PNAS
www.pnas.org

Supplementary Information

Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes

Antonio Rausell[1,2,§,*], Yufei Luo[1,2,§], Marie Lopez[3], Yoann Seeleuthner[2,4], Franck Rapaport[5], Antoine Favier[1,2], Peter D. Stenson[6], David N. Cooper[6], Etienne Patin[3], Jean-Laurent Casanova[2,4,5,7,8], Lluis Quintana-Murci[3,9] , Laurent Abel[2,4,5,‡,*]

1. Clinical Bioinformatics Laboratory, INSERM UMR1163, Necker Hospital for Sick Children, 75015 Paris, France, EU

2. University of Paris, Imagine Institute, 75015 Paris, France, EU

3. Human Evolutionary Genetics Unit, Institut Pasteur, UMR2000, CNRS, Paris 75015, France, EU

4. Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM UMR1163, Necker Hospital for Sick Children, 75015 Paris, France, EU

5. St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY, USA

6. Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XN, UK, EU.

7. Howard Hughes Medical Institute, New York, NY, USA

8. Pediatric Hematology and Immunology Unit, Necker Hospital for Sick Children, 75015 Paris, France, EU.

9. Human Genomics and Evolution, Collège de France, Paris 75005, France, EU

§ Joint first authors, equal contributions

* Correspondence to

antonio.rausell@inserm.fr
casanova@mail.rockefeller.edu
laurent.abel@inserm.fr

This PDF file includes:

Figures S1 to S8
Table S1
SI References

Other supplementary materials for this manuscript include the following:
Datasets S1 to S8

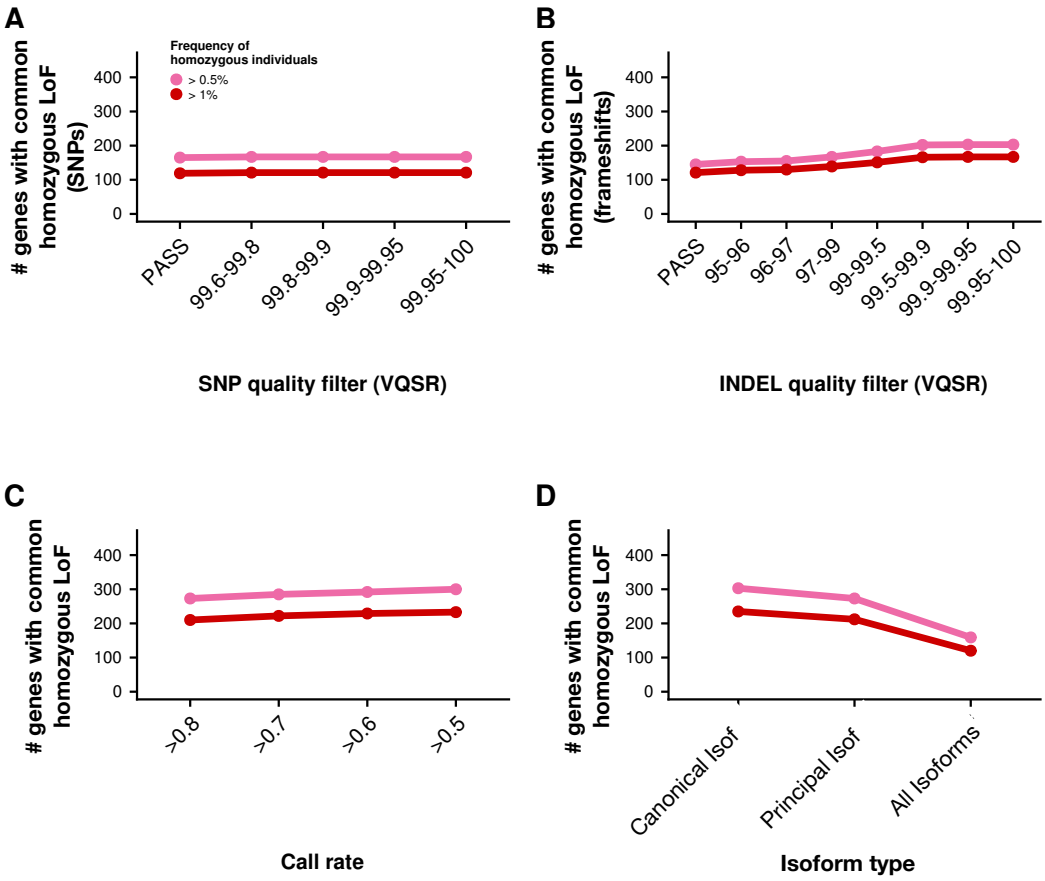

Fig. S1. Impact of the different filtering criteria on the final number of dispensable genes detected in the ExAC database.

A. Number of genes with common homozygous LoF caused by Single Nucleotide Polymorphisms as a function of the variant quality score recalibration (VQSR) threshold. **B.** Number of genes with homozygous LoF caused by frameshifts as a function of the variant quality score recalibration (VQSR) threshold. **C.** Number of genes with common homozygous LoF caused by SNPs and frameshifts as a function of the call rate threshold. **D**. Number of genes with common homozygous LoF caused by SNPs and frameshifts depending on whether LoF variant affects (i) the canonical isoform of a gene (as defined by Ensembl pipeline), (ii) the canonical isoform of a gene that represents the principal isoform of a gene, as defined by APPRIS system (corresponding to the selected criteria in our study), and (iii) all isoforms of a gene (the LoF variant is constitutive of all alternative transcripts; **Methods**).

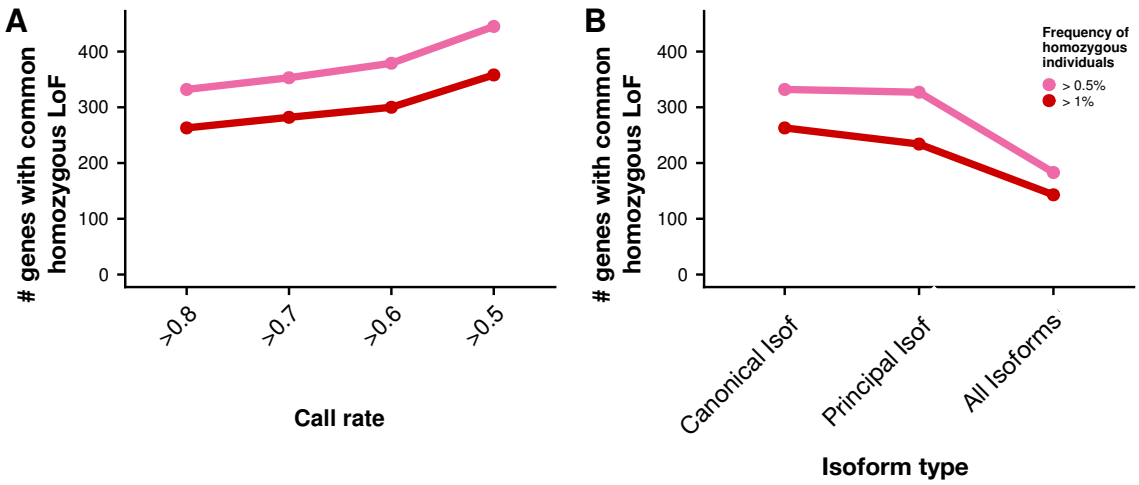

Fig. S2. Impact of the different filtering criteria adopted on the final number of dispensable genes detected in the GnomAD database.

A. Number of genes with common homozygous LoF caused by SNPs and frameshifts as a function of the call rate threshold. **B**. Number of genes with common homozygous LoF caused by SNPs and frameshifts depending on whether the LoF variant affects the canonical isoform, the principal isoform or all isoforms (variant constitutive of all isoforms). It should be noted that VQSR scores were not used in the GnomAD database, thus panels analogous to **Fig. S1 A** and **B** could not be drawn.

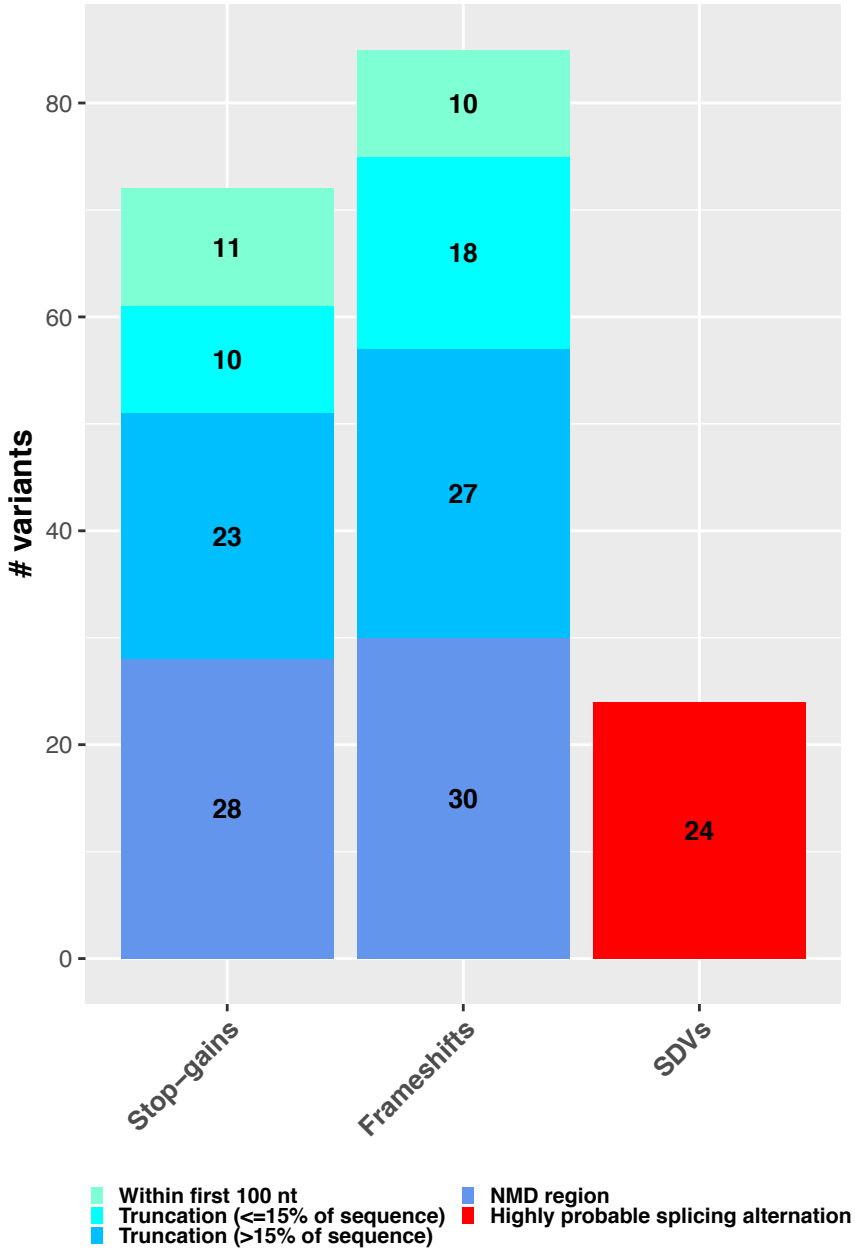

Fig. S3. Predicted functional impact of LoF variants defining the set of dispensable protein coding genes.

Bar plots show the distribution of LoF variants that define the set of dispensable genes according to their molecular consequences (stop-gains, frameshifts and splice-disrupting variants, SDV) and the predicted type of functional impact, according to the following categories: In the case of stop-gains and frameshifts, LoF variants are classified among those i) mapping to the first 100 nucleotides of the associated transcript, ii) potentially triggering NMD, or iii) truncating more, or less or equal than 15% of the affected protein sequence. In the case of putative splice-disrupting variants (SDVs), severity was computationally predicted as unknown, very low, intermediate or high impact (**Methods**).

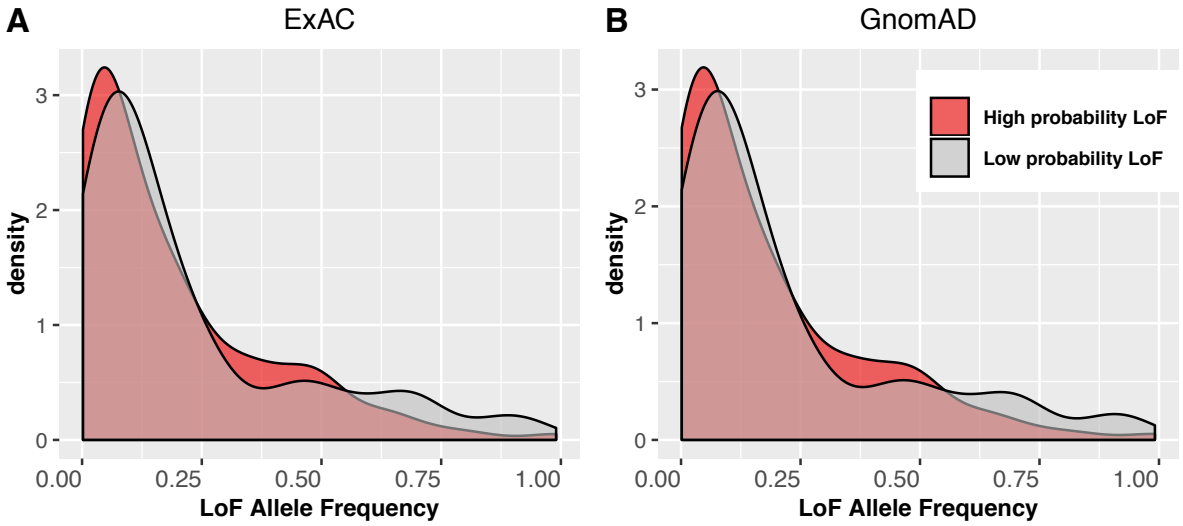

Fig. S4. **Allele Frequency distribution of LoF variants defining the set of dispensable protein-coding genes.**

Allele Frequency of LoF variants is represented separately for low probability LoF (light grey) and high probably LoF (dark red) variants. Allele frequencies from ExAC were used in **Panel A**, whereas GnomAD data were used in **Panel B**.

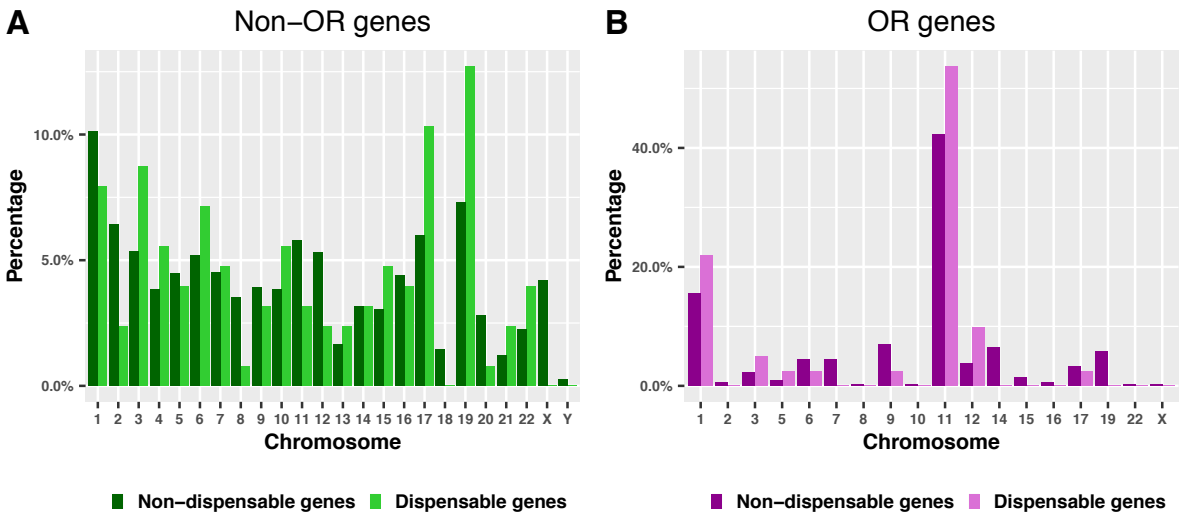

Fig. S5. Distribution of dispensable and non-dispensable genes across chromosomes.

Barplots display the percentage of genes across human chromosomes of the following 4 gene sets, each adding to 100%: (**A**) dispensable non-OR genes (light green) and non-dispensable non-OR genes (dark green), and (**B**) dispensable OR genes (light purple) and non-dispensable OR genes (dark purple).

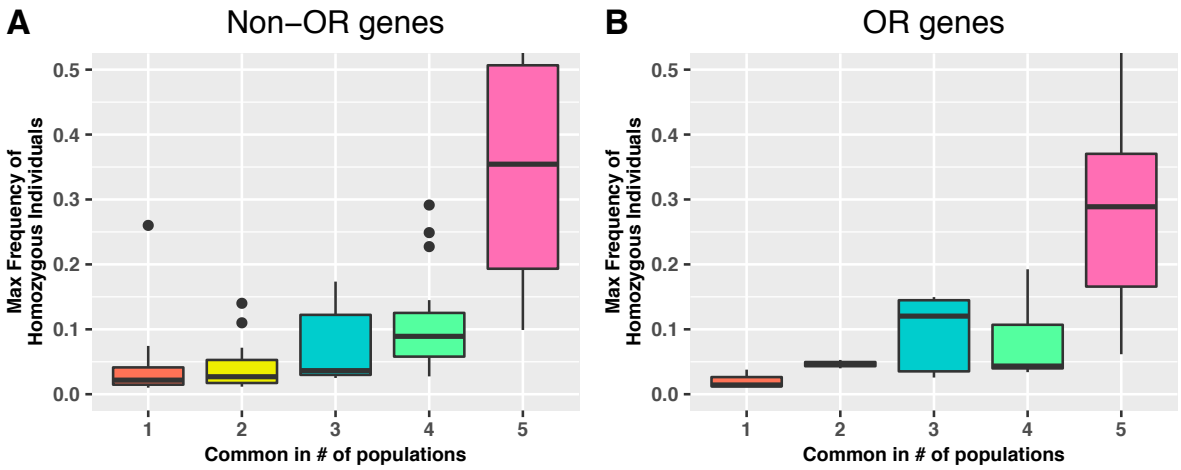

Fig. S6. Distribution of the maximum frequency of homozygous individuals for dispensable non-OR genes (**A**) and dispensable OR genes (**B**) as a function of the number of populations in which they were found to be dispensable. As in Figure 3, the homozygous LoF variant frequencies were taken from the GnomAD dataset.

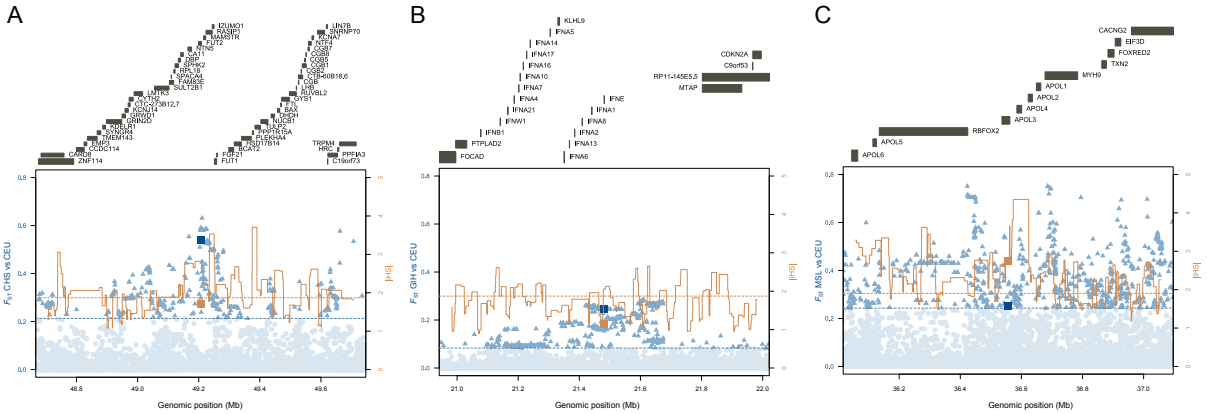

Fig. S7. Selective sweep signals. Local genomic signatures of positive selection in 1Mb regions around LoF mutations located in (**A**) *FUT2* (chr19:49206674) for CEU, (**B**) *IFNE* (chr9:21481483) for GIH and (**C**) *APOL3* (chr22:36556768) for MSL. Blue and orange squares indicate F_{ST} and |iHS| values respectively at the LoF allele. Blue dots and triangles indicate SNP F_{ST} percentiles and the blue dashed line indicate 95th percentile of F_{ST} values genome-wide. Orange solid line indicate the maximum |iHS| value in sliding windows of 50 SNPs and the orange dashed line indicate the 95th percentile of |iHS| values genome-wide.

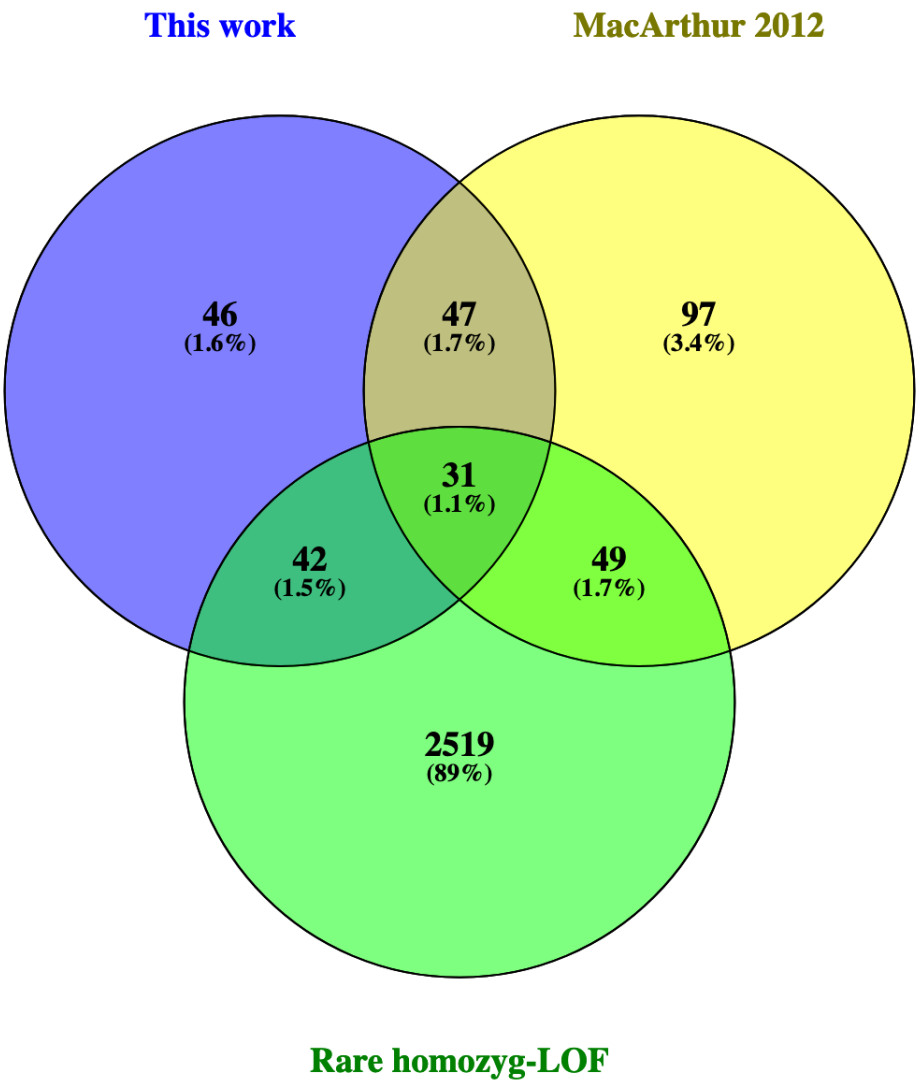

Fig. S8. Overlap of the dispensable genes detected in this work with those identified in previous studies. The figures show the Venn diagrams representing the overlap of the 166 putatively dispensable genes detected in this work with 253 genes apparently tolerant to homozygous rare LoF variants reported in MacArthur et al. (1) and a total list of 2641 presenting homozygous rare LoF variants reported from bottlenecked or consanguineous populations (2-5).

Table S1. Populations from the 1000 Genomes Project used in the positive selection analysis of common LoF variants.

Population Code	Population Description	Sample_size	Super Population Code
ACB	African Caribbeans in Barbados	96	AFR
ASW	Americans of African Ancestry in SW USA	61	AFR
ESN	Esan in Nigeria	99	AFR
GWD	Gambian in Western Divisions in the Gambia	113	AFR
LWK	Luhya in Webuye, Kenya	99	AFR
MSL	Mende in Sierra Leone	85	AFR
YRI	Yoruba in Ibadan, Nigeria	108	AFR
CLM	Colombians from Medellin, Colombia	94	AMR
MXL	Mexican Ancestry from Los Angeles USA	64	AMR
PEL	Peruvians from Lima, Peru	85	AMR
PUR	Puerto Ricans from Puerto Rico	104	AMR
CDX	Chinese Dai in Xishuangbanna, China	93	EAS
CHB	Han Chinese in Bejing, China	103	EAS
CHS	Southern Han Chinese	105	EAS
JPT	Japanese in Tokyo, Japan	104	EAS
KHV	Kinh in Ho Chi Minh City, Vietnam	99	EAS
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	99	EUR
FIN	Finnish in Finland	99	EUR
GBR	British in England and Scotland	91	EUR
IBS	Iberian Population in Spain	107	EUR
TSI	Toscani in Italia	107	EUR
BEB	Bengali from Bangladesh	86	SAS
GIH	Gujarati Indian from Houston, Texas	103	SAS
ITU	Indian Telugu from the UK	102	SAS
PJL	Punjabi from Lahore, Pakistan	96	SAS
STU	Sri Lankan Tamil from the UK	102	SAS

References

1. D. G. MacArthur, *et al.*, A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* **335**, 823–828 (2012).

2. E. T. Lim, *et al.*, Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. *PLoS Genet.* **10**, e1004494 (2014).

3. P. Sulem, *et al.*, Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448–452 (2015).

4. V. M. Narasimhan, *et al.*, Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474–477 (2016).

5. D. Saleheen, *et al.*, Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239 (2017).