Brief Report

DOI: 10.5582/irdr.2022.01063

When LUCA met gnomAD: genetic constraints on universal genes in humans

Alexandre Fabre[1,2,*], **Julien Mancini**[3,4]

[1] Aix Marseille Univ, INSERM, MMG, Marseille, France;
[2] APHM, Multidisciplinary Pediatrics Department, Timone Enfant Hospital, Marseille, France;
[3] Aix-Marseille Univ, INSERM, IRD, ISSPAM, SESSTIM, Marseille, France;
[4] APHM, BIOSTIC, Hop Timone, Marseille, France.

SUMMARY LUCA, the last universal common ancestor, is the hypothetical most recent common ancestor of the three domains of life which share the universal genes (UG). It seems interesting to evaluate whether the UG phylogeny has had an impact on current Human gene constraints. A list of human homologs of UG was retrieved from the eggNOG database. We analyzed this LUCA gene (LG) group, and a random sample of 500 genes from the gnomAD database (RG group). Gene constraint metrics were retrieved from gnomAD and associations with Mendelian diseases and modes of inheritance were retrieved from OMIM. The LG group consisted of 277 genes and the RG group, 492 (8 genes were in LG group). 38.6% of the genes in the LG group and 25.2% of the genes in the RG group were associated with a Mendelian disease ($p < 0.0001$). The mode of inheritance was more often autosomal recessive (69.0 *vs.* 50.5%), and less often autosomal dominant (19.0 *vs.* 31.3%), or mixed (6.0 *vs.* 12.1%) for those associated with the LG group ($p = 0.048$). The LG group was significantly more constrained for missense variants (MOEUF, 0.919 *vs.* 0.997, $p < 0.0001$) and was borderline significantly more constrained for loss-of-function variants (LOEUF, 0.872 *vs.* 0.947, $p = 0.051$). These results suggest that the UG in humans differs from the rest of the genome in terms of constraints and associated Mendelian diseases. It suggests that phylogenic data can explain some of the characteristics of human genes and could help in interpreting variants.

Keywords LUCA, gnomAD, last universal common ancestor, gene constraints

1. Introduction

LUCA (the last universal common ancestor) is the hypothetical most recent common ancestor of the three domains of life, archaea, bacteria and eukarya (*1*). LUCA's nature is elusive – presumed to be hyper-thermophilic to mesophilic for example (*1-3*) – because it is a reconstruction, based on comparisons of the genomes of current living organisms. While the size of LUCA's genome is unknown, it is assumed to at least contain the limited set of genes found across all three domains of life, sometimes referred to as the ancestral genetic core of cells or universal genes.

The genome aggregation database (gnomAD) contains more than 100,000 human exomes and genomes, along with annotations including constraint metrics that quantify the relative intolerance to variation of each protein-coding gene. These constraint metrics are calculated as the ratio of observed to expected synonymous, missense and loss-of-function variants, lower scores indicating more constrained genes. They have been used to interpret next generation-sequencing data, notably in the context of Mendelian diseases (*4*), and are presumed to be a reflection of natural selection (*5*)

In this study, we investigated whether the phylogenetic characteristics of LUCA genes are reflected in specific levels of genetic constraints or frequencies of associated Mendelian diseases.

2. Material and Methods

The gnomAD constraint metric by gene table (*6*) was downloaded from the gnomAD website (*https://gnomad.broadinstitute.org/downloads*, file "pLoF Metrics by Gene TSV"). The list of human homologs of universal genes (the LUCA gene (LG) group) was retrieved from the eggNOG database (*7*) (*http://eggnogdb.embl.de/*) using the clusters of orthologous groups (COGs) described by Harris *et al.* (*8*), Ciccarelli *et al.* (*9*), et Puigbò *et al.* 2009 (*10*) as being representative of universal genes. The functional categories and number of human homologs per COG were recorded (Supplemental Table S1, *http://www.irdrjournal.com/action/getSupplementalData.php?ID=103*). The LG group was analyzed in comparison with the random gene (RG) group, a random sample of 500 of the 19,704 genes in the gnomAD table.

The variables considered for each gene were the genetic constraint metrics (the synonymous, missense and loss-of-function observed/expected upper bound fractions, the SOEUF, MOEUF and LOEUF, respectively) and chromosome localization. Manual searches were performed for each gene on the Online Inheritance in Man (OMIM) website (*11*) between 15 October 2019 and 5 May 2020 for each of the included genes. The data retrieved were the existence of an associated Mendelian disease (non-diseases and multifactorial disorders were not considered), and for each disease, the recorded mode of inheritance (autosomal dominant, autosomal recessive or X-linked). For genes associated with multiple phenotypes, the number of associated Mendelian diseases was also recorded. and the mode of inheritance was recorded as mixed if it varied between phenotypes. All statistical analyses were performed with SPSS.

No ethics approval was required under French law as the study only involved data analysis. Database data were used in accordance with the corresponding data use agreements.

3. Results and Discussion

Among the 80, 36 and 102 COGs respectively described by Harris *et al.*, Ciccarelli *et al.*, and Puigbò *et al.* (*8-10*), as being representative of universal genes, 120 were unique and 33 were common to the three lists (Figure 1). Fourteen had no human homolog (COG0073, COG0250, COG0540 and COG0071 from Harris *et al.* (*8*) and COG0136, COG0195, COG0492, COG0575, COG0358, COG0455, COG0527, COG0528, COG1080 and COG2812 from Puigbò *et al.* (*10*) and three human genes were identified as homologs for two COGs (YARS, COG0143 and COG0162; NME8, COG0105 and COG0526; and EPRS, COG0008 and COG0442). The mean number of homologs per COG was 5.2 (SD, 5.9; range, 1-23), giving a total of 277 genes in the LG group. For the RG group, eight of the 500 initially selected genes were discarded because they also appeared in the LG group. The final number of genes analyzed was therefore 769 (277 in the LG group and 492 in the RG group).

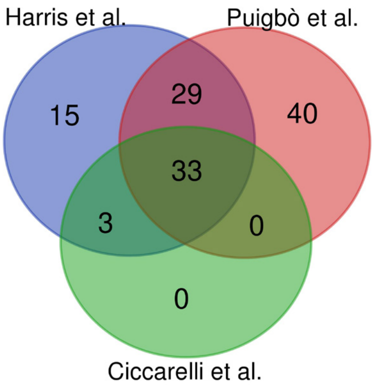

Figure 1. Venn diagram of the clusters of orthologous groups (COGs) retrieved from Harris *et al.*, Ciccarelli *et al.*, and Puigbò *et al.* Diagram prepared using the website *https://bioinformatics.psb.ugent.be/webtools/Venn/*

The OMIM database is 17.5% smaller than gnomAD (16,253 *vs.* 19,704 genes). Eighteen (6.5%) of the genes in the LG group and 99 (20.1%) of the genes in the RG group were not listed in the OMIM database ($p <$ 0.0001). Among genes present in the OMIM database, 100/259 (38.6%) of those in the LG group and 99/393 (25.2%) of those in the RG group were associated with a Mendelian disease ($p < 0.0001$). The mode of inheritance was more often autosomal recessive and less often autosomal dominant or mixed for diseases associated with the LG group (69.0 *vs.* 50.5%, 19.0 *vs.* 31.3%, and 6.0 *vs.* 12.1%, respectively, $p = 0.048$; Table 1). The LG group was significantly more constrained for missense variants (MOEUF, 0.919 *vs.* 0.997, $p < 0.0001$) and was borderline significantly more constrained for loss-of-function variants (LOEUF, 0.872 *vs.* 0.947, $p =$ 0.051). Limiting the analysis to COGs with five or fewer homologs (because only one COG each had 6, 7, 9, 12 and 23 homologs), the number of homologs per COG was not significantly correlated with the SOEUF ($\rho =$ −0.09, 95% CI [−0.22, 0.04], $p = 0.17$), MOEUF ($\rho =$ 0.13, 95% CI [−0.01, 0.25], $p = 0.061$) or the LOEUF ($\rho = 0.11$, 95% CI [−0.02, 0.24], $p = 0.095$) of the genes (Figure 2).

The analysis was repeated for the LGmin group, consisting of 62 genes associated with the 31 COGs common to all three lists (mean number of homologs per COG, 2.5; SD, 1.1; range, 1-5; details in Table 1). Comparisons with the RG group showed the same, if slightly stronger trends as observed for the full LG group, with a higher proportion of genes present in the OMIM and associated with a Mendelian disease than in the RG group. The LGmin group was significantly more constrained for synonymous variants and missense variants but was not significantly more constrained for loss-of-function variants (Table 1). The number of homologs per COG was significantly but weakly correlated with the MOEUF ($\rho = 0.38$, 95% CI [0.1249, 0.5713], $p = 0.004$), and the LOEUF of the LGmin genes ($\rho = 0.30$, 95% CI [0.05, 0.52], $p = 0.02$), and non-significantly correlated with the SOEUF ($\rho = -0.2153$, 95% CI [−0.45, 0.04], $p = 0.1$).

One possible explanation for these results is that the genes in the two groups belong to different functional categories. For example, 45.7% of those in the RG group are of unknown function (218/ 477 as 15 gene of RG group are not present in eggNOG database), whereas none of those in the LG group are; and conversely, while

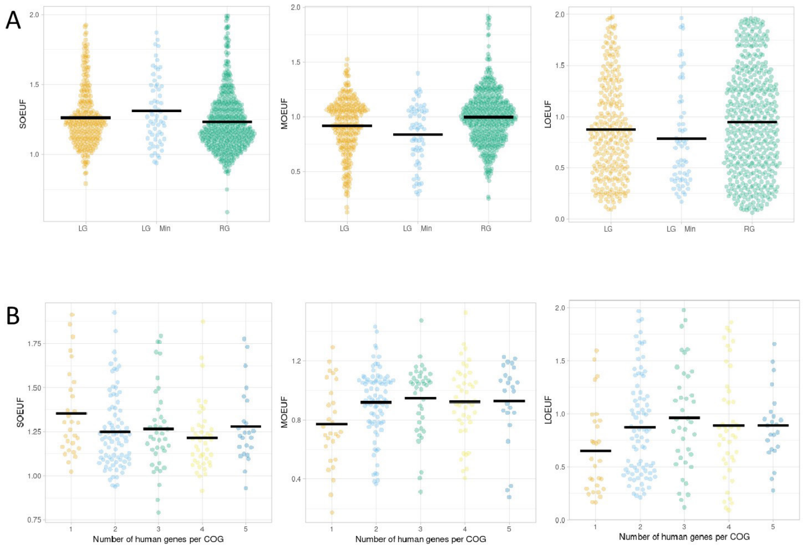

Figure 2. (A) Distributions of synonymous, missense, and loss-of-function observed/expected upper bound fractions (respectively SOEUF, MOEUF and LOEUF for LUCA genes (LG group, n = 277; orange), consensus LUCA genes (LGmin group, n = 62, blue), and a random selection of human gene (RG group, n = 492, green). **(B)** SOEUF, MOEUF, and LOEUF scores of LUCA genes as a function of the number of genes in the corresponding COG (cluster or orthologous groups). Figure prepared using the website *https://huygens.science.uva.nl/PlotsOfData*

Table 1. Gene characteristics according to groups

Items	LUCA genes (LG group)	Consensus LUCA genes (LGmin group)	Random selection of genes (RG group)	LG *vs.* RG	LGmin *vs.* RG
Genes	277	62	492		
Present in the OMIM database	259 (93.5%)	58 (93.5%)	393 (79.9%)	$p < 0.0001$	$p = 0.009$
Associated with Mendelian disease in the OMIM database	100 (38.6%)	27 (46.6%)	99 (25.2%)	$p < 0.0001$	$p = 0.001$
Autosomal dominant inheritance	19 (19%)	7 (25.9%)	31 (31.3%)	$p = 0.048$	$p = 0.782$
Autosomal recessive inheritance	69 (69%)	17 (63%)	50 (50.5%)		
Autosomal recessive and dominant inheritance	6 (6%)	2 (7.4%)	12 (12.1%)		
X-linked inheritance	6 (6%)	1 (3.7%)	6 (6.1%)		
Mean number of OMIM phenotypes per gene associated with a Mendelian disease (SD)	1.25 (0.626)	1.19 (0.396)	1.34 (0.641)	$p = 0.3$	$p = 0.225$
Mean SOEUF (SD)	1.263 (0.209)	1.312 (0.241)	1.235 (0.214)	$p = 0.079$	$p = 0.009$
Mean MOEUF (SD)	0.919 (0.252)	0.837 (0.277)	0.997 (0.248)	$p < 0.0001$	$p < 0.0001$
Mean LOEUF (SD)	0.872 (0.481)	0.787 (0.493)	0.947 (0.512)	$p = 0.051$	$p = 0.021$

LUCA, last universal common ancestor; OMIM, Online Inheritance in Man; SD, standard deviation; SOEUF, synonymous observed/expected upper bound fraction; MOEUF, missense observed/expected upper bound fraction; LOEUF, loss-of-function observed/expected upper bound fraction.

only 9 genes in the RG group (1.9%) are involved in translation, ribosomal structure and biogenesis, 131 (47.3%) of those in the LG group are. We therefore performed the same analysis considering each functional group separately.

Subgroup analysis was performed for the 12 functional categories found in both groups and containing more than 10 genes (C, CO, E, F, G, H, J, K, L, M, O, U). The MOEUF and LOEUF values for the LG group were lower than those in the RG group in 8/12 functional categories (C, G, J, K, L, M, O and U), with statically significant differences for M and U in terms of MOEUF and LOEUF and for K in terms of MOEUF. The results for the SOEUF metric were more variable, with values obtained for the LG group being lower in 6 categories but higher in the 6 others. The number of associated Mendelian diseases was non-significantly higher in the LG group for 7 functional categories (G, H, I, J, K, M, O), and significantly higher for the L category, and the same as in the RG group for the U category (Supplemental Table S2, *http://www.irdrjournal.com/action/getSupplementalData.php?ID=104*).

This is, to our knowledge, the first study of genetic constraint in the putative ancestral core of the human genome. We found that these LUCA genes were slightly more constrained than a random sample of genes for missense and loss-of-function variants, and less constrained for synonymous variants. Whereas LUCA genes were found to be more frequently associated with Mendelian diseases, strangely, the mode of inheritance was more frequently autosomal recessive (69.0% *vs.* 50.5%) and less frequently autosomal dominant (19.0 *vs.* 31.3%) than it was for diseases associated with the randomly selected genes. Genes with lower LOEUFs tend to be haploinsufficiency genes and less commonly autosomal recessive (*6*). However, the mean LOEUF of the LUCA genes (0.872) is well above the threshold below which genes are usually considered constrained (0.35) (*12*). The fact that a higher proportion of universal genes are associated with autosomal recessive diseases, suggests that ancient genes are more constrained but have become more tolerant of heterozygous loss-of-function.

The fact that the analysis in terms of eggNOG functional categories produced the same results suggests that our results are not an artefact due to the large proportion of LUCA genes linked to translation, ribosomal structure and biogenesis or due to the ~50% of randomly selected genes being of unknown function.

Unsurprisingly, since gene duplication has been an important force in evolution (*13*), most COGs were associated with several human genes. It could have been assumed that constraints on the two genes would differ after duplication, one being more constrained and the other less as a new function is acquired (*14*). However, the variations in MOEUF, LOEUF and SOEUF values were huge even when the corresponding COG was only associated with a single gene, and the number of homologs per COG was only weakly correlated with these metrics, and thus the effect does not seem to be important.

In conclusion, these preliminary results suggest that the ancestral core differs from the rest of the human genome in terms of genetic constraint and associated Mendelian diseases. An interesting line of research may be to use phylogenic data to uncover whether these universal genes can explain some of the characteristics of human genes and help in interpreting variation in a clinical setting.

Acknowledgements

We thank Pr. Céline Brochier-Armanet for useful discussion which lead to this work. We thank Paul Guerry (GreenGrow Scientific) for editing the article.

Funding: None.

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. Forterre P, Gribaldo S, Brochier C. Luca: the last universal common ancestor. Med Sci (Paris). 2005; 21:860-865. (in French)
2. Palacios-Pérez M, José MV. The evolution of proteome: From the primeval to the very dawn of LUCA. Biosystems. 2019; 181:1-10.
3. Catchpole RJ, Forterre P. The evolution of reverse gyrase suggests a nonhyperthermophilic last universal common ancestor. Mol Biol Evol. 2019; 36:2737-2747.
4. Bennett CA, Petrovski S, Oliver KL, Berkovic SF. ExACtly zero or once: A clinically helpful guide to assessing genetic variants in mild epilepsies. Neurol Genet. 2017; 6; 3:e163.
5. Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. Measuring intolerance to mutation in human genetics. Nat Genet. 2019; 51:772-776.
6. Karczewski KJ, Francioli LC, Tiao G, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020; 581:434-443.
7. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019; 47:D309-D314.
8. Harris JK, Kelley ST, Spiegelman GB, Pace NR. The genetic core of the universal ancestor. Genome Res. 2003; 13:407-412.
9. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science. 2006; 311:1283-1287.
10. Puigbò P, Wolf YI, Koonin EV. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. J Biol. 2009; 8:59.
11. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), Online Mendelian Inheritance in Man, OMIM®. *https://omim.org* (accessed December 6, 2021)
12. Francioli L, Tiao G, Karczewski K, Solomonson M, Watts N. gnomAD v2.1 *https://macarthurlab.org/2018/10/17/gnomad-v2-1/* (accessed December 6, 2021).
13. Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. Gene duplication as a major force in evolution. J Genet. 2013; 92:155-161.
14. Wagner A. Selection and gene duplication: a view from the genome. Genome Biol. 2002; 3:reviews1012.

Received June 1, 2022; Revised June 24, 2022; Accepted July 11, 2022.

**Address correspondence to:*
Alexandre Fabre, Pediatric Multidisciplinary Department, Timone Enfant Hospital, APHM, 264 Rue Saint Pierre 13005 Marseille, France.
E-mail: alexandre.fabre@ap-hm.fr

Released online in J-STAGE as advance publication July 22, 2022.