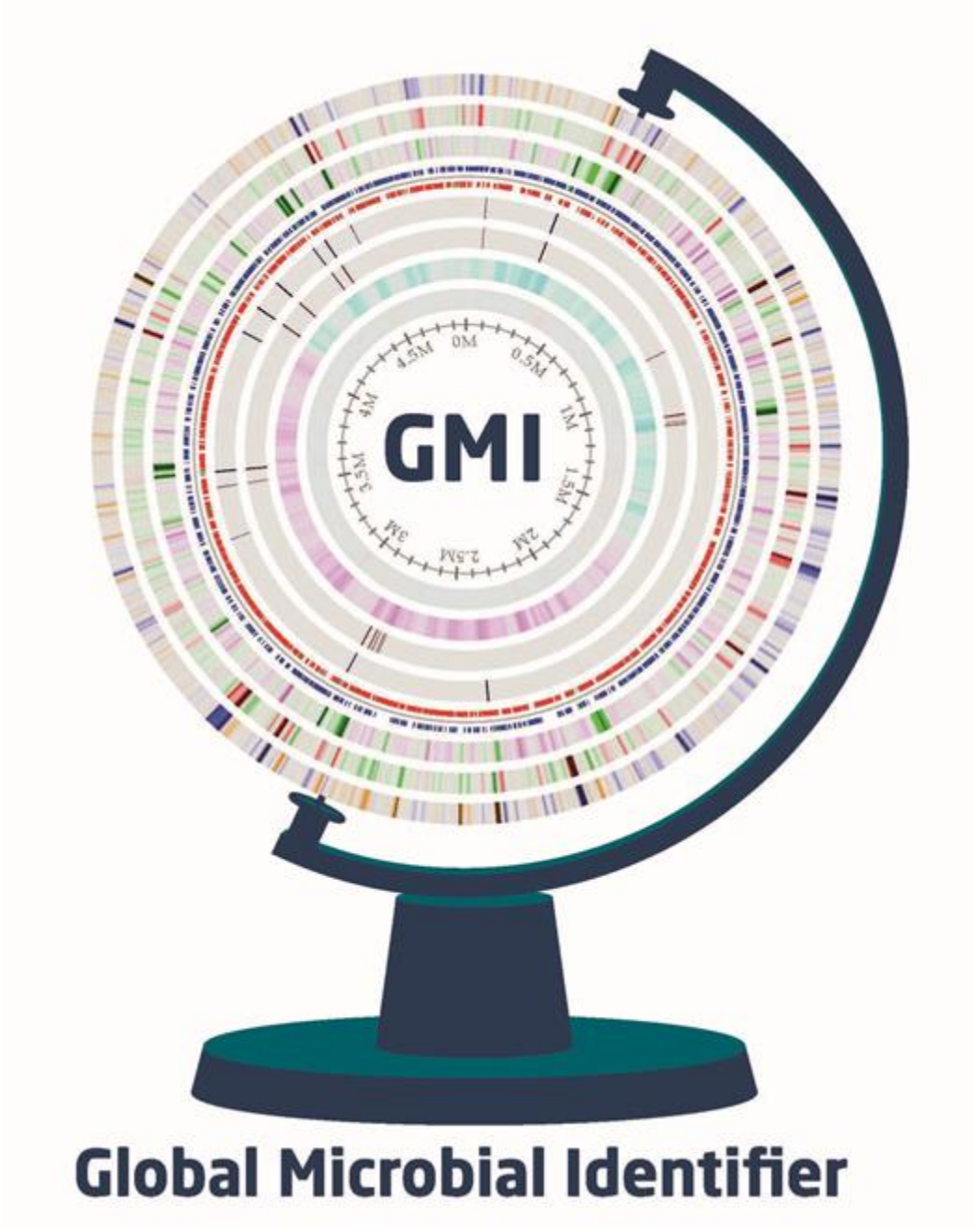

fera

Report of the 7th GMI meeting, 11-12 September 2014

Meeting Report from the 7th GMI Meeting, 11-12th September 2014, held at the Food and Environment Research Agency, Sand Hutton, York, UK

Introductory Remarks

Welcome from the meeting hosts

Phil Newton, Director of Science, Food and Environment Research Agency (Fera)

Delegates were welcomed to the Sand Hutton campus, and Fera's work and microbiological and bioinformatic capabilities were described. GMI is about pulling together capabilities to solve large, real-world problems. This reflects Fera's ethos.

Welcome from the GMI Steering Committee

Alisdair Wotherspoon, FSA

There is much to be gained from getting this approach right, and enabling equity of access. There are difficult issues to tackle, and this is the purpose of the working groups. Working groups will be larger than at previous meetings, and new people are encouraged to contribute. GMI is trying to coordinate global efforts, but we don't want to force people into a way of working

GMI Update

Jørgen Schlundt, DTU

Currently there may be confusion about what GMI is. There have been six international meetings since 2011, with a work plan devised at the 5th meeting, and a charter at the 6th. The GMI is becoming more global. In 5-20 years, the genome of every pathogen in the world could be in a cloud database. Answers to oral questions reflected the inclusivity of the GMI model, and the fact that most working groups are currently behind schedule.

Opening Talks

EFSA's Scientific Colloquium on the Use of Whole Genome Sequencing (WGS) of food-borne pathogens for public health protection

Ernesto Lieban, EFSA

Molecular typing methods for food-borne pathogens are now frequently applied in the European Union (EU) for public health protection purposes (e.g. investigating food-borne outbreaks, identifying strains of food-borne bacteria with high virulence potential or resistance to antimicrobials). This is the result of continuous advances in the understanding of the molecular characteristics of bacteria and their genetics, and very much linked to technological developments,

which ultimately are leading towards the use of bacterial whole genome sequencing (WGS) methods for food safety applications. There is currently limited experience in the use of WGS methods in microbial food safety in the EU. Most of that experience comes as a result of retrospective studies that have followed from outbreak investigations. Still, the potential application of WGS to predict pathogenicity of the strain under investigation may provide risk assessors with a powerful tool. Full integration of WGS of food-borne pathogens in food safety will only be possible after successful collaboration and coordination among scientists, paving common pathways to overcome key challenges.

The BIOHAZ Panel in EFSA has recently issued two opinions on the evaluation of molecular typing methods for major food-borne microbiological hazards and their use for attribution modelling, outbreak investigation and scanning surveillance. In addition there is an ongoing mandate from the European Commission to EFSA and ECDC on the building of a molecular typing database and regularly providing scientific analyses in collaboration with EURLs. The momentum seemed appropriate for EFSA taking the initiative and thus organise a Scientific Colloquium on the use of WGS of food-borne pathogens for public health protection.

This Scientific Colloquium took place on 16-17 June 2014, and focused in the review of the latest scientific information. It was a golden opportunity to strengthen alliances with relevant EU and international bodies necessary to initiate discussions on the use of WGS methods for food safety applications. Conclusions from this event will inform and drive EFSA's ongoing efforts in the collection of molecular typing data by proactively anticipating the specific requirements of WGS data.

The specific objectives of the three established discussion groups during the Scientific Colloquium were to discuss in an open debate: (i) the current issues, benefits and future challenges of WGS of foodborne pathogens in public health protection in comparison with current methods; (ii) the analysis and the interpretation of WGS data in the ascertaining of diversity, similarity and relatedness of foodborne bacterial pathogens and to predict pathogenicity and other relevant characteristics (e.g. virulence, AMR); (iii) the curation and analysis of WGS data, bioinformatics solutions, and (iv) the coordination of efforts between the food, veterinary and human health sectors in order to obtain maximum benefits from the use of WGS for food safety and public health protection.

A summary report of the Scientific Colloquium is being drafted under guidance of the Chairs and Rapporteurs of this event and should be published by the end of this year.

Questions included inquiries on whether the databases developed would be solely for food pathogens (EFSA databases would be), and whether any discussions had yet taken place regarding bacteriophages (something not yet discussed since the colloquium).

Future lessons from large-scale biological data management

Paul Flicek, EBI

Advances in DNA sequencing and other high-throughput technologies are leading to wide-spread changes in the practice of biological research. At the core of these changes is an emphasis on the informatics infrastructure required for large-scale data management, processing and analysis. By drawing on experience from the 1000 Genomes project and other projects focused on distributed data collection and analysis, common approaches and best practices are beginning to emerge that will enable the next generation of still larger projects.

The talk focussed on general lessons from large-scale data management, specifically from a data-driven experiment; the 1000 Genomes project. The project generated an order of magnitude more than the original estimate of data. Now all infrastructure can handle the higher rate of data flow, and we only notice infrastructure when it goes wrong. Informatics is infrastructure, in terms of; network transfer protocols; data compression; standards; archives.

The 1000 Genomes project is enormous, currently 80 Terabases of sequence. EBI data goes out to more places than it comes in from, and EBI has to manage this. Metadata stored in an uncoordinated way becomes difficult to manage. EBI has become good at solving problems, for example using standards compliant data, such as reporting minimums for metadata.

EBI is also involved in the Global Alliance for Genomics and Health, the ELIXIR project, and EMBL-EBI Embassy Cloud.

Working Group Progress

WG 1: Political challenges, outreach and building a global network.

Jørgen Schlundt

Working Group 1 is developing a long-term plan to shape political level involvement in GMI development at the global, regional and national level.

The first goal is to establish a functional link to political level decision makers in several countries or regional, international organizations.

The second goal is to initiate a coherent system for international discussion of relevant themes, e.g. global health diplomacy, coordination between different sectors, sensitivity of metadata, open access database of genome sequences, sharing of strains over borders, intellectual property rights (IPR), and funding.

Progress of WG1 since the last meeting includes a membership database, and outreach to the Bill & Melinda Gates foundation (and other foundations). We still need national commitments to GMI. The COMPARE project recently secured EU Horizon 2020 funding to work on GMI methods in Europe, and was very highly rated.

WG2: Repository and storage of sequence and metadata.

Bill Klimke, NCBI

Working Group 2 strives towards developing a format to capture ”Minimum Data for Matching (MDM)”, consisting of reads and minimum metadata.

The MDM should be deposited and made globally and universally accessible as soon as available. The MDM may or may not be accompanied by assemblies and/or annotation and/or additional metadata. If not provided with initial submission, these may be added later by the submitter, or by some agreed upon third party. Ideally, any MDM provided for purposes of searching the GMI databases should immediately also become a deposit available for searching by later submitters.

Any matches from the MDM search should be reported to the searcher and to the relevant GMI Participants. The data layer is provided by The International Nucleotide Sequence Database Collaboration INSDC and is therefore both international and public.

The search and analytical layers may be provided by INSDC members or by other parties. For research purposes it is fine to have a variety of tools and searches. But in order to provide a coordinated GMI there must be a more centrally controlled searching and reporting protocol that official sites adhere to, and to whom the food safety agencies submit, which is much more limited.

Standards are needed for analysis. The existing infrastructure has worked for decades, and should be usable for the GMI approach. Xml-based submission is being developed for rapid, high throughput submissions. A table format is available for more ad hoc submissions.

WG3: Analytical approaches

Marion Koopmans, RIVM

Working Group 3 is providing guidance for the development of analytical tools for optimal positioning and functioning of the GMI platform.

GMI is bringing together scientists, public health experts, policy makers, etc. to develop a global platform (database, or linked databases) that facilitates the application of NGS in research, clinical and public health settings worldwide.

Working group 3 aims to define requirements for GMI functioning from the perspective of end-users (clinical, public health, research) in terms of applications (identification, outbreak detection etc.) and priority targets/diseases. This working group wants to map current analytical options and solutions against the needs of GMI end-users, to identify possible R&D and implementation gaps and to identify projects that may fill those gaps.

Working group three is a very large working group, which may benefit from being broken down into themes (e.g. clustering, phylogeny, resistance genes etc). Virology also needs to be incorporated.

WG4: Ring trials and quality assurance

Rene Hendriksen, DTU

Working Group 4 is aiming for all laboratories globally to conduct NGS on bacteria and viruses to the highest degree of quality, initially to organise a pilot proficiency test for the work group participants and secondly to offer this test to GMI members working with both bacteria and viruses. This working group is working on how to build an infrastructure within the partners of GMI that has the capacity to undertake the facilitation of the proficiency testing.

The group is considering how to develop or provide the reference material and documents needed to initiate the proposed pilot proficiency test scheme. Reference material will be distributed to enrolled laboratories. Reference material, documentation and analysis will be adjusted based on previous experiences.

Another theme for this group is to find out how to conduct the analysis of submitted genomes and how to execute a fully operational proficiency test based on bacteria and viruses to GMI members. The working group will also evaluate RNA purification methods / protocols and pilot sequencing on multiple platforms to initiate the proposed parallel viral pilot proficiency test scheme.

WG5: Pilot project

Marion Koopmans

Working Group 5 is developing projects that provide progressively challenging technical demonstrations of NGS for local and global tracing of pathogens within the GMI Network.

The immediate goals of WG5 are twofold. The first goal will be to establish a viable and functional working group communications and governance structure and define how the this WG will interact with the other working groups in GMI.

The second major goal is to define the purpose and nature of a pilot project and determine the properties of a pilot project that will satisfy the requirements of the broader GMI effort.

Pilot projects are planned, but lack funding. This working group should think about and suggest suitable pilot projects.

Gold Sponsor Presentation

Rapid single-colony whole-genome sequencing of bacterial pathogens

Louise Fraser, Illumina

OBJECTIVES: As a result of the introduction of rapid benchtop sequencers, the time required to subculture a bacterial pathogen to extract sufficient DNA for library preparation can now exceed the time to sequence said DNA. We have eliminated this rate-limiting step by developing a protocol to generate DNA libraries for whole-genome sequencing directly from single bacterial colonies grown on primary culture plates.

METHODS: We developed our protocol using single colonies of 17 bacterial pathogens responsible for severe human infection that were grown using standard diagnostic media and incubation conditions. We then applied this method to four clinical scenarios that currently require time-consuming reference laboratory tests: full identification and genotyping of salmonellae; identification of blaNDM-1, a highly transmissible carbapenemase resistance gene, in *Klebsiella pneumoniae*; detection of genes encoding staphylococcal toxins associated with specific disease syndromes; and monitoring of vaccine targets to detect vaccine escape in *Neisseria meningitidis*.

RESULTS: We validated our single-colony whole-genome sequencing protocol for all 40 combinations of pathogen and selective, non-selective or indicator media tested in this study. Moreover, we demonstrated the clinical value of this method compared with current reference laboratory tests.

CONCLUSIONS: This advance will facilitate the implementation of whole-genome sequencing into diagnostic and public health microbiology.

Case Study Talks

Establishing a Whole Genome Sequence-Based national network for the detection and traceback of foodborne pathogens

Steve Musser, FDA

The Center for Food Safety and Applied Nutrition at the FDA along with our partners, CDC and NCBI, are implementing an international distributed sequencing network of public health labs. Herein we describe the components of the NGS pathogen diagnostic network that includes 6 state public health laboratories (AZ, FL, MN, NY, VA and WA), 9 US laboratories, and our first international lab in Argentina. Details of the success and failure will be provided concerning communication, coordination, data acquisition, assembly, storage, and analysis. Herein, we report enhanced molecular epidemiological insights gained by comparative analysis of *Salmonella* and *Listeria* genomes previously deemed indistinguishable by conventional subtyping methodologies. These results demonstrate an important investigative role for NGS tools within a regulatory environment while highlighting the novel additional insights provided to epidemiological investigations.

The development of the NGS diagnostic network arose from a one-time opportunity to get a large amount of funding, so sequencers were distributed to a number of state labs very early in the process (i.e. before all the data management systems etc. had been solidified). FDA have decided that distributed sequencing facilities work better than centralised core facilities, and therefore will no longer be funding the 100k genome project. Ways are now being devised to broaden the system beyond the United States. For example, funding is now being acquired from WHO for a facility in Argentina.

See also; genomelc.jifsan.org.

Bioinformatics applications: microbial comparative genomics and gene expression studies - integrated data analysis

Robert Stones, Fera

Next generation sequencing (NGS) platforms offer a rapid way in which to sequence complete bacterial genomes and gene expression profiles, important for pathogenesis studies. Harnessing NGS technologies and the development of new bioinformatics methods offers a powerful set of tools in food safety and public health.

We have developed bioinformatics tools to utilize the hypothesis driven mechanism of comparative genomics and rapid sequencing technologies of NGS data, to be screened against existing metabolic pathway data and the reconstruction of metabolic pathways.

Developed methods to compare gene expression data and integrate small molecule metabolites detected from LC-MS data. Enabling correlation with specific environmental conditions/phenotypes, which may also be hypothesized from our comparative genomics studies.

Currently developing methods to detect epigenetic modifications, with the aim to identify epigenetic signatures associated with gene expression profiles. Also using new techniques such as non-targeted mutagenesis experiments, we are identifying genes, which can be classed as essential or non-essential genes, when bacteria of interest are grown in culture under different environmental growth conditions. Knowing which genes are essential to the survival of pathogens will help researchers to better understand the ability of pathogens to grow under such unusual environmental conditions.

Future development of novel bioinformatics methods from NGS data to compare and reconstruct metabolic pathways with the integration of heterogeneous datasets, including gene expression data, metabolite datasets and new epigenetic biomarkers; will enable a better understanding of pathogenesis and provide an important tool set in food safety and public health.

Use of Next-Generation Sequencing for microbial strain characterisation. What is the same?

Leen Baert, Nestlé Research Center

Next-Generation Sequencing (NGS) is used by authorities and may soon replace conventional typing to link foodborne pathogens such as *Salmonella* and *Listeria* to illness outbreaks. NGS analysis includes several steps such as sample preparation (i.e. DNA extraction, library preparation), a sequencing run and bioinformatic analysis (with or without assembly, with or without annotation, comparison of whole genomes or a selection of genome regions). Currently, the impact of these steps on the outcome is not known. Data obtained for *Salmonella* and *Cronobacter* showed that the outcome of NGS greatly depends upon the selection of genome regions used for the bioinformatic analysis. The impact of all the steps involved in the NGS workflow for microbial strain characterization needs to be fully understood in order to achieve generally accepted, and possibly standardized, analysis approaches.

Outbreak: fully integrated, real-time detection, diagnosis and control of diarrhoeal disease clusters in the community

Sarah J O'Brien, University of Liverpool

Seventeen million people suffer from diarrhoeal disease in the UK annually. It commonly causes outbreaks - norovirus outbreaks alone are the third most costly healthcare-associated infection in the NHS. Severe outbreaks of *Escherichia coli* O104 in Germany and O157 at Godstone Farm in Surrey are reminders that rapid detection, through surveillance, is key to control. We need to identify the nature, distribution and scale of outbreaks quickly enough to lessen impact. The "hidden" burden of undiagnosed disease is growing as fewer people are seen by General Practitioners, reducing outbreak detection sensitivity.

We are targeting primary care telephone consultations (so-called physician assessment by telephone consultations) for obtaining clinical samples. We are replacing traditional diagnostics with rapid modern technologies, to identify and characterise responsible pathogens rapidly. We are applying targeted microbial metagenomics, using latest-generation genome sequencing for putative pathogen discovery, and to determine phylogeny and evolutionary trends to understand how new agents have arisen.

Many diarrhoeal diseases are zoonotic yet medical and veterinary surveillance systems are poorly integrated so we are incorporating and developing veterinary surveillance. Our target is an interdisciplinary, integrated, real-time, surveillance/diagnosis/investigation system, placing the patient at its centre, that will enable earlier detection of community outbreaks, lessening harm because Health Protection professionals and Environmental Health Officers can intervene faster.

Harnessing Microbial Genomics for Epidemiological Surveillance

Dag Harmsen, University of Münster

A variety of bacterial epidemiological typing methods have been developed to generate isolate-specific fingerprints for following transmission and to detect outbreaks. However, until recently there was no single typing method available to address different bacterial population structures (monomorphic vs. panmictic) and all study types (e.g., evolutionary/phylogenetic, population genetics or transmission chain/outbreak investigation). With fast and affordable benchtop microbial whole genome shotgun (WGS) next generation sequencing (NGS) and automatized software analysis, microbiologists can use now one method that fits all bacterial species and study types. Therefore, the next big challenge will be to harmonize analysis of WGS NGS data so that microbiologists speak one language world-wide, i.e. 'molecular Esperanto', that allows them to compare their data quickly with data not generated by them. Such a nomenclature must be additive expandable at low compute cost to handle the ever increasing data-streams in an efficient manner.

Currently two non-compatible widely used approaches are employed for genomic epidemiology, i.e. mapping followed by single nucleotide polymorphism (SNP) calling or *de novo* assembly coupled with a genome-wide gene by gene (MLST+ or core genome MLST; cgMLST) analysis. Both approaches have their pros and cons that will be discussed. Ways in which they could be used in the future complementary rather than mutually exclusive will be presented.

Gene-by-gene analysis may be a preferable approach because, even though it is reductionist, loses complexity and only focusses on coding sequences, it has a nomenclature. This means all analyses will be comparable, which is not the case for SNP calling.

Staphylococcus aureus as an exemplar for translational Whole Genome Sequencing

David Aanensen, The Wellcome Trust Sanger Institute and Imperial College London

The Center for Genomic Pathogen Surveillance is a new initiative from The Sanger Institute in conjunction with Imperial College London which aims to provide data and tools for the global surveillance of pathogens and determinants of antimicrobial resistance using whole genome sequences. This talk presented, through a series of exemplar web applications, data utilizing traditional typing methods in the context of large scale structured surveys and through the application of WGS, a new BETA web application was presented for the collection and analysis of sets of microbial whole genomes. While data presented focused on *S. aureus*, the system and tools are generic and applicable for any species datasets.

Previous examples of tools produced by this group include EpiCollect. Now involved in projects spearheaded by the EARS network.

See also; www.spatialepidemiology.net, www.wgsa.net

Visualization and analysis of the emergence, evolution, and spread of pathogens

Dan Janies, University of North Carolina at Charlotte

We have used molecular phylogenetics in many ways to understand the spread of pathogens and their features. Features can include the place of isolation, host or tropism of a pathogen, or genotypes that confer phenotypes such as drug resistance. One issue that often arises is that features mapped on to a large phylogeny are difficult to interpret. Several workflows were demonstrated designed to enhance the understanding of phylogenetic data including: 1) projection of a phylogeny and its features in a geographic information system; and 2) calculation of a betweenness graph for a phylogeny and its features. The projection is akin to a weather map for infectious diseases. The betweenness graph allows users to understand the relative flow of related pathogens among several geographical places or various hosts. Use cases with public health significance were presented.

Genomic Portrait of the Evolution and Epidemic Spread of recently Emerged Multidrug-Resistant *Shigella flexneri* Clone in China

Jianguo Xu, Chinese Center for Disease Control and Prevention

Shigella flexneri is the major cause of shigellosis in China as other developing countries. A new *S. flexneri* serotype Xv emerged in 2000 and replaced serotype 2a as the most prevalent serotype in Henan province of China. Serotype Xv is a variant of serotype X with a phosphoethanolamine (PEtN) modification of its O-antigen mediated by a plasmid encoded O-antigen phosphoethanolamine transferase gene *opt* (formerly *lpt-O*) . Serotype Xv isolates belong to sequence type 91 (ST91).

Whole-genome sequencing of 59 *S. flexneri* isolates of 14 serotypes (serotypes 1 to 4, Y, Yv, X, and Xv) indicated that ST91 arose around 1993 by acquiring multidrug resistance (MDR) and spread across China within a decade. A comparative analysis of the chromosome and opt-carrying plasmid pSFXv_2 revealed independent origins of 3 serotype Xv clusters in China, with different divergence times. Using 18 cluster-dividing single-nucleotide polymorphisms (SNPs), SNP typing divided 380 isolates from 3 provinces (Henan, Gansu, and Anhui provinces) into 5 SNP genotypes (SGs). One SG predominated in each province, but substantial interregional spread of SGs was also evident. These findings suggest that MDR is the key selective pressure for the emergence of the *S. flexneri* epidemic clone and that Shigella epidemics in China were caused by a combination of local expansion and interregional spread of serotype Xv.

Following the introduction to some of the whole genome sequencing work of the Chinese CDC, an introduction was given to the planned eighth GMI meeting to be held in Beijing in 2015.

Outcomes of Working Group Breakout Sessions

WG1: Political challenges, outreach and building a global network

The meeting agreed an agenda:

- Discussion about WHO/PIP experience of importance for GMI
- Issues related to the free exchange of microbial genomic data (Harenghuizen et al. paper)
- Legal aspects relative to GMI structure (Schlundt et al. paper)
- A potential analysis of the NGS/GMI landscape
- Collaboration with the Global Alliance for Genomics and Health
- Future FAO/WHO Expert meeting on NGS
- Process to confirm the composition of the Steering Committee (as stipulated in the Charter)

Discussion about WHO/PIP experience of importance for GMI

(PIP: Pandemic Influenza Preparedness framework)

Prof D. Houssin provided through videolink answers to six pre-defined questions, developed in collaboration with Prof. Houssin. The outcome of the related debate was as follows:

- Would the international sharing of WGS data – e.g. in a system as described under the GMI Charter – cause similar problems as the international sharing of virus strains, which the WHO PIP (Pandemic Influenza Preparedness) Framework document intended to deal with?

Yes it could! The global influenza surveillance network (GISRS) was a network of laboratories in the developed countries established to collect, ship and characterize influenza viruses, the end products being vaccines used mainly in developed countries. In 2005 and 2006, H5N1 avian flu and the pandemic threat led to increased focus on this network and above all that virus from developing countries would through the network end in vaccines which could be difficult for developing countries to buy. It was perceived that developing country research was not a recognized part of the work to produce vaccines. This led to the 'Indonesian revolt' and to the WHO PIP framework organizing virus sharing through benefit sharing. Therefore, it will be important in the GMI construction to include a transparent, fair system to include proper scientific recognition, especially

– but not only – when the results may have important product outcome (drugs, diagnostic tools, etc.). GMI should attempt to form equitable partnerships early on with all participating countries

- Is there any consideration within the WHO Technical Expert Working Group (TEWG) on Genetic Sequence Data to consider issues related to sharing over borders of microbiological WGS data?

Yes, but only with consideration to influenza viruses. In the PIP Framework Agreement the use of Gene Sequence Data (GDS) was envisaged but not treated. Recently, it appeared that influenza vaccines could be produced almost without virus strains, using only GSD. The issue is now how to take advantage of the progress allowed by the use of GSD (speed) without impairing the benefit sharing part of the agreement. The TEWG document was a first step to approach this question, but issues related to sharing over borders of other microbiological WGS data have not been discussed, due to the mandate of the TEWG. In next year's assessment of the PIP agreement, the scope of the framework will likely be enlarged to include other viruses (e.g. coronavirus), but the question of the sharing of other microbiological data has not yet been discussed

- Do you see any interesting perspective in using a common microbiological WGS database – and a common set of algorithms for identification – in support of international disease outbreak and response frameworks (incl. the WHO IHR)?

Yes, considering the intense globalization of human activities and the major risks created, for example, by antimicrobial resistance, such a perspective appears particularly important, precisely in the framework of the international health regulation.

- Is there a need to apply a 'One Health' paradigm to the future use of WGS technology, and would that mean that the WHO and the OIE need to consider common work in this area?

Yes, this would certainly be a necessary approach, as is already the case with influenza viruses through the PIP Framework Agreement. For the implementation of this agreement, the advisory group encouraged strengthened cooperation between human and veterinary labs. Bureaucratic hindrances (e.g. standard material transfer agreements) should be avoided for avian flu viruses used as controls in order not to discourage vet. labs to send viruses to human WHO labs.

- With regard to the GMI initiative, considering the present PIP Framework experience, what aspects should particularly be considered?

The advantages – to Global Health and Health Security - of having such a global sharing system under the umbrella of WHO, should be clearly delineated. The full cost of the initiative needs to be estimated, as well as different funding constructions, bearing in mind the potential of support funding from industry sources (who contribute 50% of the annual funding for GISRS ($28m)). Issues around the potential for industry to enter in the process of elaboration of a standard material transfer agreement should also be investigated. The many questions raised about the physical positioning of future databases (and the legal issues this could result in) and the mechanism of data sharing (publicly accessible (with or without restriction), public domain) and their relation to the overall objectives of preserving global and national health security and potential benefit sharing needs to be thoroughly investigated, analyzed and debated in a global forum.

Action: WG1 should prepare input, which would be relevant to both a potential Landscape Analysis of the WGS Landscape and to potential future WHO/FAO Expert meetings, outlining relevant lessons for GMI development from WHO PIP and TEWG experience.

Free exchange of microbial genomic data (Harenghuizen et al. paper)

At GMI5/6 the need for an overview of legally-related implications of free genomic data-exchange was clearly expressed. In the preparation for GMI7 a draft paper was prepared dealing with a number of important aspects relative to such implications. (George Haringhuizen, Chris Braden, Sharona Hoffman, Palmer Orlandi, Marguerite Pappaioanou: Towards a white paper on challenges and solutions to developing conditions that support and promote global, open access to and free exchange of microbial genomic information and related metadata among GMI members).

While this paper was generally considered a comprehensive account of existing issues, it was agreed that the most efficient use of the allocated time to debate at GMI7 would be to focus on the presently most important issues, leaving further issues to be addressed in the further development of the paper.

The meeting agreed that this was a well prepared paper, covering a very wide array of relevant issues. While such issues could in most cases be said to reside within the GMI debate, it was pointed out that many of the solutions would have to be found outside of GMI proper. While it was agreed that issues and problems of GMI introduction would have to be the major focus of the paper, it was suggested that benefits should also be described in the text. This could for instance be in the form of outlining the need for general cost–benefit calculations of the effect of introducing WGS.

Special attention should be given to describe the position/benefits/challenges of 3rd world countries in relation to GMI. It was suggested that a brief discussion of which challenges developing countries would face in this area within the next 5 years, would be helpful. Specifically, there is a need to focus on the issue of IT connectivity in developing countries.

What is the expected benefit of sharing metadata? Would it really be a necessary prerequisite to start sharing sequence data also to share metadata? It was mentioned that in the USA some States have asked for exceptions to be allowed to upload sequence data without even minimal metadata; when such exception was not given, ultimately States have agreed to share. Many problems seem to depend of the size of metadata. There would at least seem to be a need for a framework for communication between countries when more information on metadata/sequences will be needed. It was noted that the discussion of risks should be seen to be necessary in order to achieve protections against misuse that would in the end be needed to achieve overall trust. Thus, the necessary debate about risks would actually become a prerequisite for acceptance and therefore in the end contribute to allowing a final flow of information.

In preparation of a tentative list of arguments for - or situations leading to - refusing globally sharing genomic data and/or meta data, the following issues were mentioned:

- Concerns about national security and safety
- Scientific and technical barriers

- Institutional and management barriers
- Economic risks and IP-rights; financial barriers
- Ethical issues and concerns
- Socio-cultural and normative differences between populations
- Restrictive national regulations and laws
- Commercial confidentiality and Corporate protection
- Political/ economic impacts:
- Metadata could lead to identification of companies
- In outbreak situations data won't be public because of potential for prosecution
- Not always obvious who is the owner of the data – company, authority, laboratory?
- How to share when one would only have part of the data (food / human / animal / ??)
- Potential for misuse by competitors at national or international levels (trade)
- Potential for misuse by others to attain new markets (drugs, diagnostic methods, vaccines…)
- Reluctance to release data before a Ref lab has confirmed it
- Technical requirements : sharing data is an additional job
- Fears of sharing with foreign countries
- Fears of sharing with countries with different legal frameworks
- Uncertainty of whether data will be used for risk assessment or management or research
- Freedom of information act (USA) – or similar legislation in other countries

Action: The above list of issues should be used by WG1 as a starting point to prepare a questionnaire to GMI participants on this subject. The outcome should be used to focus further attention to the process of communication/discussion between countries and institutions on these political issues, to be included in the Harenghuizen et al. paper, and to be shared in analyses of the WGS Landscape and with potential future WHO/FAO Expert meetings. Likewise it was suggested in this process to use more concrete examples of data-sharing: It could for instance be considered to prepare MERS and/or Ebola case-studies, evaluating the actual sharing of data in such cases.

Legal aspects relative to GMI structure (Schlundt et al. paper)

At GMI5/6 the need for an overview of legal implications related to the GMI structure as such was clearly expressed. In the preparation for GMI7 a draft paper was prepared dealing with such implications (Jorgen Schlundt, Ane Sandager, Peter Wielinga, George Haringhuizen: GMI discussions on the juridical and political aspects defining conditions for a global database). The paper describes a number of legal aspects in relation to further development of GMI, but the meeting agreed to focus discussion on which construction types would be necessary corresponding to the development of global WGS initiatives.

In general, the meeting did not see an immediate need to change the legal structure of GMI at this stage. The flexible network under the current 'charter and structure' is opportune relative to the task at hand.

However, while there is no immediate need for immediate change, currently the GMI construction is clearly just an informal network structure, and it is important to consider the implications and necessary speed of moving forwards towards some sort of a legal entity. In simplified form there would basically be three possibilities for a GMI structure:

Informal network

Formalized collaboration

Legal entity

Some of the positive and negative aspects of the three types of constructs were debated:

Informal Network (e.g. present GMI construction):

Positive: swift decisions, independent, all inclusive, flexible, un-bureaucratic

Negative: Messy/intransparent, difficult interaction, funding problematic, potential problems with partner contacts, credibility, accountability and developing countries participation

Formalized Collaboration :

Positive: long term sustainability, perceived independent, borders clear but contents still flexible

Negative: not ready for coherent global action, institutions taking over responsibility from individual researchers, exclusivity

Legal/entity :

Positive: transparent, fundable, democratic (?), solid, communicable

Negative: slow, dependent, potentially non innovative and bureaucratic, exclusive, complicated to construct and maintain, potential departure from technical roots

A mélange construct was mentioned (Network + Bureau) as a solution to maintain flexibility and construct a legal entity at the same time.

The important issue would now be to consider criteria which should be governing decisions to proceed from one organizational structure to the next. Within these considerations it should be noted that, while moving too slowly could in general be perceived as lack of progress, moving to the next level too early could in itself present an impediment and interfere with achieving the goals of the GMI.

The following considerations/situations were mentioned, relative to a perceived need to become a formalized legal entity:

A situation where GMI will be the owner of the database: Legal entity needed

A situation where GMI would receive (significant) funding: Legal entity most likely needed

A situation where GMI is developing standards/guidelines: Legal entity could be needed.

Moving from networks to partnerships etc. can be a problematic process, even when such partnership are firmly anchored in strong, intergovernmental institutions: the WB Global food safety partnership (GFSP) mentioned as one such example. GFSP is a public-private initiative dedicated to supporting global cooperation for food safety capacity building, and it serves as a platform in which concerned international organizations, public sector agencies, private sector producers, leading academic institutions, etc. can convene. While the purpose of the GFSP is certainly a noble one, the partnership construction does in this case not seem to have supported the development of the initiative to the degree needed. Examples such as this can be instructive when preparing for analysis of GMI legal development potential

Action: The above considerations should be included in the Schlundt et al. paper, to be shared for comments to the GMI network before GMI8. These issues should, if possible, also be included in analyses of the WGS Landscape.

A potential analysis of the NGS/GMI landscape

J. Schlundt informed in generic terms about the outline of an analysis of the WGS Landscape. This would include an analysis of strategic options for planning and building an open global platform based on NGS and associated bioinformatics & IT solutions to comprehensively diagnose, monitor, control and prevent infectious diseases as well as identify and characterize microbiological isolates from all sectors (human, food, animals & environment). The analysis will include an investigation of the necessity and benefit of involving the developing countries already in the planning phase of the platform.

The urgent need for a WGS Landscape analysis is supported by the fact that many NGS initiatives are now underway worldwide: bioinformatics and its subfields is developing, new WGS software tools are put online every day, and also at governmental level, NGS is being regarded as a tool in the continuous quest for public health efficiency improvements. These developments have made NGS grow from a basic research tool into a mature general purpose tool with translatable advantages in both global health & agricultural development; however, the problem is that these separate initiatives lack interoperability compounded by the fact that likely all technical development takes place in the western world and therefore does not involve (nor benefit) developing countries.

An outline of a WGS Landscape analysis has been prepared with eleven 'Issue Analyses' spanning areas from existing database analysis to business case modelling. Funding is being sought to initiate the project.

Action: The outcome of funding initiatives should be communicated to WG1 through October 2014, and a plan for further action with and without external funding should be finalized with input from relevant partners.

Collaboration with the Global Alliance for Genomics and Health

J. Schlundt informed about the initial contact to the Global Alliance for Genomics and Health, which was created in January 2013, when a group of international leaders met to discuss the current challenges and opportunities in genomic research and medicine. The Alliance was created based on the White Paper "Creating a Global Alliance to Enable the Responsible Sharing of Genomic and Clinical Data." (see website www.genomicsandhealth.org).

The overarching goal of the global alliance is to accelerate progress in medicine by encouraging widespread access to genomic and clinical data by developing a common framework of international technical, operational and ethical standards needed to ensure the interoperability of genomic research platforms in a secure and responsible manner. In order to achieve these goals, the global alliance will work to (i) bring together the research, clinical, and disease advocacy communities and the private sector to support and promote the responsible sharing of genomic data and (ii) collaborate with interested parties to create an information platform that is open and accessible, and provides common standards, formats and tools to stakeholders in the genomic research community.

The initial contact has shown that the two initiatives share a vision of opportunities for alignment on current and future work between the Global Alliance and GMI.

Action: The GMI Steering Committee was urged to continue deliberations with the Global Alliance.

Future FAO/WHO Expert meeting on NGS

Amy Cawthorne informed on behalf of WHO and FAO about a potential future Expert Consultation on Whole Genome Sequencing. The focus of the consultation would be the application of WGS to foodborne diseases surveillance (including foodborne outbreak detection and response) and food contamination monitoring. However many of the outputs will be relevant to application of the technology in all areas of human, animal and plant health. The Consultation(s) would have a specific focus on the situation in developing countries There is at this stage no funding for the Consultation; WHO and FAO will develop a further description of the Consultation and the process leading to it, with a view of obtaining funding.

The full process could include the development of several 'white papers' informing specific expert meeting(s), and followed by one or two Expert Consultations.

Action: The GMI Steering Committee was urged to continue the interaction with WHO and FAO with a view of promoting the planning of international expert meetings in the area

Process to confirm the composition of the Steering Committee

It was agreed to suggest to the plenary meeting to a) continue with the present composition of the Steering Committee, but also to consider expanding the Committee with a member from the potential organizer of GMI8 in 2015: China.

Action: The GMI Steering Committee was urged to consider the inclusion of an additional member from the GMI8 organizers in China

WG2: Repository and storage of sequence and metadata

New objectives were discussed, to work towards before the next GMI meeting.

These objectives included;

- Creation of a mailing list of GMI users (potential and currant)
- Inform GMI submitters of how to submit data at EBI and NCBI, and help them through technical hurdle.
- Poll the mailing list on where they are having difficulties in submission, either technically, or socially/politically. GMI would like to know what are the objections to the GMI concept as a core model.
- Find out from the GMI community what data errors are impacting their analyses and educate said community on how they can help the archives to clean up the errors
- Continue to improve submissions and make it easier to do so. Inform the group established in point one when new submission capabilities become available. This point is currently ongoing.
- EBI/NCBI will work out the translations of metadata fields between the two systems and report back to the community. Publish paper on the standard. How to use it and what it can do (EBI/NCBI). On the GMI site describe the standard, and point to the archive descriptions and implementations (with version – at NCBI/EBI)
- Put together a publication on the standard, how to use it, and what it can do.
- Educate the community on how to access the submitted data (search, retrieval, find samples and sequences collected in 2014, for example). Inform them that some data is protected/not submitted and end users will likely need to contact the submitter in order to access the protected data.
- API to retrieve compliant data.
- Determine who is willing to contribute under the GMI concept and label their data with the GMI keyword (current contributors can be found at: https://www.ncbi.nlm.nih.gov/bioproject/?term=GMI[keyword])
- Contact these subgroups and determine what optional fields they feel are useful to add to the template in order to fill out the spec for a proposal to GSC (ex: ST = sequence type)
- Contact these submitters to get feedback on what reporting information they would like to receive – ex: programmatic way to retrieve the analysis results without having to go to the website/ftp and download them (GMI Reportng Std)

WG3: Analytical approaches

For WG3, the focus of GMI discussions was to assign different subgroups in order to enhance discussion. A survey done for inventory of methods used showed that the majority of GMI members

work on bacterial pathogens (>90%). One conclusions therefore was that a specific virus working group is needed to simplify discussions. Based on feed-back during the GMI meeting, this was indeed started, with the aim to compare performance of available approaches using existing datasets.

A pre-meeting survey was done to assess the level of expertise in the working group. Around 80% of participants had experience with cluster analyses and phylogeny and 63% with mapping of virulence markers or genes. Approximately 25% of participants were willing to present their data. Participants listed key outstanding questions that were used to prepare working group discussions.

Cluster detection:

- What tools or pipelines are available for the analysis?
- How can we evaluate and compare those tools?
- How can we reach a consensus on how NGS should be used for cluster analysis and comparison across datasets?
- How to take evolution into account.
- Which part of the sequence is most suitable for cluster analysis.
- Impact of different clustering methods on different types of pathogen.
- How do we build distributed comparison systems that are rapid, accessible approaches and software?

Phylogeny and phylogeography:

- What are methods and best practices, and how do we compare and evaluate them?
- Confidence in the reliability of the data produced.
- Dealing with recombinant regions.
- Conflicts between phylogeny and polyphasic taxonomy.
- Which part of the sequence is most suitable for phylogenetic/phylogeographic analysis?
- Robust ML methogs for phylogeny that scale.
- How do we place new sequences into a global phylogeny?

Virulence markers:

- Inventory and comparison of methods, best practices, how to reach consensus?
- Identification of resistance markers in metagenomics.
- Proper database of AMR genes and their SNPs.
- Marker choice.
- How to provide quality assured results? How to upload data to online databases?
- Is anyone else developing WGS attribution techniques?

Based on the pre-meeting survey, working group discussions were organized around each of the themes, to exchange current practices and hurdles, define approaches for comparing performance of methods, and decide on who would take the suggested topics forward.

Breakout 1:

A first breakout session was done with all participants of WG3 jointly. As there were several persons new to GMI, part of the discussion was dedicated on the vision of GMI before addressing the specific breakout questions. Some participants felt that there already was sufficient information for tool comparison, others were less convinced. The group concluded that an inventory of tools would be useful if targeted to specific stakeholders (clinical, food, wildlife, emerging), and combined with a review option to provide user feedback. Datasharing hurdles, the need for proper metadata, and the need for funding in order to further the ambitions of GMI were stressed, although participants also agreed that with increasing data availability more and more signals would become data-driven, triggering follow-up studies by other parties. Discussion also showed the need for pathogen based activities, also it was concluded that that is capturing common themes across pathogens is a stated ambition of GMI, and – while challenging- should be maintained.

Breakout session 2:

In the second breakout session, the participants were separated over three groups to provide input for a survey to provide a detailed map of available software and pipelines, come up with ideas for ways of comparing performance, and identify a group of colleagues to do this in the coming year. A first draft of an in depth survey was developed, which will be sent out before the end of the year. Ideas were developed for comparison of datasets, both for bacterial pathogens and for viruses.

WG4: Ring trials and quality assurance

The breakout sessions for WG4 were quite productive where new members volunteered to devote time to help the PT succeed. Two particularly noteworthy outcomes were:

1. An online forum would be created to allow pilot PT participants to post questions and feedback; the online forum would also be available during the full PT.
2. A new subgroup was started to develop a PT for viruses that includes a number of people.

More specific notes are below:

- Much discussion was held on how the culture and DNA samples were obtained e.g. using known reference material from e.g. ATCC or the food industry and whether they met the standards that are typically associated with a ring trial. However, it was concluded that they are satisfactory for this PT and for the moment.
- It was mentioned that we could also provide fastq files associated with a simulated dataset where the topology would be known. Due to the lack of volunteers to carry out those simulations and acknowledging that we can use the variance among methods as the criteria to evaluate them, no simulated dataset will be created.
- A discussion of the metrics used to describe the results included the following points:
    - The need to use platform (Illumina/IonTorrent) specific bioinformatic tools
    - Potentially look for plasmids also given that we will have closed PacBio chromosomes

    - Can differences be explained by GC content
- The WG4 should strive to make recommendations based on the results of the PT that include best practices to help standardize the process across laboratories.
    - A small group of participants would assess the final quality markers of the full PT roll out to set minimum criteria which potentially could fit all bacterial organisms.
- The dry lab component should avoid providing a lot of metadata as the interpretation of such data is beyond the scope of the PT at this time. As a result, questions about how to interpret the topology (e.g., how many clusters are there) should be removed from the questionnaire.
    - However, this may be something that could be pursued in the future – identifying the discordance among labs in the interpretation of the same topology/SNP differences in light of the same metadata.
- There were two participants of the pilot PT present and identified some difficulties with two of the datasets, which are being investigated to make sure such a problem does not surface in the full PT.
- Jorge de la Barrera Martãnez from the Institute of Health Carlos III in Spain will establish an online forum for participants of the pilot to post comments/questions. The forum will also be part of the full rollout.
- Anthony Underwood from Public Health England will assist in the analysis of the dry lab data and coordinate with David Aanensen to have whole genome MLST profiles created for the wet lab sequence data.
- It was agreed upon that we would remain with the taxonomic groups we are currently working on.
- We need to make sure that we have enough culture and DNA samples to send to future participants of the PT.
- Recruitment of laboratories for the full PT will begin once we have summarized the results of the pilot; discussions will also be had at that time to determine what the best avenues (twitter/message boards/word of mouth) are to advertise and recruit laboratories.
- WG3 is aware of the dry lab datasets we are working with and may use them in some of their future work.
- A subWG4 on virus PT was established. The sub-group discussed how to establish a virus PT, the components, leads and timeframe.
    - Initially, different organisms and matrixes were discussed as well as what to test for e.g. quality scores and / or diagnostics. It was concluded that the diagnostic part was the most important one and that focus should be on three tracks; human clinical, plant health and food virus. Furthermore, that the PT for the future should include two components; dry and wet lab running in parallel to the bacterial PT.
    - To speed up the process, the initial phase of the virus PT should focus on solely the dry component (analysis of four data files, Illumina and Life Technology).
        - A metagenomic file consisting of a plant matrix and plant viruses will be developed by Neil Boonham and Rachel Glover, Fera.
        - A metagenomic file consisting of a food matrix and food viruses will be developed by Nigel Cook, Fera.
        - A metagenomic file consisting of a human matrix and clinical viruses will be developed by Andreas Nitsche, Robert Koch Institute and Annelies Kroneman, RIVM
        - An artificial metagenomic file containing viruses will be developed by Mette Voldby Larsen, DTU.
        - The data files would be posted on GitHub.com by Rachel Glover, Fera.
    - The leads; Neil/Rachel, Nigel, Andreas/ Annelies, Mette will draft the accompanying documents for the different virus files based on the material developed for the

bacterial PT. This will be disseminated by Rene Hendriksen, DTU. The final versions will be developed in plenum by email exchange and conference calls. Andreas drafted a timeline which need to be discussed by the leads. DTU, Rene Hendriksen will ensure documents and invitations are posted on the GMI homepage.

- Due to complexity the virus leads; Neil/Rachel (plant), Nigel (food), Andreas, Annelies (clinical) would outline plans; 2 pager for each sector for PT with the target-viruses included in a matrix. The outline will be discussed in a future conference call.

WG5: Pilot Projects

For logistical reasons, WG5 discussion groups were merged with those of WG3. It was agreed that a renewal of focus in WG5 was required, and processes are ongoing to appoint new WG leaders.

Plant Health Satellite Meeting

A satellite meeting of the 7th Meeting of the GMI was held, sponsored by the UK Department for Environment, Food and Rural Affairs (Defra) and organised by (Fera), focussed on applying the GMI approach to the study of plant pathogens. Plenary sessions involved talks about applying NGS methods to plant and honey bee pathogens, and about dealing with large NGS datasets. Breakout groups then discussed the benefits, requirements, and overcoming obstacles to a GMI approach to plant pathogens.

Acknowledgements

The 7th GMI meeting was organised by many staff at Fera and FSA, with particular thanks owed to;

FSA:

Alisdair Wotherspoon

Fera:

Paul Brereton
Elena Fesenko
Sue Sainty
Robert Stones
Richard Thwaites
Edward Haynes
Elki Shaw
Hannah Austin

The meeting report was compiled by Edward Haynes, FSA-Fera joint fellow in molecular epidemiology.