Global Microbial Identifier

Report on the 6th meeting 11 - 12 September 2013, UC Davis, Sacramento, California

Global Microbial Identifier

Updates from GMI6 meeting
UC Davis, Sacramento, California
September 11-12 2013

Global Microbial Identifier

Report of the 6th 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 infectious disease 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.globalmicrobialidentifier.org to find out more about the project, summaries of previous meetings, and useful background information.

This is the report of the 6th meeting taking place at UC Davis, Sept. 10-11, 2013.

Agenda for the 6th Meeting

Day #1 (September 10th, 2013)

Overview, Updates, Initiatives

08:30 AM Welcome/Overview	Marc Allard and Bart Weimer
08:50 AM Status and perspective of GMI meeting 5	Frank Aarestrup

Charter & Structure of GMI

09:05 AM GMI Charter and Structure - An Introduction to the Discussion on Shared Principles -	George Haringhuizen
09:20 AM Discussion in Break-Out Sessions	
10:00 AM Discussion in plenum	
10:30 AM BREAK	

Status and Perspective of Each Working Group

11:00 AM Political and financial challenges	Jorgen Schlundt
11:10 AM Work Group 2 Update: Repository and Storage of Sequence and Meta-data	William Klimke
11:20 AM Analytical approaches	Marc Allard
11:30 AM Work Group 4 Update: Ring Trials and QA/QC	Rene Hendriksen
11:40 AM An Update on Pilot Projects	Mark Wilson
11:50 AM LUNCH BREAK	

GMI Resources

1:00 PM FDA Genome Trakr database: details on dataflow from field labs to a public *Salmonella* reference database at NCBI	Ruth Timme

Focused Discussion to Review and Update Strategic Plan for Each Working Group

Working group 3 will continue with talks on software tools and standardization

1:15 PM Focused Break-Out Sessions in working groups Side Rooms

3:30 PM BREAK

Produce Written Update Strategic Plan for Each Working Group

3:34 PM Focused Break-Out Sessions in working groups	Side Rooms
5:30 PM QUESTIONS	
6:00 PM ADJOURN DAY #1	

Day #2 (September 11, 2013)

08:30 AM Welcome	Marc Allard and Bart Weimer

GMI Resources

08:40 AM Tomorrow's Genome: Complete Bacterial Genomes in <24 hfor outbreak resposne	Ken Dewar
09:05 AM U.S. Nation-wide Genome Sequencing-based *Listeria monocytogenes* Surveillance	John Besser
09:30 AM 100K Pathogen Genome Project	Bart Weimer
09:55 AM Establishing Whole-genome based approaches as a Routine Tool in Reference Microbiology	Jonathan Green
10:20 AM BREAK	

Identify Milestones and Responsible Volunteers to

Accomplish Strategic Plan for each Working Group

10:35 AM Focused Break-Out Sessions in working groups	Side Rooms
12:00 AM LUNCH BREAK	
1:00 PM Enabling Sequence-Based Technologies for Microbial Diagnostics	Heike Sichtig/Uwe Scherf

1:25 PM DTU software tools for NGS GMI data analysis Ole Lund

Finalizing and Drafting Milestones to Accomplish Strategic Plan for Each Working Group

1:50 AM Focused Break-Out Sessions in working groups Side Rooms
3:30 PM BREAK
3:45 PM (Generate final report from focused groups) MAIN HALL/Side Rooms
4:45 PM QUESTIONS and Discussions

6:00 PM ADJOURN DAY #2

Opening Remarks and Introduction

The meeting was opened by Drs. Marc Allard and Bart Weimer of the US Food and Drug Administration and UC Davis, respectively.

Dr. Frank Aarestrup of the Danish Technical University presented an overview of the progress of the GMI consortium up to the start of the 6th Meeting, including the evolving vision of how GMI will develop into a global network that will contribute to infectious disease surveillance.

Main purpose of the 6th meeting:

The purpose of the 6th meeting was to continue the work on structuring each of the working groups and establishing a common framework for the operation of GMI. Each of the working groups established prior to the 5th meeting were to continue the development of a strategy for operation within the context of the larger GMI framework, and were asked to come with a strategic plan for integration with the other working groups.

WG1 Update (Political challenges, outreach and building a global network)

Discussion in the WG1 break-out group and plenum discussion

It was decided to change the vision into a short statement and move the remaining contents to the relevant sections of the charter. During the plenum discussion it became clear that the vision might need to be further altered, however, no new suggestions came up.

> The new version text suggested by the WG1 was:
> - A world where high quality microbiological genomic information can be freely shared among nations to improve global human, animal and plant health

During the WG1 discussion it became clear that the charter needed to be organized with a more clear structure:

- Introduction that is not as narrow as in the draft. Taking the new vision into account
- Vision and goals – clear statement of the benefit for the society. important to make sure all sectors feel included
- Values -
- Mission – has to be broadened according to the new vision which include all microorganisms and not only health - **Carl Sciacchitano will send some sentences about this to John Besser**
- How to get to our goal
- Statement that the data base is for all microorganisms and difference between raw data and metadata due to different issues like confidentiality
- Description on how Steering Committee members are elected
- 1 page summary of the central information

John Besser volunteered to go through the original vision and statements and update according to our discussion. A new version of the charter will be send around in the WG1 for comments

Points to take into account in the **new charter text to be prepared by George Harringhuizen**

- It is important that the charted is not about 3world aid.
- The importance of having access to metadata in order to use results from genome sequencing should be mentioned. Could probably be included in the “Ethics sections”
- High level strategy section to make a statement about how to use data
- Some of the criteria for participations needs to be more clear: when are you a partner, participant, data submitter? What are the different expectations according to the type of participation?
- How to differentiate between associated projects and scientific projects. How should they contribute to the values and goals

Declaration of interest

SC member and chairs of the working groups need to declare potential conflict of interest, e.g. direct connection to the industry for themselves or their organization.

Organization form

Steering Committee structure

There was some concern about the election of Steering Committee (SC) members as it is not clear in the charter. It was pointed out that at this moment in the process it is very important to have continuity in the SC. Members are appointed to the committee due to their knowledge to ensure coverage of relevant disciplines and not because they represent an organization or institution. Presently one of the main tasks of the SC is to push forward the process and take care of administrative issues. Members of GMI can comment the work of the SC during the GMI meetings. Once the charter has been agreed upon, the SC should agree on the combination of members. To ensure involvement of all regions, it is important to ensure that all regions are represented in the SC. At a later stage it should be considered to change the structure of the SC to e.g. appointment of members for a two year period, including international organizations.

Organization of WG's

It was suggested to have subcommittees focusing on developing of all tools in order to move forward faster. These groups should have e.g. monthly meetings and report back to the SC.
However, the WG are free to have subgroups when relevant. The charter does not define this**.**
It is important that all WG leaders and members are well aware of the direction the GMI is moving in WG leaders are nominated by the WG members, but at the moment it is not clear how members or leaders are nominated. It should be considered to define this more clearly in the charter, although we should always remember that we are not an organization and therefore the organization of GMi needs to be more loose at the moment
Recommendation: A dedicated working group in communication and dissemination should be under WG1.

Stakeholder analysis

Provide a stakeholder analysis to identify key political and economic entities, health, animal and plant related institutions/organizations, and civil societies. There has to be focus on the end users.
Some obvious stakeholders are already defined such as laboratories, companies producing diagnostic tools and medical devises, tech companies.
A land grant University like UC Davis may be able to do it. If the university takes up the task, the WG has to be a close collaborator as we have to identify what kind of questions we need to include. WE have to define a clear objective for the stakeholder analysis.
Canadian Food Inspection Agency (CFIA): Mentioned a networks of networks
Rozann Saunier from ANSES (France) said that they can probably help with the analysis as there should be an European institution involved in the European analysis.

Political level:

The analysis should look at who is ready to join now. To get information on the political level in EU, the European Commission might have information . The EU Commission has included WGS in the Horizon 2020 (EU Framework Programme for Research and Innovation). WHO might have information on other countries.

Plant stakeholders:

Contact should be taken with the plant protection organizations and epidemiology groups to have them understand the possibilities and usefulness of GMI. Maybe have some of them participate in the next GMI meeting.

Communication strategy

We need to raise awareness, understanding and ensure high visibility. GMI should be known by 65% of all stakeholders and include all pathogens; virus, parasites etc. We need stakeholders to understand that by collaboration we can lift things to a higher level. Our success criterion is to convince stakeholders about the open access to data, sharing of information.

One-page information sheets

Several 1-page information sheets will be made focusing on the most important target groups which will be designated in the stakeholder analyses. The 1-pagers should include information on what stakeholders can get out of it and what the GMI can get from the stakeholders.

Stephanie Defibaugh-Chavez will draft these with help from people with expertise in the different areas and send them to the WG1 for commenting

Political level:

We want GMI connected with national political levels to make them aware of the importance of participation

Newsletter

The newsletter will be send to everybody on the mailing list. If there is anything in the letter that will need a response from the members, the newsletter will also be sent to the WG leaders as a back-up.

Sharepoint

There is a sharepoint, but very few use it. A reminder will be sent out.
University of Minnesota has software (Freeware??) that might be useful. **Carl Sciacchitano will send something about this.**

Mini symposia at conferences

We will try for a mini symposium or a both at some big international symposia like the American Society for Microbiology (ASM) and International Associated for Food protection (AFP)
Eric Brown will be in charge of an application to ASM and AFP. He will draft an agenda and send to the SC for approval
We need to draft charter instructions fast for such event quickly

Before the next meeting we should involve International Conference on Emerging Infectious Diseases (ICEID). We have to find meetings in regions so we can become more visible.
Peter Gerner will prepare a document about how to implement the epidemiological site and Amy Cawthorne will comment.

Media

Once we have the 1-pagers and the stakeholder analysis we should have media strategy. Science had an article, maybe this could be updated. **Rene Hendriksen has a contact**

Other issues

We should invite more epidemiologist (human health, veterinary and plant), bioinformatics, big-data people for the meetings.

Funding

A strategy on how to approach the possible entities has to be developed. We need 1-pager to be used as an introduction.
It is important to try to find founding so members from developing countries can participate in the meetings.
There is a need for founding of the GMI platform and process. GMI will not compete with research institutions about research money except when trying to set up research in developing countries based on money from e.g. the Danish DANIDA.
List of possible entities: Bill and Melinda Gates Foundation, Google foundation, Wellcome trusts, Wollmart and other big supermarket chains, Rockefeller. We could also apply for discount on machines to developing countries. The list should be extended
The food defense and nuclear threat foundation in the US have money for Global Health Security. They define this as security in the food chain. Food safety has not traditionally tried to get money from these foundations.

Stefano Pogngolini will share some experience about funding from private banks

Jørgen Schlundt will be responsible for an application to the Bill and Melinda Gates Foundation focusing on the links between IT and food safety and how their technology made it possible.

Involvement of developing countries

Involvement should be through international organizations to deal with political issues in an appropriate way. The WHO GFN platform is a possibility and The WHO GFN steering committee has discussed the issue and is not negative. We should send a request for inclusion in the training to the GFN Steering Committee. OIE reference laboratories have training courses for developing countries and should be involved as well. When going through the GFN and OIE we will get information to the technical levels not the political levels. This is a good opportunity to get technical staff involved and build awareness. Although GFN and OIE training courses cannot cover both as the involvement has to be started at the political level.

There was a general notion that this has to be a slow process as there are many issues of great concern, e.g. trading, law suites, intellectual properties. Developing countries can keep the data and have the IT power help from other countries that already developed the system.

George Haringhuizen and Carl Sciacchitano will draft a white paper on how to involve countries with no prior knowledge about GMI. This will be a discussion paper for the next GMI meeting in UK

Metadata model

We should ask the other group what is the minimum data requirements, what data could be hidden. We need to decide on how the proceed with this to come to some kind of consensus. The Brussels and Copenhagen reports have some of the issues listed. Once we have decided on a list of issues we have to describe how to come around this. We might not be able to take care of all.
Some countries have high restriction on sharing any metadata. What do we cope with this so these countries get involved? Can the GMI databases work without metadata?
George H, Palmer, Marguerite Pappainoanou, Eric Brendon will draft a suggestion and send to the group

Review and develop the communications strategy of outbreak investigations

How to react and who to contact in case there is a suspicion of an outbreak? This issue should be on the list for the metadata list.

Advocacy

To identify advocates we will need stakeholder analysis and communication strategy

Global roll-out

We are sort of doing that all the time

Governance and ownership

At present we have not identified parallel initiatives.
We can reach out the countries through WHO by engaging decision makers into the process. Hopefully we can get some of them to participate at the next meeting.
We also need to collaborate more with the Canadians

Importance of national level engagement

We need to identify front runners at national levels. One way of getting national contact could be through international institutions (e.g WHO, OIE, FAO), EU institutions (COMM, ECDC, EFSA) and US Institutions. **Stephanie Defibaugh-Chavez will send a list from the US.**
We might also be able to get lists of who the companies selling sequencing machines to.

Know the environment

We need to raise money for ring trails. Through the European and international reference labs mandates (mainly for animal health), ANSES organize regularly ring trials (proficiency tests) with competent authorities of developing countries. If needed **Rozenn Saunier will send some information**

Legal issues

George Haringhuizen will contact WTO and ask about possible legal issues

Model framework

Model framework is part of the metadata analysis

Risk benefit

The industries are welcome in GMI and should be included in the communication strategy.

For the program in York, UK:

The program should include epidemiological presentations and examples of collaboration between technical and epidemiological groups.
We have to contact the human and veterinary epidemiologist preferably leaders to invite them to the next meeting. **Amy Cawthorne will ask the regional WHO offices and Vincenzo Caporale will ask for veterinary epidemiologist within the OIE system about this**

Date	Milestone	Responsibility
2013 Q2	Map and engage stakeholders, catalogue regulations and international agreements	George H will contact WTO about legal issues
2013 Q2	Define GMI management funding group	
2013 Q3	Advocacy paper for end-users	George H
2013 Q4	Agreement on organization form and communication strategy	
2013 Q4	Develop minimum optional metadata model	George H, Palmer, Marguerite Pappainoanou, Eric Brendon
2013 Q4	Risk/benefit. Identify / develop communication strategy to industry, academia, governments	Stephanie Defibaugh-Chacez (one pagers)
2013 Q4	Resource needs report. Coordinate funding applications	Jørgen Schlundt (Application to Bill Gates foundation)
2014 Q1	GMI should be known by 65% of professionals	Eric Brown (stands/seminars at ASM and AFP)
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)	

WG2 Update (Repository and storage of sequence and meta-data working group)

GMI5 WG2 Timelines			
Year	*Time*	*Goal*	*Responsibility*
2013	*Q2*	First flow of data into GMI repository (NCBI/EBI)	*NCBI/EBI*
2013[1]	*April*	Discussion of GMI and MixS std. harmonization at GSC15.	*NCBI/EBI*
2013[5]	*May*	Discussion at INSDC for two new tags for pathogen data:	*NCBI/EBI*
2013[2.3,4]	*Q3*	*GMI reporting standard*	*NCBI/EBI*
2013[2.3,4]	*Q4*	*Working repository infrastructure, prototype GMI data discovery programmatic interface and generic web interface*	*NCBI/EBI*

GMI5 WG2 Timelines			
Year	*Time*	*Goal*	*Responsibility*
2014[2,3,4]	*Q1*	GMI presentation standard	*NCBI/EBI*
2014[2,3,4]	*Q2*	Feedback from GMI analysis groups for refinement of reports and standards	*NCBI/EBI*
2014[6]	*Q3*	Enhancements to APIs and web interfaces (too early to start)	*NCBI/EBI*
2014[6]	*Q4*	*GMI toolkit specification* (too early to start)	*NCBI/EBI*

1. Review optional metadata fields for inclusion into specification for GSC GMI standard in order to make presentation at GSC16 or GSC17 (a full spec of the required and optional fields need to be made)
    a. Existing minimal template for foodborne traceback as formulated at previous GMI meetings:

Table 1. Minimal Pathogen Sample Metadata Template

sample name	*unique ID for the sample*
attribute package	*Indicate the type of pathogen.* *Allowed values are "clinical or host-associated pathogen" or "environmental, food or other pathogen".* *Value provided in this field drives validation of other fields.*
organism	*scientific name of the organism that provided the sequenced genetic material- expect genus species*
strain	*strain/isolate from which sequence was obtained*
collection_date	*Date of sampling, in "DD-Mmm-YYYY", "Mmm-YYYY" or "YYYY" format (single instance, eg., 05-Oct-1990, Oct-1990 or 1990) or ISO 8601 standard "YYYY-mm-dd" or "YYYY-mm-ddThh:mm:ss" (eg. 1990-11-05 or 1990-11-05T14:41:36)*
collected-by	*Name of the person or lab who collected the sample.*
isolation-source	*Describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived.*
geo_loc_name	*Geographical origin of the sample*
lat_lon	*Report values in decimal degrees and in WGS84 system*
specific_host	*Required for 'clinical or host-associated pathogen' sample type- Taxid or organism name of host*
host-disease	*Required for 'clinical or host-associated pathogen' sample type- Name of relevant disease, e.g. Salmonella gastroenteritis. Controlled vocabulary, http://bioportal.bioontology.org/ontologies/1009 or http://www.ncbi.nlm.nih.gov/mesh*

Examples of optional info that may be useful to ask for include:

- host age
- host age range
- host sex
- outbreak code
- PFGE pattern codes
- Serovar

Therefore, review existing qualifiers that can be included in the optional qualifiers for pathogen metadata.

a. Existing INSDC feature table documentation includes source qualifiers http://www.insdc.org/files/feature_table.html#7.3
b. Existing MiXs http://gensc.org/index.php?title=Main_Page
c. Prototype template for antimicrobial resistance phenotypic assay data

Antimicrobial	Interpretation	Test Method	Vendor	Platform	Reagent	Numeric Result	Units	Comment

One prototype example exists: http://www.ncbi.nlm.nih.gov/biosample/SAMN01163409.
One additional request is that compound groups be added/searchable (ie. carbapenems as groups in addition to individual compounds).

Once a list of contributors under the GMI initiative is established (see #2 below), request a list of optional qualifiers that should be considered by those contributors

2. Create a survey for WG2 to ask GMI members who are willing to provide data under the GMI WG2 standards set forth at GMI5 (that minimum data for matching including sequences and metadata are submitted to the INSDC members NCBI or EBI). Those who indicate they are willing to contribute data should form the members who will contribute feedback for reporting and presentation standards. Those contributors who need actionable items for public health and safety will need to collaborate with the analytical working groups. Two immediate projects would be the ring trials and pilot study working groups (WG4 and WG5) to provide publicly accessible datasets as proof of concept but also as repositories for testing analytical pipelines in WG3.

3. Match standards for cluster detection and thresholds. Instead of setting absolute thresholds, if we use the weather analogy, provide a rank of critical items vs. important warnings: ex: a tornado warning means there is a tornado on the ground vs. a tornado watch, which means conditions are perfect for a tornado to form, but a tornado has not been spotted. Matches could be ranked accordingly. The data contributors established via survey and the analytical working groups (WG3) should aid in the investigation of these thresholds and to determine what is the minimal information to be actionable. This should also include submission standards for data quality.

4. Archive of the data produced by the analytical pipelines, reporting formats. NCBI is working on the technical data formats from the NCBI Pathogen Detection Pipeline. Are there new data formats required? For example, is there an exchangeable format to capture variation in the context of biological impact. Once the list of data contributors is established survey them for additional views on the data and formats in collaboration with WG3.

5. INSDC Bioprojects can be labeled by the keyword 'GMI' if they are to be considered are part of the GMI initiative and are searchable with that term (NCBI Example: http://www.ncbi.nlm.nih.gov/bioproject/?term=GMI[keyword]). A measurement of the contributions under the GMI initiative can be made by finding the data linked to these Bioprojects and reported for the GMI newsletter and the data contribution rate reviewed at GMI7. NCBI is reporting preliminary outputs at ftp://ftp.ncbi.nlm.nih.gov/pathogen/.

6. These phases are too early to start and are dependent on other stages to finish.

WG3 Update (Analytical approaches)

1st breakout group discussion WG3 on charter

WG3 participants acknowledged that there is likely to be no single solution to the tools/informatics/standards problem. While NCBI or any other centralized analysis platform will produce some type of analysis, raw data must be available for re-analysis using more manual methods. In addition, the group concluded that there are multiple analytical pathways to obtain the same results. WG3 concluded that a focus on analysis standards and methods would be its major role, as quality assessment and quality control functions rightly belong with WG4; thus further discussion focused on the development of standardized analytical pipelines and approaches to analysis of the GMI data.

In the first breakout session, WG3 focused on establishing a clear governance structure, on formulating requirements for tool validation and standardization. It was decided that as many informatics tools as possible be open-source and/or off-the-shelf, with minimal input required from professional bioinformaticians. The GMI could function as a tool validator and promoter of standards for use throughout the initiative. Recommendations were made to establish liaisons with other standardization bodies (e.g. the Global Alliance http://oicr.on.ca/globalalliance) to harmonize reporting and data format standards with other initiatives. In addition, methods for handling newly emerging or previously unrecognized pathogens would need to be established.

WG3 Administration/Governance:

- Subdivision of the discussion groups are needed, because WG3 is too broad: for instance covering clinical / public health/ research, tools versus standards.
- Information needs to be gathered on all pipelines and resources: links should be incorporated on the GMI websites to tools and reviews.

Recommendations/observations:

- A mechanism to handle 'new' pathogens may be required
- Any tool should be analyzed and validated, commercial and open source. Public health labs would need off the shelve bio-informatics tools: these are version controlled, no bioinformatician needed to use them.
- GMI could act as the central access point for decentralized databases and tools → federated approach. NCBI could offer centralized approach.
- GMI could form a group of bioinformaticians which test new software or new versions of existing software.
- It may be desirable to make a distinction between pipelines and basic pathways and algorithms. The latter should be tested and validated by the research community
- Bioformats and data interchange standards are important. GMI should give guidance here. Make a choice of standards we want to support. Needs harmonization with existing genomics standards. GIS data need to be handled. Standardization bodies already exist.
    - There are already several initiatives for standards. GMI should liaise with these:
        - Global alliance http://oicr.on.ca/globalalliance : to develop standards, both technical and regulatory for sharing genomic data (human and cancer genomics. In a first white paper a small part on infectious disease genomics. http://oicr.on.ca/files/public/White%20Paper.pdf
- The bioinformatic needs are different for public health and clinical setting.
    - Public health: tracing back contaminant from source.

    - Clinical: key is: how does this improve the treatment of the patient.
- GMI should keep the eventual end-user in mind; while currently bioinformaticians are primary end-users of the data, in the future biologists will be the end-users of the data and will need user-friendly interfaces and data/information portals with the GMI infrastructure.

Yongwei Cao: rules from government bodies part of this WG? Shipping, reporting publishing etc.
Mark Allard: certainly GMI, other WG
First assemble WG relevant units and find volunteers via GMI questionnaire.

Summary of Breakout Session 1:

- WG3 will form active subgroups
- WG3 will evaluate and/or validate potential tools and pipelines, both commercial and open source
- WG3 will develop / facilitate standards for analytical approaches and data processing (as opposed to data QC).

2nd breakout discussion of WG3, 11 Sept 2013: CGE service, sequence typing, genome assembly, tree building. Webservice for international use.

The second WG3 breakout session focused on sharing information about existing tools and approaches to disease tracking using genomics. Several institutes presented existing initiatives (e.g. FDA's GenomeTrakr, UC Davis/BGI's 100k Microbial Genome Project) that may feed data into GMI. In addition, presentations included suites of genomic epidemiology tools under development by commercial vendors (e.g. Illumina and CLC Bio) that were being evaluated as possible methods for routine analysis of GMI-generated data.

Part 1: presentations

Ruth Timme, FDA, Genome Trakr database and analysis pipeline - At the moment the database is only available for federal health labs, and only for salmonella. In a later stage open access for all public health labs. Metadata are put in the database first, later the reads via Illumina BaseSpace cloud. Works via a CLCplugin. Now all via FDA, in a later stage this will not be necessary anymore. Biggest source of errors is not the sequence errors but human errors: mistakes in metadata, accession nrs etc. FDA + CLC have developed a plugin to automatically update to SRA. Has not yet been used because of internal issues.

Martin Shumway, NCBI, Automated pathogen detection at NCBI
A pipeline is currently available for all steps of NGS analysis, starting with read trimming.
Argo: reference assisted assembly or de novo assembly or combined assembly
Coming months all reference genomes will be reannotated (only bacterial?)
Vaiant analysis, block blast, muscle

BaseSpace, Illumina Inc.
HiSeq and Miseq raw read can be streamed directly from the sequencer.
BaseSpace is hosted by Amazon. Data security: same data security environment as financial data ; cloud service supports data encrytion. Data can be private or shared and a growing number of analysis and viaualization programs (apps) free both in beta- and full versions.

Bart Weimer, 100K Microbial Genomes Project update
Dr. Weimer provided an update on the progress of the 100k genome project, with particular emphasis on the logistical details of isolate/DNA receipt and the process of moving from isolate to sequenced genome. 100k genome project prefers receipt of isolates, as customer-provided DNA is often of too low quality. Many isolates arrive contaminated too. Consequently, each isolate is tested first and purified. At the moment most of the work

is: authentication, checking completenes of metadata and getting pure isolates. Only the raw sequence reads are submitted to NCBI; all analysis is to be done by the submitter. Current work focuses on the automation of the DNA prepearation from isolates in 96 wells system. They can produse max. 600 libraries per day using the Illumina HiSeq 2500; other platforms will be added in the fututre.

Bruno Pot, Bionumerics
No centralized approach, but linked local services. Every user can decide what to share. BioNumerics 7.5 has a NGS analysis pipeline with many options.

CLC Cecilie Boysen
CLC Bio provided an update on its development of the Genomics Workbench (GWB), a common off-the-shelf analysis platform. Program can download from BaseSpace, and CLC has developed an SRA upload tool along with a growing number of analysis tools. It is also possible to build in external application into the program.

Gary van Domselaar, Canada's IRIDA
IRIDA: integrated, rapid infection disease analysis http://www.IRIDA.ca
Federated database design, adopt existing ontologies, open standards and publicly available genomics tool and pipeline, using Galaxy

Ole Lund (DTU) - Web based methods for genomic epidemiology
Dr. Lund presented an overview of tools available from the DTU web-servers that provide centralized limited analysis of antibiotic resistance gene profiles (ResFinder) and the *in silico* multilocus sequence typing tools that can utilize sequencing reads and assembled sequences uploaded from anywhere with internet access and a reasonably quick internet connection.

Part 2: discussion

The second breakout session was devoted to discussion of requirements for Genomic Epidemiology tools to enable robust, defensible data analysis within the GMI framework.

- Requirements were discussed regarding potential CGE services, which included:
    - High-quality reference sequences
    - Quick, high-throughput data analysis with minimal analyst input required
    - Easy-to-use software or automated pipelines
    - Inclusion of clinical and epidemiological data in analysis functions
    - Desire for packaged PCR primer design tool for new outbreak strains
- Balances to be weighed:
    - Centralized versus distributed analytical solutions – both needed to accommodate complexity of NGS data and the diversity of user experience with bioinformatic methods
        - "Quick and dirty" automated pipelines
        - Labor- or informatics-intensive, user-directed analysis
        - Crowd-sourcing of analysis
    - Standardizatoin versus flexibility – exclusive focus on standards leads to ossification of the pipelines and lack of flexibility in the face of changing technology

WG3 Plan for the upcoming period:

- Collect list of contributors plus their role in standards development and develop a further survey among end-users for CGE needs. Ideas and volunteer opportunities can be published in the GMI newsletter
- Volunteers needed for standards survey and tools evaluation

3rd Breakout session WG3: updating strategic plans, milestones coming 6 moths, define highest priorities.

Fiona Brinkman: in this stage of the project the organization form needs to be changed.

For this breakout session WG3 is sub-divided in three subgroups: 1:. . moving data ?? Ole Lund. . 2 : . . . 3: core pipeline

Subgroup chared by Gary van Domselaar: definition of core pipeline.

- Cecilie Boysen: inventory needed amongst members; What NGS questions, which software is used. Can this be done by DTU?
- Martin Shumway: epigenitc modifications need to be included (f.i. methylation)
- Ruth Timme: has a simulated data set, almost ready, can be shared.
- Mark Allard: has compared 454, ion torrent and hiseq data. Small difference between the platforms. GenomeTrakr has also been built in order to get testing datasets. There is a beta version at NCBI.

Mark Allard: volunteers are needed for simulated exercises. People willing to sequence, people willing to test data outcome
Set of MR organisms with known genotype and phenotype, volunteers to submit to NCBI
Fiona Brinkman wants to be involved in testing algorithms. First training, then testing. This can be done with the same set of data if necessary. Things to be tested: virulence, horizontal gene trtansfer.
Set up different sets of testgroups.
Some GMI participants need to go back to their groups and get permission tto participate further.
Gary van Domselaar: wants to make a questionnaire for end users: who wants to join what initiative, with which method. Elena Bolchakova wants to be involved in this.
Gary: core NGS analysis: QC, assembly variant detection genome annotation, contamination detection, phylogeny
Ole Lund: important focus: moving data, combined with formats for standards, SRA uploads wiki

Summary:

Key next move is to form subcommittees that will meet more frequently to deal with specific tasks. Volunteers need to be contacted and regular reports need to be communicated through the website or newsletter.

WG4 Update (Ring trials and quality assurance working group)

Attendees: Errol Strain (FDA), Rene Hendriksen (DTU), William Hsiao (UBS), Frank Aarestrup (DTU), Andreas Nitsche (RKI), Darcy Hanes (US FDA), Eija Trees (US CDC), James Pettengill (US FDA),

Discussion Topics:

1) The survey results are in agreement with the targets planed prior to GMI5 meeting in Denmark
   a) To ensure all of the WG are on the same page and acknowledge the decisions according to the updated action plan and details below, the results (Excel) of the survey will be disseminated to the WG members in week 38, Sep 2013 by DTU.
   b) WG members should assess and ensure that vital partners / collaborations have submitted data to the survey.
   c) Based on the survey results, it was decided to target the following organisms; Salmonella, Escherichia coli, and Staphylococcus aureus
   d) It was discussed whom would supply isolates for the ring trial. FDA and DTU have suitable WT isolates in stock but agreed that Errol (FDA) would contact ATCC (Brian) to elaborate about their standpoint / interest to supply isolates. Isolates should be selected by the end of Sep 2013.
   e) The trial will as previously decided consist of 2 components; a wet component of live culture and DNA and a dry component of electronic data sets. The wet component will consist of 2 isolates of Salm, E.C, and S.A, respectively of both DNA and live culture – a total of 12 samples representing 6 strains If closed genomes for these strains are not available then FDA will finish the genomes using PacBio. The dry components will consist of 3 data sets, one per organisms, comprised of 20-40 isolates for cluster analysis and SNP calls. FDA will draft data sets by the end of Sep 2013.

2) Speed up the process in relation to the current action plan
   a) It was decided to send out a last reminder for participation in the survey in week 39, Sep 2013 by DTU.
   b) The survey will be terminated by the end of Sep 2013 by DTU.

3) Appointed indicators, milestones and responsibility to the first part of the plan.
   a) The action plan was updated – see file or above.
   b) If the electronic data sets are too large for some sites to download it was suggested to ship the data sets of the dry component on USB sticks.
   c) The invitation letter should direct recipients to a web site where options for participation in component 1 and 2 and organisms should be indicated – however, we'll seek a more rapid solution for the pilot while the invitation letter will be sent out by mail. The invitation letter would contain information to capture participants shipping account numbers. However DTU will cover additional shipping costs to max 10.000 USD in case participants have issues paying for the parcel eg. developing countries.
   d) DTU will establish a sub-WG to draft all relevant documents for both ring trials. This would include the development of invitation letter including a ring trial description and note about provision of import permits etc.., development of a pre-notification, development of instructions: what to do (also what reference to use)/ how we would measure the performance / questionnaire to capture the methodologies, pipelines used.
   e) The ring trial wiki will be postponed until the full roll out of the ring trial.
   f) DTU with provide a private FTP site for the ring trial – ready by end of Sep 2013
   g) DTU will create a sign up web site for the full roll out of the ring trial.

4) Arrange conference calls with the entire WG to ensure commitment and delegate tasks / responsibilities
   a) Conference calls in Sep 2013 to discuss management tasks and responsibilities.
   b) Conference calls in Q1-2 2014 to discuss analysis. It will be manually conducted by several WG units!

5) Establish a virus sub-group to initiate discussion on how to conduct a virus ring trial
   a) Establishment of a virus sub PT group - link by Email Andreas Nitsche from Robert Kock to Marion and Anna Charlotte etc.

WG 5 Update (Pilot Project Working group)

Communications structure

- Sharepoint for docs
- Conf calls
- Email distribution list

Governance / Steering structure

WG5 debated two models for an organizational structure, keeping in mind that the GMI work is not yet funded. Ideally, an exercise would be funded, with dedicated resources)

a) Steering group
Liaisons to WGS
Project desgin
Executive control

b)
Working Group steering committee
Liaisons to other working groups
Ad hoc project groups
Group lead for each sub-project (exercise leader/monitor)

- Technical monitor;
- Trusted agent (they would be the one that has all the answers for blinded study integrity/chain of custody
- Sample log or electronic notebook to feed back to technical or exercise monitor or both

Option b seemed better received and something to strive for. Even though to do to this degree, option b will need to start small with volunteers and then expand later with details and dedication with funding
Concern raised about the expansion of GMI to 'all microbiology' – dilutes our effort and our message.

WG 5 purpose discussion?

- Visible/transparent pilot design(s) - agnostic approach, "anyone can play"
- Platform-wide pilot of the other GMI WGs outputs - bring the exercise to the other WGs. Where we are working up the first part, and the other WGs are working up the latter parts
- WG5 needs to design exercises with appropriate questions being asked, measurable metrics for success, and attainable results up front: Elements of a successful exercise include clear definition of goals, design, metrics, decision points, and elements of "theater."
- Design notional scenarios to the GMI as a whole- sample prep; platform SOP; data analysis pipeline - goal is to be science-based, evidence-based, fair and transparent
- WG5 participants will set up the measurement criteria, gap analysis, design criteria in consultation with the other WGs
- WG5 will also document the exercise output to include communicate the exercise outcome to GMI and to stakeholders
- Take proof-of-concept success to funding body to acquire funding (WG1 – we will liaise with them)
- Genomic signature detection using NGS and appropriate pipelines to convert to screening tool and red flags for cluster detection - - the demonstration aspect requires the co-analysis of the WGS data sets with the meta data
- At later stages, validation of the approach may involve the secondary developers of diagnostics – (see discussions in WG3) to include conversion of genomic signatures to screening tests/molecular diagnostic candidates or strategies for intervention
- Exercise development will initially follow the KISS principle (Keep It Simple, Stupid): start with easiest analysis and then move to the more complex quasispecies and then mixed samples
- For each exercise, WG5 should provide lessons learned, feedback to the other WGs – work with them to make it successful – identify the weakest links/modes of failure to flush them out and solve by providing the GMI infrastructure with recommendations and well-designed scenarios
- How will we do it: liaise with the other WG's

- WG 5 is the lab side wherein we acquire the test data to make it robust – integration of the micro, test labs, field labs, sequencing, tools - - cooperative approach

Sept 11th discussion:

- Pilot project has an element of "theatre" which is required to impress upon funders that GMI works and is fundable
- An ideal scenario would include developing countries in partnership to increase perception of GMI as perceived as a "global" initiative – best if all countries can be included in pilot; but not realistic owing to funding limitations
- WGS is not currently the focus of need in developing countries and they likely do not have sufficient resources,even though they are interested in participating in GMI
    - Compromise: stage the pilot project(s) – first coordinate isolate collection from volunteer global partners (especially lower resourced entities who may only be able to provide isolates rather than WGS)
    - These partners could then be staged in later to do sequencing once GMI has access to funding to provide = Pilot project invitation to anyone that can meet the minimal criteria that the Ring Trials group establishes and is willing to do the sequencing
- Identified two existing pilots are already being done by the USA:
    - *Listeria monocytogenes* real-time prospective surveillance project already being done
    - FDA Salmonella 12-state mock exercise expanded to include one external partner
- **WG5 goal**: design a pilot with no funding, while encouraging global participation – model design on a laboratory response network (get the right answer (cluster ID) is the penultimate goal). Theatre to demonstrate that GMI genomic epidemiology approach that is standardized is better than what currently exists
- WG5 can design a pilot - but will likely will need WG1 to help execute the pilot if some partners are causing issues that prevent its success
- Identified potential trust/political/international issues – A retrospective (rather than prospective/real-time or "live fire") study may be more easily accepted by global partners. For example the global analysis of a small number of isolates from one strain from more than 3 years ago or older would likely assuage some of the potential landmines of doing analysis of ongoing outbreaks.

Catalogue of ideas for pilot need

- *In silico pilot exercise, with a limited amount of WGS data.* Digital data exchange validation study wherein whole-genome sequences are deposited to a central repository and each partner has access to the data to analyze with their own means (O104-like crowdsourcing exercise). For example, modeled after the Assemblethon wherein everyone is analyzing the same data sets (inclusive approach, open to all who are willing) – upload the data and the scenario for download by participants; analyze; pre-establish metrics to evaluate whether does everyone comes up with the same conclusion for cluster detection. Could be based on an artificial (or real?) outbreak scenario. Q: Select an unpublished but controlled scenario that is realistic – poll for volunteers for read data sets? Ideal scenario would be isolates sequences that are temporal and some closely related and some more outlying, more distantly related.
    a. Questions being addressed: How well does data transfer work? Establish minimum standards. How well does data analysis/Cluster ID work? Notification of status – enhancement of the notification of a common, emerging profile in clusters? Turnaround?
    b. Needs only a small task group such that not too many people know about the scenario in advance
    c. Biggest challenge will be coordinating the epidemiological analysis – the pilot needs to include this epidemiology component to demonstrate the importance of meta data, controlled vocabularies, (epi & NGS ontologies) and standardized inputs/output means the interoperability for successful, rapid turnaround cluster identification
- *Wet-scale pilot exercise.* Smaller-scale pilot that goes from wet lab to full on data analysis with partners that can afford
- Larger-scale exercise bringing in more partners (might depend on funding)-

Pilot Tasks breakdown:

- Define the metrics and standards (the objectives) to which the demonstration pilot will be measured-poll the other WGs to identify the minimum metric or stds they wish to prioritize first [tasked to: WG Liaisons]
- Design a retrospective pilot project/scenario according to the above exercise criteria (ie. Appropriate exercise designed to measure the efficacy of the standards/metrics put forth by the other working groups) [tasked to: scenario or project task group]
- Coordinate, assist with the pilot exercise [tasked to: project coordinator]
- Distribute the exercise data sets/isolates to the participants
- Execute the pilot exercise [project team and participants]
- Measure the pilot exercise in consultation with the other working groups [project team with WG liaisons]
- Prepare pilot evaluation report to summarize the pilot [WG5]
- Provide feedback exercise back to the other GMI working groups WG1,2,3,4 to assist with the evolution of the GMI and the next exercise design

By next 6th GMI meeting – September 2014

1. In silico pilot done?
    - Q: How quickly can it be done?
    - Mark Wilson – to provide summary of the FDA first exercise
2. Wet lab pilot experiment designed to demonstrate that the modules under development are interoperable and compatible and all the pegs in the pipelines are truly functional. Need to design the pilot based on the answering the need questions – with input from the other GMI working groups/subject matter exports

Global Microbial Identifier