Proposal Title: A Chilean population database of exon-level genetic variability.

Young Investigator contact information.

Evelin Valeria González Feliú

e-mail: pancancer_eg@udd.cl / evefeliu@gmail.com

phone number: +56949160632

Introduction.

The massive growth (and use) of sequencing technologies has led to unprecedented developments over the last decade. This is the result of different international collaborative projects that contribute to an extraordinary increase in the knowledge of the mutational spectrum of diseases. This knowledge generation is especially significant in high morbidity and mortality disorders caused by highly penetrating variants (typically protein-coding). Among the strategies used to discover new disease variants, especially in monogenic disorders, population frequency-based filtering has proven a helpful tool. In addition, an important caveat to interpreting Latin American cancer patient’s genetic data is the under-representation of Latin American individuals in the global resources characterizing the frequency of both germinal and cancer genome variants. Therefore, it seems clear that the availability of healthy controls is a decisive factor to advance the discovery of new determinants of the disease. These observations reveal the need for catalogs of genetic variation of specific populations. However, most studies have been carried out in individuals of Northern European ancestry. Only a few initiatives to study genetic variation in the Chilean population have been performed to date. In 2015, a study estimated the local ancestry of 313 Chilean admixed samples from two case-control studies on Hantavirus infection (n=112) and 22q11 microdeletion syndrome (n=201) genotyped on an Affymetrix 6.0 GeneChip Array. On the other hand, the Chilegenomico project inferred the ancestry in 3,349 Chilean mestizo population using a small panel of 150 ancestry-informative markers (AIMs) and a study relies on genome-wide single nucleotide polymorphism data from 2,039 admixed Chileans. These efforts to characterize the Chilean population are focused on relatively known variants commonly present in microarrays, and the data is not readily available to the scientific community.

To date, there is no publicly accessible WES (whole exome sequence) or WGS (whole genome sequence) data for the Chilean population. Despite their recognized utility, large-scale sequencing projects of local healthy population cohorts require expensive consortium-based projects to obtain a representative sample of the target population. In this scenario, a crowdsourcing strategy, which allows adding existing data produced by local genomic projects, can provide a viable alternative to traditional work schemes. In fact, so far, we have already identified and committed WES data from two hundred Chilean subjects for this project.
Vision.

To build a publicly available database of genomic variation of the Chilean population, generated as a collaborative crowdsourcing effort to collect, harmonize and aggregate sequencing data produced by local biomedical genomic projects through common ethical and bioinformatic standards.

Aims.

This project aims to create an aggregated database of the genomic variation of the Chilean population at the exon-level.

1. Harness the power of existing WES or WGS obtained for different projects and add another 100 healthy population control individuals to increase representativeness.
2. Calculate the allele frequencies of all variants found in the Chilean population, after the use of a standard bioinformatics methodology for all the samples to create an aggregated database of the genomic variation at the exon level.
3. Generate a web based public aggregated variants database to promote access to data following GA4GH guidelines.

Impact.

This repository will be used as a pseudo-control population for finding new disease-causing variants and genes. It will allow the identification of low-frequency variants (MAF below 5%) that are specific to the Chilean population, to potentially establish the risk of many diseases that differ in different populations according to their genetic background and the importance of considering different genetic ancestries.

Executive Plan.

<table>
<thead>
<tr>
<th>Aim/activities</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Obtain data from NGS (WES or WGS) of 200 people with Chilean nationality who have been sequenced in different projects (for other purposes) through collaboration and a 100 “healthy” controls subjects.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1 Prepare Ethics Committee submission for reuse of genomic data from other projects and to perform WES in healthy control subjects</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2 Procure 100 germline DNA samples from U. de Chile Biobank suitables for WES analysis</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3 Submit 100 samples for WES analysis as a service (such as Novogene)</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
2. Calculate the allele frequencies of all variants found in the Chilean population, after the use of a standard bioinformatics methodology for all the samples to create an aggregated database of the genomic variation at the exon level

2.1 Analyze WES or WGS data through a standardized bioinformatic methodology

2.2 Perform quality control over genomic variants to select proper data for further processing

2.3 Aggregate variants information into an allelic frequency table

2.4 Program a nonSQL database and a front engine variant/gene search and visualization web tool

3. Generate a web based public aggregated variants database to promote access to data following GA4GH guidelines.

G2MC Partners:

<table>
<thead>
<tr>
<th>G2MC members</th>
<th>work position</th>
<th>country affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabriela Repetto</td>
<td>Professor and researcher</td>
<td>Facultad de Medicina, Universidad del Desarrollo, Chile.</td>
</tr>
<tr>
<td>Boris Rebolledo</td>
<td>Professor and researcher</td>
<td></td>
</tr>
<tr>
<td>Ricardo Armisén</td>
<td>Professor and researcher</td>
<td></td>
</tr>
</tbody>
</table>

Non-G2MC

<table>
<thead>
<tr>
<th>Non-G2MC</th>
<th>work position</th>
<th>affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evelin González</td>
<td>Bioinformatics Engineer</td>
<td>Facultad de Medicina, Universidad del Desarrollo</td>
</tr>
<tr>
<td>Alejandro Blanco</td>
<td>Bioinformatics Engineer</td>
<td></td>
</tr>
<tr>
<td>Katherine Marcelain</td>
<td>Professor and researcher</td>
<td>Facultad de Medicina, Universidad de Chile</td>
</tr>
</tbody>
</table>

Future funding: Plans for follow-on funding for continued support.

With the population genetic variation database using 300 unrelated Chilean samples, we will find polymorphisms specific to the population. However, increasing the number of samples will be necessary to identify low-frequency variants and rare (MAF below 0.5%) specific to the Chilean population. With the results of the first version (300 subjects), we will apply for the national grant CORFO and/or FONDEF for public interest projects to obtain extra funding to achieve this goal (ca. 1000 subjects).
 Budget with brief [1 page] budget justification

- Miscellaneous costs: Services for the shipping and NGS WES analysis of 100 samples.
  1. WES Novogene USD 224/sample x 100 samples, USD 22400/year
  2. International Shipping + Sample processing USD 3600/year

Total: USD 50,000.-

Annex

References