Meeting report series

Report of the 2nd DSC Working Group on Population Controls Variant Datasets teleconference

28 October 2014

Organization

Organized by: IRDiRC Scientific Secretariat
Teleconference

Participants

Dr Xavier Estivill, Barcelona, Spain (chair)
Prof Fahd Al-Mulla, Kuwait City, Kuwait
Dr Kym Boycott, Ottawa, Canada
Dr Michael Brudno, Toronto, Canada
Dr Justin Paschall, Cambridge, UK
Dr Jianguo Zhang, representative of Jun Wang, Shenzhen, China

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

Apologies

Dr Hidewaki Nakagawa, Tokyo, Japan
Prof Tayfun Özçelik, Ankara, Turkey
Prof Woong-Yang Park, Seoul, Korea
Dr Peter Robinson, Berlin, Germany
Prof Pak-Chung Sham, Hong Kong, China
Prof Jun Wang, Shenzhen, China

Agenda

1. Update on the EVA and GEEVS databases to upload exome sequencing data
2. Exploration on population control data sets that could join or link with this initiative
    - Within the rare disease community
    - Within other communities

REPORT

1. Update on the EVA and GEEVS databases to upload exome sequencing data

The European Bioinformatics Institute (EBI, UK) and the Genetic European Variation in Disease consortium (GEUVADIS, Spain) aim to create a global network to exchange exome sequencing data, having a shared dataset which would also exchange with the Single Nucleotide Polymorphism Database (dbSNP).

The GEUVADIS European Exome Variant Server (GEEVS) standard format with its recommended settings will certainly be adopted by the European Variation Archive (EBI project) as no guidance is provided at the moment on how to generate allele frequencies from sample level data.

A summary email on the recent release of the EBI European Variation Archive (EVA) was sent. Members of the Working Group (WG) were asked to give their feedback on the EVA. GEEVS and EVA will continue to communicate over the next months.

The question on how to approach GEEVS and EVA with a bunch of exome data that could be analyzed in an aggregated way was raised during the teleconference.

For submission:

- At sample level data, a bunch of individual VCF files or an aggregate VCF file with samples columns need to be turned into allele frequencies/genotype frequencies. To do so, the GEEVS guidelines should be followed. They explain how to generate an aggregate VCF file with allele frequencies. Then, people will only need to submit this to GEEVS or EVA.
- If allele frequencies are already available but submitters do not want them compute or cannot compute them, or if submitters computed allele frequencies with GEEVS guidelines and want to submit the aggregate VCF file directly to EVA, EBI will make a copy available to GEEVS.
- It would be better not to submit aggregate VCF files to both EBI and GEUVADIS as it would be confusing and that data submitted in one of the databases will be made available to the other database.

Aggregated data do not impact publication efforts, and could then be submitted before publication, when they are generated. Allele frequencies are the main focus of this effort. Individual level consented data that could be shared should also be submitted.
Issues of privacy and ethics are the primary reason of collecting aggregate data as opposed to detailed data. People should submit **aggregate data and/or individual data.**

The more granular data can be in terms of specific counts within a specific disease subset, the better. Privacy issues will determine granularity.

The point of this entire process is actually to get population level control datasets. A few sentences describing the population cohort would be enough:

- Critical annotation: precise if control or disease population, ethnicity.
- The disorder annotation is much less important as there will also be disease causing mutations in the general population.
- At individual level, it is recommended to indicate the phenotype information.

Some members of this WG will submit data to GEEVS and EVA:

- Genome Arabia: aggregated data will be submitted. Explanations on how to calculate allele frequencies from their data is required. ~30 "normal" Arabs, ~50 breast cancer, ~50 multiple sclerosis and ~150 for whole genome.
- FORGE Canada/Care for Rare Canada: calculation of the allele frequency needs to be realized before the submission of VCF file. Be aware that these French Canadian data will have a lot of causative mutations.
- BGI: ~200 samples could be submitted.

⇨ A document (i.e. Google doc) should be created to start including WG members datasets that could be submitted to GEEVS or EVA, in order to generate information about the number of samples and different populations that could be available for submission to the databases.

2. Exploration on population control data sets that could join or link with this initiative

Within the rare disease community

The Exome Aggregation Consortium (ExAC) could joint this initiative. It has analyzed over 61000 exomes and has announced the public release of allele frequency data.

Within other communities

The WG should link with other communities such as cancer, autism, psychiatric disorders, etc with initiatives in genome sequencing for optimal numbers of controls:

- In Canada, the Autism Speaks Ten Thousands Genomes program (AUT10K), which is working in collaboration with BGI.
- The International Cancer Genomic Consortium is searching for controls samples that could come from other projects.
- The Epilepsy Project in Canada will be doing 25 000 whole genome over the next two years.

Main deliverables

- Give some feedback on the EVA
- Create a Google doc with the list of WG members datasets available for submission
- Explain how to calculate allele frequencies to some members for them to submit their data to GEEVS or EVA

- Get in touch with some members to explain them how to submit data to GEEVS and EVA
- Submit aggregate data to GEEVS and EVA
- Approach International Cancer Genomic Consortium
- Approach the Epilepsy Project and AUT10K initiative
- Send a list of suggestions for other initiatives needing control data on genome sequencing, to initiate a contact with them
- Schedule another teleconference for next month