Global Microbial Identifier

Report on the 5th meeting 27-28 February 2013, Copenhagen, Denmark

Global Microbial Identifier

Global Microbial Identifier

Report of the 5th meeting

The Global Microbial Identifier GMI is currently an informal global, visionary taskforce of scientists and other stakeholders who shares an aim of making novel genomic technologies and informatics tools available for improved global patient diagnostics, surveillance and research, by developing needs- and end-user-based data exchange and analysis tools for characterization of all microbial organisms and microbial communities.

Vision of GMI

The GMI vision is to shepherd analysis and sharing of genomic data in real time that enables faster, cheaper and more accurate microbiological identification, tracing, disease control and epidemiological and biological research; locally as well as globally. The use of new whole genome sequencing technology in combination with global sharing and analysis of data will complement and partially substitute traditional microbiology and enable a giant leap for health systems in all countries, especially developing countries. GMI will also open a new avenue of collaboration between different sectors in health, agriculture and environmental research and management.

GMI mission

The GMI mission is to build a platform linked to an interactive global network of databases for standardized identification, characterization and comparison of microorganisms through the storing of whole genome sequences of all microorganisms and provision of analytic facilities and standards for all. The database may be used by different end-users for the identification of all types of microorganisms, both for single clinical tasks (simple microbiological identification) as well as for national and international public health surveillance and outbreak investigation and response. The databases will include all genera of microorganisms: bacteria, viruses, parasites and fungi, and be accessible through user-friendly interfaces for end-users in academia, industry and government (e.g. clinicians, veterinarians, epidemiologists, microbiologists). The use of the platform and linked databases would significantly improve health systems, as well as systems aiming at a safe food supply, and environmental control systems.

Who we are

The GMI visionary taskforce is composed of approximately 200 experts from at least 30 countries, including clinical-, food-, and public health microbiologists and virologists, bio-informaticians, epidemiologists, representatives from funding agencies, data hosting systems, and policy makers from academia, public health, industry, governments. The Initiative was started in September 2011 at the first meeting convened in Bruxelles. During the 4th meeting in Bethesda in September 2012 an interim steering committee (SC) was formed and it was decided to create a web-page and initiate a process leading to a more formalized way of moving forward. Visit our project website at www.g-m-i.org to find out more about the project, summaries of previous meetings, and useful background information.

This is the report of the 5th meeting taking place in Copenhagen, February 27-28th, 2013.

Agenda for the 5th meeting

Wednesday, February 27
08.30 - 09.00: Registration
09.00 - 09.15: Official Welcome by Provost Henrik Wegener
09.15 - 09.30: Welcome and introduction to the initiative 'Global Microbial Identifier'
Alisdair Wotherspoon, FSA, United Kingdom
09.30 - 09.45: Process of the meeting / 'much too creative'
09.45 - 10.15: Status and perspective of each working group – 1 (WG chairs)
10.15 - 10.30: Coffee break
10.30 - 12.00: Status and perspective of each working group – 2 (WG chairs)
12.00 - 13.00: Lunch
13.00 - 15.00: WG themes - 1
15.00 - 15.20: Coffee break
15.20 - 16.30: WG themes - 2
16.30 - 18.30: Transport and reception at the Copenhagen City Hall
18.30: Transport to the hotels

Thursday, February 28
09.00 - 10.40: Action plan - part 1
10.40 - 11.00: Coffee break
11.00 - 12.00: Action plan - part 1
12.00 - 13.00: Lunch
13.00 - 14.20: Action plan - part 2
14.20 - 14.40: Coffee break
14.40 - 15.50: Overall Road map
15.50 - 16.00: Presentation of digital platform
16.00 - 17.00: Future aspects / Frank Aarestrup, DTU, Denmark

Opening remarks and introduction

The meeting was opened by the Provost of The Technical University of Denmark, Henrik C. Wegener. He highlighted the major perspectives in implementing and using novel technologies for microbiology potentially creating the largest change in microbiology in >100 years.

Member of the GMI Steering committee Alisdair Wotherspoon, Joint Head of the Chief scientist team from Food Standards Agency, UK, addressed the audience and gave a summary of the initiative and the vision and objective from the point of the steering committee. He also highlighted the importance of working together in a global perspective and to think alternatively to progress through any problems which emerge on the road to success.

Main purpose of the meeting

The main purpose of the 5th meeting was to develop a number of roadmaps for the future. In addition, decisions on the organizational name and a process for the future work and structure were to be discussed.

Meeting organization

Prior to the meeting the SC had formed five working groups (WGs), each with a chair and a co-chair (mentioned below). All participants to the 5th meeting were requested to register to a specific WG and the WGs should prior to the meeting initiate e-mail or telephone conference discussions and exchange

information within their area. Based on this initial exchange of information a number of themes to be discussed during the meeting were identified.

The meeting itself (agenda above) was structured with very short presentations of the status and work to be done in each WG followed by most of the time in the specific WGs working on the themes identified prior to the meeting. These themes were discussed and specific actions and milestones identified, further discussed and drawn into an overall roadmap for each WG.

Status presented by WG chairs and in documents circulated prior to the meeting

The power point presentations are available on www.g-m-i.org.

Work group 1: Political challenges, outreach, building a global network and funding

Long-term vision:
The Political Working Group (WG1) will develop a long-term plan to shape political level involvement in GMI development at the global, regional and national level.

Short-term vision:
The first goal is to establish a functional link to political level decision makers in several countries or regional, international organizations. The second major goal is to initiate a coherent system for international discussion of the relevant themes listed.

Numerous **political themes** to be addressed, some mentioned here:

Global health diplomacy: Building the expectation that providing national data to an international system similar to that envisioned by the GMI is in accordance with the WHO International Health Regulations, which require all signatories (i.e. all 194 WHO Member States) to share relevant data about 'public health events of international concern'.

Coordination between different sectors: Stakeholders and governments need to find ways to collaborate and agree on issues such as standardization, ownership and security of sensitive data.

Sensitivity of metadata: There will need to be some sort of international agreement about how certain metadata can be included in a protected way, and who will have the right to use such data.

Open access: To make full use of the capacity of WGS, a global and open-access database of genome sequences has to be built. This will only be possible through close cooperation internationally, across sectors (*e.g.* human, animal, food and environment), as well as, between different stakeholders (*e.g.* commercial and not-for-profit). Systems for quality assurance need to be considered also.

Sharing of strains over borders: In recent years, developing countries have been disappointed when pathogenic strains were shared with the global community. In some cases the consequence has been that developing countries have made money out of making and selling vaccines, diagnostics etc. related to such strains.

IPR: For some companies and some governments there is a perceived need to maintain DNA sequences as a patentable commodity; therefore an open data-base could present problems.

Ownership of initiative: In a case where a GMI system were to be developed only between OECD countries, there would be little buy-in from developing countries, who need to have buy-in from initiation of process.

Information Technology (IT) and Internet needs: The backbone of the database will have to be robust, such that enormous datasets can be stored, sent across the world and compared in real-time. This will require a major investment in an IT infrastructure and requires cooperation between the world's leading soft- and hardware engineers. Additionally complex is the question of how 'raw' the genomic data can be to perform a diagnosis. Algorithms to handle data in different states of completeness have to be developed, which also may require investments by the commercial sector.

Funding: Although a number of initiatives can be initiated with skeleton funding at some stage major investment will be needed to finalize construction of a real system. Likewise there will be a need for significant funding for the continued development and maintenance of the system. There is experience from other sectors (e.g. cancer genomic databases) where countries have made specific agreements of funding a common database, then accessible for all funders.

Work group 2: Repository and storage of sequence and meta-data

General Goals

1) Minimum Data for Matching (MDM), consisting of reads and minimum metadata, should be deposited, and made globally and universally accessible as soon as available.
2) 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 3rd party.
3) Ideally, any MDM provided for purposes of searching the GMI databases should immediately also become a deposit available for searching by later submitters.
4) Any matches from the MDM search should be reported to searcher and to the relevant GMI Participants.
5) The data layer is provided by The International Nucleotide Sequence Database Collaboration INSDC and is therefore both international and public.
6) 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.

Implementation 2013

1) NCBI and EBI agree to accept deposit of MDM from participating laboratories. DDBJ has not been included.
2) NCBI and EBI agree to exchange MDM within 24 hours of receipt at either site.
3) NCBI offers an analysis pipeline that will:
    a. Based on molecular methods (currently K-mers and SNPs), compute the relationship of MDM to other isolate genomes in the database. If there is an accepted standard for clonality determined by the GMI, the system will report whether MDM is “clonal” (potentially part of the same outbreak) to previously submitted isolates.
    b. Report the matched genomes and show a tree of their relationships to the submitter and to other relevant GMI Participants.
    c. Update and maintain a tree of such relationships and make it publically visible to anyone.
    d. NCBI also has the capacity to assemble and annotate the submitted genome and is willing to offer that as well if appropriate and useful. Assembly is required for the SNP analysis, so NCBI will automatically assemble any genomes for that purpose.
4) If EBI or some other designated group chooses to offer a similar function as part of GMI, NCBI offers to work with EBI to compare and share the conclusions from such analysis and maintain a consistent view of this.

Discussion Points for the WG

1) In this model, MDM is part of INSDC, is fully public, distributed internationally, and available for research and analysis by any scientist in the world.
2) The GMI analysis and notification service offered by NCBI depends on the data in (1), but is focused specifically on serving the needs of the international food safety community, specifically GMI participants.
3) Sequence data from sources other than GMI participants would be included in the dataset used for the GMI analysis pipeline.
4) The expectation is that GMI participants using the GMI analysis pipeline would be depositing data and triggering the analysis as effectively a single process. Other relevant GMI participants would automatically receive the analysis reports as well.
5) However, since the data is public, in principle some other group could use the same data and create a rogue site which searched GMI data, but did not contribute the MDM to the global corpus. It is suggested that if enough large players participate in GMI, in the long run others would choose to join the majority, rather than going their own private way. By making the data public, science can proceed unfettered to apply creative new approaches to analyzing world health data, no matter what country they come from, and the world benefits.
6) It is also possible to imagine a much more restrictive approach, but this would exclude using INSDC for distribution and archiving, and would greatly reduce the scientific scrutiny applied to the data and analysis methods. While there may be some reasons why some players might prefer such an approach, it is expected this should be avoided or very much minimized.

Jim Ostell initiated his presentation with an analogy to weather forecasts: Models of the hurricane Sandy's movement before it hit New York were on the News, even though the models were not perfect. People appreciated the models just the same.

- We already have collaborations on data sharing between USA (NCBI), Europe (EBI) and Asia (DDBJ). Four terabytes of data comes in a day, while 26 TB/day are downloaded.
- Metadata structure: What (sample name, organism, strain, 1a= clinical/host-associated 1b = environmental/food/other), When (collection date), Where (place or lon/lat), Who (Collected by). If it is sensitive data, only the "Who" is allowed to see, e.g., the "Where".
- Jim provided example on Salmonella monitoring: Kmer methods highlights that there is a problem - a lot of similar Salmonella strains are popping up (Montevideo outbreak). From here, the Kmer method is not of high enough resolution, and SNP analysis is needed. All can be done within 4 hours.
- Commercial company is working on developing software for automatic upload to the system. Otherwise very resource-intensive to upload to repository.

Work group 3: Analytical approaches

Long-term vision:
To provide guidance for the development of analytical tools for optimal positioning and functioning of the GMI platform.

Short-term vision:
The GMI initiative aims at bringing together scientists, public health experts, policy makers, etc. to develop a global platform (database, linked databases) that facilitates the application of next generation sequencing technology (NGS) in research, clinical and public health settings worldwide. This is an ambitious outlook, and the work in the GMI working groups aims to develop a work plan towards reaching this overall objective.

Themes for the WG:

1. 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.

2. To map current analytical options and solutions against the needs of GMI end-users.
3. To identify possible R&D and implementation gaps.
4. To identify projects that may fill those gaps.
5. (If necessary) to develop pilot projects to fill those gaps.

During her presentation Marion Koopmans stressed that we must focus on speed, robustness, and the end-user. The output reports must be simple.

- We should consider: Who are the stakeholders and the end-users and what are their needs 5-10 years from here?
- We are here to develop roadmap and free release of data – what would it take to make people comfortable sharing data?
- There are already a lot of analytic approaches going on – a lot to learn from, we are not starting from scratch.
- We have to broaden up the analysis and not only focus on bacteriology. Virologists are getting involved, now we miss some parasitologists. Challenges differ depending on bacteria<>virus<>parasite, but there are also things in common.
- How should we decide which approaches are the best. Should we start comparing pipelines. How do we keep it up to date?

Comments from audience:

- A timeframe of 5-10 years is too much. Things are moving much faster.
- Data sharing is no problem when it comes to the sequence, but much more sensitive when it comes to metadata.
- Analytical support is one thing, but treatment guidelines are very problematic. Then you might be legally responsible.
- The weathercast forecast analogy was very good - we need people to accept that we are also just providing forecasts.
- As users, we need guidelines on which tools to use.

Work group 4: Ring trials and quality assurance

Long-term vision:
That all laboratories globally conducting NGS on bacteria and vira to the highest degree of quality.

Short-term vision:
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 vira.

Themes for the WG:
Infrastructure: How can we build an infrastructure within the partners of GMI that has the capacity to undertake the facilitation of the proficiency testing.

Reference material: How should we develop or provide the reference material and documents needed to initiate the proposed pilot proficiency test scheme. Disseminate reference material to enrolled laboratories. To adjust the reference material and documents as well as the analysis based on previous experiences.

Genome analysis: How should we conduct the analysis of submitted genomes.

Proficiency test: How should we execute fully operational proficiency test based on bacteria and vira to GMI members. To evaluate RNA purification methods / protocols and pilot sequencing on multiple platforms to initiate the proposed parallel viral pilot proficiency test scheme.

During his presentation Rene Hendriksen stated that this group is a bit further (more concrete) than the other groups. Since the DC meeting in Sep. 2012, a working group has by e-mails and telephone conferences discussed how to make a proficiency test. To date, four teleconferences have been conducted to discuss how to approach establishing the PT and what reference material to include. Additionally, a mini review has been completed to assess what quality markers to include in inter-laboratory comparisons based of whole genomes by other scientific groups. The data published have assessed parameters related to platforms / technologies and data analysis. Those platform specific parameters included DNA input requirements / Library preparation, Comparison of read technologies / Read length assessments, Platform specific errors - no of Error-free reads rates. Sequence coverage depth and GC bias, and assessment performance metrics at lower coverage. The following parameters for assessing data quality were included the studies such as Comparison of assembler algorithms, Anomalies in assembly accuracy through eg. N50, Coverage, Contig size, etc., Ability to single nucleotide base variant calling, Detection of indels and differences in size, Technology-dependent variants - Unmapped regions / missed variants, and Assessment of mappability to specific genes.

The discussion in the WG has come to the conclusion that the biggest problem will be which parameters to test for. Selection of suitable reference material (bacteria) to evaluate quality of sequencing and the platforms has also been discussed. Three groups: Salmonella, Vibrio (not cholera) and campy. Also discussed was selection of data for evaluating bioinformatics pipeline. We need virologists in this working group. This is more practical work there is labor in it. Although discussions have already been made in this WG they are only meant as a starting point for further discussion.

Based on the objectives, the WG decided to target two goals;

- To initially organize a pilot proficiency test for the work group participants.
- To secondly offer this to test to GMI members working with both bacteria and vira.

Prior to the meeting, five themes were identified for the meeting discussion:

1. Infrastructure: To build an infrastructure within the partners of GMI that has the capacity to undertake the facilitation of the proficiency testing.
2. Reference material: To develop or provide the reference material and documents needed to initiate the proposed pilot proficiency test scheme. Disseminate reference material to enrolled laboratories. To adjust the reference material and documents as well as the analysis based on previous experiences.
3. Genome analysis: To conduct the analysis of submitted genomes.
4. Virus experiences: To evaluate RNA purification methods / protocols and pilot sequencing on multiple platforms to initiate the proposed parallel viral pilot proficiency test scheme.
5. Proficiency test: To execute fully operational proficiency test based on bacteria and vira to GMI members.

Work group 5: Pilot projects

Long-term vision:
The Pilot Project working group (WG5) will develop discrete projects that provide progressively challenging technical demonstrations of NGS for local and global tracing of pathogens within the GMI Network.

Short-term vision:
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 PPWG 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.

Themes to cover at Feb 2013 meeting:

1. Working group governance and communications structure
2. Define synergisms between the different PPWG members
3. Discussion of topics and purpose of pilot projects/demonstration/exercise
4. Development of a preliminary draft mission statements and road map
5. Define mechanisms by which PPWG interacts with other WG's
6. Examination of precedents for pilot projects or demonstration exercises
7. Establish action plan for work prior to next GMI meeting
8. Collect a list of ongoing or starting pilot project on different areas of NGS applications

Outcome and conclusions of the WGs

Work group 1: Political challenges, outreach, building a global network and funding

The major issues debated with a view of preparing a roadmap and suggesting specific action items in this area were:

a) Global health diplomacy
b) Ownership of the initiative
c) Open access, sensitivity of meta-data and IPR issues
d) Information technology and internet needs
e) Funding

In the discussion the following important points emerged:

- It is extremely important to provide a clear description on the vision, mission and intentions of the GMI initiative. This should include a clear description of the governance structure and this should be clearly communicated to scientists, policy-makers, politicians and the general public. It should be made clear that GMI is a support to and not in any way intended as a replacement of current efficient public health structures.
- It is also important to very soon develop clear and short advocacy papers for specific end-user groups, as well as a formal publication on legal implications of a potential GMI construct.
- It was considered essential that a broad stakeholder analysis together with the development of a model GMI framework informs a broader GMI strategy, which should also include a strategy for outbreak response.
- It was likewise considered important to clarify both risks and benefits related to the construction of a GMI framework.
- A roadmap for resources needed for the initiative should also be developed.

A roadmap with action items and milestones was developed and is presented in table 1. Annex 1 and 2 present the action plans at a more detailed level.

Work group 1 will continue to develop these plans leading up to GMI-6.

Work group 2: Repository and storage of sequence and meta-data

Discussions revolved around informatics relating to the provision in the public domain of a global and rapid pathogen genomic surveillance system that will support such activities as outbreak detection, tracking, modelling, prediction, clinical decision-making and scientific research. While many issues were tackled, several themes were discussed at length and are reported here.

Given the scale and ambition of GMI, the recognition, reuse and repurposing of existing informatics infrastructures was seen as being critical to success. Amongst these existing resources is the International Nucleotide Sequence Database Collaboration (INSDC), provided by its partner institutions, NCBI, EBI and DDBJ. This long-standing collaboration unites the sequence databases to provide a

globally comprehensive data resource for genomics data. Importantly, the infrastructure exists at these institutes for the global storage, organisation and dissemination of comprehensive genomics data and is offered freely for leverage by GMI. In addition, the group identified existing standards and ongoing standardization work that relate very closely to the needs of GMI data reporting. Here, the Genomics Standards Consortium (with the MixS family of standards), members of the INSDC and the US FDA are already aligning on a set of minimal reporting standards that will be appropriate for GMI data. This work is also offered freely for GMI usage.

An early attention to prototyping and proof-of-principle projects was preferred by the group. This arose from an awareness that commitment to GMI, in particular to data provision to the repository, will be greater when the benefits of GMI are clearly demonstrable. The group decided that a 'founder' or 'pioneer' consortium of GMI data providers, defined as those groups willing to dispatch at least one GMI genome data set into the system, would provide data to seed analysis and demonstrate utility.

The establishment of appropriate and workable standards was a particular focus of discussion. The group felt that there was a distinction between 'reporting' information - those data and metadata elements that are essential for utility and must be available to all in some way when GMI data are being shared - and 'presentation' information - includes all reported information but may also include additional inferred elements. This second component to the GMI standards is expected to bring requirements for such fields as summary information/indices for sequence quality metric matching and sequence similarity look-ups respectively and may bring requirements for deeper inferences, such as pathogenicity potential and host specificity.

Issues relating to formal languages used to described genomic data and their generation were discussed in the context of reporting and exchange standards. While opinions were many on the opportunities that formal ontologies and other structured language provide for genomics data utility, a feeling that pragmatism must prevail in order for data to flow was evident. In particular, it was felt that unstructured descriptive language would be appropriate in the early stages of GMI, perhaps leading to harmonization and more formal ontology development activities as data flow becomes more established.

The group felt that the level of genomics data of interest, and the level that should be the primary focus for the data reporting and sharing, was raw sequence data – unassembled read information. While assembled sequence and deeper functional information should be supported in the data reporting and sharing layers of GMI, GMI data providers should see these levels as optional. Since centres will exist that will implement pipelines to provide these kinds of interpretation (indeed, NCBI already has such a system running), there is no requirement that the reporting/sharing layers are provided by the same centers that provide these analyses.

Given the taxonomic and environmental breadth of GMI – covering all pathogenic taxa from patient, animal, food and other environmental settings - it was clear in the discussions that a multi-layer system for contextual metadata (such as host disease status, location and environmental description) was required. For those metadata elements that must travel with genomic data and must be available for all GMI data to render utility to the data, there needs to be some central organization, and this set of fields must be defining of the minimal reporting standard. For less generic descriptive information, specialist and more dispersed resources must connect into the centralized system. Examples will include private patient-related information, veterinary health records relating to zoonoses, environmentally specialized descriptions, local information needed to interpret georeference information (such as floor plans in the case of hospital infections), etc.

The group identified the need to provide tools and services to render useful the core data. Low-level tools, such as those that support the discovery of data sets in the GMI collection, will be provided by INSDC partners (at the least). Higher-level tools, such as for taxonomic identification and typing will

also be provided – US FDA and NCBI are building these tools while other institutions may follow. Further tools for outbreak detection and functional analysis must also follow.

It was clear from the discussions that the user community for the GMI repository is large and varied. Indeed many different stakeholder groups can be identified, each of which will have different perspectives on GMI, including public health authorities, clinicians, vets, those in the food industry, epidemiologists, researchers, pharmaceutical companies, etc. It is important that the infrastructure provisions for this breadth. While not all of these stakeholders will be supported immediately with specialist interfaces, a modular architecture in which programmatic interfaces can be built upon freely by the community will be of value here.

Much of the work to be done to deliver a core GMI repository relates to the reconfiguration of existing infrastructure. One area of attention for which new development (and hence new resources) must be targeted is that of data capture from the extremes of the global network of GMI data providers. While good mechanisms exists for those facilities with good network connectivity and informatics infrastructure that already allow data to flow, the growing dispersal of small-scale sequencing capacity to ever more remote and less well connected locations will need new software engineering activity. Furthermore, an ongoing need clear to the group was that of the development of analysis tools and interfaces that will sit upon the GMI repository to make it ever more useful.

A roadmap with action items and milestones was developed and is presented in table 2. Annex 1 and 2 present the action plans at a more detailed level.

Work group 3: Analytical approaches

In the discussions the following points emerged:

There was overall agreement that we need to move forward in a collaborative way.

- We need to actively figure out how to engage more potential end-users in clinical and public health laboratories, particularly from developing countries, because their needs are not entirely clear yet. We decided on surveys, and engaging some social scientists for this.
- There is a big area to span, and we may need for a two (or more)-tiered approach: the groups that are " already there", and groups that expressed an interest on working together to get things implemented. There is some tension between those because of speed and capacity differences, possibly also some interests. This is not something we can change, but need to realize and somehow work with (related to the point above).
- It is important to share what did NOT work and have an activity around that.
- When discussing priority diseases, it was clear that this topic can be divided in themes with different data analysis needs. Lists of diseases to prioritize already exists on, e.g., WHO's website. But of course it might differ depending on region of World.
- We need to discuss and set guidelines how to deal with leadership in working groups from commercial partners. It might be preferred not to put someone from the private sector in charge, unless unavoidable, to not get into trouble with our wider agenda (global, sharing, WHO etc.). Some group discussions were dominated by this issue.
- There are quite a few examples for pilot projects. It was suggested to review the workshop summaries and see if we can actively identify some that include the not-so-upfront participants. (There were participants less fluid in English, who had difficulties in jumping in some of the discussions).
- Common theme: the current setup of uploading sequences was mentioned by many as a huge bottleneck. People are not only afraid to share data because they might miss out on publication opportunities, but uploading of data to central repositories is also very time consuming and tedious., It would be advantageous if turn-key solutions could also be developed for this step. And once data is shared with central repository it should not be necessary to re-upload for further analysis with additional tools. It would be advantageous if a possibility existed to upload the short

reads directly from the sequencing platform to, e.g. NCBI's SRA (Illumina is working on providing this solution). More in general: the ICT is not there yet. Also, a sequence database alone is not enough, there is a need for translational activities that facilitate work form clinical and public health laboratories. We need a central site for which to upload the data. After preliminary analysis, which will typically be species identification, the data should be directed to expert groups for strain identification and characterization. The central site should direct the user to the relevant further analytic tools.

- There is also a need for a site where genomic, phylogenetic, geographic information is integrated to aid epidemiological analysis. There are currently no tools that can be used for this.
- Methods need to be simple, but be aware that black-box solutions should be documented and some users might want to adjust parameters. Methods also have to be user-friendly and standardized. Standard Operating Procedures are needed. Probably one per bug will be necessary.
- There is also a need for a quality control of the incoming data. Initially, we must define what good quality is and which parameters should be used to evaluate it.

A roadmap with action items and milestones was developed and is presented in table 3. Annex 1 and 2 present the action plans at a more detailed level.

Work group 4: Ring trials and quality assurance

The WG had identified five themes but due to the size of the WG, two themes; reference materials and quality markers were selected for discussion during the meeting days. Along the theme discussions, it became evident that both groups discussed more or less the same issues including all aspects of proficiency testing (PT) which is why the groups decided to merge the themes into one overall theme; planning a PT.

In preparing the roadmap for PT, several issues were elaborated such as:

End users:
We need to clarify who the end users are for the PT. This will be important to address as there are different needs for the individuals performing the actual laboratory testing compared to those individuals who perform the analysis, upload and report the results. The PT could target diagnostics/typing assessing variant calling, sequencing assessing DNA and library preparation and different platforms, and raw read assembly data quality assessment using various quality markers. It was decided develop a questionnaire to survey the potential end users of the PT.

Quality markers and target organisms:
The WG had prior to the meeting, selected a few target organisms but during the state of art presentation many questioned if those organisms were be the best candidates as the end users weren't yet identified. The same issues were related to the quality markers. Due to this, the WG decided to also include questions about what target organisms and quality markers to address in the questionnaire.

Documentation and guidance:
In order to conduct the PT in a smooth way, the WG suggested developing a web site where all documents etc could be posted for the end users. This would also include some kind of a wiki where users could post questions for assistance. Likewise, an upload portal will be developed for users to upload data files for the quality assessment.

Reference materials:
Originally, the WG had discussed to send out DNA and sequence data for PT. However, this was thoroughly discussed in terms of the overall purpose – what to measure. It was therefore agreed to include three reference material matrixes in two components; component 1 (assessment of sequencing

and quality markers): culture and DNA and component 2 (variant calling, phylogenetics): data from a flow-cell.

Accreditation and certification:
The WG did not foresee the PT moving into certification or accreditation of the users. This will not be within the objective as the PT will be provided as a self-correcting / evaluation process.

A roadmap with action items and milestones was developed and is presented in table 4. Annex 1 and 2 present the action plans at a more detailed level.

Work group 5: Pilot projects

The group worked on 5 major points for which milestones have been developed:

1) Communication and Governance
2) Previous Pilot Projects
3) WG interactions
4) Synergisms
5) New Pilot projects

In the discussions the following points emerged:

- A WG steering committee appears to be necessary to drive the development of the WG and follow the different milestones defined.
- There is a need to create a web-site for sharing ideas, information, and suggestions for pilot-projects. This could be done though e.-mails, web-pages and newsletters.
- A procedure for initiating pilot-projects should be established. This should comprise a process for selection, performance, analysis, and publication of the pilot projects.
- One way might be that the pilot projects can be submitted from any GMI member to the WG steering committee. Then, the projects are sent around to all GMI members for consideration. If enough will participate, pilot project can be initiated at best with group of leaders that will drive it.
- An important point is that we the GMI is named in resulting publications to form a “GMI” brand as a “logo”.
- A global pilot project on *Listeria* has been defined as first possible pilot project.
- There is also a great need for sharing bioinformatic and visualization tools.
- A review on already ongoing projects should be performed and shared between all participants.
- There is a need for determining/showing why it is worthwhile to share data. It might be good for public health, but bad for individual researchers. Trust is very difficult to obtain in open source. One option might be to ensure that project-leaders/participants keep sufficient of the meta-data to ensure control and QC.
- There is a need to perform a cost-benefit calculation of using WGS in the clinical/global setting.
- There is a need to go from retrospective to real-time.
- Since we need to build from existing budgets, it should be examined whether we can combine data already available.
- A decision on target pathogens should be taken. Global pathogens (TB, influenza, HIV) or local (food borne MRSA).
- Also for the pilot projects there is a large need for agreement on meta-data.
- The end-users for each pilot should be defined.
- The same ontology should be used.

A roadmap with action items and milestones was developed and is presented in table 5. Annex 1 and 2 present the action plans at a more detailed level.

Future aspects and final remarks

Following the work in the WGs there was a vote on a logo for GMI and a final decision on the name of the initiative. The logo used for the present report was chosen and it was unanimously decided that the name should be “Global Microbial Identifier”. There was no discussion on the future process for the structure and rules of engagement for GMI, since that was an outcome of the work performed in WG 1.

It was considered highly important the following meetings should build on and further develop the work conducted during the 5th meeting, including updates of the roadmaps.

The greatest challenge in the next year was considered to be keeping the momentum going, also when we go back to normal work.

We need to make sure that the work groups continue their work. This is the responsibility of the chairs and co-chairs. They should keep the ball rolling between the meetings.

While there despite some concerns is almost general consensus regarding the benefit of storing sequence data in central repositories, there are major concerns regarding meta-data. There is a major need to discuss which meta-data to store where and what should be made publicly available and when. There is also a need to exemplify how this can be done in a safe way.

There is a major need to focus on end-users needs and priority organisms. This was discussed in several WGs and included on different roadmaps. This should be an area of high priority.

Furthermore, there was clearly a need to make existing knowledge and already available analytic tools easily available on a central web-interface, an issue that also was included in several WGs.

Another important issue raised was how we can keep track of what is going on? One possibility is the GMI website, but we also need something more interactive for the work groups. Google groups were suggested. At CDC access is, however, blocked to google groups and similar sites, and we need to be aware of that. A novel platform for interacting realtime was presented: Innovation Embassy – Cocreatorx. It enables matchmaking, sharing, IPR rights, newsfeeds, alerts, notifications, exhibitions, crediting. The platform is not quite finished yet, but will be in a matter of weeks – months.

Next meeting(s)

It was suggested that the next meeting should be in California at UC-Davis in September 2013, followed by the 7th meeting in UK in 2014. The desire to also have meetings in Asia, Africa or South America was expressed. The SC will soon take decisions on the future meetings.

TABLE: GMI WG1 ROADMAP

Date	Milestone	Responsibility
2013 Q2	Map and engage stakeholders, catalogue regulations and international agreements	
2013 Q2	Define GMI management funding group	
2013 Q3	Advocacy paper for end-users	
2013 Q4	Agreement on organization form and communication strategy	
2013 Q4	Develop minimum optional metadata model	
2013 Q4	Risk/benefit. Identify / develop communication strategy to industry, academia, governments	
2013 Q4	Resource needs report. Coordinate funding applications	
2014 Q1	GMI should be known by 65% of professionals	
2014 Q1	Present stakeholder analysis and recommendations	
2014 Q2	Develop approach to release data	
2014 Q2	Overall strategy involving global funding	
2014 Q3	GMI information points in 50 countries	
2014 Q3	Technical expert MTG	
2014 Q4	Survey model acceptance	
2014 Q4	Get money	
2014 Q4	Risk / benefit. Stakeholder outreach to illustrate benefits of open access.	
2015 Q2	Publication on legal implications of GMI	
2015 Q2	Global level political MTG	
2015 Q2	Review and develop communication strategy for outbreak response	
2015 Q4	Side event at governing bodies (WHO, OIE, FAO)	
2015 Q4	Global agreement	
2016	Resolution at governing bodies (WHO, OIE, FAO)	

TABLE: GMI WG2 ROADMAP

Date	Milestone	Responsible
2013 Q2	First flow of data into GMI repository from 'founder group'	NCBI/EBI
2013 April	Discussion of GMI and MixS standard harmonization at GSC15 meeting	NCBI/EBI
2013 May	Discussion at INSDC meeting of introduction of two new tags for pathogen data, to indicate 'provided as part of GMI' and 'compliant with GMI reporting standard'	NCBI/EBI
2013 Q3	GMI reporting standard	NCBI/EBI
2014 Q4	Working repository infrastructure, including first prototype GMI data discovery programmatic interface and generic web interface	NCBI/EBI
2015 Q1	GMI presentation standard	NCBI/EBI
2015 Q2	Feedback from GMI analysis groups to indicate further information to be included in GMI presentation and/or reporting standard	NCBI/EBI
2015 Q3	Enhancements to programmatic interface and user group-focused web interfaces, including support for updates	NCBI/EBI
2015 Q4	Specification of a 'GMI Toolkit' – a set of analysis tools and services to be made available as part of GMI	NCBI/EBI

TABLE: GMI WG3 ROADMAP

Date	Milestone	Responsible
2013 Q1	Survey for methods in use and data to be stored.	
2013 Q2	Establish WG between academia and industry	
2013 Q3	Common pipeline to prepare data to be shared	
2013 Q3	Compile BoD estimates	
2013 Q3	Survey to ID enduser needs prepared.	Fiona Brinkman, Simon Fraser University, Canada?
2013 Q4	Milestone forum created	DTU
2014 Q1	End-user needs identified.	Fiona Brinkman, Simon Fraser University, Canada?
2014 Q1	Reports on: Tool availability and gaps. Previous successes and failures. Epidemiology and bioinformatics integration.	
2014 Q2	Regional priority of organism database created.	
2014 Q2	GMI session on tool availability and ontology.	
2014 Q2	Decision tree for standardized sample preparation.	
2014 Q3	Model for genotype to phenotype prediction.	DTU
2014 Q3	General SOP for pilot projects	
2014 Q3	12 countries upload to public databases.	
2014 Q3	Friendly user interface for analytic tools.	DTU
2014 Q3	Genomic diagnostic traits identified.	
2014 Q3	Top applications where NGS is relevant identified.	
2014 Q4	SOP for novel pathogen discovery	
2014 Q4	Approved ontologies	
2014 Q4	All NGS upload to central repositories.	
2014 Q4	Species and strain characterization running	DTU
2014 Q4	Industry buy-in and shared ownership.	
2015 Q2	Centralized repository for novel strains.	
2015 Q2	Databases and outreach modules linked.	
2015 Q4	Transparency of methods used.	
2015 Q4	Interpretation to public warning running	
2015 Q4	Data standards implemented.	
2015 Q4	All microbial ID is digital.	

TABLE GMI WG4 ROADMAP

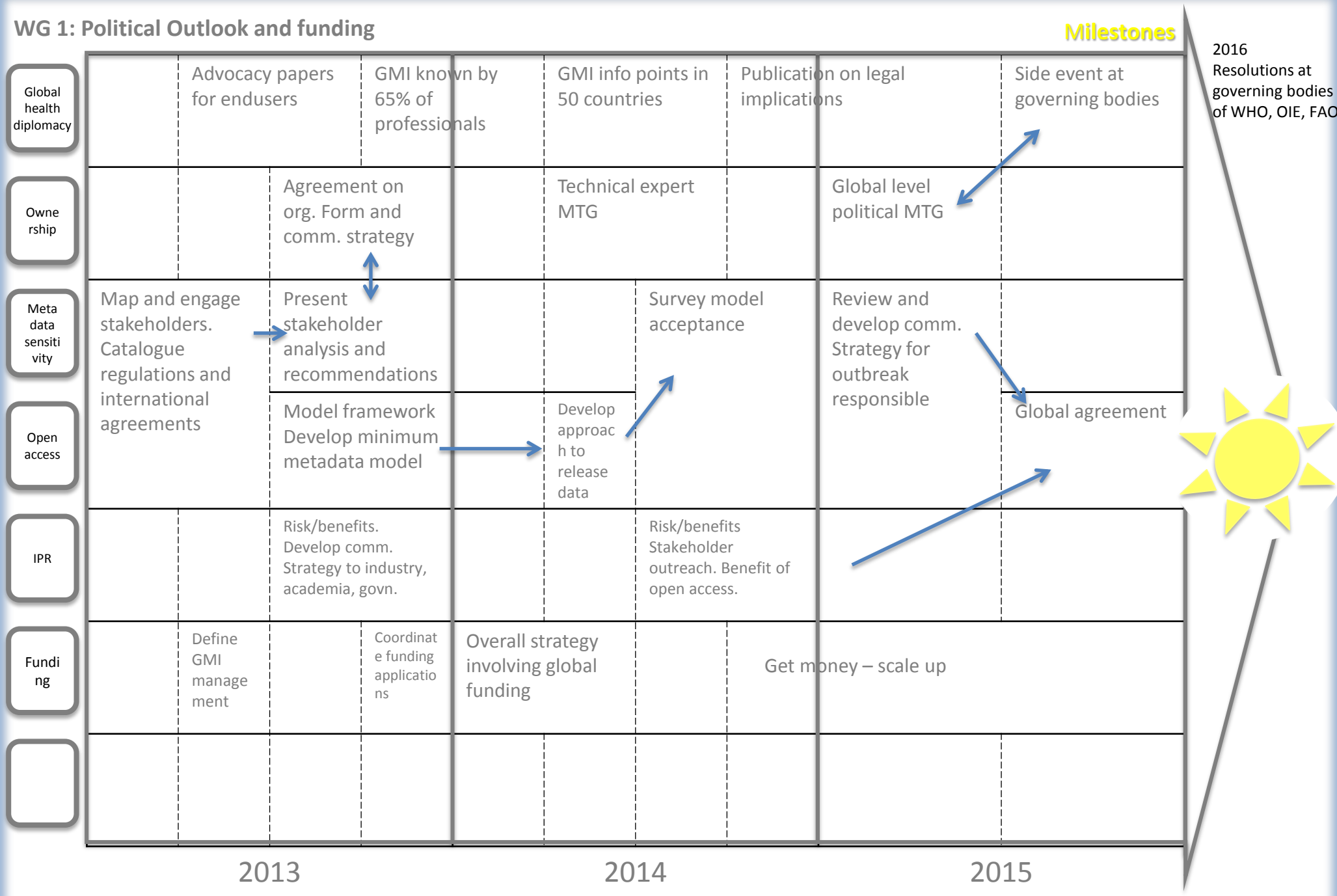
Date	Activities	Milestone	Responsible
2013 Q2	Develop a questionnaire to identify end user, target organisms and quality markers Create a website / wiki		FDA
2013 Q3	Develop a submission portal	Perform the survey	FDA
2013 Q4	Assess the outcome of the questionnaire Identify target organisms and quality markers Develop documentation, instructions, and guidance		FDA
2014 Q1	Develop documentation, instructions, and guidance	Final list of targets and markers	FDA
2014 Q2	Preparation of reference materials	Website and Submission potal goes live	FDA
2014 Q3	Dispatch of reference materials	Reference materials ready	FDA
2014 Q4	Evaluate PT. Provide feedback		FDA
2015 Q2		Complete first round of PT	FDA

TABLE: GMI WG5 ROADMAP

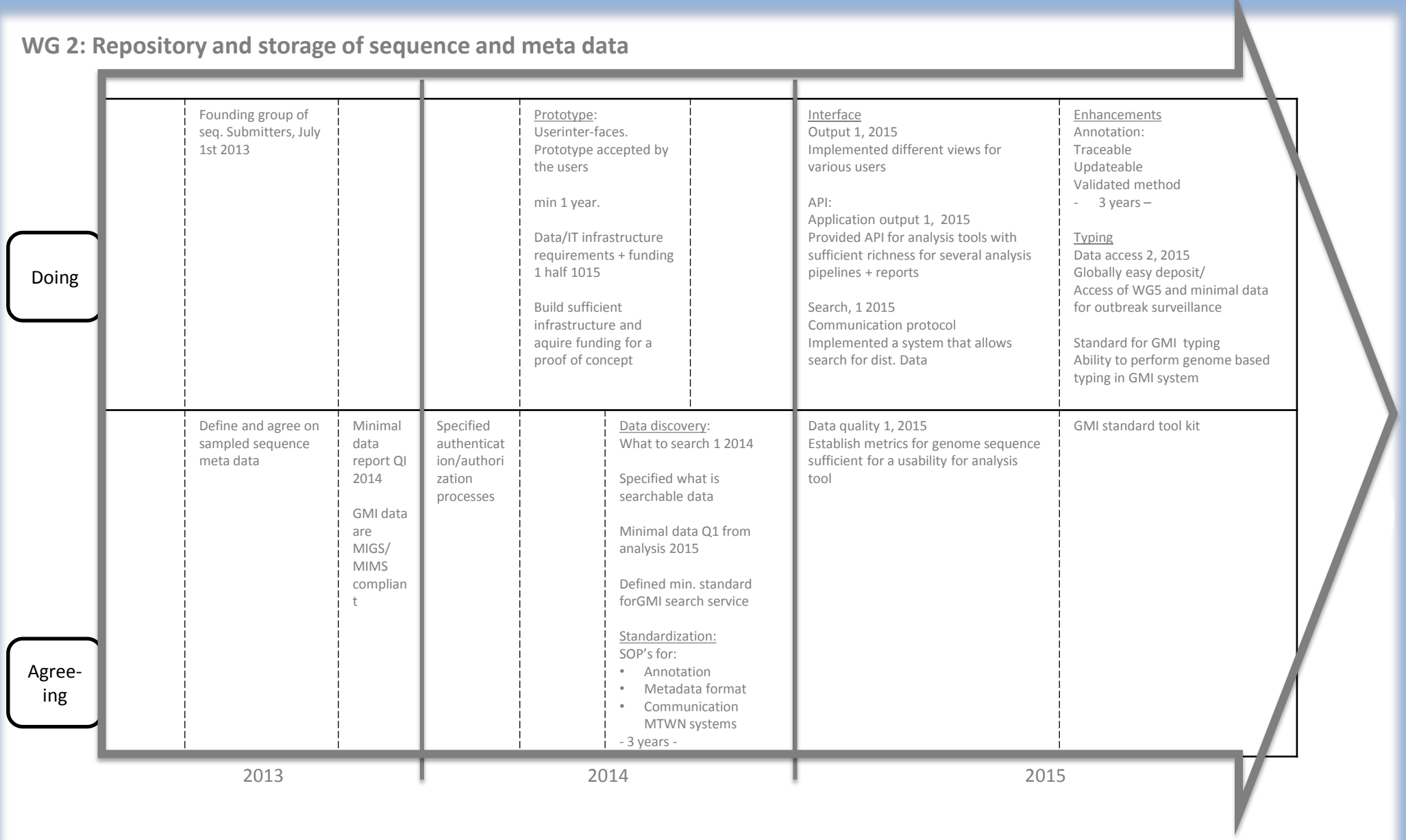
Date	Milestone	Responsible
2013 Q1	Establish steering governance committee	
2013 Q2	Communication platform established, including platform for existing knowledge	
2013 Q3	Listeria pilot launched	Geoff Hogg
2013 Q4	List of existing networks and pilot projects. Matrix of technologies and markers. Vision benefit and milestones formulated.	
2014 Q1	Guidelines for GMI pilots developed. Rules of engagement developed.	
2014 Q1	Listeria data transferred to public databases.	Geoff Hogg
2014 Q2	Listeria data analyzed.	Geoff Hogg
2014 Q2	Review written on gaps and lessons learned.	
2014 Q2	Fit for purpose pilot projects with realistic timelines.	
2014 Q4	Listeria project evaluated.	Geoff Hogg
2014 Q4	Accessibility and utility for end-users demonstrated. Universal pathogen ID and typing tool.	
2015 Q2	Pipeline for realization of GMI. Network structure defined.	
2015 Q4	Standard report formats and communication templates developed.	

Roadmaps

WG 1: Political Outlook and funding



WG 2: Repository and storage of sequence and meta data



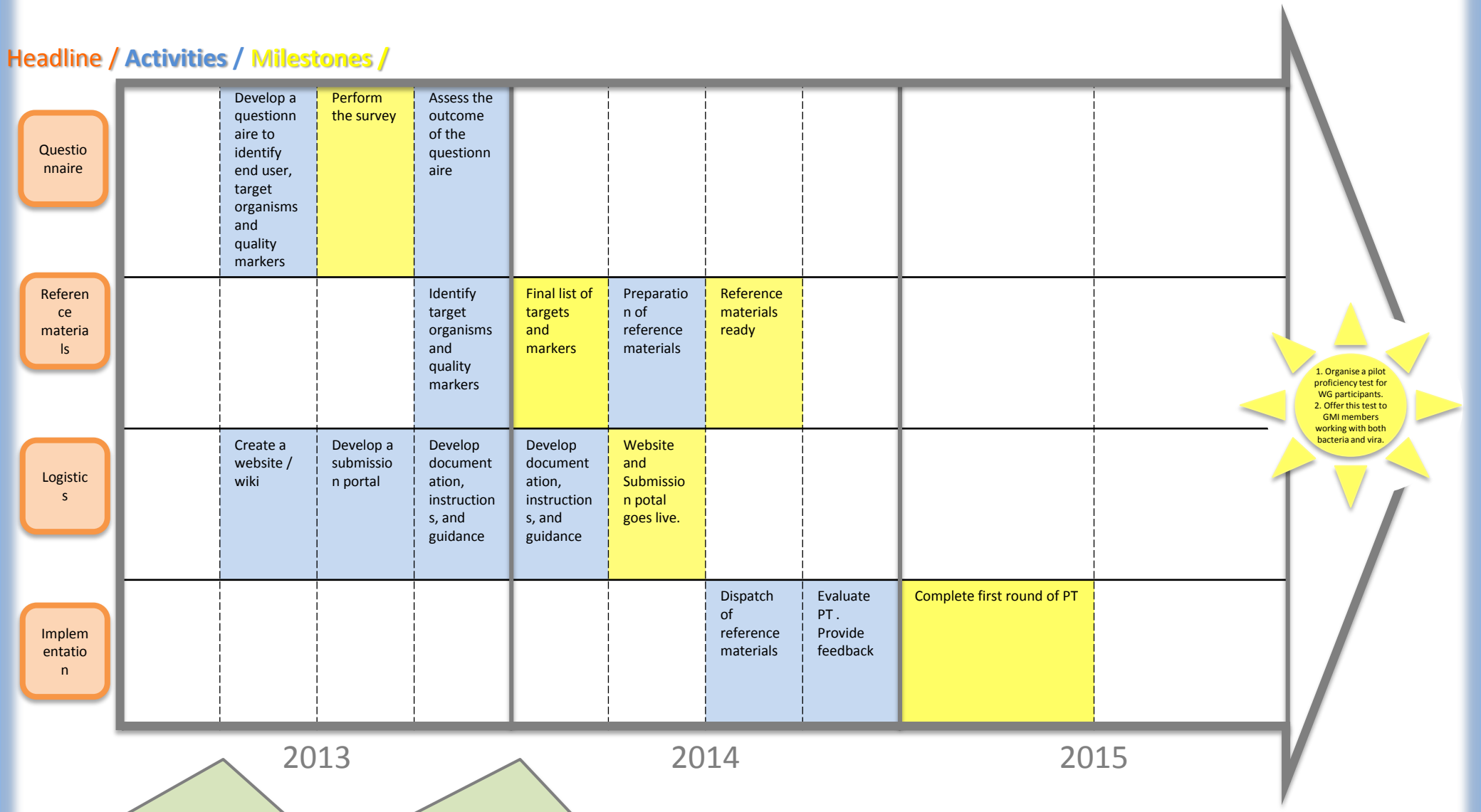
WG 3: Analytical approaches

Milestones

Priority targets			Compile burden of diseases estimates			Regional/ world priority of organisms database	Model systems for geno to pheno prediction	General SOP for novel pathogen discovery	Centralized repository for novel strains	
R&D				Report on: Tool availability and gaps Prev. success & failures Epi. And bio. Integration.		GMI session on tools & availability , and ontology.	General SOP for pilots	Approved ontologies. .		
Enduse r req.			Survey to determine endusers.		Defined endusers and their needs		12 countries upload to public databases	all NG seq. upload to repositorie s		Transparancy of the methods used.
Applica tions							Friendly user interface	Species and strain characteris ation running.		Interpretation to public warnings running.
Analyti c options	Survey for methods in use & data to be stored	Establish WG between industry/a cademia/cl inics	Common pipeline to prepare data to be shared				Genomic diagnosti traits identified.	Industry buy-in shared ownership	Link databases and out-reach modules.	Datastandards implemented.
Analyti c approa ches				Milestone forum created.		Decision tree for standardiz ed sample prep.	Define top apps where NGS is applicable			All microbial identification is digital

2013 2014 2015

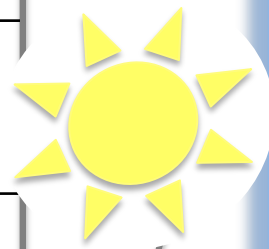
WG 4: Proficiency Testing



WG 5: Pilot projects

Milestones

	2013				2014				2015	
Comm. & govern.	Establishment of steering govn. Comm.	Comm. Platform established			Guidelines for GMI pilot project selection				Pipeline for realisation	
Prev. Pilot projects			List of existing networks. Matrix of technologies and markers. Comprehensive list of prev. Projects.			Gaps and lessons learned. Review written				
WG interactions			Managing expectations. Formalisation of vision, benefits and milestones		Fit for purpose pilot projects with realistic timelines.				Levels of access. Policy document. Define network structure, user roles and levels of access.	Communication content. Standard report formats and communication templates.
Synergism		Existing knowledge. Establish web-site.			Rules of engagement for people invilved.		Demonstrate accessibility and utility for enduserrs. Universal pathogen ID and typing tool.			
New pilots			GMI pilot for Listeria launched.		Listeria data transferred io public databases.	Implemented GMI project for Listeria.		Listeria project evaluated.		



Themes

WG 1

Theme: **Ownership**

POST-ITS ARE COLOR CODED

Supportive factors

Keep it simple

Supporting evidence of impact
-Proof of principle
-Simulations

Headline / Activities / Milestones /

	2013				2014				2015	
Gover-nance	Hire people dedicated to administer market and funding GMI	Draft descr. by steering comm. Broadening representation in steering comm. Geographical - Political	Consul-tation -org. form -Future work structure	Agreement on organizational form Annual review needed sept 2013						
One health	Develop stakeholder ana-lysis plan. Meetings (skype etc.) to engage all stake-holders	Create country consultation template (One Health approach)	Stake hol-der ana-lysis. - who? Geogr. and polit. repr. -Public health …- ben-efits & chalenges	in-put & output – levels of involvement dealing with ”antagonists	Present stakeholder analysis & rec. of actions Marh 2014					
Advo-cacy	Newsletter GMI session at stake-holder conferences	Draft comm. Strategy - Sci-entific publ. –Social media – internal review process	Identify advocates for different target groups	Communication strategy sept. 2013		Target comm. Strategy based on stakeholder analysis	Updated communication strategy Sept .2014			
Global roll-out	Coll. Req-uirements and que-stions from publ. Health inst. for pi-lot proj.		Def. Roll-out pro-cesses from well-de-veloped and less-developed countries		Connect to G8 Bigdata	Technical expert meeting -global med., 2014			Global level political meeting to get buy-in for GMI 2015	

We will develop a long-term plan for open sharing of genomic data in real time at the global, regional and national level. Also, we will suggest ways to fund the GMI project.

Challenges

Funding
Keep it simple

Obtain high impact examples of the utility of genomics in time. High impact; saved the taxpayers X$ ”prevented 20.000 illnesses . Generating credible high-impact

Uncertainties

Theme-roadmap

Theme: Global Health and Diplomacy

POST-ITS ARE COLOR CODED

Supportive factors

- Survaillance =FREE. The argument of benefit for diagnostics FIRST
- EU - European agenda
- -Political country champions.
- -Funding
- IHR

Headline / Activities / Milestones /

	2013	2014	2015
WHO Benefits? -Impact -Pruducts	GMI website diplays -Results -Outcomes -Call for partici-pants Advocacy Papers for each of 5 different endusers. Dec 2013	GMI known by (????) of profes-Sionals 2014	
Gover-nance and owner-ship of GMI initiative	EU political support meeting agenda 2020 2013 Involve-ment of stake-holders of parallel initiatives	Communicati on strategi-cally.! Use of (recent?) pilot studies	-Involve and support patient groups into political demands -Preparation of backgr. Docs. for WHA basrd on results of pilot studies + other activities -Commitment of 30 country (mimtes?) reading WHA large (scale?) event 2016 Resolutions WHA, OIE, (FAO?) May 2016
Import-ance of national level engage-ment	Identifty front-runner Stakehold er mapping and mating	National GMI information point 2014; 50 countries	
Dealing with legal impli-cations	Finding/ making legal (??????) Working groups	Research legal implications – lessons learned + avenues for problem-Solving	Complications on legal implications related to GMI Global health diplomacy 2013 Form a governance structure

We will develop a long-term plan for open sharing of genomic data in real time at the global, regional and national level. Also, we will suggest ways to fund the GMI project.

Challenges

Internal coordination?!
GMI should be in the end a a public responsibility.

Human rights issues?!
Privacy
1. Metadata
2. Seq + metadata

Theme-roadmap

Theme: Fundraising $

POST-ITS ARE COLOR CODED

Supportive factors

Roadmap, gap analysis to inform budget – pilot (+scale).

Clarify mit NCBI, EBI, DOBJ, on funding to support "core" repository/analytics

External agensies -co-orditaion -what's required

Headline / Activities / Milestones /

	2013			2014				2015	
Know the invironment	Define GMI an establish funding commitee <Survey>	Talk to: - Databases infrastructure -Pilots about support –GMI central body infrastructure	Identify existing resources that can be leveraged -REPORT > PROPOSAL		Define funding areas -R+D -Pilots -Organisation				
Fund-raising			Initial co-ordinating funding application -organisation -Pilots -Infrastr.	Meet with funders CONVERGE ON NEEDS	Funders EU – CDC – ECDC – NIH – WT – NGO's – Illumina f's Gate	Big Data submission $$	2-5 pilot projects -existing capacity -demonstrate PoC –funding easier -Identify hidden costs	SCALE-UP?	

We will develop a long-term plan for open sharing of genomic data in real time at the global, regional and national level. Also, we will suggest ways to fund the GMI project.

Challenges

Uncertainties

Theme-roadmap

Theme: Sensitivity of Metadata/IPR issues/Open access

POST-ITS ARE COLOR CODED

Supportive factors

Headline / Activities / Milestones /

Legal	Catalog current requirements by finding agencies to report microbial genomics data -Engage with who to determein interest level discuss the model framework	Catalog what are the current multi-national agree-ments to microbial data	Stakeholder mapping -who? - purpose? - needs? Catalog national legalities to open access to microbial genomic data	Survey to model framework to bio-medical reporting agencies to explore need/requirements for making data available	Survey industry and countries to explore acceptance of model framework to ac-ceptance of open access			Global aggreement
Model frame-work	Developing a model of reporting framework for addressing legal obsticles	Deter-mining "option-al"/minimum metadata Use existing models		Develop teared approach to release of meta-data Timed - trusted				Develop communication protocols for notification of parties for outbreak response International coordination
Risk benefit	Develop strategy that identifies stakeholders and most effective way to communicate	Invite/involve industry partners to GMI meeting or seperate congress	Develop communication strategy for model messaging industry/ Govt./aca-demia		Proactive education to illu-strate be-nefits of open access /how it midigates/ minigates risk	Identify industries that would have the greatest most apparent benefit to risk ratio. What is it? Making genomic microbial data available -minmize risk -Brand equity/protection deference on retrobutrion for too quick Intreput		

We will develop a long-term plan for open sharing of genomic data in real time at the global, regional and national level. Also, we will suggest ways to fund the GMI project.

Uncertainties

What are the concerns arround IPR, open access, sensitivity af metadata.mechanism (G8 bigdata)
Who determines min. Metadata
Who can agree to metadata/open access policy / protocols

Challenges

Whose data is it

Theme-roadmap

WG 2

Theme: Annotation/Metadata 2.2

POST-ITS ARE COLOR CODED

Supportive factors

- Funding for Developing countries & Infrastructures
- Setup a concrete action plan (pilot study) after the meeting
- Funding for setup the action plans
- Good political Relations ntwn involved countries
- Broad geo-political & scientific expertises

Headline / Activities / Milestones /

	2013	2014	2015
User/ interfaces	Define user groups (how many types of interfaces Identify what user groups want to do Define requirements for different users Build beta prototype	User test phase (3 months) users use the interphase and send feedback Final "prototype" accepted by the users (min 1 year	
Standardization	Look at the existing metadata format	Pick standard(s) automatic annotations pipeline	Define the communication btwn local and general databases SOP's for: •Annotation •Metadata firnat •Communication btwn systems - 3 years -
Anno-tation		Def kind of annotation: •Seq-lab M data •Automated M data •Other labs assays Create a method for upgrading the annotation that must be traceable	Traceable updateable validated method - 3 years -
Meta data privacy & distribution		Define which metadata can be used and dist. Globally or to specific countries (lawyers) Globally accepted policy on M-data privacy and ddistributio n - 2 years -	

We will develop a plan for a GMI repository, building on experience and ideas from similar efforts.

Challenges

- Globally shared protocol for M-data update
- Agreement on policy & M-data at un equal levels
- Define what data & where/how to store it
- **SECURITY**
- **Uncertainties**

Theme-roadmap

Theme: Data availability

POST-ITS ARE COLOR CODED

Supportive factors

NCBI + FDA Existing project

INSDC existing international standards, exchange, communication

Headline / Activities / Milestones /

	2013				2014				2015	
Data access	Define outbreak. Minimal M-data. Subset of MIGS/MIM fir prototype		•Develop easy and automated upload tools (fastQ, BAM) •Define tagging for GMI activities/records across INSDC (May 2013)			Prototype Review •Easy enough? •Does INSDC tagging work?				Globally easy deposit/access if WGs and minimal data for outbreak surveillance
Data/IT infrastructure, requirements, funding	Prototype: Ongoing funding. FDA/NCBI Food safety program		Expand prototype to include international partners	Define userrequirements: •??? •Time to upload and retrieve		Whats already there? Internet? What needs funding? Explore alternative models to store data in the cloud/???		Find and build infrastructure	Build sufficient infrastructure and aquire funding for a proff of concept	
Applications, output				Prototype API supports limited analysis •Assembly •Annotation •Kmer cluster •SNP Cluster	Analysis group: Define analysis types important ofr GMI /surveillance Define requirements of analysis pipeline •Inputs to pipeline •Vary by analysis type		Define what retrieval API look like (based on metric data)		Provided API for analysis tools with sufficient richness for several analysis pipelines + reports	
Data quality				Assesment of sequence quality needed for prototype analysis		Assesment of quality needed for all analysis	Define computation and labels for each metric	Repository formats and records metrics sequence record	Establish metrics for genome sequence sufficient for usability for analysis tool	Tools use + validate metrics

We will develop a plan for a GMI repository, building on experience and ideas from similar efforts.

Challenges

INSDC Int collaboration for tagging GMI

What countries will join prototype + funding

Private metadata
1. Level of sharing
2. How to share private metadata international

Funding infrastructure to max participation

Changing technologies

Different levels of expertise. Infrastructure, Funding

Uncertainties

Theme-roadmap

Theme: Searching protocol

POST-ITS ARE COLOR CODED

Supportive factors

Excisting systems

Major epidemic

Headline / Activities / Milestones /

	2013				2014			2015		
How to search communication protocol	Determine priorization of programmatic interactive interface		•Definitin of standard for human interface •Definition of standards for programmatic interface		Define a standard of interproces communications		Code it....	Implement a system that allows search for distributed data		
Output		Define the need for analytic tools besides the sequence data output	Definition of outputs/views for specific end user needs	Outouts must report •Ownership of data •Provenance of data •Attribution of data		Evaluation of possibilities to conjugate outputs as new inputs for searching		We have implemented different views for various users		
What to search	Determine acceptable MDM to se includes in DBs for search		Determine the level of data pre-processing required to enable sequence searches (reads, contiges, SNPs, annotation)		We have specified what is searchable data					
Autho-rization	Define control autorazition system to define unique identifications Define cascading access to database for searching		Determine which DB are to be placed under control authorization Determine who grants and controls access to control and authorize oxiliary		We have specified an authentication / authorization process for control vs distribution DB.					

We will develop a plan for a GMI repository, building on experience and ideas from similar efforts.

Challenges

Funding

Legal issues

International / institutional disagreements

Computational aspects of distributed searches

Uncertainties

Computational needs for pre-indexing searches

User acceptance and proper data submission

Theme-roadmap

Theme: MDM

POST-ITS ARE COLOR CODED

Supportive factors

Excisting systems

Major epidemic

Headline / Activities / Milestones /

	2013				2014				2015	
Minimal data to report	Establish pilot groups. Data providers who send at least 1 sequence in 2013.	Define and agree on sample + sequence metadata	Create integrated database extension of SRA/ENA	Test prior to NGS pilot completion	Milestone: MIGS/MIMS compliant vy January 2014					
Minimal data from analysis			Consider options for searching data inferred from sequence		Understand how to use species specific genes / conserved genes (proteins) when searching data		Run prototype	GMI data counter		
Stand-ard for match-ing			Consider options for sequence similarity search		Implement phylogeny within database – heirachkal? •Kmer based •SNP based	Understand contamina-tion in context of search	Run prototype			

We will develop a plan for a GMI repository, building on experience and ideas from similar efforts.

Raw sequence search

Challenges

Capture raw data or contigs (optionally)

Uncertainties

Theme-roadmap

WG 3

Theme: Priority targets/diseases

POST-ITS ARE COLOR CODED

Supportive factors

Money
Private industries (machines)

Collaboration of experts, countries, institutions

Headline / Activities / Milestones /

	2013				2014				2015	
Defining priority organisms	Litterature review	Experts consultation	- Compiling burden of disease - Regional differences		-FERG -ECDC -CDC -RIVM (Koopmans) -PHAC (van Damstetar) -Will BCCDC	Regional/ world priority Organism database				
	Does Genomics make a difference?									
Novel pathogen discovery	More metagenomics sequencing (CBS-MG, DTU)	Automated pipelines for pathogen discovery				Dirk	General SOP			
Viruses: Resistance & virulence	ATB-R^{ce} Predictor Review + GHP ID	Genotype to Phenotype Expert system	More reference sequences (ref strains) + plasmids Will-BCCDC RIVM (Marion Koopmanns) FDA-CVM Sanger Flora Brinkman			Model systems for accurate Geno to pheno predictio n CGE - Mette Voldby Larsen, DTU			Centralized repository	
Bacteria: Resistance and virulence										

We will provide guidance for the development of analytical tools for optimal positioning and functioning of the GMI platform.

Challenges:
Time
Cost

Uncertainties

Theme-roadmap

Theme: R&D – implementation/gaps

POST-ITS ARE COLOR CODED

Supportive factors

We need funding for all this

Need more assessment of current tools and R&D gaps (split between groups) => allocate assessment work to different groups

Next GMI meeting -people report on assessments day 1. -formulate development plan, day 2

Headline / Activities / Milestones /

	2013			2014				2015	
Pilot for realtime monitoring control (for analysis tools)	Assess knowledge from academic/research pilots		Meeting of those interested in real-time pilot (GMI fall session?)	1st real-time pilots chosen Q3	Pilot(s) run	Generalized SOP QS	2nd round of pilots		
Integration of genomic & epi data	Assess gaps in 'metadata'/epi integration, datastandards&ontologies		Report/talk on gaps (GMI session?)	Target development of key standards &ontologies Iterative review with domain experts	1st proposal of new data standards&ontologies (document)	Reveiw of developed standards&ontologies	Approved standards &ontologies (PHAC, van Domselaar)	DTU Fiona Brinkman BCCDC	
Assessing Pulsenet & other genotype systems – successes & failures	Meet with PulseNet & other genotypic systems key stakeholders Itemizesuccess and failures from focus groups		Develop solutions for failures Propose model(s) for a new platform	Report on successes &failures in PulseNet & other systems -propose model for new platform	EURLs ECDC WHO RIVM PHAC (van Domselaar) DTU				
Make analysis tools adaptable		Assess analysis tools for flexibility/adaptability (review by one or two labs/student centres)	Report/paper on tool adaptability Q5 Factor in what's learned from pulseNet study (and others)	DTU to find someone	GMI session on tool adaptability Q6	Develop orgnizational structure that supports change Develop platform (IT etc.) structure that supports change	Short report on tool/platform adaptability	Factor report into platform development	

We will provide guidance for the development of analytical tools for optimal positioning and functioning of the GMI platform.

Challenges:

Social sci study reporting what

akeholders need to be comfortable with

Constructive tention between academic and service users

Overcoming human factors which resist change

Uncertainties

Theme-roadmap

Theme: General or specific

POST-ITS ARE COLOR CODED

Supportive factors

Headline / Activities / Milestones /

Technology adoption		Which pathogen at first?	Which method to replace? When?				Simplified reporting	Standardized analysis defined		All microbial identification is digital
Applications		Establish a platform for information sharing (PHAC , van Donselaar)		Identify areas where there is most added value & where it is most desirable	Select a set for pilot study		Define top apps where NGS is applicable (Steve/Marim)			
Sample complexity	Pathogen experts involvement in extraction methods definition (PHAC , van Donselaar)	Extraction methods defined	Defined reads+read length per application	Defined 'unknown'sample prep rules/SOP	Simplified reporting per agent/pathogen (Steve-RWM)	Decision tree for standardized sample prep				
Exploration [New areas]	Brainstorm Share experiences State of the art	Define where NGS can be used	Pilot study	Estabilshment of a milestone for reporting forum (DTU)						
	2013				2014				2015	

We will provide guidance for the development of analytical tools for optimal positioning and functioning of the GMI platform.

Challenges

Uncertainties

Theme-roadmap

Theme: End-user requirements

POST-ITS ARE COLOR CODED

Supportive factors

Headline / Activities / Milestones /

	2013	2014	2015
Transparency Integration	Bringing vendors together Produce SOPs to make uploads to public databases	Meeting with industry on transparency &integration Define the Metadata to be uploaded	Integration of Metadata Shared software @GMI Open source pipelines documented and available Tech Support Increaesed transparency of the methods
Shared ressource and tools	Clusters of data available for testing the pipeline $ (FDA/NCBI)	Subcommittee of GMI's delivering analysis pipeline List of used and available software for pipelines at GMI with user feedback (HPA Bioinformatics)	Modulise SOPs (EU PathoWG trace.Dag Harmsen; FDA-Errol Strain) Upload to public repositories from all commercial available NG sequencers GMI Meeting Fall 2014
Quality control validation	Comparison of de novo assembly Comparison among labs -Paper -EU pathogen trace (EU PathoWG; Dag Harmsen) Publicise SOPs with GMI Lis t of training available on GMI $	Sharing the data and QC values Writing SOPs At least 12 countries upload data in public database	
Identification of end-users	Questionnaire/Survey through participants of this meeting Create user-specific sub-committees (FDA, RIVM, HPA) Next GMI meeting results of survey	Defining users -research -clinical -public health -industry (IRTA 'Food'; ASM; ESCMID) GMI Meeting spring 2014	

We will provide guidance for the development of analytical tools for optimal positioning and functioning of the GMI platform.

Challenges:
Need to collect more end-user specific datasets $

-International engagement
-Get volunteers to write SOPs/review them
-Commecial interest via ID
- solving political sharing

Uncertainties
- Availability of turnkey software
- Availability of turnkey servers

Theme-roadmap

Theme: Applications

POST-ITS ARE COLOR CODED

Supportive factors

Budget $'s

IT-support

Education

Society awareness

Headline / Activities / Milestones /

	2013			2014				2015	
Interpretation and epidemiology						Metadata	Expert groups ont he server for each pathogen - Database update	Geographic mapping G(MI)oogle Earth	Public Warning
Identification & characterization		16S rRNA & additional genes	K-mer for species ID	Species-specific database pipeline	AMR -subtypes -MLST -MLVA -etc	In-silico genome mapping	Species &strain characterization (M Voldby L) Public Launch (RIVM, AnneLies)		
Facilities for end-users	No bioinformatics needed for end-user	Robust easy interpretation	Stable and fast	Real-time database		Friendly user interface Public Launch	CBS-DTU NCBI		
Validated WGS data	Pilot studies of complete genomes RON	Use WGM to confirm validation of platforms & samples		Quality control for sequencing data		Validated WGS data			

We will provide guidance for the development of analytical tools for optimal positioning and functioning of the GMI platform.

Challenges:
Knowledge & skills

Financial
Manpower

Political support

leadership

Institutional support

Uncertainties

Theme-roadmap

Theme: Analytical options

POST-ITS ARE COLOR CODED

Supportive factors

Funding

Surveys

Collaboration

Headline / Activities / Milestones /

	2013			2014				2015
Household vs. Commercial analysis		Establish working group to communicate between industry/clinical/academic research	User <> Developer communication		Report back into next GMI meetings/publish findings/action			Industry buy-in collaboration Shared Ownership+goals DTU Steve Picton (PacBio)	
Compatibility and sharing			Common pipeline to prepare data to be shared/published **Illumina**		Taxonomy Identification Methods should be more defined	Different Methods Performance comparison			Link databases Establish standard drive adoption NCBI?
Dick Modules	Design a survey to map methods in use	Integrate modules which are done already	Defining Interfaces + Analysis of survey data			Different methods performance comparison	Implement standard interface		Implement data standards (module specific) Output oriented modules
Storage – What & when to share	Design a survey to understand what info should be stored		Analysis of survey data		Metadata + establish sequence quality (WG needed) NCBI/EBI/DDBJ	Expand databases suited for application	Genome signature for diagnostic traits		

We will provide guidance for the development of analytical tools for optimal positioning and functioning of the GMI platform.

Challenges

Get people to shar all agreed standard data

Funding
Time frames
Uniforming mindset

WHO Coordinates global database + monitoring

Uncertainties

Theme-roadmap

WG 4

- All included on the overall roadmap

WG 5

Theme: 5:3 Launch a GMI Demonstrator Project for Listeriosis

POST-ITS ARE COLOR CODED

Supportive factors

Do-able now

Does not ne… much extra funding

Build it and they will conme

Headline / Activities / Milestones /

	2013			2014				2015	
Asses stake holder utillity	Identify stake holders	Dialogue with stake holders			Technical evaluation	"So what" evaluation	**Stakeholder feedback collatad + circulated; final report for GMI**		
Imple-men-tation phase	Identify collections	Sequences + QA	Begin data transfer	Report back to collaborators	**Implementation completed**				
Listeria datatransfer into public data bases	Platforms identified; Datatransfer proto-cols estab-lished ; Analasys + display decided	Transfer of data starts with demo		**Platforms ready**					
Launch GMI demonstrator project for listeriosis	Difine succes criteria	Invit + establish collaborators	Write project plan	**International collaborators identified + recruited**					

1. Establish a WG communications and governance structure and define how we will interact with the other WG's. 2. Define pilot project that will satisfy the requirements of the broader GMI effort.

Challenges

I.P.

Data confidentiality

Uncertainties

Funding for project

Theme-roadmap

Theme: Synergisms

POST-ITS ARE COLOR CODED

Supportive factors

Knowledge + tools aggregator

Headline / Activities / Milestones /

	2013			2014			2015	
Make it simple		Decide on the perfect pilot project		Profici-ency test		Capacity building user education		Demonstrate asseability and utility for enduser
Cross discipli ne inter action		Collect bio informatic tools			Graphical output – visualisation tool		Ease of use	Universal pathogen ID + geno-typing
What am ”I” going to get out of it	Create a culture for exchange of ideas + data	Create a value propo-sition		Data sharing plan (rights) Attributors rights for collaboration	MoU rules of engagement for people and entities			
Lever age excis ting know ledge		**Design a website and ”make it fly”**		Incorporate abillity to upload + share data		Create ability to to existing databases	Focused website	

1. Establish a WG communications and governance structure and define how we will interact with the other WG's. 2. Define pilot project that will satisfy the requirements of the broader GMI effort.

Challenges

Webside needs to be flexible design (NOT

Funding

Uncertainties

Theme-roadmap

Theme: WG interactives

POST-ITS ARE COLOR CODED

Supportive factors

Headline / Activities / Milestones /

Comm-uni-cation content					Written review		Define and write standardized report format and com-municatio template	Collect and review reports that are in use today	Standardized report formats Communication templates
Levels of access		Appoint group/ Commitee To write a policy documenmt	Defining access level con-straints by stakeholders . Generate document	Survey for access needs	Define roles by survey results	Develop user agreemen t policy	Level of access training		Policy document Implementation of access levels Definition of network structure an user roles	
Fit for purpose pilot project	Immediate E-mail up-date when funding oppor-tunities are available	Video describing pilot project A centralized file of on-going proj-ects with gant charts and other detailed Proj. Info	Progress report and making results public		Concrete pilot projects with realistic timelines					
Mana-ging expec-tations	A video for advo-cating GMI that high-Lights ob-sticles and andvisi-on for overcommi	Quarterly podcast tailored for different audiences (from general to more detailed)	Formalizati on of vision, benefits and milestones							

1. Establish a WG communications and governance structure and define how we will interact with the other WG's. 2. Define pilot project that will satisfy the requirements of the broader GMI effort.

Challenges

Funding

2013

What's in it for me

2014

Uncertainties

Theme-roadmap

Theme: Prev. Pilot Projects

POST-ITS ARE COLOR CODED

Supportive factors

Headline / Activities / Milestones /

Gaps + lessond learned			-Synthesis Collabora-tive groups			Written review				
Compar ability and Compat ability		-Survey -List technologi es usen in pilots	-Use draft review for matrix	Matrix of technologi es vs types of markers		**Update**			**Update**	
Inter-actions		Identfy large scale networks +parti-cipants -Ident. Scope Registry (invite new parti-cipants)		List of existing networks + partici-pants		**Update**			**Update**	
List		-Survey Literature review Website (GMI) Funding agencies		Compre-hensive list of previous projects		**Update**			**Update**	
		2013				2014			2015	

1. Establish a WG communications and governance structure and define how we will interact with the other WG's. 2. Define pilot project that will satisfy the requirements of the broader GMI effort.

Challenges

Resources

Uncertainties

Theme-roadmap

Theme: Communication and governance

POST-ITS ARE COLOR CODED

Existing collaborators

Quality monitoring of systems – transparency – worldwide – rec.keeping

Supportive factors

Horizon 2020 Funding – cost action?

Legal support

Existing experience and initiatives: The Int. - Consortium for the Barcoding of Life -Pulsement – Pathogentrace, CGE, etc.

Headline / Activities / Milestones /

	2013				2014				2015	
Pilot project execution	Define expert group, capacity group (crowd financing)		Define standard for reporting		Guidelines for execution & evaluation	Pipelines ready				
Pilot project initiation	Define prodedure for exe-cution & evaluation Creation of pre-mium "brand"	Define technical criteria	Define procedure for P.P. suggestions	Define procedure for P.P. selection Guideline for GMI pilotproject selection						
Communication	Mailing list, web-site, social media (linked-in, research gate)	News-letter	Define network			Efficient communi-cation platform estblished				
Structure and governance		Decide on decision body Evaluate structure	Decide on structure, Establish structure	Establish-ment of steering comitee						

1. Establish a WG communications and governance structure and define how we will interact with the other WG's. 2. Define pilot project that will satisfy the requirements of the broader GMI effort.

Challenges

Find persons to pass

Competition / trust

Uncertainties

Confidentiality/transparency

Ubiquituous –

Intellectual property -

Sustainability

Theme-roadmap

Global Microbial Identifier