Supercentenarian project. Rare variants analysis.

Contents

1 Introduction 1

2 Annotation 2
2.1 VEP annotation . . . 2
2.2 ANNOVAR annotation . . . 2
2.3 CSVS annotation . . . 2

3 European rare variants, defining categories of interest 2
3.1 European rare variants . . . 2
3.2 Rare variants of interest . . . 3
3.3 European rare variants with different AF in CSCV Healthy . . . 3
3.4 Gene-level metrics: burden of rare variants of interest . . . 3

4 Functional Analysis of M116 rare variants 4
4.1 Over Representation of Aging/Longevity (previously described) genes with potentially altering rare variants . . . 4
4.2 Over Representation of genes with association to supercentenarian phenotype in Gierman HJ et *al.* 2014[1] . . . 5
4.3 Overrepresentation of functional categories in genes with rare variants of interest . . . 5

5 Hypothesis-free analysis: in search for novel target genes 7

6 Remarkable findings 7
6.1 Immune genes . . . 7
6.2 Bitter taste receptors . . . 7
6.3 Mitochondrial . . . 7

7 Conclusions 7

1 Introduction

We part from VCFs from WGS data of Olot individuals of interest, including three samples from the supercentenarian (M116): blood, saliva and urine, and one sample from each of her daughters (R79 and T90), from blood and saliva respectively. The VCFs have a total of 3.8 M variants across all chromosomes. All variants are PASS according to previous QC performed in the genomics facility, all have QUAL > 20. No further QC was performed from our side. The average depth of coverage is 21.9.

The goal is to identify, characterise (section 3), and functionally analyse (section 4) rare variants, defined as AF < 0.015 in European populations (subsection 3.1), and potentially interesting rare variants (subsection 3.2) in particular (attending to their potential to be "damaging", "altering" or "moderate" modifiers of protein behaviour (subsection 3.2) and their differential AF in healthy individuals (subsection 3.3)), and to further derive gene-level metrics of supercentenarian's (M116) genome including the number and proportion

of said rare variants as used in [1] to be used to compare them with those of other European individuals of same demographics: IBS women from Phase 3 1000G in order to ascertain the extremeness of M116's genes metrics and identify "extreme genes" potentially liked to the extreme phenotype (hypothesis free analysis, section 5).

2 Annotation

VCFs were annotated with the annotation sources below, then converted to tables with VariantsToTable function from GATK [2].

2.1 VEP annotation

The VCFs where annotated with Ensemble Variant Effect Predictor (VEP) [3] version 112 (latest). This annotation adds transcript location, class, and other attributes, IMPACT and Consequence annotations, prediction scores for Polyphen (2.2.3) , SIFT (6.2.1), as well as AF of existing variants for populations including: 1000G NFE (Phase 3 (remapped), gnomAD genomes (r3.1.2, genomes only), gnomAD exomes (r2.1.1, exomes only), ClinVar (2023-10), among other information. It also includes SNP identifier annotations dbSNP (156), and gene/features identifier ENSG IDs (v102), which are more reliable and will be used to merge on lists later on (section 4), as well as HGNC symbols, useful for reporting.

2.2 ANNOVAR annotation

After figuring out that sometimes the AF field value in the VEP annotations slightly differ from the ANNOVAR annotation previously being performed by the genomics facility on the last sequenced sample (M116 urine), and some times a variant would appear in only one of them, ANNOVAR [4] annotations were further added to the VCFs, in order to maximise the cover variants being annotated and to account for possible mismatches among database versions.

This annotation adds CADD, and different versions of Polyphen, SIFT, as well as different versions of AF for populations including: 1000G NFE, Gnomad genome EUR, Gnomad exome EUR, ExAc exome, etc (GWAS hit, Tissue specificity) that serve to cover as many variants as possible..

2.3 CSVS annotation

The Collaborative Spanish Variant Server [5] provides AF of variants found in Spanish population CSVS. Datasets "All" and "Healthy" variation allele frequencies in all and healthy individuals in Spanish population were added were added as an additional annotation to the VCFs (as fields CSVS_all and CSVS_healthy).

* Note: CSVS also contains datasets of cases for several diseases, but the aggregate data for these datasets are not publicly accessible.

3 European rare variants, defining categories of interest

The annotations for AF in Europeans from all sources (1000G, Gnomad genome, Gnomad exome, CSVS all) and the fields IMPACT, Consequence from VEP, Polyphen and SIFT fields from VEP and ANNOVAR, and CADD field from ANNOVAR, were used to identify **rare variants** and to classify them into three different (partially overlapping) categories of interest for further analysis (section 4).

3.1 European rare variants

European rare variants were defined as variants with no instance of AF ¿ 0.015 in any EUR dataset from any of the two annotations sources (section 2).

European novel variants, with AF of exactly 0 in all datasets, are additionally labelled as so.

Additional filter was made to consider variants called from 2 or 3 of the M116's samples. These are the ones referred to in the analyses in section 4. Likewise, the replication in 1 or 2 of the daughters' samples was registered for a further filter, for the most reliable germline variants.

3.2 Rare variants of interest

The European rare variants identified (subsection 3.1), were classified into three categories attending to their potential impact on protein structure/expression, as follows:

- DAMAGING: IMPACT is HIGH (disruptive variants probably causing truncation, loss of function or triggering nonsense mediated decay) or Polyphen/SIFT predictions are damaging/deleterious or CADD > 15
- ALTERING: IMPACT is MODIFIER and Consequence is not intron_variant, synonymous, non_coding_transcript or intergenic_variant (added to DAMAGING variants, some types of variants from IMPACT category MODIFIER, with less harmful or difficult to predict impact that are in or close to protein-coding sequences or in regulatory regions).
- MODERATE: IMPACT is MODERATE (a non-disruptive variant that might change protein effectiveness)

The categories are based on Ensembl Variation - Calculated variant consequences, Polyphen [6][7], SIFT [8] and CADD [9].

Results:

Table 1 shows the number of variants and genes associated per category. Full lists are available here: REF.

Table 1: EUR rare variants and variants of interest: 3 categories

Category	No. Variants	No. Genes
rare	99331	22663
altering	24968	16274
moderate	564	444
damaging	1691	1603

3.3 European rare variants with different AF in CSCV Healthy

An additional intersection was made between the variants of interest (3 categories, subsection 3.2) and CSVS non-zero AF rare variants ($0 <$ CSVS_all AF < 0.015, CSVS_healthy AF ¿ 0) showing a difference of 1.5X between AF in CSVS_all and CSVS_healthy.

These variants might be worth a closer look as the difference might be suggestive of them having an association with disease diagnoses, and hence potentially with disease susceptibility ("potentially differentiating" variants).

The lists variants are divided in "higher_healthy" if the variant has higher AF in CSVS_healthy dataset than in CSVS_all, and "lower_healthy" otherwise.

shows the number of variants and genes per category alongside with the number of "higher_healthy" and "lower_healthy" variants among them.

3.4 Gene-level metrics: burden of rare variants of interest

Gene-level number of rare variants of interest and proportion rare variants of interest/all rare variants were calculated for all genes with European rare variants (subsection 3.1).

Table 2: EUR rare variants and variants of interest: 3 categories

Category	No. Variants	No. Genes	No. Variants Higher	No. Variants Lower	Genes with Differential Variants
rare	99331	22663	481	6549	3433
altering	24968	16274	100	1852	1806
moderate	564	444	15	31	35
damaging	1691	1603	19	130	166

Genes in the upper quantile of such metrics with regard to other genes in same chromosome were labelled as so, as suggestive of extreme genes (although this is rather arbitrary and not taking into account genes' features such as gene size or exome size)

The aim of this gene-level metrics is to compare them with other control individuals (section 5).

4 Functional Analysis of M116 rare variants

The aim is to characterise the rare variants of interest found in M116 (3 categories, subsection 3.2) attending to their enrichment in genes of known functions that might shed light on involved mechanisms on the extreme longevity phenotype.

For this we use Functional Annotation Databases (subsection 4.3), curated lists of longevity/aging gene sets (subsection 4.1) and a set of differentiating genes in a supercentenarian cohort (subsection 4.2)

To harmonise gene names BioMartR R package [10] was used with Ensemble v102, and all original gene identifiers (EntrezID, HGNC symbols) were converted to ENSG IDs to match the IDs in our VEP annotated VCFs.

4.1 Over Representation of Aging/Longevity (previously described) genes with potentially altering rare variants

Seven longevity/aging gene sets suggested by Manel were downloaded from Human Ageing Genomic Resources, listed below, plus one gene list provided directly my Manel (Manel Excel).

- Manel excel - excel (74)
- GenAge (human) - hagr_genage_human (339)
- GenAge complementary dataset Genes Commonly Altered During Ageing (from a microarray meta-analysis study) - hagr_ageing (683)
- CellAge: The Database of Cell Senescence Genes - hagr_cellage (952)
- CellAge: The Database of Cell Senescence Genes - hagr_cellsignatures (1368)
- NGDC Aging Atlas Aging-related geneses (human) - ngdc (554)
- Longevity Variants Database (LongevityMap), a database of human genetic variants associated with longevity - longevitymap (996)

The number of longevity genes with potentially altering rare variants is 986 out of 11272 genes with potentially altering rare variants in total. The number of longevity genes is 2194. The total background of genes in the VCF is 34265.

Hypergeometric test for longevity/aging genes in genes with rare variants of interest (any category) shows a significant enrichment (p-value 4.615873e-34), indicating that the longevity/aging category is overrepresented among the genes with rare variants of interest in M116.

Table 3 shows hypergeometric test results for all longevity genes in lists above. P-values are not corrected.

Table 3: EUR rare variants and variants of interest: 3 categories

Category	p-value
altering	0.86
damaging	2.098835e-09
moderate	2.362362e-10

Table 4: EUR rare variants and variants of interest: 3 categories

Geneset	altering p-value	moderate	damaging
excel	0.2168677	0.4478633	0.6366342
hagr_genage_human	0.06979495	0.1593044	0.1092659
hagr_genage_ageing	0.3496352	0.06811434	0.03516871
hagr_cellage	0.4597072	0.02839254	**0.004753071**
hagr_cellsignatures	0.7056075	**4.482411e-07**	**0.0007486425**
ngdc	0.1468118	0.2080856	0.03632849
longevitymap	0.7601565	**0.0004565057**	**6.264597e-05**

4.2 Over Representation of genes with association to supercentenarian phenotype in Gierman HJ et *al.* 2014[1]

A previous study with a supercentenarian cohort [1] calculated the burden of rare protein-altering variants per gene in supercentenarian individuals and controls (RVT1). The list of top genes in RVT1 gene burden test* (uncorrected p value RVT1 $<$ 1E-02) in a cohort of 13 supercentenarian vs 34 PGP Europeans (controls) was intersected with the genes with potentially altering rare variants in M116. * the genes statistically suggestive of being differentiating between supercentenarian and controls based on the proportion of damaging rare variants/all rare variants.

Certain genes with potentially altering rare variants in M116 were among those.

?? shows hypergeometric test results for all longevity genes in lists above. P-values are not corrected.

Table 5: EUR rare variants and variants of interest: 3 categories

Category	p-value
altering	0.0836528
damaging	7.481644e-05
moderate	6.134779e-05

Table 6: EUR rare variants and variants of interest: 3 categories

Geneset	altering p-value	moderate	damaging
Sc17	0.2475203	0.1361437	**0.007319936**
Chinese	0.1422479	**3.431131e-05**	**0.002027067**

4.3 Overrepresentation of functional categories in genes with rare variants of interest

Overrepresentation Analysis (ORA) was performed using WebGestaltR [11] to determine the enrichment of certain biologically relevant categories in the gene sets harbouring rare variants of interest in the supercentenarian genome (M116). WebgestaltR ORA uses hypergeometric test to calculate p-values of the observed

number of genes in one gene set versus the expected number of genes in that set from the reference. FDR is p-values corrected from multiple testing with BH method.

Categories analysed included:

- functional: Gene Ontogogy (GO) Biological Process, Molecular Function and Cellular Component
- phenotype: Human Phenotype Ontology (HPO)
- pathway: KEGG, Reactome and Panther
- disease: OMIM, GLAD4U and Disgenet

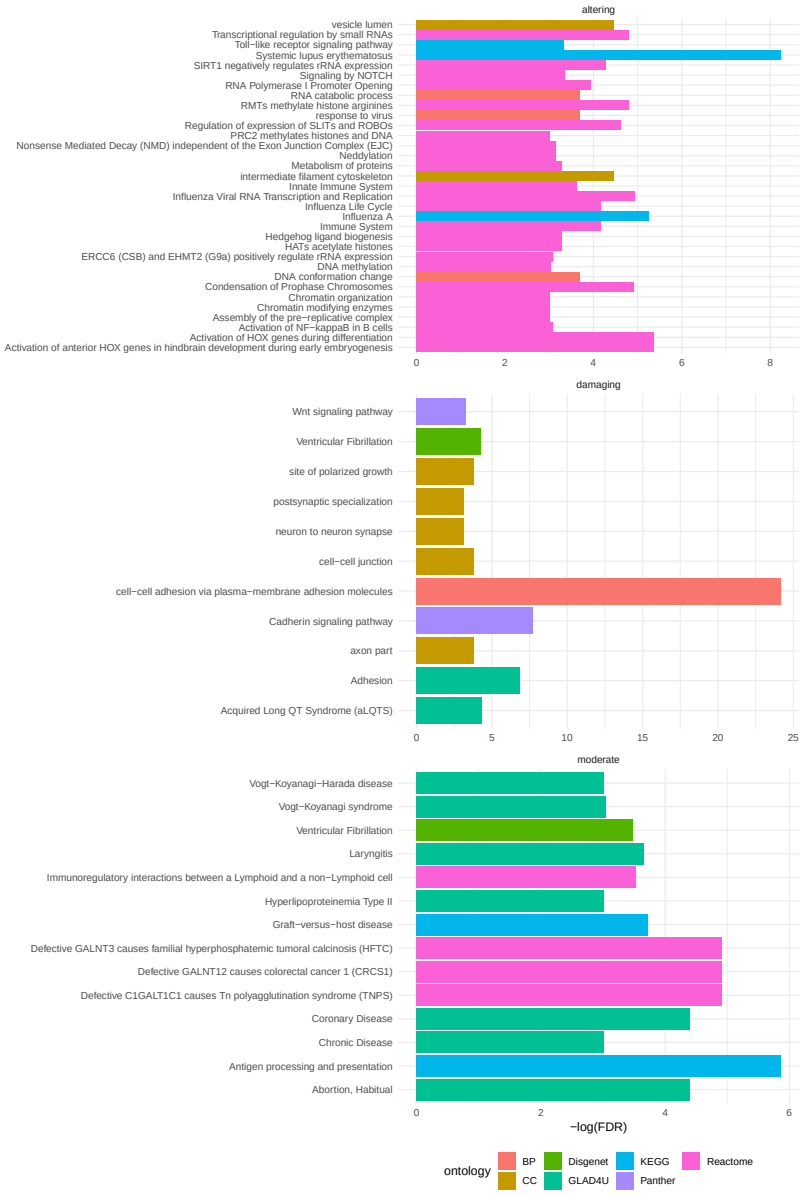

Figure 1: ORA results FDR < 0.05.

5 Hypothesis-free analysis: in search for novel target genes

A "control set" of 76 1000G IBS women was considered. 1000G high coverage (30X) VCFs REF were filtered to SNVs and restricted to the "control set" samples. All SNVs whereQUAL > 20 and FILTER is PASS were considered and the filtered VCFs were treated the same way as the described above for the Olot VCFs.

The gene-level metrics of number and proportion of rare variants of interest in M116 (**??**) are compared with that of the "control set" (TO BE FINISHED).

6 Remarkable findings

6.1 Immune genes

Showed in ORA (section 4) the Immune System Immune System Innate Immune System Immunoregulatory interactions between a Lymphoid and a nonLymphoid cell Antigen processing and presentation Toll-like receptor signaling pathway Systemic lupus erythematosus Influenza A response to virus

DNA conformation change RNA catabolic process

6.2 Bitter taste receptors

Ventricular Fibrillation Signaling by NOTCH

Bitter taste receptors have been associated with longevity REF and to cardiovascular morphology/function [12][13], with a possible role in cardiac contractility and overall vascular tone.

The finding of variants of interest in TAS2R16 (only one statistically associated to longevity) and TAS2R5 is interesting in the light of the ORA results.

6.3 Mitochondrial

One mitochondrial rare variant of interest,, is associated to gene XXX. This gene is part of XXX machinery, associated with aging REF.

7 Conclusions

References

[1] Hinco J Gierman, Kristen Fortney, Jared C Roach, Natalie S Coles, Hong Li, Gustavo Glusman, Glenn J Markov, Justin D Smith, Leroy Hood, L Stephen Coles, et al. Whole-genome sequencing of the world's oldest people. *PloS one*, 9(11):e112430, 2014.

[2] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[3] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17:1–14, 2016.

[4] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[5] María Peña-Chilet, Gema Roldán, Javier Perez-Florido, Francisco M Ortuno, Rosario Carmona, Virginia Aquino, Daniel Lopez-Lopez, Carlos Loucera, Jose L Fernandez-Rueda, Asuncion Gallego, et al. Csvs, a crowdsourcing database of the spanish population genetic variability. *Nucleic acids research*, 49(D1):D1130–D1137, 2021.

[6] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.

[7] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. Predicting functional effect of human missense mutations using polyphen-2. *Current protocols in human genetics*, 76(1):7–20, 2013.

[8] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003.

[9] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47(D1):D886–D894, 2019.

[10] Hajk-Georg Drost and Jerzy Paszkowski. Biomartr: genomic data retrieval with r. *Bioinformatics*, 33(8):1216–1217, 2017.

[11] Yuxing Liao, Jing Wang, Eric J Jaehnig, Zhiao Shi, and Bing Zhang. Webgestalt 2019: gene set analysis toolkit with revamped uis and apis. *Nucleic acids research*, 47(W1):W199–W205, 2019.

[12] Conor J Bloxham, Simon R Foster, and Walter G Thomas. A bitter taste in your heart. *Frontiers in Physiology*, 11:536822, 2020.

[13] Conor J Bloxham, Katina D Hulme, Fabrizio Fierro, Christian Fercher, Cassandra L Pegg, Shannon L O'Brien, Simon R Foster, Kirsty R Short, Sebastian GB Furness, Melissa E Reichelt, et al. Cardiac human bitter taste receptors contain naturally occurring variants that alter function. *Biochemical Pharmacology*, 219:115932, 2024.