ganova-Raeva LM, Geigenmüller U, Gunselman SJ, Hegde MR, Johnson PL, Kasarskis A, Kulkarni S, Lenk T, Liu CS, Manion M, Manolio TA, Mardis ER, Merker JD, Rajeevan MS, Reese MG, Rehm HL, Simen BB, Yeakley JM, Zook JM, Lubin IM. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol. 2012;30(11):1033–6. doi:10.1038/nbt.2403.
33. Moran-Gilad J, Sintchenko V, Pedersen SK, Wolfgang WJ, Pettengill J, Strain E, Hendriksen RS. Proficiency testing for bacterial whole genome sequencing: an end-user survey of current

capabilities, requirements and priorities. Global Microbial Identifier initiative's Working Group 4 (GMI-WG4). BMC Infect Dis. 2015;15:174. doi:10.1186/s12879-015-0902-3.
34. Wielinga PR, de Vries A, van der Goot TH, Mank T, Mars MH, Kortbeek LM, van der Giessen JW. Molecular epidemiology of *Cryptosporidium* in humans and cattle in The Netherlands. Int J Parasitol. 2008;38(7):809–17.