**WES Representative Sample of the Portuguese Population**

To create a representative sample of the Portuguese population we gathered all the available WES samples from the Center for Predictive and Preventive Genetics (CGPP).

Given that previous research was initiated in CGPP about the frequency of variants located in genes associated with recessive diseases (ReproScreen project) the minimum sample to have a representation of the Portuguese population was calculated, based on the equation 1.

* n - number of samples needed to identify carriers with a population frequency at or above 1/1,000.
* Z - standard (Normal) distribution that reflects the level of confidence chosen (1.96 for a confidence interval of 95 %).
* *p* - population proportion, in this case 1/1,000.
* E - acceptable margin of error, established as 0.001.

By replacing the variables with the referred values, the minimum number of samples required to have representativity of a carrier frequency higher or equal to 1/1,000 in the Portuguese population is 3,838. To achieve this goal three major steps were conducted: i) filtering process of all the available CGPP’s patients’ information; ii) gathering all the information of all the patients in regard to localization (first 4 digits of the postcode, municipality and district) and iii) down-sampling to ensure representation of the Portuguese population.

**Data filtering**

The initial dataset consisted of 12,167 WES samples, and the information about this dataset was in a CSV file (exported from the CGPP laboratory database). The patient-related information included the ID, date of birth, gender, type of test, situation, postcode, municipality, district, clinical information, family information, request information, relatives, and other collections. For data normalization and curation, we followed these steps:

1. Ensure uniqueness.
2. Exclude Pre-Natal diagnosis (PND) samples.
3. Restrict the dataset to unrelated individuals – first method.
4. Exclude consanguineous couples.
5. Restrict the dataset to unrelated individuals – second method.

Firstly, we eliminated the duplicates by looking at other collections from the same patient, with an exclusion criteria based on the capture kit, and 12,060 WES entries remained.

From this step forward, we added flags to the unwanted data based on the exclusion criteria, specified in Table S1-1.

**Table S1-1.** Flags used and respective label.

|  |  |
| --- | --- |
| Flag | Label |
| 0 | To keep |
| 1 | PND sample |
| 2 | Consanguineous sample |
| 3 | Family member |
| 4 | Other sequencing kit |
| 5 | GIAB sample |

PND samples were flagged based on the column gender equal to fetus or, when the column ‘date of birth’ was equal to “-” and “test type” was equal to “pre-natal”. This resulted in data from 11,858 patients to deal with.

The next step was to find the information about family relationships within patients. We extensively analyzed our data to find family-related samples. Firstly, we filtered the column relatives to see the samples with no data (11,195 samples). We checked other columns in search of any word or patient ID that could be related to the patient and family relationships (mother, father, sibling, parents, grandparents, grandchild, uncle, aunt, cousins, nephew, niece, spouse). This left us with 2,514 samples with relevant information for further scrutiny. After looking at the data, there were some patterns, so we decided to consider them to automatically fill the column “Family”. The exclusion criteria for patients that contained 2 IDs for the same relationship, was by capture kit used. After all the processing, 353 samples remained and were assorted manually, since they had no pattern, or the time spent to automate the process would not be worth it.

After having the family relationships, we identified the families and got one family representative for each family. A new column to store a flag for Family was created, with 0 being patients with no family and the other families were attributed a number per family.

The consanguineous cases were identified through a list of consanguineous patients’ IDs previously curated. The prioritization rules were healthy member(s) and women(an). However, when there was no healthy member, the next one was to prioritize flagging the male member in these couples. Initially, this choice was driven by the future use of WES samples in other projects, emphasizing the need for sufficient sensitivity in detecting rare variants in X-linked genes. Furthermore, the decision was influenced by the gender distribution in Portugal, where the population predominantly consists of females as illustrated in Table S1-2. Contrarily, in the context of CGPP, there was a higher representation of males.

**Table S1-2.** Gender proportions comparison between the Portuguese population (PORDATA) and CGPP WES samples.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Population | Men | Women | Men (%) | Women (%) |
| Portugal | 10343066 | 4920220 | 5422846 | 47,57 | 52,43 |
| CGPP | 11858\* | 6217 | 5641 | 52,43 | 47,57 |

\*CGPP samples not including the PND samples and duplicates.

The number of Unis after selecting the non-consanguineous cases was 11,848.

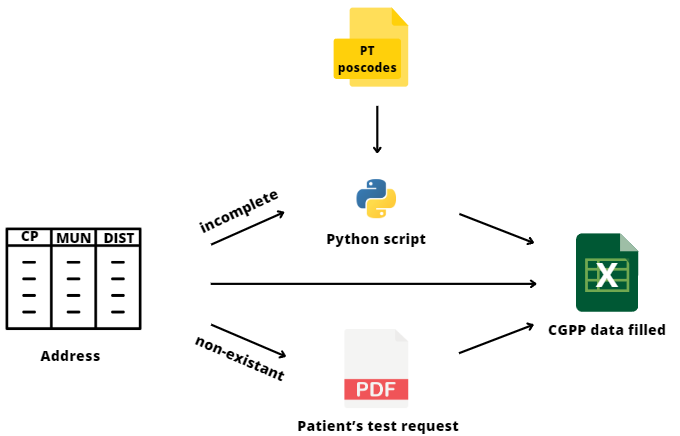
Then, we identified the WES Trio families by searching the words “WES” and “TRIO” in the clinical information column and also filtered the column test type for “*confirmation/exclusion*”. Our prioritization rules were the following:

1. Healthy member(s);
2. Women(an);
3. Capture kit;
4. Address information.

Using the grouping strategy, we were able to find patterns and decide, accordingly, which cases to exclude from our analysis by adding the flag 3. After dealing with the WES Trio cases, we dealt with the other families’ cases, following the same process of identifying the patterns and flag with number 3 the samples to exclude. In the end of this process, we had 11,328 samples with no flag.

**Gathering information for samples’ selection**

The address information about the samples was either completed, incomplete, inconsistent or non-existent for a subset of samples, thus a more complex process was performed (Figure S1-1).



**Figure S1-1.** Scheme for address information gathering.

We started by identifying the samples with all the information in the correct format (first 4 digits of the postcode, municipality and district) and then try to find data inconsistencies (e.g.: postcode that did not correspond to the written municipality or district).

A postcode is composed of a numerical part and a postal address. The numerical part has 7 digits and the first 4 are hyphen-separated from the last 3. The postal designation is the name of the locality, for example: 4200-135 PORTO. The first digit indicates one of nine postal regions, the following two digits indicate the postal distribution centers, and the fourth digit is the municipality. The last three digits are the specific address. To represent the municipalities, only the first 4 digits are needed.

The following steps were done using Python. The first step was to find the list of the Portuguese postcodes and respective municipalities and districts (available at: <https://en.youbianku.com/Portugal>, accessed June 2023). The file with the all the Portuguese postcodes, municipalities and districts.

Our samples’ data was processed and a list of the patients with no data was generated. After consulting the PDF files for genetic tests’ requests, the missing data was manually filled in the Excel file. After that we re-ran the previous Python script to place everything in the right format. The Python script was used to format the postal codes to a 4 digits number. However, several situations were encountered:

1. Postcodes with less than 4 digits;
2. Postcodes with format “XXXX-XX’' Locality”;
3. Postcodes with format “Locality XXXX Locality”;
4. Postcodes with format “XXXX Locality”;
5. Names of districts or municipalities in the Postcode column.

All these possible scenarios were addressed.

In the end, a new column “flag\_pc” was added, where 0 means that the patient has no suitable localization and 1 means that the patient has suitable localization. The final file contained information about gender, health status, postcode, municipality, district, family group, flag and flag pc.

**Data down-sampling**

From the 12,167 patients, only 8,885 had the information needed to find representativity (postcode, municipality and district). As previously mentioned, a representative sample of the Portuguese population requires at least 3,838 individuals. Since the data available for localization has the first 4 digits of the postcode, municipalities and districts, the distribution of the Portuguese population per municipality was used. This demographic data was sourced from Pordata, a Portuguese statistical database covering Portugal and Europe. For this analysis, the total number of residents from the most recent 2021 Census was used.31

An Excel file was created to store information about the total number of residents for Portugal, the name of the municipality and the total number of residents for each municipality, according to the Census 2021. The CGPP data with the total number of patients (8,885) and the total number of patients from each municipality was added in another column.

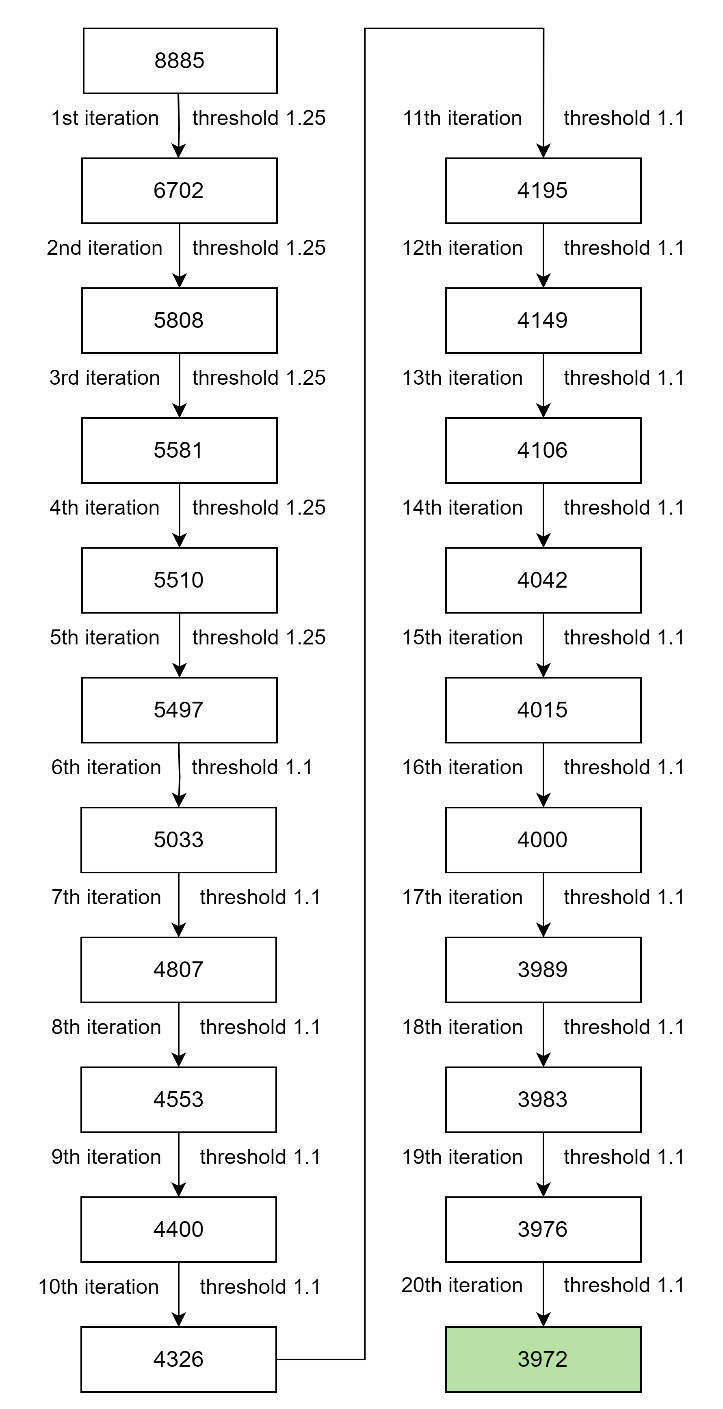
After having the raw data, we followed the steps here described:

1. Calculate proportions for the Portugal (PT) data using equation 2.
2. Calculate proportions for the CGPP data using equation 3.
3. Calculate ratio between CGPP and PT proportions using equation 4.

In order to have a representative number of patients in a determined municipality, the ratio needs to be as near as possible to 1. Municipalities with ratios higher or lower than 1 are overrepresented or underrepresented, respectively.

1. The next step was to calculate the CGPP population needed to have representativity for each municipality using equation 5.
2. Create a bar graph to visualize the distribution of municipalities based on intervals of the calculated ratio for each municipality.
3. Filterthe ratio Excel column from point 3 to **ratio above 1.25**. Marking the start of the iteration process.
4. Create a new column for the CGPP population (number of CGPP samples) updated for the municipalities filtered in step 6. The update is based on the calculations from Equation 5. This iteration will reduce the total number of CGPP samples compared to the previous iteration.
5. The steps are repeated from points 2 to 7, only for the CGPP data.
6. When we reached the 5th iteration, the number total CGPP population was stabilizing, so we changed the ration threshold to 1.1, which was used in the next 15 iterations (Figure S1-2).

After performing 20 iterations to the data, we ended up with 3,972 samples.



**Figure S1-2.** Scheme of the number of samples resulting from the 20 iterations.

We had a total of 8,885 samples with information, from which we needed to select 3,792 samples. Using Python, we followed the steps described below, prioritizing the capture kit used (kit1 > kit2 > kit3 > kit4).

1. **Import** patient IDs and respective municipalities from the **8,885 samples available.**
2. **Import** the **number of samples per municipalit**y **needed** for representativity.
3. Create a **Python dictionary** with the **municipalities as keys** and the respective **patient IDs as values**.
4. Create listsof all **patient IDs** **for each capture kit**.
5. Separate **patient IDs** per list of capture kits.
6. **Create list of municipalities** (of the dictionary created) where the number of samples needed for representativity is **equal to** the number of patients available for that municipality (**93 municipalities**), independent of the capture kit. The **patient IDs** of these municipalities will be automatically used the next steps of this work.
7. **Repeat point 3** for the **remaining 200 municipalities**.
8. For each municipality of point 7, check if the number of kit1 patient IDs was enough for the number of samples needed. If this number is equal to the number of samples needed, the list would be complete. If the number is greater than the number of samples needed, the selection of the patient IDs for that municipality is randomized. After that 76 municipalities remain for sample selection.
9. **Repeat point 3** but **only** for the **remaining 76 municipalities**.
10. For each municipality of the last dictionary created, check how many patient IDs are kit1 and right after check how many of the remaining patient IDs of that same municipality are kit2. If the number of kit2 patient IDs covers equally the samples needed, the selection is straightforward (patient IDs kit1 + patient IDs kit2). If the number of kit2 patient IDs is greater than the number of samples remaining needed, these are randomly selected.
11. **Repeat point 3** but **only** for the **remaining 47 municipalities.**
12. **Repeat point 10**, but now begin by considering the availability of kit1, kit2 and kit3 samples by this order. The last capture kit checked is always the one under consideration.
13. **Export** all **select patient IDs** (3,972) to an **Excel file**.

After these steps, an additional step was performed: the exclusion of municipalities with a ratio below 0.5, ending up with 3,941 samples.

This methodical approach ensures the representativity and diversity of our sample, providing a solid foundation for the subsequent stages of this research project.