Big Data Supporting Public Health Policies

- SC1-PM-18-2016

Project **H2020-727658 / IASIS**

http://project-iasis.eu/

Genomic Use Case report

Maria Esther Vidal, Samaneh Jozashoori, Gian G Tartaglia, Magdalena Arnal Segura, Dietmar Fernandez Orth, Fotis Aisopos

Status: Final (Version 1.0)

November 2020

Project

Project Ref. no	H2020-727658
Project acronym	IASIS
Project full title	Integration and Analysis of Heterogeneous Big Data for Precision Medicine and Suggested Treatments for Different Types of Patients
Project site	http://project-iasis.eu/
Project start	April 2017
Project duration	3 years
EC Project Officer	Dr. Jose Valverde Albacete

Deliverable

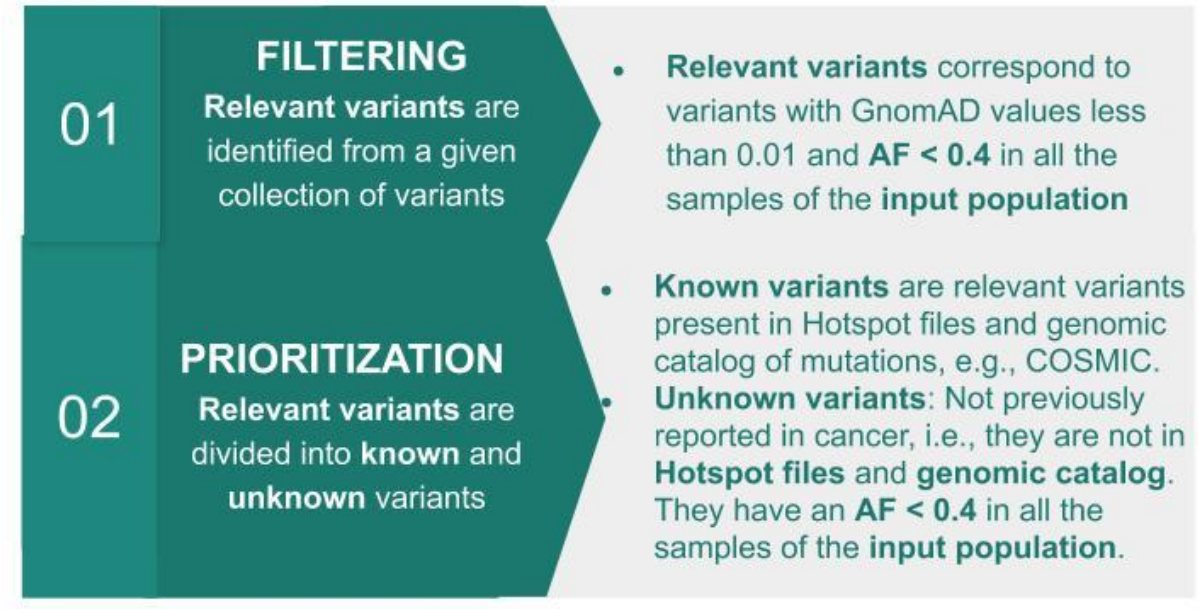
Deliverable type	Report
Distribution level	Public
Deliverable Number	D1.3 (Appendix of the Final Progress Report)
Deliverable Title	Genomic Use Case report
Contractual date of delivery	
Actual date of delivery	
Relevant Task(s)	WP3/Task 3.3 WP5/Tasks 5.2
Partner Responsible	LUH
Other contributors	CRG
Number of pages	13
Author(s)	Maria Esther Vidal, Samaneh Jozashoori, Gian G Tartaglia, Magdalena Arnal Segura
Internal Reviewers	Fotis Aisopos
Status & version	Final
Keywords	Genomic Data, mutations, Genomic API

Executive summary

This document reports on the genomic use case that develops a computational pipeline, able to identify the most frequent mutations observed in the populations of lung cancer and dementia.

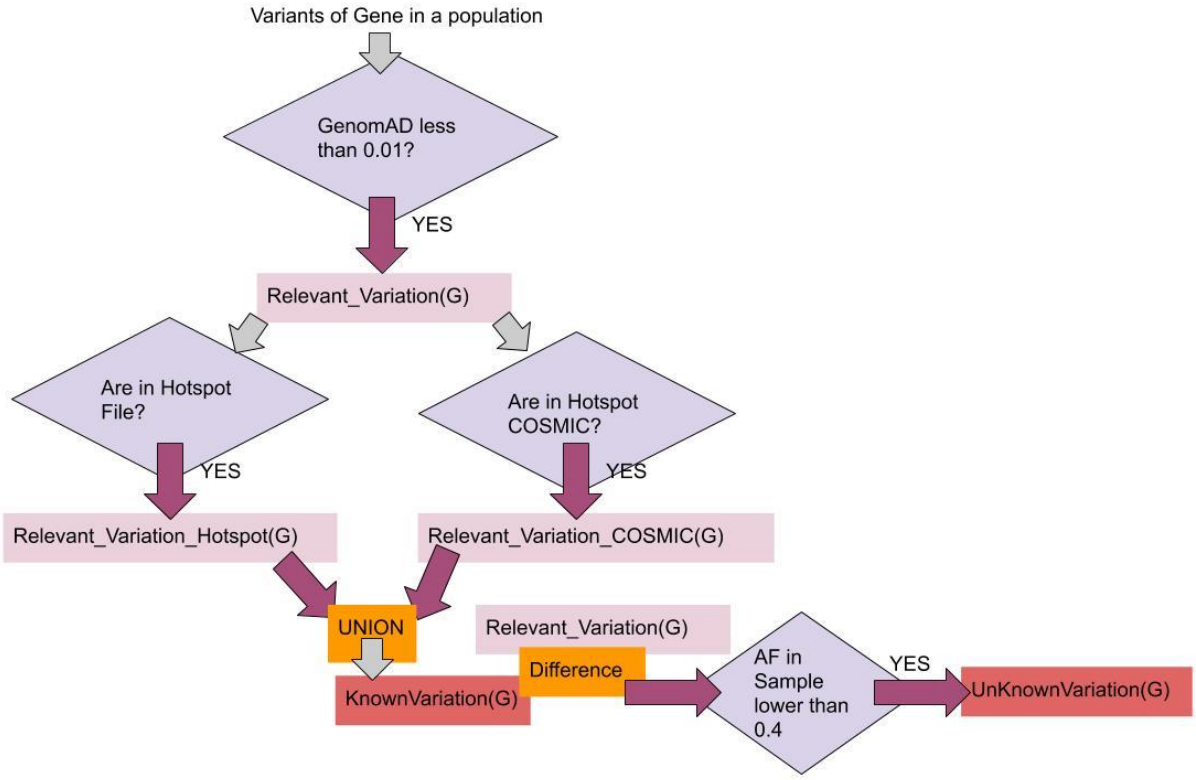
Pipeline for classifying mutations in lung cancer patients

A two-fold process has been followed to discriminate the relevant variants from a group of variants observed in a population of lung cancer patients. First, the **filtering step** is applied to discard the relevant and irrelevant variants. Then, the **prioritization step** classifies the relevant variants into known and unknown. The following Figure depicts the two steps of the pipeline. In the **filtering step** allele frequency of a mutation reported in the Genome Aggregation Database (GnomAD[1]) is used to discard variants commonly present in healthy populations. Thus, only variants with **aggregate allele frequency** in GnomAD less than 0.01 and **with values of allele frequency lower** than 0.4 in all the samples of an input population are considered **relevant**. Then, once irrelevant variants are filtered out, the relevant variants are classified into **known** and **unknown** variants. **Hotspots and genomic catalogs** like COSMIC, are used to identify which variants have been already reported; these variants correspond to **known variants.** Those which are not present in either **Hotspot files or genomic** catalogs are considered **unknown**.



The flowchart in the following Figure implements this pipeline.

[1] https://gnomad.broadinstitute.org/



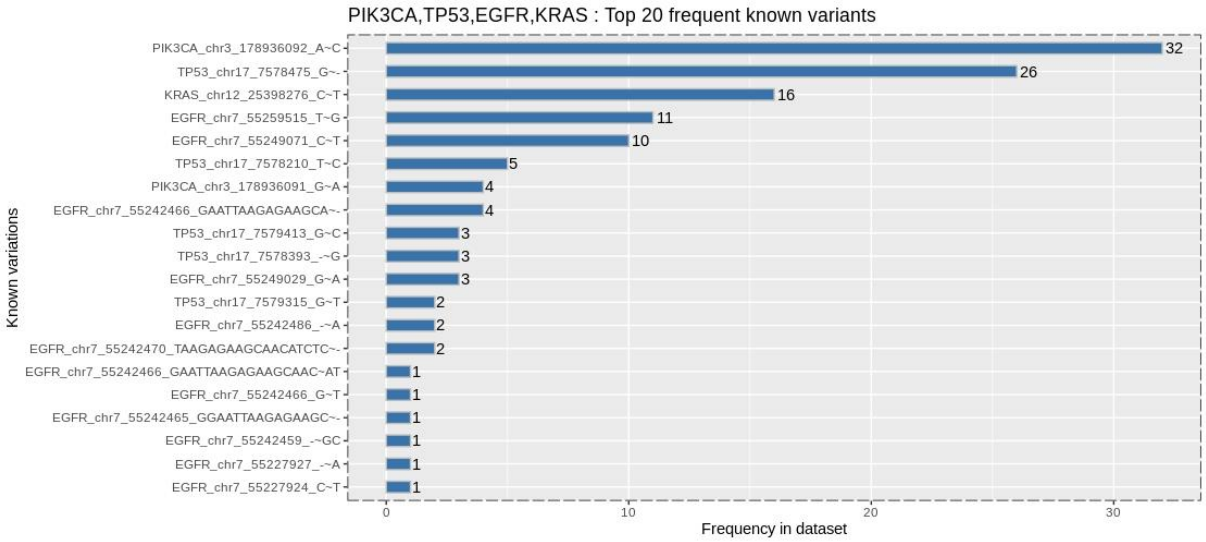
Using this flowchart, a Web API has been implemented; it is executed over the iASiS knowledge graph and characterized the variants of a particular group of genes, identified in the samples of a population of lung cancer patients. Then, for the patients associated with the relevant known and unknown variants, their main features are also collected from the knowledge graph. As a proof of concept, the mutations associated with the genes EGFR, KRAS, PIK3CA, and TP53 have been analyzed. The results of these analyses reveal promising insights about the most frequent variants observed in the population of lung cancer patients stored in the iASiS knowledge graph.

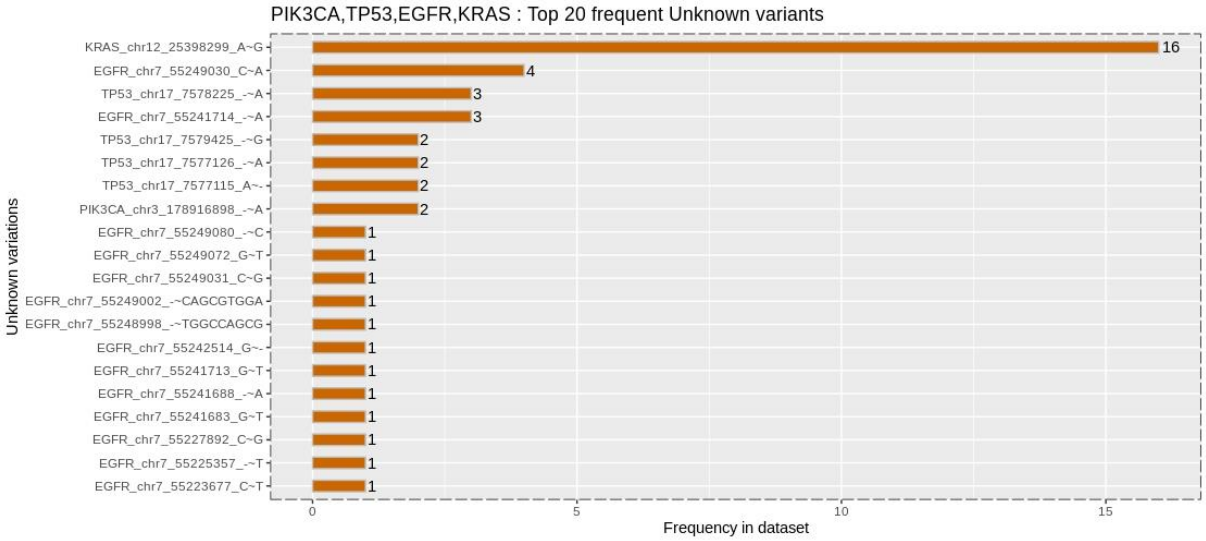
Method:

A cohort of lung cancer patients is considered as input. They received chemotherapy in line 1 of treatment. A pipeline implemented by the partners of CRG, and more specifically by Dietmar Fernandez and Magdalena Arnal, was followed to extract **somatic genomic variants** from samples of liquid biopsies of **55 patients**. The liquid biopsies of the **51 are taken after line** 1 of treatment and only 4 patients had the liquid biopsy before the treatment line 1. The following Figures depict the top 20 frequent relevant variants that are known and unknown according to the execution of the Web API described above, respectively. It is important to highlight that the top 3 known variants, i.e., *PIK3CA_chr3_178936092_A~C, TP53_chr17_7578475_G~*, and *KRAS_chr12_25398276_C~T* are not commonly observed in lung cancer patients. According to COSMIC, *PIK3CA_chr3_178936092_A~C* has been mainly observed in breast, oesophagus, and ovary as primary tissues. Furthermore, the variant *KRAS_chr12_25398276_C~T* is reported in the primary tissues Haematopoietic:and_lymphoif_tissue, ovary, and thymus. Additionally, in the genomic database Pan-cancer[2], TP53 mutations are the most common in the studied cancer-patients, and PIK3CA mutations are also frequently observed. Nevertheless, in Pan Cancer, *PIK3CA_chr3_178936092_A~C* is observed in samples of 27 donors and only one of those has lung cancer, while *TP53_chr17_7578475_G~* is only observed in our population.

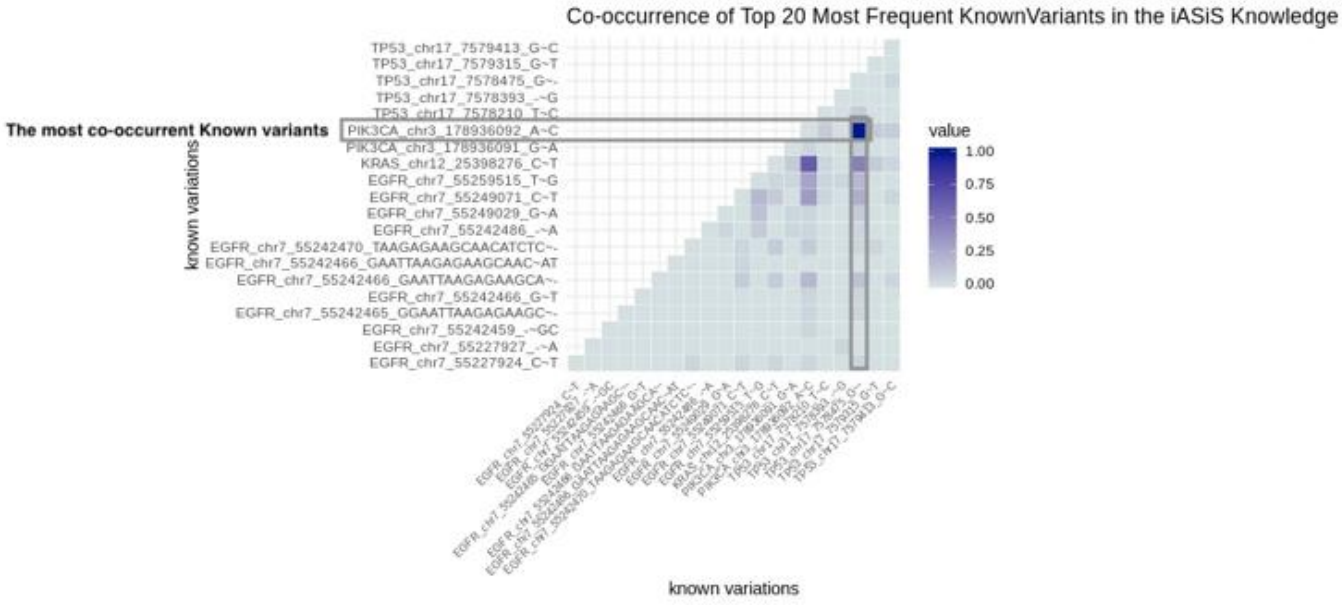
[2] https://www.embl.de/campaigns/pancancer/index.html

Additionally, the co-occurrence among variants V_i and V_j is computed as the number of patients who are associated with both V_i and V_j.

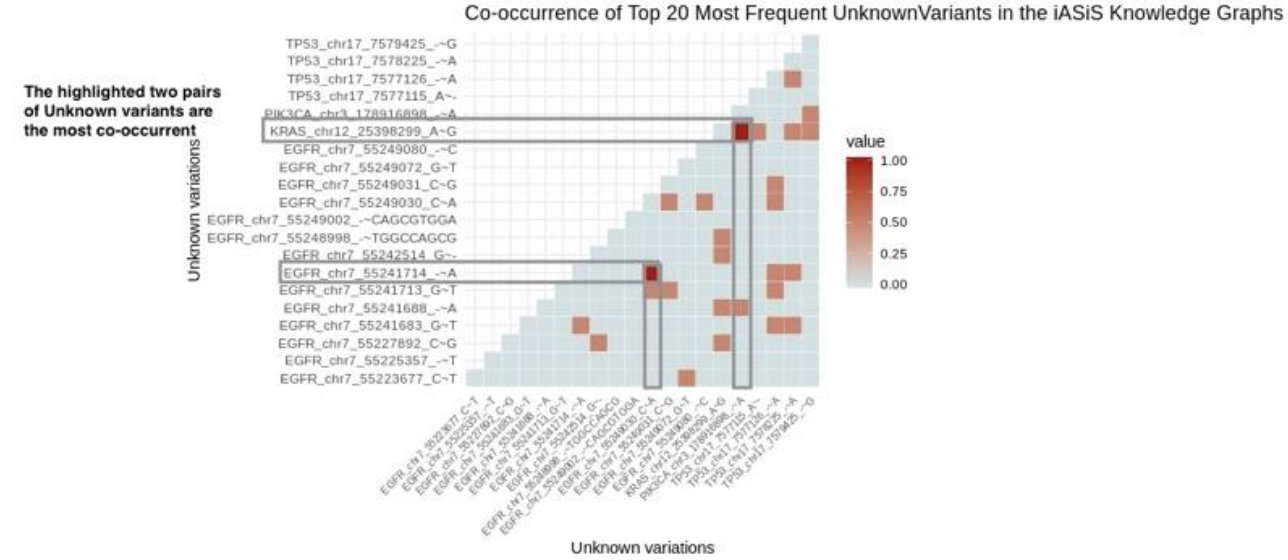




The following Figure illustrates - with a heatmap - the results of the computation of the co-occurrence in known variants. *PIK3CA_chr3_178936092_A~C*, and *TP53_chr17_7578475_G~* are the genetic variants that co-occur most together, and also individually, they co-occur with other genetic variants, e.g., with *KRAS_chr1225398276_C~T*.



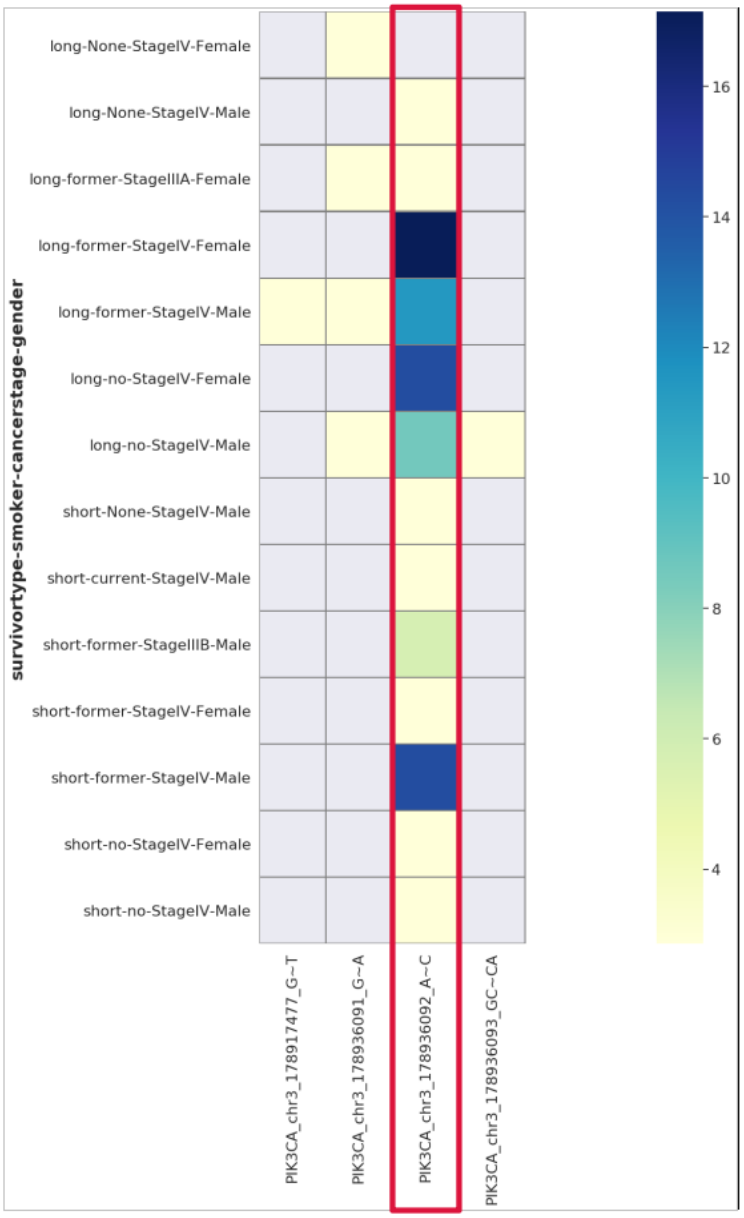
Similarly, the co-occurrence matrix is computed for the unknown variants; The next Figure reports on these results. The pairs of unknown variants: a) *KRAS_chr12_25398299_A~G* and *PIK3CA_chr3_178916898_ . ~A* b) *EGFR_ch7_55241714_ . ~A* and *EGFR_ch7_55249030_C~A* are the ones that co-occur the most. The existence of the variant *PIK3CA_chr3_178936092_A~C* in the population of lung cancer with non-small lung cancer (NSLC) has been recently reported by Romero et al.[3] for patients taking Osimertinib. However, contrary to the study reported by Romero et al., the patients in the cohort studied in the iASiS genomic use case, are taking diverse chemotherapy drugs.

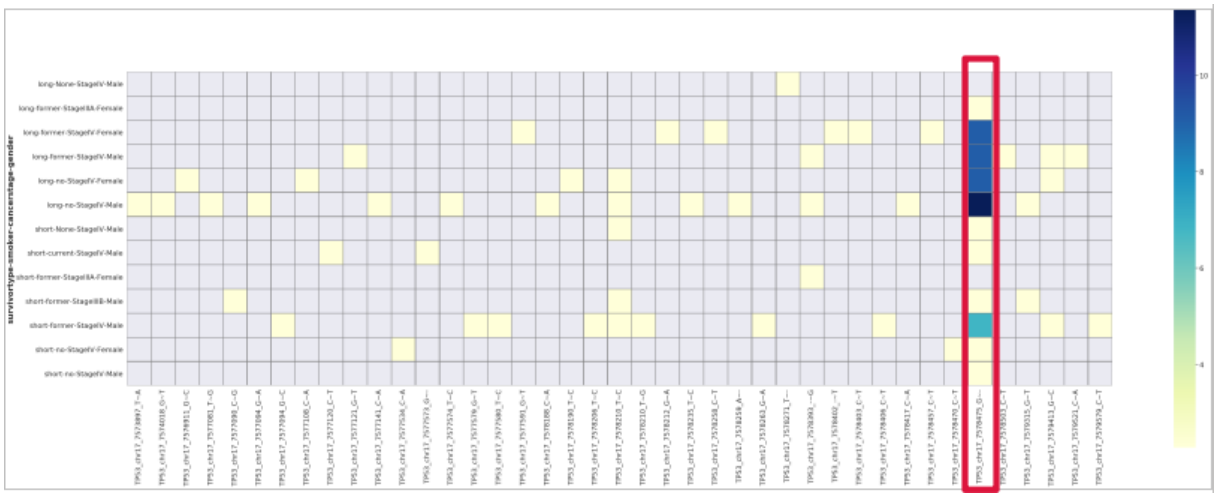


The patients associated with the studied variants are also analyzed with the aim of finding patterns that could be a potential indicator of significant implications in survival time. A patient is considered a long survivor if his/her survival time is greater than 24 months, otherwise, the patient is considered a

[3] Romero A, Serna-Blasco R, Alfaro C, et al. ctDNA analysis reveals different molecular patterns upon disease progression in patients treated with osimertinib. Transl Lung Cancer Res. 2020;9(3):532-540. doi:10.21037/tlcr.2020.04.01

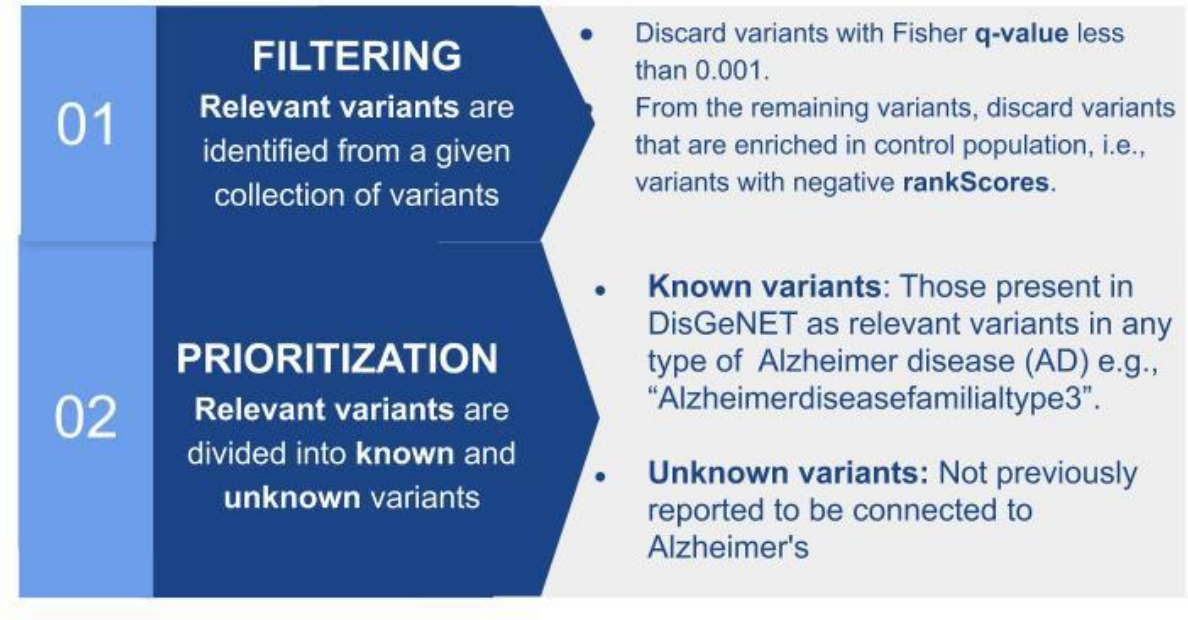
short survivor. The *PIK3CA_chr3_178936092_A~C* and *TP53_chr17_7578475_G~* are mostly observed in groups of long survivors with cancer in Stage IV. 57.2% of the patients who have a **known variant** have the variant ***PIK3CA_chr3_178936092_A~C*** and are **long survivors,** while 41% of the patients who have a **known variant** have the variant ***TP53_chr17_7578475_G~*** and are **long survivors**. Since these variants have not been observed in lung as primary tissue, their existence in conjunction with the fact that many patients with these variants are in Stage IV, suggests that they may be associated with metastasis. Nevertheless, there exist short survivors who also have these two variants, so a further analysis of all the main characteristics of this cohort of lung cancer patients is needed. The next Figures present in a heatmap frequency of patients with the variants of genes PIK3CA and TP53, respectively.




survivortype-smoker-cancerstage-gender
long-None-StageIV-Male
long-former-StageIIA-Female
long-former-StageIV-Female
long-former-StageIV-Male
long-ne-StageIV-Female
long-ne-StageIV-Male
short-None-StageIV-Male
short-current-StageIV-Male
short-former-StageIIA-Female
short-former-StageIIB-Male
short-former-StageIV-Male
short-ne-StageIV-Female
short-ne-StageIV-Male
10
8
6
4

Pipeline for classifying mutations in dementia patients

A two-fold process has also been followed to discriminate the relevant variants and irrelevant variants in patients with Alzheimer's Disease. First, the **filtering step** is applied to discard the relevant and irrelevant variants. Then, the **prioritization** step classifies the relevant variants into known and unknown. The following Figure depicts the two steps of the pipeline.



In the **filtering step** relevant variants correspond to those present in regions previously reported in GWAS catalog[4]. A control subset with people > 70years without any neurodegenerative disease, and the variants equally present in cases and controls (**adjusted Fisher qVal**) are discarded, i.e., **with a Fisher q-value less than 0.001**. From the remaining variants, we discard variants that are enriched in the control population, **i.e., variants with negative rankScores.** Variants highly enriched in AD and Dementia (using a **Fisher qVal < 0.001**) with respect to the controls are candidates to be **predisposing to the disease.** Then, once irrelevant variants are also filtered out, the relevant variants are classified into **known** and **unknown** variants. **Genomic catalogs** like DisGeNET, are used to identify as relevant variants those associated with any type of Alzheimer Disease. Otherwise, the relevant variants are considered **unknown**.

Method

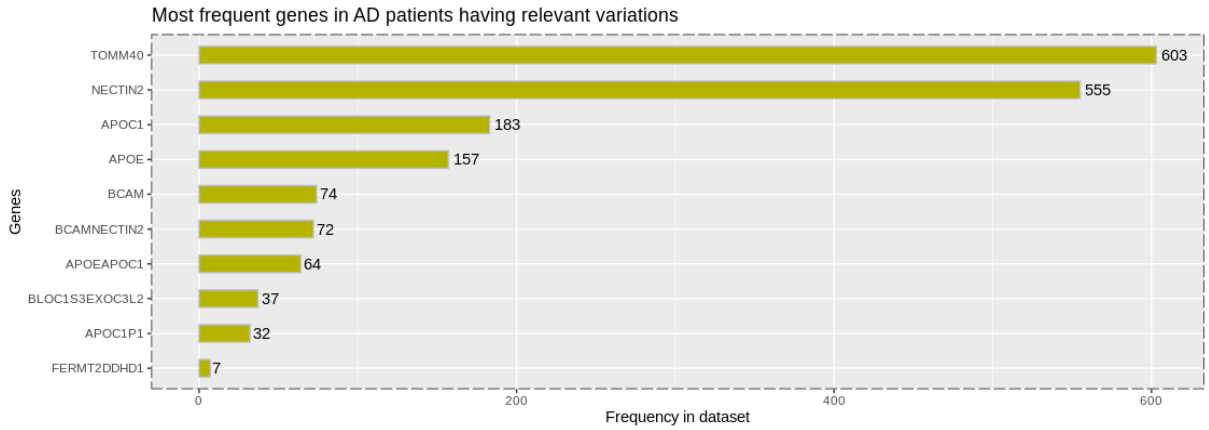
A **pipeline implemented** by the partners of CRG, and more specifically by Dietmar Fernandez and Magdalene Arnal, was followed to **extract genomic variants from the genomic data available in UK-Biobank** for patients who are associated with Alzheimer Disease (AD) as primary condition. These 1730

[4] https://www.ebi.ac.uk/gwas/

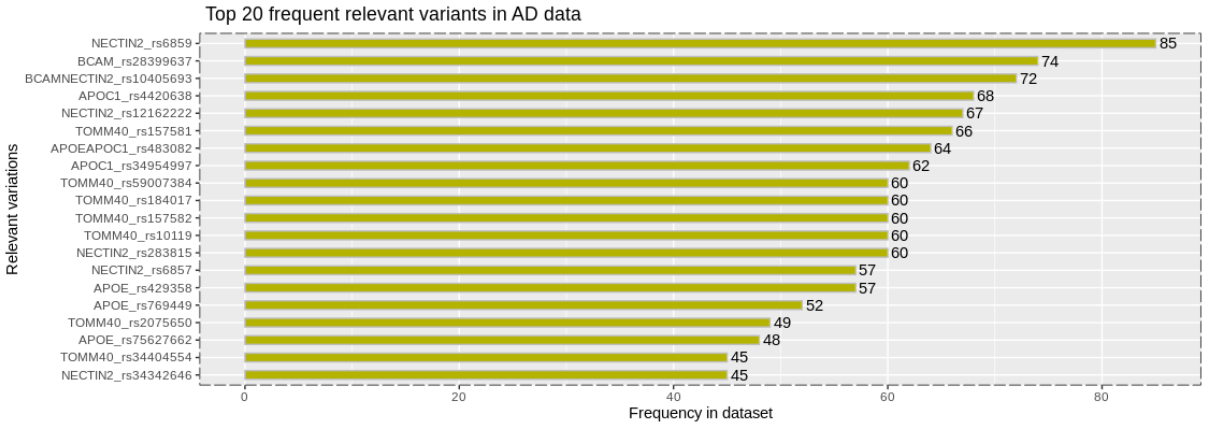
patients are part of the iASiS knowledge graph and their demographic and clinical features were collected from UK-Biobank. The study cohort is created based on the following criteria:

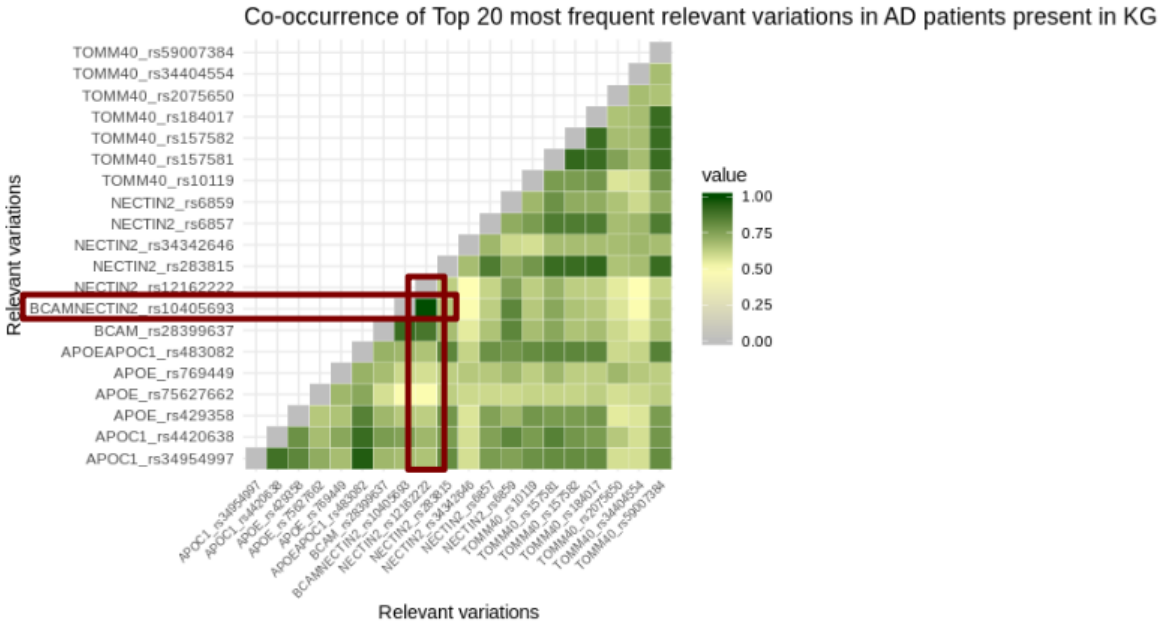
1. Only records of patients who have non-empty value for 'gender', 'longestPeriod' and 'bipolar_majorDepression_status' are considered in the analysis.
2. Negative values of 'longestPeriod' are replaced by -1 to indicate *"Do not know*" or "*Prefer not to answer*" values.
3. Meaning of the features that describe the patients in the cohort:
    a. longestPeriod: represents the longest period of depression in weeks (if <1 or >999, rejected), otherwise; -1=Do not know/Prefer not to answer. Values of this variable have been discretized to include ranges of length 20 & closed to left. For example, range [1,20) implies including 1 and excluding 20.
    b. Bipolar and major depression status: 0=No Bipolar or Depression
    1=Bipolar I Disorder
    2=Bipolar II Disorder
    3=Probable Recurrent major depression (severe)
    4=Probable Recurrent major depression (moderate)
    5=Single Probable major depression episode

The analysis of the most frequent genes in the patients with AD reveals that the genes TOMM40, NECTIN2, APOC1, APOE, and BCAM are the most frequent genes, with variants classified as relevant for the pipeline described. Also, these genes are the ones that co-occur the most in the studied cohort of AD patients. This observation is consistent with other studies in the literature that reveal the co-occurrence of these five genes in patients with AD (e.g., Kulminski AM et al. 2018[5]). The following Figures also support this statement in the cohort analysed in this study.

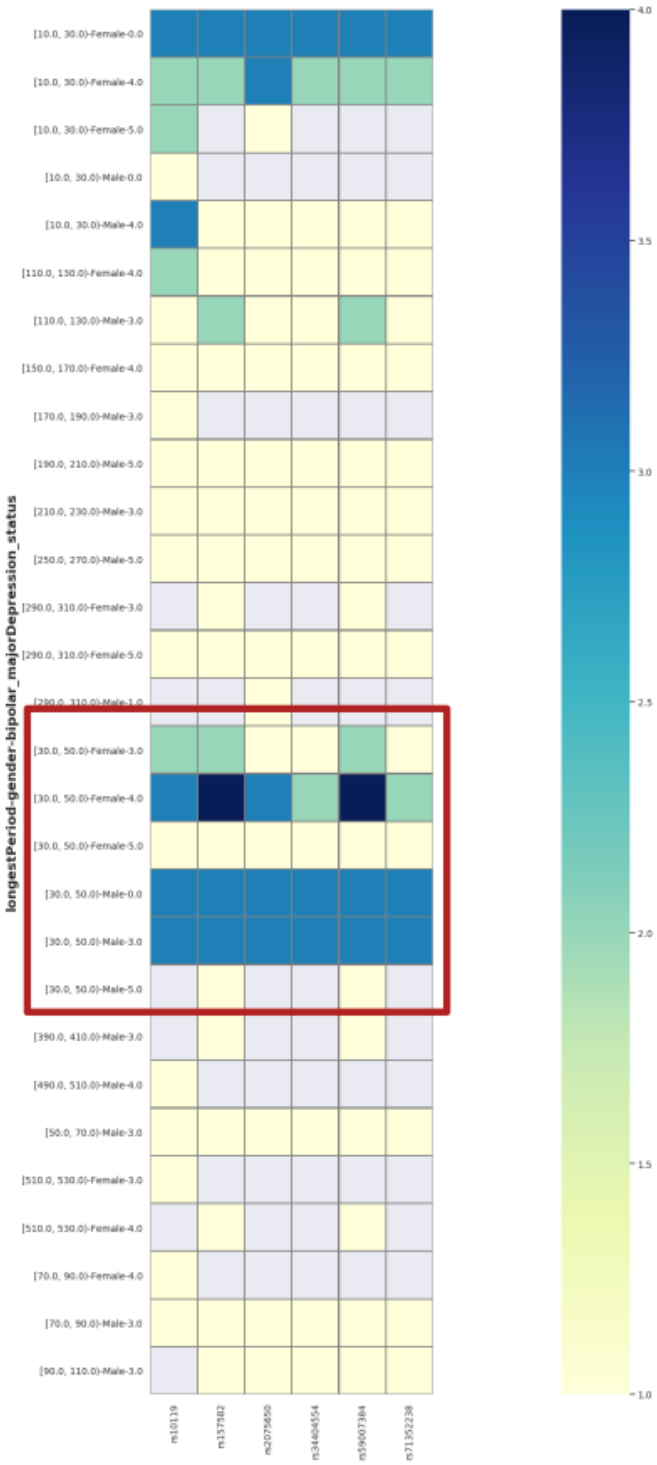


[5] Kulminski AM, Huang J, Wang J, He L, Loika Y, Culminskaya I. Apolipoprotein E region molecular signatures of Alzheimer's disease. Aging Cell. 2018;17(4):e12779. doi:10.1111/acel.12779





The patients associated with the studied variants are also analyzed with the aim of finding patterns that could be a potential indicator of significant implications in diverse types of mental problems (e.g. bipolar or depression, as well as events of depression and the duration of these events. As observed in the heatmap of next Figure, the results suggest that 15 out of 48 patients with variants in the gene **TOMM40 have had periods of dementia between 30 and 50 weeks**.


longestPeriod-gender-bipolar_majorDepression_status
[10.0, 30.0)-Female-0.0
[10.0, 30.0)-Female-4.0
[10.0, 30.0)-Female-5.0
[10.0, 30.0)-Male-0.0
[10.0, 30.0)-Male-4.0
[110.0, 130.0)-Female-4.0
[110.0, 130.0)-Male-3.0
[150.0, 170.0)-Female-4.0
[170.0, 190.0)-Male-3.0
[190.0, 210.0)-Male-5.0
[210.0, 230.0)-Male-3.0
[250.0, 270.0)-Male-5.0
[290.0, 310.0)-Female-3.0
[290.0, 310.0)-Female-5.0
[30.0, 50.0)-Female-3.0
[30.0, 50.0)-Female-4.0
[30.0, 50.0)-Female-5.0
[30.0, 50.0)-Male-0.0
[30.0, 50.0)-Male-3.0
[30.0, 50.0)-Male-5.0
[390.0, 410.0)-Male-3.0
[490.0, 510.0)-Male-4.0
[50.0, 70.0)-Male-3.0
[510.0, 530.0)-Female-3.0
[510.0, 530.0)-Female-4.0
[70.0, 90.0)-Female-4.0
[70.0, 90.0)-Male-3.0
[90.0, 110.0)-Male-3.0
rs10119
rs157582
rs2075650
rs34604554
rs59607384
rs71352238
4.0
3.5
3.0
2.5
2.0
1.5
1.0