IRDiRC
INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

Meeting report series

Report of the 5th DSC WG Genome/Phenome teleconference

30 September 2014

Organization

Organized by: IRDiRC Scientific Secretariat
Teleconference

Participants

Prof Anthony Brookes, Leicester, UK, chair
Dr Kym Boycott, Ottawa, Canada
Peter Bauer, Tuebingen, Germany, in replacement of Prof Olaf Riess
Prof Ada Hamosh, Baltimore, USA
Prof Jim Lupski, Houston, USA
Prof Yves Moreau, Leuven, Belgium

Dr Barbara Cagniard, IRDiRC Scientific Secretariat
Dr Sophie Höhn, IRDiRC Scientific Secretariat

Apologies

Prof Han Brunner, Nijmegen, the Netherlands
Dr Xavier Estivill, Barcelona, Spain
Dr George Patrinos, Patras, Greece
Prof Olaf Riess, Tuebingen, Germany
Dr Wei Wang, Shenzhen, China

Agenda

- Feedback from the Diagnostic Scientific Committee
- Discussion on possible topics for the WG
  - Metadata requirements and standardization
  - Minimum content open data conventions
  - How to share knowledge without exposing data
  - Pathogenicity classes and related matters (expressivity, penetrance)
  - Query without direct access to primary data
- Summary and conclusion

IRDiRC
INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

REPORT

Feedback from the Diagnostics Scientific Committee

The Diagnostics Scientific Committee met by teleconference on 2 October 2014 and reviewed the progress of each DSC WGs.

Regarding the topics of the WG on Genome/Phenome

- An update was given on the Matchmaker project.
- Metadata requirements and standardization: DSC approved the idea of organizing a workshop on the topic. It will be necessary to find funding for it.
- Minimum content open data convention: DSC is interested by the topic and would like the WG to write a short statement on it.
- Other topics mentioned by the WG on their 4th teleconference (July 2014): DSC agreed that the WG should further discuss these topics (topics of discussion of this teleconference).

However, as there is no funding available, the WG will not be able to work on all of these topics.

Discussion on possible topics for the WG

Metadata requirements and standardization

Datasets need to be complemented with metadata (information that contextualizes and describes the data) to increase the effectiveness of data use, discovery and sharing.

Metadata are related to items such as:

- How data are generated
- Who generated the data
- Who can access the data
- What consents are behind the data
- What form of sharing is possible
- What quality metrics apply

For metadata to be interoperable between environments, there is a need for standardization of structure and minimum contents of the metadata.
Members of the WG are not aware of any metadata standardization project, but other initiatives such as Global Alliance for Genomics and Health, and RD-Connect may be interested in the topic.

Members agreed that this WG will not develop metadata standards but should work on:

1) Finding teams/projects working on metadata
2) Compiling and comparing work of others
3) Fostering the organization of a workshop on the topic [already agreed by DSC]

Minimum content open data conventions

If more data are made openly available, then those data can more easily and more quickly be used by others. Concerns about subject identification can be countered if the openly shared data comprises a very shallow layer of data (fully anonymous information, such as HGVS variant name and headline phenotype term from HPO).

- ⇨ The concept of having some level of minimal open data as a convention or standard practice needs to be discussed and promoted, and this WG should play an important role in this.

Geno+Pheno Query API

A common computer-to-computer 'language' (API) for searching across genotype plus phenotype data is neede. IRDiRC members have previously devised genotype plus phenotype query strategies, and between this WG and other WG members, a lot of experience in how to approach the challenging 'phenotype similarity' question is available.

- ⇨ This WG should organize its past work into a v1.0 geno+pheno query API, deploy it in its projects, and hook up with GA4GH to compare and consolidate efforts. This would also be a good topic for an invitation-only workshop.

Model/Standard Consent Clauses

The lack of consistency and clarity in consent clauses is preventing the automated management of permissions in data discovery and data sharing situations. Ideally, standard set of core consent clauses, with unique, immutable, computer-readable IDs should be generated.

This is particularly needed for consent for discovery uses, as few consent procedures have previously considered this. Likewise, there is no universal opinion on the necessity of consent depending on the type of data.

- ⇨ As the WG Ethics of the Global Alliance for Genomics and Health is working on this topic, this WG agreed to focus on metadata and query APIs rather than consent, but to stay informed of any evolution in the field. It should also be noted that the IRDiRC WG on Ethics and Governance is now planning to work with GA to develop standard consent clauses specific to Rare Diseases.

How to share knowledge without exposing data

Conversion of data to knowledge would provide another resource for discovery. However, standards and tools are not available for this purpose. When considering the type of knowledge that would be useful to help IRDiRC reaching its objectives, members of the WG agreed that there are many, and the topic is large and complex. One exemplar would be variant frequency information, with or without phenotype (i.e. in control populations). The Exome Variant Server for American populations is one such resource. FINDbase (www.findbase.org) established in 2006, has a large collection of aggregate level data on clinically relevant genomic variation allele frequencies from almost 100 populations worldwide,

while ALFRED is yet another resource worth being looked at. FINDbase was an integral part of GEN2PHEN and now of RD-CONNECT and work is currently being done to extract allele frequency data from whole exome/genomes (work in preparation).

The WG was informed that the new IRDiRC WG on Population Controls Variant Datasets is working on assembling a set of standards on how to aggregate set of data for data sharing of several control populations around the world.

- ⇨ The WG on Genome/Phenome will define how it can help the WG on Population Controls Variant Datasets.

Pathogenicity classes and related matters

Pathogenicity is for the moment mostly described through simple scales. However, concepts such as expressivity, penetrance, etc. are not included in the definition of pathogenicity. A more complete definition of pathogenicity – beyond just a number - would require more data and experience and is perhaps not realistic at the moment for most RDs. Several groups are now working on the topics: Clin Gene, LOVD, Enigma.

It emerged from the discussion that clarity on scales and process for pathogenicity is needed. A publication is being written on OMIM's approach on the topic. A discussion paper on the topic, that could become a reference paper, would be useful.

- ⇨ WG will monitor this topic.
- ⇨ WG will review the publication on disease-associated gene for consideration for recommendation by IRDiRC

Finding variants in federated datasets

The idea of being able to find (discover) variants in federated datasets lies at the heart of the GA4GH 'Beacon' project (http://genomicsandhealth.org/our-work/initiatives/beacon-project). This now plans to develop further to discover more information and to provide different levels of control of access. In this context, the WG was informed of the tool 'NGS-logistics' (https://ngsl.esat.kuleuven.be/), which is a prototype system to query positions in the genome for frequency of variance across multiple centers without need direct access to primary data. It is planned to include phenotypic data in the tool in the future.

Summary and conclusion

Seven areas of required standardization were discussed (above), and this WG has also previously discussed 4 other areas (Globally Unique Patient IDs, Bio-Resource Metrics System, Matchmaker Exchange, Data Standards Clearinghouse). Going forward, this WG should continue to work with GA4GH on the Matchmaker Exchange project, and initiate new efforts into:

- Metadata requirements and standardization

- Geno+Pheno Query API
- Minimum content open data conventions

Additionally, this WG should closely monitor and promote the work of the WG on Population Controls Variant Datasets as it relates to sharing knowledge without exposing data. Other needs and opportunities in tune with that may emerge.

The development of the International Consortium of Human Phenotype Terminologies (ICPHT) is a perfect model for moving forward on the 3 items below:

- ⇨ Identifying the problems by WG discussion, key players
- ⇨ Organizing a workshop
- ⇨ Keep interaction, follow up by email, etc. to produce a valuable output such as Interoperability system

Deliverables

- Preparation of a summary of the above topics and an action plan
- Discuss with the WG on Population Controls and Variant Datasets during the next teleconference of the WG on Genome/Phenome