Genetic Load of Loss-of-Function Polymorphic Variants in Great Apes

Guillem de Valles-Ibáñez[1], Jessica Hernandez-Rodriguez[1], Javier Prado-Martinez[1,2], Pierre Luisi[1], Tomàs Marquès-Bonet[1,3,4], and Ferran Casals[5,*]

[1]Institut de Biologia Evolutiva (UPF-CSIC), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Barcelona, Catalonia, Spain

[2]Present address: The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, United Kingdom

[3]Centre Nacional d'Análisi Genómica (CNAG), Parc Científic de Barcelona, Barcelona, Catalonia, Spain

[4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

[5]Servei de Genòmica, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Barcelona, Catalonia, Spain

*Corresponding author: E-mail: ferran.casals@upf.edu.

Accepted: February 18, 2016

Abstract

Loss of function (LoF) genetic variants are predicted to disrupt gene function, and are therefore expected to substantially reduce individual's viability. Knowing the genetic burden of LoF variants in endangered species is of interest for a better understanding of the effects of declining population sizes on species viability. In this study, we have estimated the number of LoF polymorphic variants in six great ape populations, based on whole-genome sequencing data in 79 individuals. Our results show that although the number of functional variants per individual is conditioned by the effective population size, the number of variants with a drastic phenotypic effect is very similar across species. We hypothesize that for those variants with high selection coefficients, differences in effective population size are not important enough to affect the efficiency of natural selection to remove them. We also describe that mostly CpG LoF mutations are shared across species, and an accumulation of LoF variants at olfactory receptor genes in agreement with its pseudogenization in humans and other primate species.

Key words: loss of function, Great apes diversity, genetic load, comparative genomics.

Stop gain and loss, splice-site and frameshift mutations, as well as coding region deletions are genetic variants predicted to disrupt gene function and are consequently denominated loss of function (LoF) variants. Because of their predicted effect on gene function, these variants are obvious candidates to originate genetic diseases and are often prioritized as putative causal mutations in genetic disease studies. For example, nonsense mutations account for approximately 20% of the single nucleotide polymorphisms (SNPs) associated with disease in coding regions (Mort et al. 2008), although they represent only about 1% of the functional SNPs in a human genome (Bamshad et al. 2011). Because of natural selection LoF genetic variants are mainly found at very low frequencies, and are mostly found in heterozygosis in healthy carriers.

Until recently, the number of lethal alleles in an individual has been estimated by indirect methods yielding different results. Early studies based on the analysis of consanguineous marriages in humans estimated the average number of lethal equivalents (which includes the effect of LoF and other functional variants) to be 1.4 (Bittles and Neel 1994) or 3–5 (Morton et al. 1956) per individual, similar to the 3.14 estimated in 38 mammalian species from captive populations with inbreeding depression (Ralls et al. 1988). However, other theoretical approaches estimated up to 100 lethal equivalents per human genome (Kondrashov 1995). Deep resequencing studies have allowed a direct estimate of the number of LoF variants per human genome. On average, any human individual has been proposed to harbor about

100 LoF variants, with around 20 of them in homozygosity (MacArthur et al. 2012), exceeding previous estimates of the genetic burden of the species. In addition, these numbers probably represent an underestimation because of the low coverage data (Consortium 2010), the stringent filters used to avoid false positives, and our incomplete knowledge on the noncoding regions of the genome that may also contain LoF variants altering gene expression.

Several reasons can be argued to explain the seemingly excessive number of LoF variants in an individual's genome, such as a high level of redundancy of gene function in genomes, different levels of gene essentiality, or incomplete penetrance. In agreement with these hypotheses, pathway network analyses have shown that LoF variants seem to be preferentially located in less central genes and with a lower connectivity in gene pathways (MacArthur et al. 2012). Also, although in a lower proportion, some of the LoF variants may have been under positive selection because of conferring an adaptive advantage, as proposes the "less is more" hypothesis (Olson 1999). Thus, in some gene function losses across species might have an adaptive effect (Olson and Varki 2003) and might have been important to recent evolution (Olson 1999). Finally, some polymorphic LoF variants have been proposed to be maintained by balancing selection in human populations, as in the case of the AB0 and the secretor status of Lewis blood group systems (Calafell et al. 2008; Ferrer-Admetlla et al. 2009).

There are no direct estimations of the number of LoF genetic variants for primate species other than humans based on whole-genome sequencing. Especially for endangered species, this information is of interest for a better understanding of the effect of declining population sizes and habitat loss and for predicting the viability of natural and captive populations. In this work, we estimate the number of LoF variants from whole-genome sequences of 79 individuals from 6 great ape species (*Pan troglodytes*, *Pan paniscus*, *Gorilla beringei*, *Gorilla gorilla*, *Pongo abelii*, and *Pongo pygmaeus*) (Prado-Martinez et al. 2013), and then apply different filters based on the nature of the mutation, position along the coding region, evolutionary conservation, or population frequency to approach the genetic burden due to LoF variants in these great ape species, in comparison with humans.

Materials and Methods

Data

Data were retrieved from Prado-Martinez et al. (2013), consisting of whole-genome sequences from 79 great ape individuals (PRJNA189439 and SRP018689, Sequence Read Archive). We restricted our analysis to populations with five or more sequenced individuals (which allows establishing a frequency threshold of MAF 0.1): *Pan paniscus* (13), *Pan troglodytes ellioti* (10), *Pan troglodytes schweinfurthii* (6), *G. gorilla gorilla* (23), *Pongo abelii* (5), and *Pongo Pygmaeus* (5), for a total of 62 individuals from six populations. All samples were sequenced in an Illumina platform (HiSeq 2000) with a mean coverage of 25× and sequence reads were mapped against the human reference assembly NCBI Build 36/UCSC hg18 using BWA 0.6.2 software to facilitate gene model interpretation (Li and Durbin 2009). SNPs and indels were called using the GATK 2.6 pipeline (DePristo et al. 2011) and annotated with Annovar (May 25, 2012) (Wang et al. 2010). The variants overlapping segmental duplications and tandem repeats, as well as those close to indels clustering within 10 bp, were filtered out. A conservative allelic imbalance filter was applied to remove those heterozygous calls that may be due to contamination. The final set of vcf files with the annotated variants can be found at http://biologiaevolutiva.org/greatape/ (last accessed January 22, 2016). From those variants, we selected the polymorphic LoF variants for further analysis (stop gain and loss introducing mutations, frameshift indels and mutations in canonical splicing sites; Huang et al. 2009b, 2009a) when there was available sequencing information for five or more individuals. For some of the analyses to estimate the number of LoFs with a drastic phenotypic effect, we filtered out those variants affecting the first and last 5% of gene coding sequences, as the selective constrains in terminal regions are more relaxed (Wetterbom et al. 2006). We also considered the list of error prone genes produced in previous next-generation sequencing studies (Fuentes Fajardo et al. 2012), to exclude possible false positives. To make a valid comparison with the human mutational load, we also included data from three African and six non-African human genomes from the same study (Prado-Martinez et al. 2013), and annotated data from the 1000 Genomes project (MacArthur et al. 2012) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/functional_annotation/annotated_vcfs/, last accessed January 22, 2016).

Validation

Eighteen genetic variants were selected for validation by Sanger sequencing following a stringent criteria and limited to individuals with DNA availability. Variants were visually inspected using IGV 2.2.4 (Robinson et al. 2011; Thorvaldsdóttir et al. 2013) to exclude those with mapping reads displaying abnormal features. Conservatively, homozygous variants were only considered when no reads with the other alleles were present. We also excluded variants where the presence of an adjacent SNP alters the annotation of the LoF variant. We also analyzed the genomic context of the candidate variant using Tandem Repeats Finder software 4.07b (Benson 1999), and excluded those LoF variants included in a tandem repeat.

Connectivity and Enrichment Analysis

For each protein, connectivity degree was computed as the total number of interactions in which it is involved in the human protein–protein interaction network reconstructed

Table 1

Average Number of Autosomic Polymorphic LoF Variants Per Individual in Each Population

	Stop Gain	Frameshift Indels	Splice Site	Stop Loss	All
Homo sapiens	45.1(11.6)	78.5(16.7)	82.6(30.2)	7.6(3.2)	214(61.8)
Pan paniscus	64.7(14)	171(58.8)	85.2(26.7)	11.1(7.1)	332.2(106.6)
Pan troglodytes ellioti	86.8(24.2)	349(104.5)	118.5(34.2)	10.8(4.6)	565.1(167.5)
Pan troglodytes schweinfurthii	76(16.6)	278.6(78.5)	100.5(24.6)	4.6(1)	459.8(120.8)
Gorilla gorilla	112.7(31.9)	368.1(150.1)	136.9(53.1)	9.3(2)	627(237.3)
Pongo abelii	101.4(21.4)	465.2(114.2)	124.6(34.8)	5(1.4)	696.2(171.8)
Pongo pygmaeus	90.8(27.4)	408(98.6)	103.8(37.6)	8.6(4.6)	611.2(168)

NOTE.—Homozygous LoF variants are shown in parentheses.

from the physical interactions described in the Human Protein Reference Database (Keshava Prasad et al. 2009). We considered the whole set of nonredundant interactions between two different proteins (thus excluding redundant interaction and self-interaction). We performed a chi-square goodness-of-fit test to detect accumulations of LoF variants at certain gene classes. We used the Web-Based Gene Set Analysis Toolkit (WebGestalt) (Wang et al. 2013; Zhang et al. 2005) for enrichment analyses, using the Benjamini and Hochberg correction (Benjamini and Hochberg 1995).

Results and Discussion

Number of LoF Variants in Great Ape Species

We applied stringent mapping and quality filters (see Methods) to estimate the number of LoF variants in 62 great ape individuals from six different populations. Similar to the pattern described for all SNPs and indels (Prado-Martinez et al. 2013), the average number of LoF variants per individual is higher in great apes than in humans, and in general increases with evolutionary distance to humans (table 1 and supplementary table S1, Supplementary Material online). This observation might be due to two possible artifacts. First, functional annotation is based on humans, which is expected to increase the number of LoF variants because of differences in splicing, transcription, and exon usage across primates (Barbosa-Morais et al. 2012). Second, the use of the human genome as a reference for mapping and variant calling can artificially inflate the number of variants in other species. Indels, which are known to show reduced concordance among sequencing platforms and bioinformatic pipelines (Fang et al. 2014), are especially sensitive to these factors. This fact may explain the excessive number of indels reported in all great ape populations in comparison with humans (table 1). Because of that, we excluded indels from some of the subsequent analyses. Even without considering indels, there is still an excess of LoF variants in great ape species compared with humans, although it has been substantially reduced and does not correlate anymore with evolutionary distance from humans.

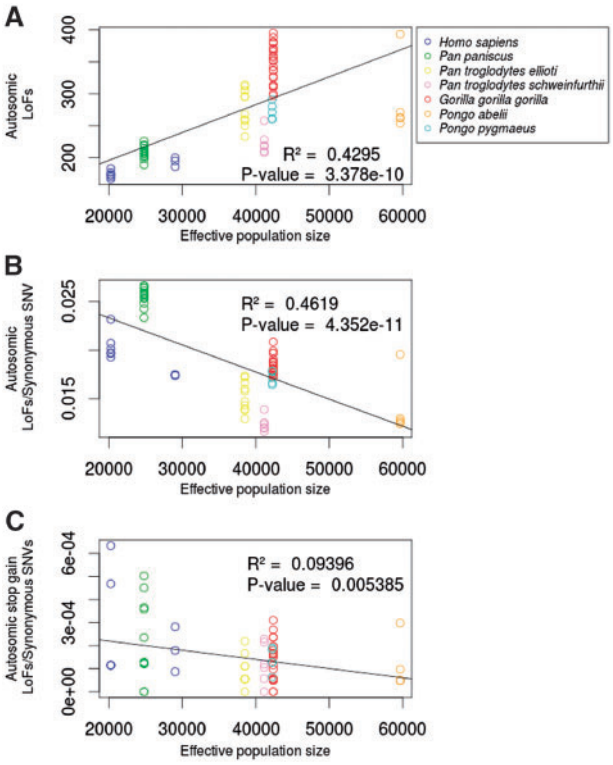

FIG. 1.—(*A*) Effective population size versus the number of polymorphic autosomic LoF mutations per individual. (*B*) Effective population size versus the ratio of polymorphic autosomic LoF to polymorphic synonymous autosomic variants per individual. (*C*) Effective population size versus the estimated number of detrimental polymorphic autosomic stop gain mutations to polymorphic synonymous autosomic variants per individual.

The number of polymorphic LoF variants per individual, after removing indels, correlates with effective population size (Ne) (fig. 1*A*), suggesting that an important fraction of the LoF variants detected could act as neutral variation. However, the correlation with Ne is negative when the number of LoF variants is normalized by the number of synonymous variants per individual (fig. 1*B*). This pattern has been

previously reported for all nonsynonymous compared with synonymous variants in the same samples (Prado-Martinez et al. 2013), and is a consequence of the higher efficiency of natural selection to remove detrimental variants at higher Ne (Petit and Barbadilla 2009).

LoF Variants Load per Individual

The number of LoF variants described per individual (table 1) is probably an overestimation of the actual number of variants with a drastic effect on the individual viability if found in homozygosis. To estimate the number of LoF variants with a more likely harmful phenotypic effect, we focused on stop gain mutations, because they are expected to be less prone to missanotation. Thus, for this analysis we did not consider splice variants because of transcript differences across primates (Barbosa-Morais et al. 2012), stop loss variants because of their less important predicted phenotypic effect (Richards et al. 2015) or indels (see above). We also established a set of filters to exclude those variants with less probability of being detrimental for the organism. First, we considered the relative position of the variants along the gene. As previously described for fixed LoF variants across species (Prado-Martinez et al. 2013), polymorphic LoF variants do not distribute randomly along the coding region but also tend to accumulate at the 5′ and 3′ gene ends (supplementary fig. S1, Supplementary Material online). This probably indicates that LoF variants at the gene ends have less effect on the protein and the selective constraint is relaxed (Wetterbom et al. 2009). Thus, LoF variants at the two 5% gene terminal regions were removed. Second, we excluded polymorphic LoF variants located at genes including a fixed LoF position in the population (Prado-Martinez et al. 2013) because the relaxation of selection after the truncation of the gene will favor the accumulation of other LoF variants. Third, we also excluded LoF variants at genes that recurrently accumulate false positives because of sequencing or mapping artifacts, as previously reported in Fuentes-Fajardo et al. (2012). Fourth, we also hypothesized that stop gain variants with an important phenotypic effect will be mostly found in evolutionary conserved positions (Davydov et al. 2010) (fig. 2). We established a threshold of GERP > 4 to keep potential damaging LoF mutations. Although positions with GERP values >2 are considered to be conserved among mammals and therefore more prone to be of functional importance (Davydov et al. 2010), we established a higher GERP (Genomic Evolutionary Rate Profiling) threshold for including mostly variants with important phenotypic effects (Amendola et al. 2015). And finally, we assumed that variants with an important phenotypic effect should be kept at low frequencies by natural selection. For that, we use a permissive threshold of 0.1 frequency, which allowed to include the six populations with $N > 5$ in the analyses.

After applying the filters described above, the number of stop gain variants in the six great ape populations analyzed is very similar across great ape species, and close to the estimates in nine human samples sequenced in the same study (table 2). Although other types of LoF variants (large deletions, frameshift indels, splice-site mutations) were not considered in the last analyses, it seems reasonable to consider that the relative amount of each type will be similar across species, and therefore the final number of LoF variants with an important phenotypic effect in each great ape population might approximately be four times the number of stop gain variants, assuming the results described in humans (MacArthur et al. 2012). Considering this and that LoF variants represent 50% of the lethal equivalents in a population (Simmons and Crow 1977), the total number would then be ~16. However, this number, higher than the 3.14 estimated by Ralls et al. (1988), is probably an overestimation because of a permissive frequency threshold originated by sample size limitations as

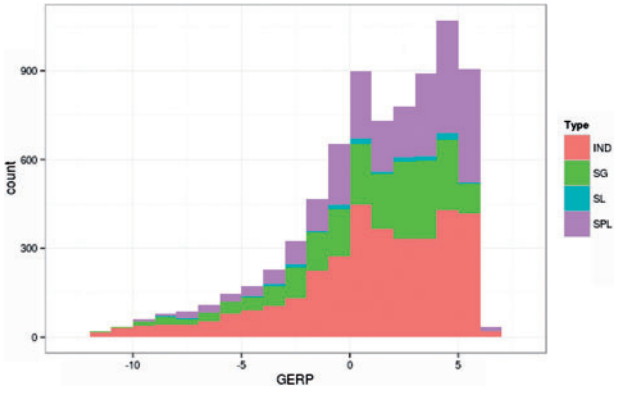

Fig. 2.—GERP score distribution for the different LoF variant categories. IND = frameshift insertion or deletion; SG = stop gain; SL = stop loss; SPL = splice-site donor or acceptor.

well as of different degrees of tolerance and gene redundancy (MacArthur et al. 2012). On the other hand, the estimation by Ralls et al. of an average value of 3.14 lethal equivalents in 38 mammalian species (4.6 of median and ranging from −1.4 to 30.3) is probably an underestimation because of being based only on one component of the fitness (survival of young) and analytical artifacts related to the higher homozygosity in captive populations (Ralls et al. 1988).

Interestingly, although the total genetic burden of these populations correlates with effective population size (Prado-Martinez et al. 2013), as well as the load of LoF variants (fig. 1), the number of disrupting variants is quite homogeneous across species (fig. 1C), suggesting that slightly deleterious mutations might be driven by effective population size but not strong deleterious mutations. The high selection coefficient of these LoF variants with a drastic phenotypic effect would greatly increase the efficiency of natural selection, reducing the differences across populations because of variation in their effective sizes. Only more critically low population sizes in primates would then affect the number of LoF variants. Interestingly, in the case of the mountain gorillas the effect of a critically small population size and population decline seems to have acted on the unexpected direction, increasing the overall genetic burden but also producing the purging of LoF variants. This observation has been attributed to consanguinity, which increases the probabilities of

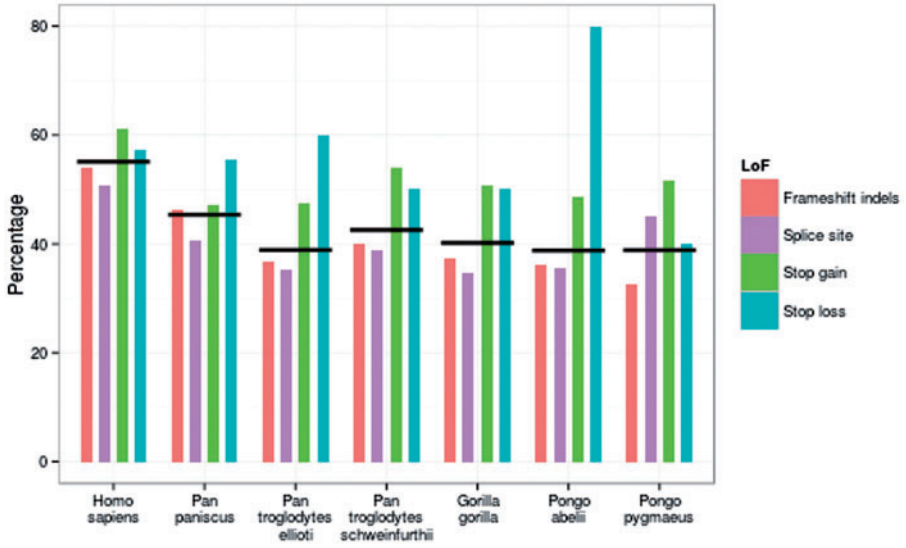

Fig. 3.—Percentage of autosomic polymorphic LoF variants in great ape species located at genes with a polymorphic LoF variant in the 1000 Genomes Project samples. Horizontal black lines represent the average value for the four types of LoF variants.

Table 2

Number of Total Autosomic Polymorphic Stop Gain Variants at Evolutionary Conserved Positions (GERP > 4) Per Population and Average Per Individual

	Population				Individual			
	Total		Not Fixed		Total		Not Fixed	
	All	≤0.1	All	≤0.1	All	≤0.1	All	≤0.1
Homo sapiens	28	23	28	23	4.1(0)	2.5(0)	4.1(0)	2.5(0)
Pan paniscus	40	22	39	21	10.1(2.1)	1.8(0)	10.1(2.1)	1.7(0)
Pan troglodytes ellioti	31	19	28	17	8(2.1)	2.5(0)	6.9(1.4)	1.9(0)
Pan troglodytes schweinfurthii	23	13	23	13	6.8(2.1)	2.1(0)	6.8(2.1)	2.1(0)
Gorilla gorilla gorilla	59	48	58	47	6.6(0.6)	2.9(0)	6.6(0.6)	2.9(0)
Pongo abelii	20	10	20	10	8.2(0.6)	2(0)	8.2(0.6)	2(0)
Pongo pygmaeus	26	14	20	10	10.2(1.6)	2.8(0)	8.2(0.6)	2(0)

Note.—Not fixed, variants at genes without a fixed LoF between species. ≤0.1, frequency of 0.1 or less. Homozygous variants are shown in parentheses.

Table 3

Number of Autosomic Polymorphic LoF Variants Found in Genes Containing a Fixed LoF across Species

Species (number of LOF)	*Homo sapiens*	*Pan paniscus*	*Pan troglodytes ellioti*	*P. troglodytes schweinfurthii*	*Gorilla gorilla gorilla*	*Pongo abelii*	*Pongo pygmaeus*	Fixed LoF
Homo sapiens[a] (548)	3	40	36	36	44	63	63	35.58%
Pan paniscus (734)	0	31	34	38	37	60	63	33.10%
Pan troglodytes ellioti (1,096)	0	69	49	63	46	75	81	33.57%
P. troglodytes schweinfurthii (719)	0	48	41	44	35	54	59	31.84%
Gorilla gorilla gorilla (1,577)	3	77	69	70	88	144	152	34.74%
Pongo abelii (859)	2	40	48	49	46	96	108	34.10%
Pongo pygmaeus (676)	3	28	29	28	41	90	80	29.73%

NOTE.—Fixed LoF, percentage of variants at a gene including a fixed LoF in one or more of the populations of the study.
[a]Nine *Homo sapiens* samples analyzed in Prado-Martinez et al. (2013).

detrimental variants being found in homozygosis and consequently removed by natural selection (Xue et al. 2015).

Genes with LoF variants

We detected that an important fraction of the polymorphic LoF variants are located at genes including a fixed LoF in the population (Prado-Martinez et al. 2013) (table 3). This proportion ranges from 4% of the total LoF in *Pan paniscus* to 11% in *Pongo abelii*, and is statistically significant in all cases ($P < 0.001$). This fraction increases to 8–17% when LoF fixed at other species from the same genus are also considered, and to 30% when estimating the number of LoF variants located at one of the 1,479 genes harboring a fixed LoF in at least one of the species of the study (Prado-Martinez et al. 2013) ($P < 0.001$). Among the rest, 2,503 (61.6%) polymorphic LoF variants are located at genes including a polymorphic LoF variant in the same or another species analyzed in this work. Six of them were successfully validated by Sanger sequencing (supplementary table S2, Supplementary Material online). When comparing with data in the 1000 Genomes Project, with 5,092 genes including a polymorphic variant (Khurana et al. 2013), the percentage of LoF variants from great ape species at genes including a polymorphic LoF variant in humans is of 38–45%, and 47–53% for the stop gain mutations (fig. 3) ($P < 0.001$).

In 187 cases, the same polymorphic stop gain variant is shared by two or more than one great ape species, resulting in 40 after stringent mapping filter. Among them, we selected 12 for Sanger validation (see Material and Methods for selection criteria), with 8 variants for potential damaging (without a fixed LoF in the same gene) and 4 located at genes including a fixed LoF in the populations under study. All 12 variants were validated by Sanger sequencing. Among the 40 shared variants, 25 of them are CpG mutations (63%), similar to the 71.5% of shared SNPs reported between humans and chimpanzee (Leffler et al. 2013) suggesting that those shared LoF variants across species are most likely to be originated by recurrent mutations. For the other 15, 9 LoF variants are shared between close species, mostly *Pongo abelii* and *Pongo pygmaeus* and in one case *Pan troglodytes* and *Pan paniscus*. Altogether, shared LoF variants are unlikely to be long-term polymorphisms maintained by selective forces as balancing selection.

We finally examined if some particular biological class or pathway is enriched with genes including LoF variants using WebGestalt (Wang et al. 2013; Zhang et al. 2005). We did detect a significant enrichment for olfactory receptors (supplementary table S3, Supplementary Material online) in all the populations ($P < 0.001$). We also detected this enrichment for olfactory receptors ($P < 0.001$) for shared LoF variants across species, although most of them are located at genes with already a fixed LoF variant. All this is in agreement with the description of a massive pseudogenization of olfactory receptors in humans (Gilad et al. 2003) and also other primate species (Go and Niimura 2008). Finally, we did not detect an association between the gene's connectivity degree estimated in the protein–protein interaction networks and the presence of homozygous LoF variants (supplementary fig. S2, Supplementary material online), as previously described in humans (MacArthur et al. 2012). In addition to our smaller sample sizes, all this may also reflect the limitation of using functional and connectivity information from humans.

Conclusions

The number of LoF genetic variants in great ape species is similar to that found in humans. Contrary to other functional mutations, LoF variants do not correlate with effective population size. Even if the actual number of variants with a drastic effect in the organism is difficult to estimate, especially when using mapping information and functional annotation from different species, our results suggest that these endangered species have not accumulated an excess of disruptive variants. The estimated value is close to previous estimation of the number of lethal equivalents of 1.4 in humans (Bittles and Neel 1994) or 3.14 on average for 38 mammal species (Ralls et al. 1988), assuming that about 50% of this genetic load is due to lethal alleles (Simmons and Crow 1977). However, all LoF variants described in ours and other sequencing projects would ultimately need functional validation experiments to confirm their detrimental effect.

Supplementary Material

Supplementary tables S1–S3 and figures S1 and S2 are available at *Genome Biology and Evolution online* (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank funding to F.C. by grant SAF2012-35025 from the Ministerio de Economía y Competitividad (Spain) and FEDER and by Direcció General de Recerca, Generalitat de Catalunya (2014SGR-866). T.M.B. is supported by EMBO YIP 2013, MINECO BFU2011-28549, BFU2014-55090-P (FEDER), BFU2015-7116-ERC and BFU2015-6215-ERC (www.mecd.gob.es), NIH U01 MH106874 grant and Fundacio Zoo Barcelona.

Literature Cited

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.
Amendola LM, et al. 2015. Actionable exomic incidental findings in 6503 participants: challenges of variant classification. Genome Res. 25:305–315.
Bamshad MJ, et al. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 12:745–755.
Barbosa-Morais NL, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. Science 338:1587–1593.
Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc. 57:289–300.
Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.
Bittles AH, Neel JV. 1994. The costs of human inbreeding and their implications for variations at the DNA level. Nat Genet. 8:117–121.
Calafell F, et al. 2008. Evolutionary dynamics of the human *ABO* gene. Hum Genet. 124:123–135.
Davydov EV, et al. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 6:e1001025.
DePristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43:491–498.
Fang H, et al. 2014. Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med. 6:89.
Ferrer-Admetlla A, et al. 2009. A natural history of *FUT2* polymorphism in humans. Mol Biol Evol. 26:1993–2003.
Fuentes Fajardo KV, et al. 2012. Detecting false-positive signals in exome sequencing. Hum Mutat. 33:609–613.
Gilad Y, Man O, Pääbo S, Lancet D. 2003. Human specific loss of olfactory receptor genes. Proc Natl Acad Sci U S A. 100:3324–3327.
Go Y, Niimura Y. 2008. Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees. Mol Biol Evol. 25:1897–1907.
Huang DW, Sherman BT, Lempicki RA. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37:1–13.
Huang DW, Sherman BT, Lempicki RA. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4:44–57.
Keshava Prasad TS, et al. 2009. Human protein reference database–2009 update. Nucleic Acids Res. 37:D767–D772.
Khurana E, et al. 2013. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342:1235587.
Kondrashov AS. 1995. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? J Theor Biol. 175:583–594.
Leffler EM, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 339:1578–1582.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.
MacArthur DG, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823–828.
Mort M, Ivanov D, Cooper DN, Chuzhanova NA. 2008. A meta-analysis of nonsense mutations causing human genetic disease. Hum Mutat. 29:1037–1047.
Morton NE, Crow JF, Muller HJ. 1956. An estimate of the mutational damage in man from data on consanguineous marriages. Proc Natl Acad Sci U S A. 42:855–863.
Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet. 64:18–23.
Olson MV, Varki A. 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. Nat Rev Genet. 4:20–28.
Petit N, Barbadilla A. 2009. Selection efficiency and effective population size in *Drosophila* species. J Evol Biol. 22:515–526.
Prado-Martinez J, et al. 2013. Great ape genetic diversity and population history. Nature 499:471–475.
Ralls K, Ballou JD, Templeton A. 1988. Estimates of lethal equivalents and the cost of inbreeding in mammals. Conserv Biol. 2:185–193.
Richards S, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 17:405–423.
Robinson JT, et al. 2011. Integrative genomics viewer. Nat Biotechnol. 29:24–26.
Simmons MJ, Crow JF. 1977. Mutations affecting fitness in *Drosophila* populations. Annu Rev Genet. 11:49–78.
Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 14:178–192.
Wang J, Duncan D, Shi Z, Zhang B. 2013. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res. 41:W77–W83.
Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38:e164.
Wetterbom A, Gyllensten U, Cavelier L, Bergström TF. 2009. Genome-wide analysis of chimpanzee genes with premature termination codons. BMC Genomics 10:56.
Wetterbom A, Sevov M, Cavelier L, Bergstrom TF. 2006. Comparative genomic analysis of human and chimpanzee indicates a key role for indels in primate evolution. J Mol Evol. 63:682–690.
Xue Y, et al. 2015. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science 348:242–245.
Zhang B, Kirov S, Snoddy J. 2005. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res. 33:W741–W748.

Associate editor: Naruya Saitou