IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Meeting report series

Report of the 1st DSC Working Group on Population Controls Variant Datasets teleconference

16 September 2014

Organization

Organized by: IRDiRC Scientific Secretariat
Teleconference

Participants

Dr Xavier Estivill, Barcelona, Spain (chair)
Prof Fadh Al-Mulla, Safat, Kuweit
Dr Hidewaki Nakagawa, Tokyo, Japan
Dr Hui Jiang, Shenzhen, China
Dr Justin Paschall, Cambridge, UK
Dr Kym Boycott, Ottawa, Canada
Prof Tayfun Özçelik, Ankara, Turkey

Dr Barbara Cagniard, Scientific Secretariat

Apologies

Dr Michael Brudno, Toronto, Canada
Prof Pak-Chung Sham, Hong Kong, China
Dr Peter Robinson, Berlin, Germany
Prof Woong-Yang Park, Seoul, Korea

REPORT

Overview of IRDiRC

IRDiRC is an initiative of the NIH and the European Commission with the purpose of reaching the following two main goals:

- Produce diagnostic tools for most of rare diseases by 2020
- Develop 200 new therapies for rare diseases by 2020

IRDiRC is a consortium of funders investing at least 10 million USD over 5 years in research projects contributing towards IRDiRC objectives and invited patient advocacy group. IRDiRC has three Scientific Committees:

- Diagnostics Scientific Committee (DSC)
- Interdisciplinary Scientific Committee (ISC)
- Therapies Scientific Committee (TSC)

The purpose of the DSC is to try to translate research in rare diseases into diagnostics for patients. The DSC oversees five Working Groups (WG):

- WG on Genome/Phenome
- WG on Model Systems
- WG on Ontologies and Rare Disease Prioritization
- WG on Population controls Variant Datasets
- WG on Sequencing

Each WG will bring new insights and strategies to be considered by the DSC. The Chair of the DSC will report the priorities identified by the WG to the Executive Committee. This bottom-up approach will allow better efficiency and will keep the patient at the heart of IRDiRC missions.

The main aim of this WG is to define how to proceed to facilitate the aggregation of anonymized variant frequency data from specific populations to obtain variant population-specific datasets that are freely available to the research community to accelerate rare disease gene discovery.
Main questions are:

- What is available?
- Which projects are ongoing?
- How to standardize data?
- How to link with other initiatives (such as the International Cancer Genomic Consortium)?

Projects presented by participants

BGI (China)

BGI collected 6000 samples (1/3 is Chinese) from collaborators, which cover more than 200 diseases. 4000 samples, including patients and their family, were already sequenced.

Control databases are separated in 2 populations:

- Han population: 1000 samples (WES)
- Caucasian population: 1000 samples (WES)

Department of Molecular Biology and Genetics, Bilkent University (Turkey)

This group has sequenced 215 Turkish individuals including around 100 controls, all exome sequencing with the exception of four WGS. Patient groups are extreme obesity, essential tremors and some very rare diseases.

In addition, Yale Center for Mendelian Genomics (Murat Gunel's group) has WES of more than 2000 individuals from Turkish population.

European Bioinformatics Institute (EBI, UK)

EBI is building a database called European Variation Archive (EVA; http://www.ebi.ac.uk/eva/), which will be launched in October 2014. This is a large public database - genotype, locus frequency and variants – populated with several work sets (Geuvadis, genome from Netherlands, UK10K 10,000 genomes project, etc.).

FORGE Canada/Care for Rare Canada (Canada)

Through the project FORGE Canada (2011-2013) followed by the project Care For Rare Canada, 1500-2000 patients with rare diseases were sequenced (WES) and data gathered in the same place to accelerate gene discovery. A patient with a rare disease serves as a control for a patient with another rare disease. Canada is a mix of several populations including Asian population and several sub-group of Caucasian population.

Genome Arabia (Kuwait)

This project, sponsored by the Qatar National Research Fund, aims to create an Arab-specific databases for the "normal" population, cancer-specific population (breast, colorectal and prostate cancers) and rare diseases-specific population by using sequence of 500 hundreds to a thousand individuals from seven countries in the region — Qatar, Bahrain, Kuwait, United Arab Emirates, Tunisia, Lebanon, and Saudi Arabia — using WGS.

At the moment, WGS were conducted on:

- 140 samples from Qatar
- 50 samples from Kuwait

Geuvadis European Exome Variant Server (Spain)

The Genetic European Variation in Disease (GEUVADIS) consortium created the GEUVADIS European Exome Variant Server (GEEVS; http://geevs.crg.eu/), that has around 2000 WES from aggregated data from European countries (Spain, Netherlands and Germany). Aggregates have a minimum of 50 analyzed samples. The server allows search data from chromosome position, references, variant frequency, etc.

Laboratory for Genome Sequencing Analysis (Japan)

This laboratory, part of the RIKEN Center for Integrative Medical Sciences, mostly conducts cancer genome sequencing analysis and has for the moment 300 WGS of liver cancer patients. In the group, 1000-2000 WES of patient with common diseases are available, which would be a good control for rare diseases.

Discussion on the possibility of data sharing

It was proposed to the WGs members to test two databases to decide if they could be good resources for the community by uploading some data.
The two databases are:

- European Variation Archive (EVA; http://www.ebi.ac.uk/eva/): A password protected FTP site can be created for the WG members to upload data. Individual VCF and BAM files will be able to be uploaded and data will be available to GEUVADIS to download
- GEUVADIS Exome Variant Server (GEEVS; http://geevs.crg.eu/): aggregated data of a minimum of 50 samples should be submitted.

The location of the uploading is thus not important as EVA and GEUVADIS, which had already been working together, will be exchanging data.

- Submitted data should be covered by appropriate consent for public sharing. It seems to be the case for most of the WG members if data are de-identified and/or already published.
- Standards on how to prepare aggregate data and submit them will be circulated so that data will conform to the same protocol and the data will be exchanged and viewable in either place (EVA or GEUVADIS). Draft guidelines already exist but may need to be completed.

Other points

Two other points were mentioned:

- The possibility to briefly met at ASHG meeting in October for those present
- Publication of a paper on this WG effort

Main deliverables

- Set up a pipeline between EVA and GEVAUDIS server
- Uploading of data as exercise
- Identify additional populations and members that we could invite to join this WG
- Provide name of available databases to be able to retrieve public data
- Schedule another teleconference for October