REPORT

Intronic CNVs cause gene expression variation in human population

Maria Rigau[1], David Juan[2], Alfonso Valencia[1,3,¶,*] and Daniel Rico[4,¶,*]

[1]Barcelona Supercomputing Centre (BSC), Barcelona, 08034, Spain.

[2]Institut de Biologia Evolutiva, Consejo Superior de Investigaciones Científicas–Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Barcelona, 08003, Spain

[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, 08010, Spain.

[4]Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, NE2 4HH, United Kingdom.

[¶] These authors contributed equally to this work.

[*] Corresponding authors

E-mails: alfonso.valencia@bsc.es, daniel.rico@newcastle.ac.uk

Abstract

Introns can be extraordinarily large and together they account for the majority of the DNA sequence in human genes. However, little is known about their structural variation in populations and the functional impact of this variation. To address this, we have studied how copy number variants (CNVs) differentially affect exonic and intronic sequences of genes. We found that intronic losses are the most frequent form of variation in protein-coding genes in human, with more than 12,986 intronic deletions, affecting 4,147 genes (including 1,157 essential genes and 1,638 disease-related genes). Surprisingly, intronic losses are significantly enriched in evolutionarily ancient genes: up to 20% of the extant human genes that were born before the ancestor of tetrapods show intronic length variation. This result was quite unexpected, as CNV losses overlapping exons almost exclusively affect primate-specific genes. Finally, an integrated analysis of CNVs and RNA-seq data showed that intronic loss can be associated with significant differences in gene expression levels in the population (CNV-eQTLs). These intronic CNV-eQTLs regions are enriched in intronic enhancers and can be associated with expression differences of other genes showing long distance intron-promoter 3D interactions. Our data suggests that genomic variation is shaping gene evolution in different ways depending on their age and function. Intronic variation of ancient genes can exert an important role in maintaining the variability of expression and splicing of conserved genes in human populations, a previously neglected source of variability that could have a meaningful impact in adaptation and disease.

Main Text

Most eukaryotic protein coding genes contain introns that are removed from the messenger RNA during the process of splicing. Although the potential functions of introns remain elusive, a number of cases support the idea that several acquired functionalities might compensate the energetic disadvantage that they represent for the cell[1–3]. In humans, up to 35% of the sequenced genome corresponds to purely intronic sequence, while exons cover around the 2.8% of the genome (based on the genome version and gene set used for this study). Human introns can have very different lengths, contrarily to exons. This difference in intron length leads to substantial differences in size among human genes, which cause differences in the time taken to transcribe a gene from seconds to over 24 hours [4]. Indeed, intron size is highly conserved in genes associated with developmental patterning [5], suggesting that genes that require a precise time coordination of their transcription are reliant on a consistent transcript length. It has been suggested that selection could be acting to reduce the costs of transcription by keeping short introns in highly expressed genes [6], which are enriched in housekeeping essential functions [7]. Genes transcribed early in development [8–10] and genes involved in rapid biological responses [11] also conserve intron-poor structures. Interestingly, Keane and Seoighe [12] recently found that intron lengths of some genes tend to coevolve (their relative sizes co-vary across species) possibly because a precise temporal regulation of the expression of these genes is required. In fact, these genes tend to be coexpressed or participating in the same protein complexes [12].

It is well known that introns contribute to the control of gene expression by their inclusion of regulatory regions and non-coding functional RNA genes or directly by their length [13–15]. Despite the importance of introns in regulating transcription levels, transcription timing and splicing, little attention has been payed to their potential role in human population variability studies. A recent analysis of the literature has revealed a substantial amount of pathogenic variants located “deep” within introns (more than 100 bp from exon-intron boundaries) which suggests that the sequence analysis of full introns may help to identify causal mutations for

many undiagnosed clinical cases [16]. Given that direct associations between intronic mutations and certain diseases have been reported [16–19], we need to characterise the normal genetic variability in introns so we can better distinguish normal from pathogenic variations.

We studied the effect of structural intronic variants on protein coding gene loci in healthy humans using five copy number variant (CNV) maps of high resolution [20–24]. Most of these CNVs were detected using whole genome sequencing (WGS) data, which allows to determine the exact genomic boundaries of these variants. CNVs may have neutral, advantageous or deleterious consequences [25] and can be classified in:

- Gains: regions that are found duplicated when compared with expected number from the reference genome (which is 2 for autosomes).
- Losses (homozygously or heterozygously deleted regions).
- Gain/loss CNVs (regions that are found duplicated in some individuals - or alleles - and deleted in others).

Each of the maps in our study was derived from a different number of individuals, from different populations and using different techniques and algorithms for CNV detection (**Figure S1** and **Table S1)**. Due to these differences, each dataset provided us with a different set of CNVs (**Figure S1**), which we analysed independently, excluding sex chromosome and private variants.

CNVs affect genes in different ways depending on the degree of overlap with them. Some CNVs cover entire genes (from now on *whole gene CNVs*), other CNVs overlap with part of the coding sequence but not the whole gene (*exonic CNVs*) and other CNVs are found within intronic regions (*intronic CNVs*, **Figure 1A**). We defined intronic CNVs as those that do not

overlap with exons from any annotated transcript isoforms or from other overlapping genes.

We found that most of the CNVs overlapping with genes fall within intronic regions (~63% of all CNVs) without any overlap with exons. More surprisingly, of the 13,561 purely intronic CNVs detected, over the 95% (12,986) are losses (12,334) or gain/loss CNVs (652). This is in stark contrast with whole gene CNVs (1,412), which tend to be exclusively gains (55% of the cases) or gain/loss CNVs (25% of the cases) (**Figure 1B-D**). The methods used to generate the other two maps (Handsaker's [21] and Sudmant-Science's [23]) tend to detect less losses and larger CNV regions, resulting in maps with fewer purely intronic deletions (**Figure S1**). There is a significant enrichment of purely intronic losses ($P < 0.001$; permutation testing) in 4 out of 5 maps, with 6 to 15% more deletions falling in introns than expected by chance, depending on the CNV map. (No significant differences with the expected values were found with Sudmant-Science's map, $P = 0.6683$). In contrast with the intronic deletions, there were 13-70% fewer deletions overlapping with exons than would be expected by chance ($P < 0.05$ in all maps). We observed that this enrichment of intronic losses was still significant when controlling for different intron sizes **(Figure S2)** and DNA replication timing ($P <$ 5e-04; permutation testing).

Given the high frequency of purely intronic deletions in the population, we restricted our next analyses on deletions (loss and gain/loss CNVs) from Sudmant-Nature's [24], Zarrei's [22] and Abyzov's [20] maps, which together represent the 86% of all intronic deletions from our datasets.

The percentage of each intron that can be lost in the population due to CNV losses is highly

variable, from 0.03% to 96.8% (51 bp to 293 kb), representing a loss of the 0.01% to 77.5% of the total genic size (51 bp to 893,4 kb). (**Figure 2A-C**). Some examples of genes with a notable change in size after a single intronic deletion in one individual are the neuronal glutamate transporter *SLC1A1 (Solute Carrier Family 1 Member 1)*, with a loss of the 37% of its genic size (**Figure 2D**) and the *LINGO2* (*Leucine Rich Repeat And Ig Domain Containing 2*, alias *LERN3* or *LRRN6C*) gene with a loss of the 34% of its size.

The combination of different intronic deletions within a gene can give place to alleles of several different sizes (**Figure 2E**). In the dataset from the final phase of the 1000 Genomes Project (1KG) [24], we found 5 different intronic deletions in *SLC1A1*. These deletions result in 8 different sizes of genes in the population, with individual losses ranging from 1,1kb bp to 48kb. In *LINGO2*, the 20 different deletions give place to 36 different gene lengths in the 1KG population, with losses of 51 to 233 kp. The gene with more different allele sizes in the 1KG population [24] is *CSMD1* (*CUB And Sushi Multiple Domains 1*), with a total of 66 intronic annotated deletions that, combined, produce 150 alleles of different sizes. This gene has been associated to diseases such as Benign Adult Familial Myoclonic Epilepsy (Malacards [26] MCID: BNG079) and Smallpox (MCID: SML019).

Remarkably, these genes are highly conserved at the protein level and are amongst the 20% of genes most intolerant to functional variation according to the ranking of the RVIS (Residual Variation Intolerance Score) gene scores, which is based on the amount of genetic variation of each gene at an exome level [27]. *CSMD1* gene is the most conserved of the above mentioned genes, with only 0.169% genes in the human genome more intolerant to variation in their coding sequence [27]. In summary, intolerance to variation in the coding sequence seems to be compatible with extreme variation in the intronic sequence.

This shows that genes with a very conserved coding sequence with a general depletion of

coding mutations can have important losses of intronic regions. Remarkably, these losses might affect their regulation without affecting their protein structure. A total of 1,638 OMIM genes carry intronic deletions in the population. For instance, diseases associated to *SLC1A1* (OMIM: 133550) include Dicarboxylic Aminoaciduria and susceptibility to Schizophrenia, while *LINGO2* (OMIM: 609793) has variants associated with essential tremor and Parkinson disease and also has an intronic SNP associated with body mass [28]. To better understand possible epistatic effects between protein-coding and intronic mutations, it might be useful to incorporate information about gene length variation in future studies of these disease genes.

Structural variants in the germline DNA constitute an important source of genetic variability that serves as the substrate for evolution. Therefore, dating the evolutionary age of genes allows the study of structural variants that were fixed millions of years ago. Whole gene CNVs are known to differentially affect genes depending of their evolutionary age, mainly involving evolutionary young genes [29]. Genes of younger ages are generally cell-type specific, while ancient genes tend to be more conserved, ubiquitously expressed and enriched in cellular essential functions. We were intrigued to see many cases where intronic CNVs were affecting highly conserved protein-coding loci, so we decided to compare the distribution of coding (including exonic and whole gene) and intronic deletions across different gene ages (**Figure 3**). We observed that most ancient genes are depleted of deletions that affect their coding regions, while primate-specific genes are enriched in coding CNVs (**Figure 3A**). This pattern was also observed when CNV gains were included (**Figure S4**), meaning that the coding region of recent genes have a higher tendency to be lost or disrupted. The generation of random background models revealed that ancient genes were significantly depleted of coding region losses (both exonic and whole gene) ($P < 0.05$), while these were enriched in young genes ($P < 0.05$, **Figure 3A**).

Surprisingly, we observed an opposite trend for purely intronic deletions: the proportion of ancient genes with intronic deletions was higher than that of young genes, and also higher

than expected by chance ($P < 0.05$, **Figure 3B**). This finding was confirmed with additional analyses considering only genes with introns and adjusting by the different size distributions of introns (**Figure S5**). In summary, the frequency of coding deletions in ancient genes is lower, but their frequency of intronic deletions is higher than the one observed for young genes (**Figure 3C**).

We would expect that essential genes, which tend to be ancient [30], could be an exception to the enrichment of deletions. Essential genes have on average shorter introns than the rest of the genes [31,32] [33,34] and relative to the genes of the same evolutionary age (**Figures S6A** and **S6B**). However, we found 1,158 essential genes carry intronic deletions, a higher proportion than expected by chance (*P < 0.05*, **S2 Table**).

Multiallelic CNVs affecting whole genes have been shown to correlate with gene expression: generally, the higher the number of copies of a gene, the higher its expression levels [21,24]. We hypothesized that intronic size variation may also impact the expression of the affected genes (without affecting the actual number of copies of the gene). Therefore, we looked into the possible effect of intronic hemizygous deletions on gene expression variation at the population level, comparing the effects with hemizygous deletions in coding (whole gene and exonic) and intergenic non-coding deletions (**Figure 4**). We used available RNA-seq data from Geuvadis [35] that was derived from lymphoblastoid cell lines for 445 individuals for whom we have the matching CNV data (Sudmant Nature's map [24]).

In order to look for differences in gene expression we selected variants for which we had at least 2 hemizygous individuals (individuals with copy number = 1) and at least 2 wild-type individuals (copy number = 2) and we compared the expression levels among these two groups to identify deleted CNV regions associated with expression quantitative trait loci (eQTL, **Table 1**). We will refer to the deleted regions associated with expression changes as

DEL-eQTLs, and the genes associated with as eGenes. For comparative purposes, we first looked at the effect of hemizygous deletions in coding regions (whole gene and exonic DEL-eQTLs). We found that 7 eGenes out of 50 genes with whole gene deletion CNVs resulted in significant downregulation of gene expression in lymphoblastoid cell lines (14%, a higher number eGenes than expected by chance, $P < 5e\text{-}4$). In addition, we found 35 eGenes out of 437 genes with partial exonic deletions that were differentially expressed (8%, a number higher than expected by chance, $P < 5e\text{-}4$). The majority of these eGenes (32/35) where down-regulated in the individuals carrying the deletion.

Although intronic deletions do not affect the coding sequence of genes, we observed significant differences in gene expression in 53 eGenes out of the 1,505 genes with intronic deletions, a number of intronic-eGenes that is also higher than expected by chance ($P < 51e\text{-}4$) **(Table 1A)**. Given the higher frequency of intronic deletions in the population, the absolute number of intronic-eGenes (53 genes) was similar to the total of coding-eGenes (39 genes, **Table 1A** and **Table S4**). Of the intronic-eGenes, 62% were downregulated and the other 38% upregulated, suggesting that intronic deletions might result both in enhancing or repressing gene expression (while coding losses mostly associate to gene down-regulation).

Since intron length can impact the inclusion of alternative exons [36], we hypothesised that there might be genes with differentially expressed transcripts (eTranscripts) in any gene containing an intronic deletion. In addition to the 53 intronic-eGenes, we found 217 intronic-eTranscripts in a total of 185 genes (this is more than expected by chance, P = 0.018, **Table 1B)**. These results suggest that deletions within introns may cause the inclusion or exclusion of exons and thus influencing the relative proportion of alternative transcripts in many genes. Changes in GC content as the result of intronic deletions might also contribute to these splicing differences, as in genes with long introns, the recognition of introns and exons by splicing machinery is based on their differential GC content [37,38] and the lower GC content in introns facilitates their recognition. We found that, in general, the deleted sequences have a

significantly higher GC content to that of the introns where they are located (P = 1.8e-28), and the loss of these sequences causes a significant decrease of the overall GC content of the introns (P = 2.23e-16) (**Figures S7 and S8**). This drop in intronic GC content would increase the difference of GC content between introns and their flanking exons, what could facilitate exon definition during splicing and might contribute to the observed differential expression of some transcripts.

Introns in human are particularly enriched in regulatory regions and frequently interact with gene promoters of other genes via chromatin looping (**Figure 4D**). Therefore, deletions in introns that show long-range interactions with promoters of other genes could potentially affect their expression (*trans* effects). We used promoter-capture Hi-C published data for B-lymphocytes [39] to link intronic regions and gene promoters. We identified 322 deletions in intronic regions that interact with gene promoters of other genes (672 in total). Next, we searched for trans-DEL-eQTLs: intronic regions that, when deleted, are associated with changes in expression of a different gene. Twelve of these genes were found to be significantly differentially expressed in the individuals presenting an intronic deletion in another gene (trans-intron-eGenes, **Figure 4F**) In addition, 81 transcripts from 65 genes were also differentially expressed (trans-intron-eTranscripts) in the individuals with a trans-DEL-eQTLs (**Figure 4G**). The number of intergenic deletions associated to significant changes in expression affected a similar number of genes (16 eGenes, 123 eTranscripts, **Figure 4F-G**).

To further study the potential impact of intronic deletions in regulatory regions, we analyzed the co-occurrence of these events with enhancers. In eGenes or eTranscripts, 15 intronic DEL-eQTLs overlap with enhancers present in B-lymphocytes (an overlap that is higher than expected by chance, P = 0.023, odds ratio = 2.04). Moreover, in intronic-eGenes and intronic-eTranscripts, the distance between the DEL-eQTL and the closest enhancer is shorter than the distance of the deletions not associated with expression changes (P = 9.2e-04). These results suggest that intronic DEL-eQTLs could also be affecting interactions between

promoter and intronic enhancers without directly disrupting the enhancer sequence.

Motivated by this findings, we investigated if there is a global tendency (independently of gene expression) for intronic deletions to affect or not affect enhancers. First, we observed that enhancers are enriched in introns ($P < 1e-04$, agreeing with previous findings in plants [40,41]). Strikingly, we find that intronic deletions and intronic enhancers co-occur in the same intron more often than expected by chance ($P < 2.2e-16$), suggesting that variable length is frequent in enhancer-containing introns. However, the direct overlap of the deletions with enhancers is significantly lower than expected ($P < 1e-04$, **Table S3**). A possible functional interpretation of these results is that it might be some degree on plasticity on the distance between intronic enhancer and promoters, but many intronic enhancers might be essential and cannot be lost. Interestingly, as we saw above, the lost of non-essential intronic enhancers can be associated to changes of gene expression.

Our results suggest that intronic CNVs might have a previously unsuspected role in shaping gene expression variability in populations with potentially important consequences in human evolution, adaptation and disease. Intronic CNVs constitute the most abundant form of CNV in protein-coding genes (**Figure 1**). This intronic length variation implies that the actual size of many genes is not yet fixed in human populations. If we specifically analyse the age of different types of eGenes, we see that whole-gene and exonic eGenes are enriched in young age classes (**Figure 5A**). This pattern is very different in intronic and intergenic eGenes: intronic cis-eGenes are enriched in old ages, while intronic trans-eGenes and intergenic-eGenes do not seem to be associated with gene age. If we compare the RVIS of the different types of eGenes, we find that whole gene and exonic eGenes are actually among the most tolerant genes to point mutations in their coding sequence. In contrast, we found that a significant proportion of intronic cis-eGenes show low RVIS percentiles, indicating that protein-coding genes that are intolerant to point mutations at the protein level can have intronic deletions associated to gene expression changes. Strikingly, trans-eGenes show the

lowest RVIS percentiles, indicating that intronic variation might impact the gene expression of interacting genes that are quite intolerant to coding mutations (**Figure 5B**).

An association between gene age and frequency of eQTL SNPs has been previously reported. Popadin *et al*. showed that primate-specific genes in human are enriched in single nucleotide variants correlated with gene expression (cis-eQTLs) with their associated SNPs tending to be closer to the TSS than in older genes [42]. Taken together, these data highlight the need of dissecting the different types of genetic variation in order to understand the complex relationships between SNPs, CNVs, gene expression and gene age. While point mutations near the TSS[42] and coding CNVs seem to have a higher effect in young genes, intronic CNVs are frequently associated with gene expression variation in old genes. Our results suggest that copy number variation is shaping gene evolution in different ways depending on the age of genes, duplicating or deleting young genes and contributing to fine-tuning the regulation of old genes (**Figure 5C**).

Previously published studies on the effect of genetic variants on gene expression have proven the effect of CNVs on expression variability [43–45]. Chiang and co-workers identified 789 SVs associated to changes in gene expression, most of them (88.3%) not overlapping with exons from the eGene[45]. DeBoever and co-workers observed that a large proportion of common CNVs associated with gene expression levels is located in intergenic regulatory regions[44]. However, research on the subject has been mostly restricted to SVs found within 1Mb from the gene and previous works did not analysed intronic regions in detail. **In contrast, we relied on Hi-C data to define deletions affecting regions in 3D contact with a gene**. In this way, we don't require the CNV to be located within any particular distance to the TSS position of a gene.We tested intergenic eCNVs that can be located at any distance from 864bp to 82Mb the nearest gene.

Despite the clear trends shown, our results are likely to underestimate the extent of the impact of intron loses in gene expression. The interaction maps change in different cell types [39,47] and many enhancers are tissue-specific [46], meaning that the loss of intronic sequence could affect the expression of genes in other cell types. In addition, the 3D contacts involving frequently deleted regions in the population will be underrepresented in the interaction map used in our study, as they are less likely to be present in the assayed samples. The availability of CNV, personal gene expression and genome interactomes from multiple tissues will allow to evaluate more accurately what is the impact of coding and non-coding deletions in the whole organism. With the results presented here, we emphasize the importance of sequencing and analysing variants located in introns as they can potentially be as consequential as regulatory elements found in intergenic regions.

Supplemental Data

Supplemental Data include eight figures and six tables, including a list of all CNVs used in this study and a list of overlapping by these CNVs.

Materials and methods

Origin and filtering of CNV maps

Whole genome CNV maps were downloaded from 5 different publications [20–23]. For our analysis we selected autosomal and not private CNVs. Some extra filters were applied to some maps: In Handsaker et al. we removed CNVs marked as low quality and all the variants from two of the individuals (NA07346 and NA11918) because they were not included in the phased map. From Zarrei's maps we used the stringent map that considered CNVs that appeared in at least 2 individuals and in 2 studies. The complete list of CNVs analysed is available in **Table S5**.

Gene structures

Autosomal gene structures and sequences were retrieved from Ensembl [48] (http://www.ensembl.org; version 75) and principal isoforms were determined according to the APPRIS database [49], Ensembl version 74. In order to avoid duplicate identification of introns, intronic regions were defined as regions within introns that aren't coding in any transcript of any gene. When analyzing real introns, in order to avoid duplicate identification of introns, the principal isoform with a higher exonic content was taken. The complete list of genes affected by different types of CNVs is available in **S6 Table**. Genomic sequences were obtained from the primary GRCh37/hg19 assembly, and were used for calculating the GC content of introns and intronic CNVs.

Essential genes

The list of essential genes was obtained by aggregating lists of genes reported as essential after CRISPR-based genomic targeting [50,51] or gene-trap insertional mutagenesis methodology [52].

Dating gene and intron ages

An age was assigned to all duplicated genes as described in Juan et al. 2013. In the case of singletons gene ages were assigned from the last common ancestor to all the genes in their family according to the gene trees retrieved from ENSEMBL. Singleton's ages can be noisy

for genes suffering important alterations as gene fusion/fission events or divergence shifts. As a consequence, these ages should not be interpreted as the age of the oldest region of the gene, but as a restrictive definition of gene age considering a similar gene structure and gene product.

The ages (from ancient to recent) and number of genes per age are as follows: FungiMetazoa: 1119, Bilateria: 2892, Chordata: 1152, Euteleostomi: 8230, Sarcopterygii: 182, Tetrapoda: 154, Amniota: 408, Mammalia: 375, Theria: 515, Eutheria: 848, Simiiformes: 233, Catarrhini: 170, Hominoidea: 106, Hominidae: 64, HomoPanGorilla: 204, HomoSapiens: 500. For some analyses, Primates age groups (Simiiformes to HomoSapiens) were collapsed. For other analyses, we only considered two extreme groups, "ancient" (collapsing groups from FungiMetazoa to Sarcopterygii) and "young" genes (Primates).

Intronic regions were assigned the evolutionary age of the gene they belonged to. In the cases when an intron could be assigned to more than one gene, the most recent age was assigned to them.

Statistical assessment of genome-wide distribution of CNVs

To estimate statistical significance of our results we performed permutation tests. In order to compare the number of overlaps of CNVs with genic functional elements we compared our observed values to a background model. This model was obtained by relocating all the CNVs in the whole genome (except for centromeres and telomeres) 10,000 times.

In addition, we generated background models correcting by DNA replication timing. For this, we downloaded DNA replication timing data from 15 cell lines from ENCODE [53,54] and assigned the median value of all cell lines to each 1 Kb window of the genome. Then, we classified the genome in 5 intervals of DNA replication timing and we relocated the CNVs within its interval of replication timing.

We compared the location of the CNVs in our datasets and compared with their distribution in the random models in order to calculate enrichments or depletions depending on the intron size and gene age and essentiality.

Regulatory features

We downloaded a genome-wide set of regions that are likely to be involved in gene regulation from the Ensembl Regulatory Build [55], assembled from IHEC epigenomic data [56]. We checked if introns are enriched in these regulatory features (promoters, enhancers, promoter flanking regions or insulators) by comparing to a random background model generated by relocating 10,000 times all regulatory features in the genome. P-values are the fraction of random values superior or inferior to the observed values.

In order to check for the significance of the overlaps between intronic deletions and regulatory features we relocated 10,000 all intronic deletions within their introns and checked for differences in overlap with regulatory features.

Gene expression analysis

We used available RNA-seq data at Geuvadis [35] that was derived from lymphoblastoid cell lines for 445 individuals who were sequenced by the 1000 Genomes Project and for whom we have the intronic deletions in the largest CNV map [24]. We focused our analyses on the 763 genes that have only one intronic deletion in the population with at least two individuals affected in the Geuvadis dataset. For each of these genes we classified the PEER normalized gene expression levels [57] in two groups: 1) gene expression of individuals homozygous for the reference genotype and 2) gene expression of individuals with one allele with the deletion and the other with the reference genoytpe. We then performed Student's t-tests to compare the expression of the two different genotypes. We corrected for multiple testing with p.adjust R function (Benjamini-Hochberg method). In addition, we randomized the individuals with the intronic deletions 10,000 times and calculated the expected percentages of significantly differentially expressed genes.

Acknowledgments and funding information

The authors thank Salvador Capella, Matthew Bashton, Venetia Bigley, Joris Veltman, Vera Pancaldi and Inmaculada Hernandez and Ruth Rodriguez Barrueco for their constructive comments on the manuscript. Alfonso Valencia acknowledges the Joint BSC-CRG-IRB Research Program in Computational Biology and Daniel Rico thanks the Newcastle University Centre for Health and Bioinformatics. Maria Rigau acknowledges a La Caixa

fellowship.

This work was supported by Project Retos BFU2015-71241-R Spanish Ministry of Economy, Industry and Competitiveness (MEIC, http://www.mineco.gob.es), co-funded by European Regional Development Fund (ERDF) to A.V. and Wellcome Trust (https://wellcome.ac.uk) Seed Award in Science (206103/Z/17/Z) to D.R.

References

1. Lynch, M. *The Origins of Genome Architecture*. (Sinauer Associates Incorporated, 2007).
2. Chorev, M. & Carmel, L. The function of introns. *Front. Genet.* **3,** 55 (2012).
3. Jo, B.-S. & Choi, S. S. Introns: The Functional Benefits of Introns in Genomes. *Genomics Inform.* **13,** 112 (2015).
4. Kirkconnell, K. S. *et al.* Gene length as a biological timer to establish temporal transcriptional regulation. *Cell Cycle* **16,** 259–270 (2017).
5. Seoighe, C. & Korir, P. K. Evidence for intron length conservation in a set of mammalian genes associated with embryonic development. *BMC Bioinformatics* **12 Suppl 9,** S16 (2011).
6. Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V. & Kondrashov, F. A. Selection for short introns in highly expressed genes. *Nat. Genet.* **31,** 415–418 (2002).
7. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29,** 569–574 (2013).
8. Heyn, P. *et al.* The earliest transcribed zygotic genes are short, newly evolved, and different across species. *Cell Rep.* **6,** 285–292 (2014).
9. Artieri, C. G. & Fraser, H. B. Transcript length mediates developmental timing of gene expression across Drosophila. *Mol. Biol. Evol.* **31,** 2879–2889 (2014).
10. Swinburne, I. A. & Silver, P. A. Intron delays and transcriptional timing during development. *Dev. Cell* **14,** 324–330 (2008).
11. Jeffares, D. C., Penkett, C. J. & Bähler, J. Rapidly regulated genes are intron poor. *Trends Genet.* **24,** 375–378 (2008).
12. Keane, P. A. & Seoighe, C. Intron Length Coevolution across Mammalian Genomes. *Mol. Biol. Evol.* **33,** 2682–2691 (2016).
13. Rose, A. B. Intron-mediated regulation of gene expression. *Curr. Top. Microbiol. Immunol.* **326,** 277–290 (2008).
14. Le Hir, H., Nott, A. & Moore, M. J. How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci.* **28,** 215–220 (2003).
15. Jo, B.-S. & Choi, S. S. Introns: The Functional Benefits of Introns in Genomes. *Genomics Inform.* **13,** 112 (2015).
16. Vaz-Drago, R., Custódio, N. & Carmo-Fonseca, M. Deep intronic mutations and human disease. *Hum. Genet.* (2017). doi:10.1007/s00439-017-1809-4

17. Agrawal, A. *et al.* An intronic ABCA3 mutation that is responsible for respiratory disease. *Pediatr. Res.* **71,** 633–637 (2012).
18. Lo, Y.-F. *et al.* Recurrent deep intronic mutations in the SLC12A3 gene responsible for Gitelman's syndrome. *Clin. J. Am. Soc. Nephrol.* **6,** 630–639 (2011).
19. Nurnberg, S. T. *et al.* From Loci to Biology: Functional Genomics of Genome-Wide Association for Coronary Disease. *Circ. Res.* **118,** 586 (2016).
20. Abyzov, A. *et al.* Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat. Commun.* **6,** 7256 (2015).
21. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nat. Genet.* **47,** 296–303 (2015).
22. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16,** 172–183 (2015).
23. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349,** aab3761 (2015).
24. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
25. Conrad, D. F. *et al.* Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat. Genet.* **42,** 385–391 (2010).
26. Rappaport, N. *et al.* MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.* **45,** D877–D887 (2017).
27. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9,** e1003709 (2013).
28. Rask-Andersen, M., Almén, M. S., Lind, L. & Schiöth, H. B. Association of the LINGO2-related SNP rs10968576 with body mass in a cohort of elderly Swedes. *Mol. Genet. Genomics* **290,** 1485–1491 (2015).
29. Juan, D., Rico, D., Marques-Bonet, T., Fernández-Capetillo, O. & Valencia, A. Late-replicating CNVs as a source of new genes. *Biol. Open* **3,** (2014).
30. Chen, W.-H., Trachana, K., Lercher, M. J. & Bork, P. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Mol. Biol. Evol.* **29,** 1703–1706 (2012).
31. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes are compact. *Trends Genet.* **19,** 362–365 (2003).

32. Vinogradov, A. E. Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* **20,** 248–253 (2004).
33. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes are compact. *Trends Genet.* **19,** 362–365 (2003).
34. Vinogradov, A. E. Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* **20,** 248–253 (2004).
35. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
36. Roy, M., Kim, N., Xing, Y. & Lee, C. The effect of intron length on exon creation ratios during the evolution of mammalian genomes. *RNA* **14,** 2261–2273 (2008).
37. Amit, M. *et al.* Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep.* **1,** 543–556 (2012).
38. Gelfman, S. & Ast, G. When epigenetics meets alternative splicing: the roles of DNA methylation and GC architecture. *Epigenomics* **5,** 351–353 (2013).
39. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167,** 1369–1384.e19 (2016).
40. Majewski, J. & Ott, J. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12,** 1827–1836 (2002).
41. Rose, A. B. Intron-mediated regulation of gene expression. *Curr. Top. Microbiol. Immunol.* **326,** 277–290 (2008).
42. Popadin, K. Y. *et al.* Gene age predicts the strength of purifying selection acting on gene expression variation in humans. *Am. J. Hum. Genet.* **95,** 660–674 (2014).
43. Haas, J. *et al.* Genomic structural variations lead to dysregulation of important coding and non-coding RNA species in dilated cardiomyopathy. *EMBO Mol. Med.* (2017). doi:10.15252/emmm.201707838
44. DeBoever, C. *et al.* Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. *Cell Stem Cell* **20,** 533–546.e7 (2017).
45. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat. Genet.* **49,** 692 (2017).
46. Pennacchio, L. A., Loots, G. G., Nobrega, M. A. & Ovcharenko, I. Predicting tissue-specific enhancers in the human genome. *Genome Res.* **17,** 201–211 (2007).
47. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep.* **17,** 2042–2059 (2016).

48. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res.* **44,** D710–6 (2016).
49. Rodriguez, J. M. *et al.* APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41,** D110–7 (2013).
50. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163,** 1515–1526 (2015).
51. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350,** 1096–1101 (2015).
52. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350,** 1092–1096 (2015).
53. Thurman, R. E., Day, N., Noble, W. S. & Stamatoyannopoulos, J. A. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.* **17,** 917–927 (2007).
54. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 139–144 (2010).
55. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The ensembl regulatory build. *Genome Biol.* **16,** 56 (2015).
56. Stunnenberg, H. G. & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167,** 1897 (2016).
57. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6,** e1000770 (2010).

Figure titles and legends

Figure 1. Types of CNVs in the different datasets. (A) CNVs can overlap entire genes or fractions of genes. CNVs overlapping with exons of a gene (exonic CNVs) and CNVs found within introns (intronic CNVs). (B-D) Number of whole gene, exonic and intronic CNV events, showing the different proportions of CNV gains, losses and gain and loss CNVs.

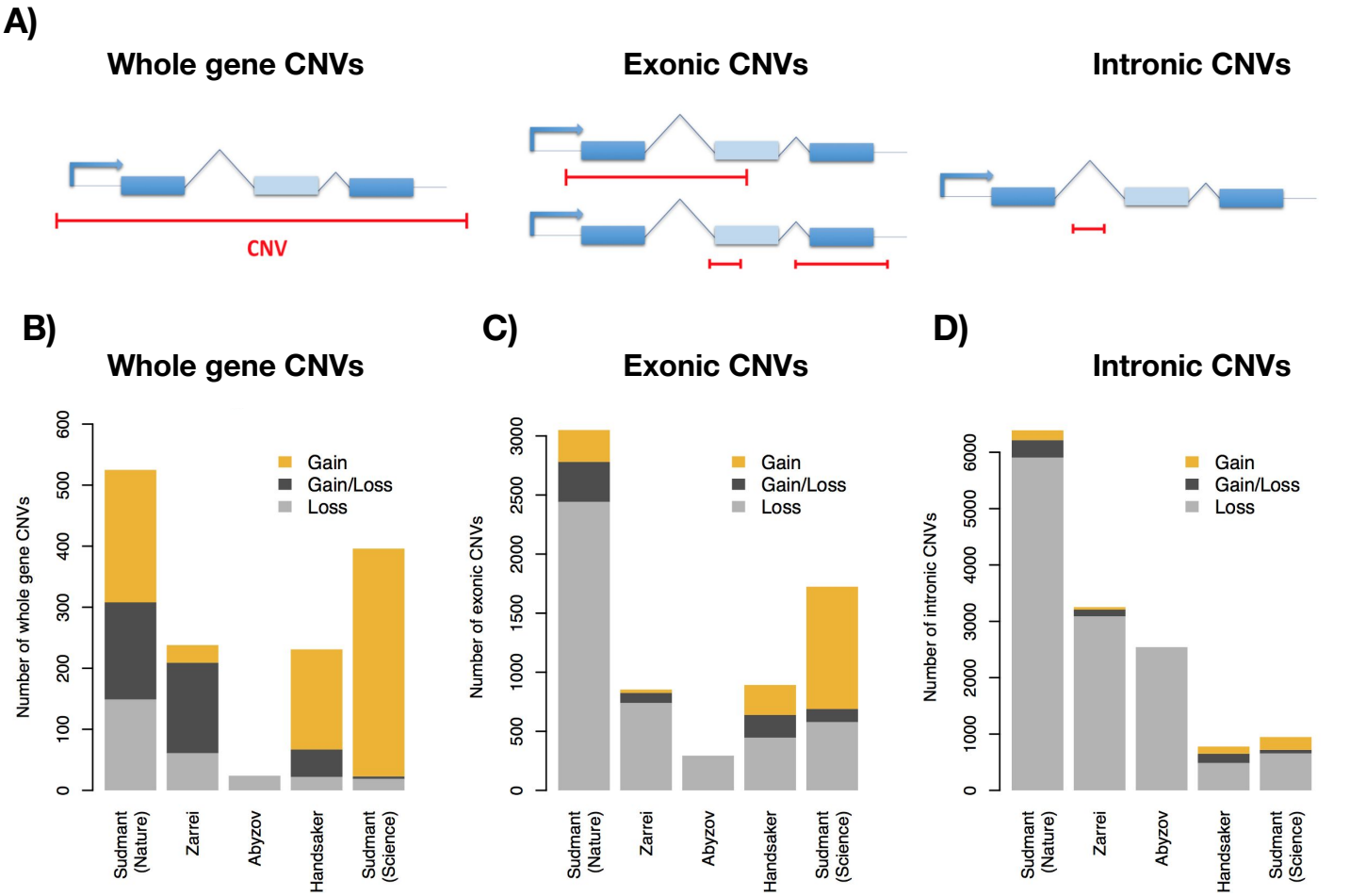
Figure 2. Changes in intron and gene size. (A) Proportion of the reference intron that has been observed as deleted in any of the studies. (B) Proportion of the whole intronic content of a gene that has been observed as deleted. (C) Change in gene length by intronic deletions. (D) Example of gene with a substantial change in gene size with a single intronic deletion. (E)

Number of different gene sizes observed in the population as a function on the number of intronic deletions detected. Genes names of the seven most extreme cases are indicated.

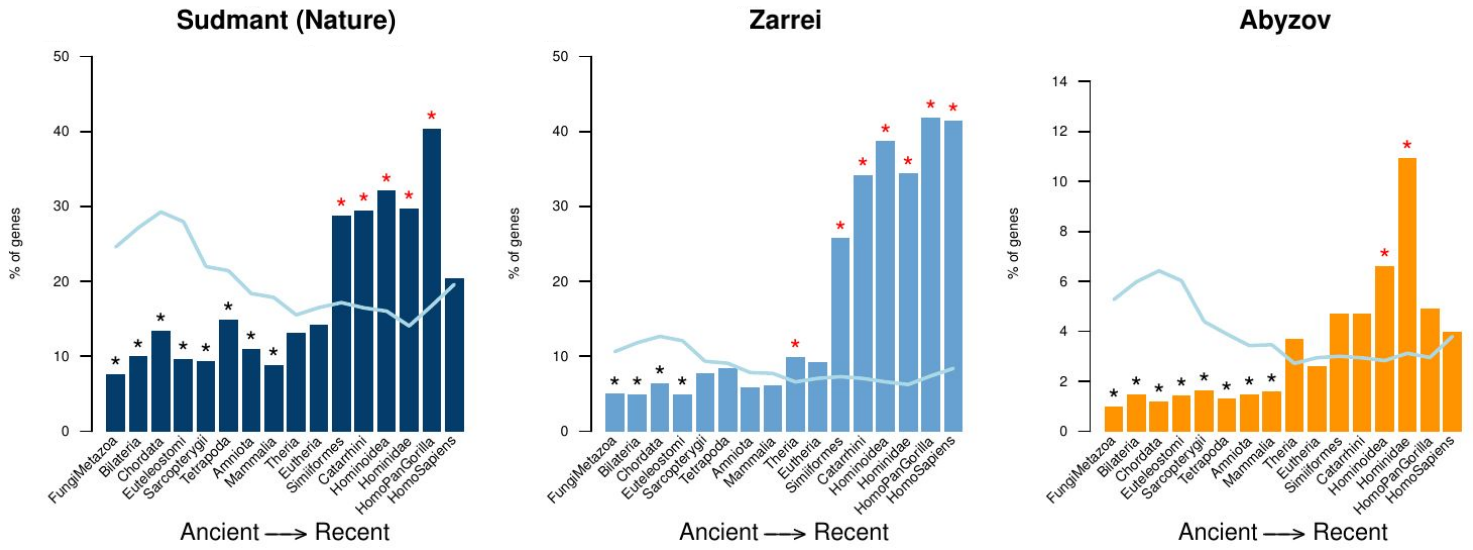
Figure 3. Evolutionary age of affected genes. Percentage of genes from each gene evolutionary age that contain deletions overlapping with exons, including partial and whole gene CNVs (A) or intronic deletions (B). The light blue line represents the expected value, calculated as the mean of the genes in the 10,000 random permutations. Red asterisks mark the significantly enriched groups of genes, while black asterisks mark gene age groups with fewer deletions than expected ($P < 0.05$). Plot (C) shows, from all the genes overlapping with deletions after aggregating the three maps, what is the proportion of genes that have all or part of their exons affected by deletions and what is the percentage of genes with intronic deletions only. Bar width is proportional to the percentage of genes from each evolutionary age that is affected by deletions of any kind, which spans from 18.5% (Mammalia) to 49.8% (HomoPanGorilla). The equivalent figure for each separate map is shown in S3 Fig.

Figure 4. DEL-eQTLs. (A-E) Five types of DEL-eQTLs analysed. Thunder symbols indicate deletion breakpoints. (F-G) DEL-eQTL results for the five types. Number of eGenes (F), eTranscripts/genes with eTranscripts (G) when comparing expression levels of individuals with a reference allele and an allele with a specific deletion versus individuals with two reference alleles. P-values obtained after performing Student's t-tests were FDR-corrected (FDR = 10%). The number of expected eGenes, eTranscripts or genes with eTranscripts was calculated after randomizations of the individuals carrying or not the deletion, and P-values were calculated by comparing the observed versus the 10000 random values. Significance: * for $P<0.05$. ** for $P<0.005$.** for $P<0.0005$.

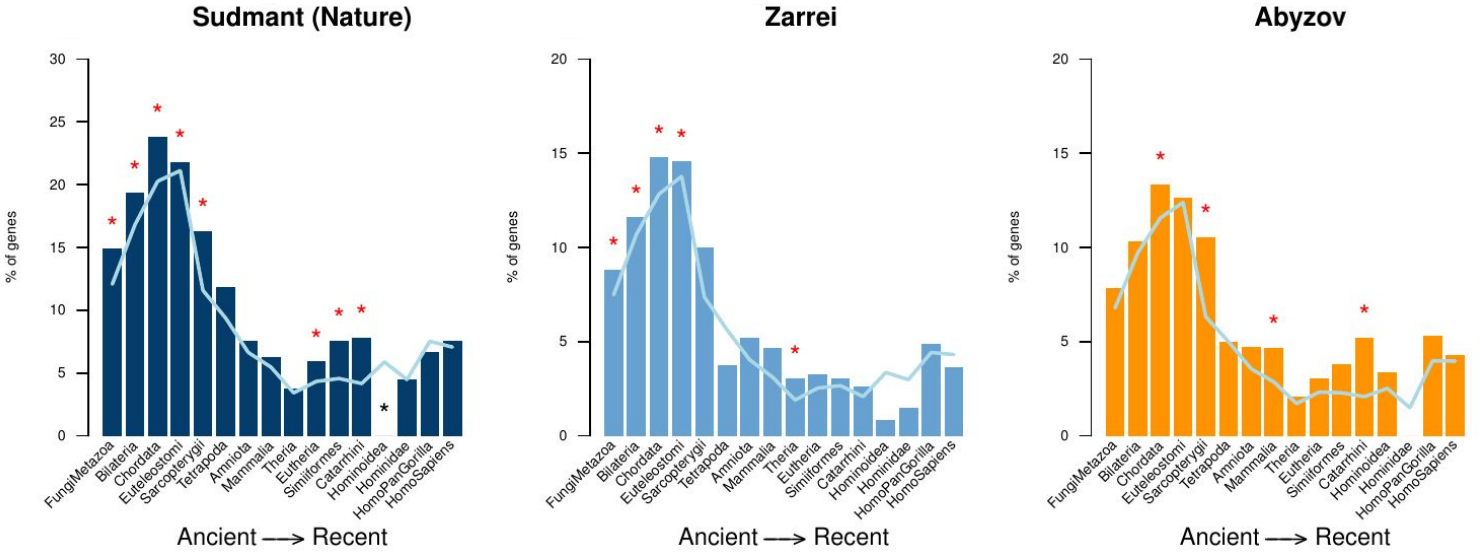
Figure 5. Impact of CNVs on genes and their evolution. A) Percentage of genes of each group of evolutionary ages that is associated to an eCNV, for each type of eCNV. **B)** RVIS percentile of the eGenes, by type of eCNV. Genes with the lowest percentile are among the most intolerant of human genes. C) Evolutionarily ancient and young genes accumulate different kinds of structural variants. While young genes are enriched in coding deletions (which alter gene dosage or disrupt the protein, sometimes affecting gene expression), ancient genes have highly conserved coding sequence but an enrichment of deletions within their introns. As we have shown, these changes in introns can be associated with changes in gene expression, showing that although the protein is highly conserved, the expression of it can change from an individual to another due to changes in regulation.


A)
Whole gene CNVs
CNV
Exonic CNVs
Intronic CNVs
B)
Whole gene CNVs
Number of whole gene CNVs
0
100
200
300
400
500
600
Gain
Gain/Loss
Loss
Sudmant (Nature)
Zarrei
Abyzov
Handsaker
Sudmant (Science)
C)
Exonic CNVs
Number of exonic CNVs
0
500
1000
1500
2000
2500
3000
Gain
Gain/Loss
Loss
Sudmant (Nature)
Zarrei
Abyzov
Handsaker
Sudmant (Science)
D)
Intronic CNVs
Number of intronic CNVs
0
1000
2000
3000
4000
5000
6000
Gain
Gain/Loss
Loss
Sudmant (Nature)
Zarrei
Abyzov
Handsaker
Sudmant (Science)

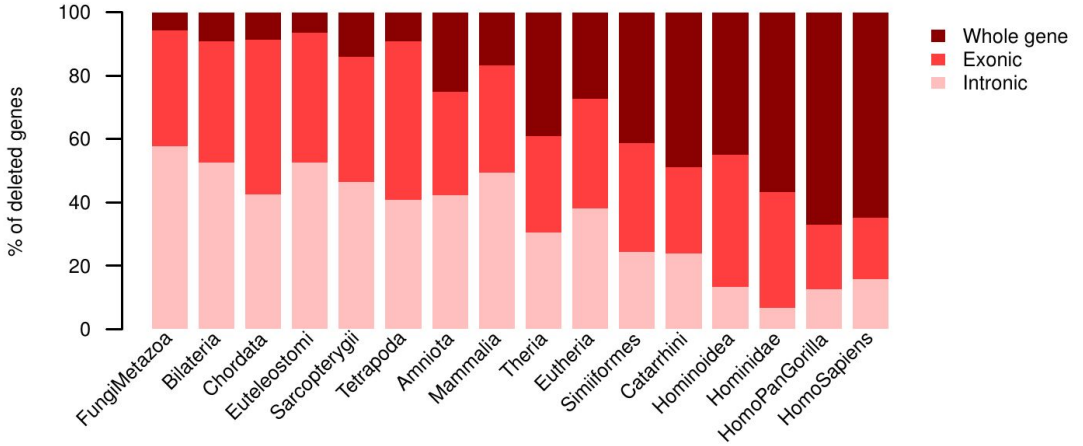
A) Coding deletions

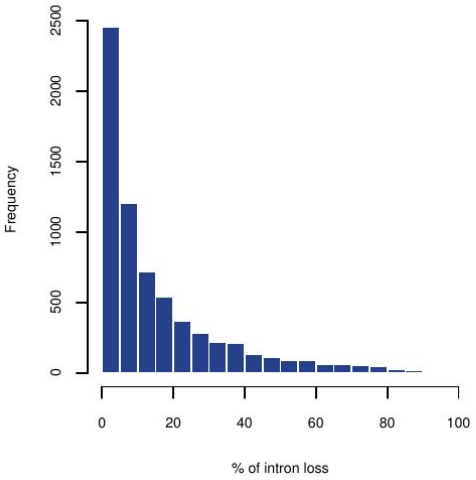


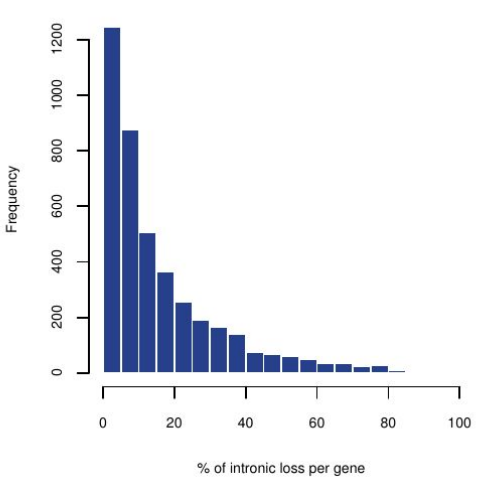
B) Intronic deletions

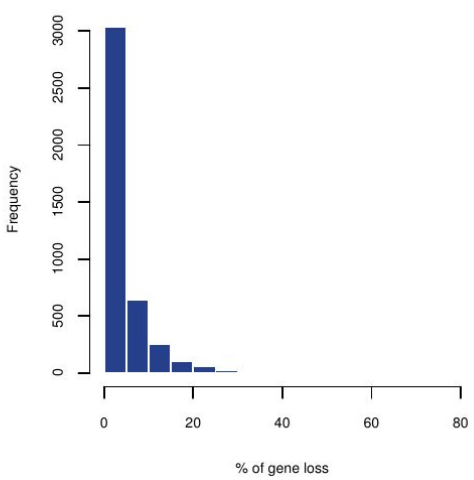


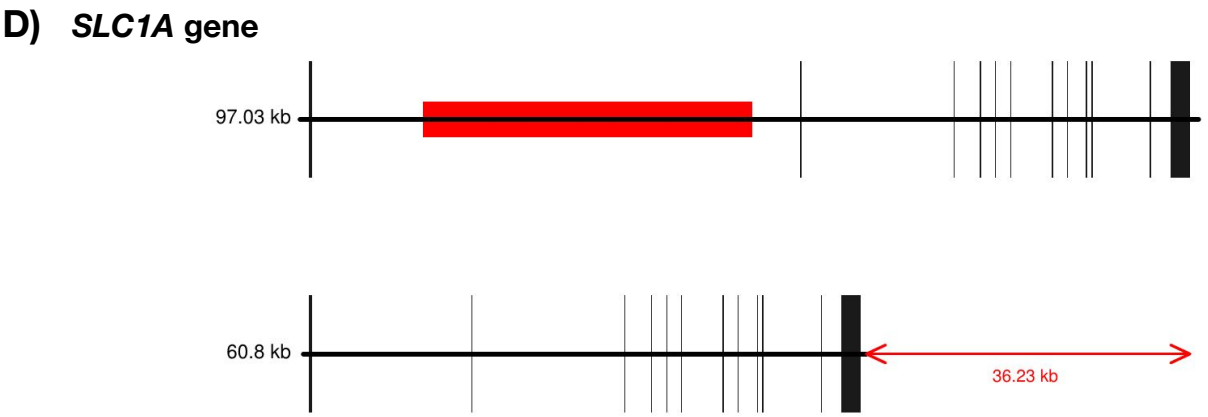
C)

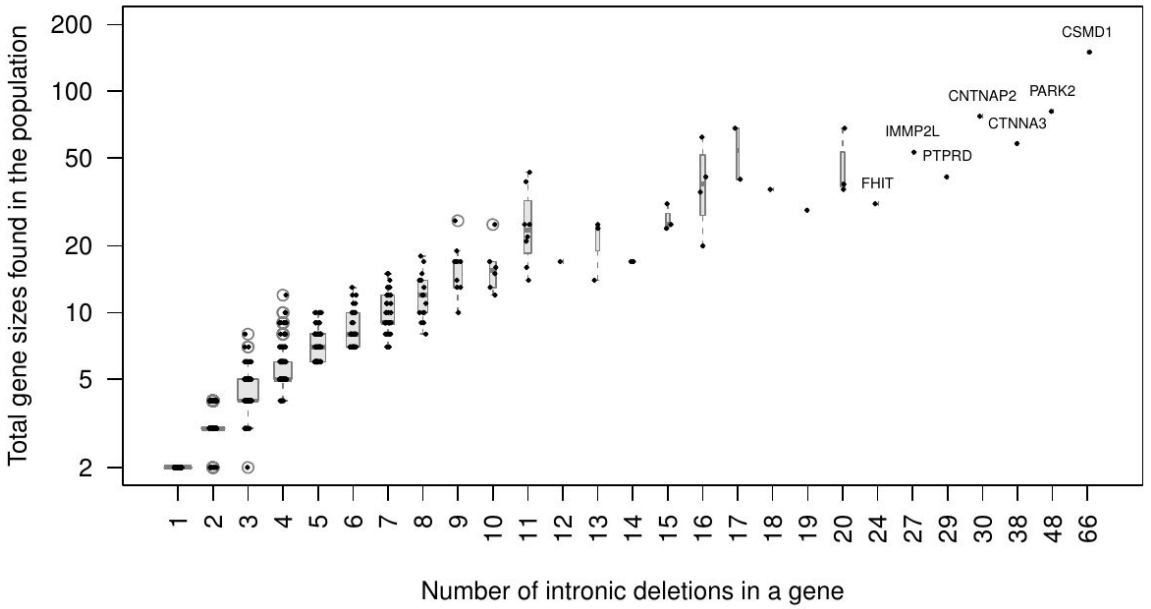



A)
Proportion of intron lost
Frequency
0
500
1000
1500
2000
2500
0
20
40
60
80
100
% of intron loss

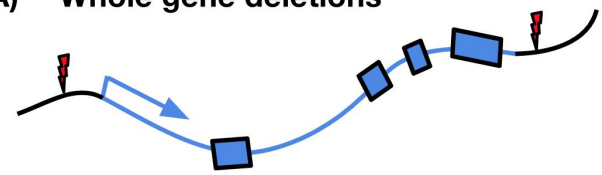

B)
Proportion of intronic content lost per gene
Frequency
0
200
400
600
800
1000
1200
0
20
40
60
80
100
% of intronic loss per gene


C)
Change in gene length by intronic losses
Frequency
0
500
1000
1500
2000
2500
3000
0
20
40
60
80
% of gene loss

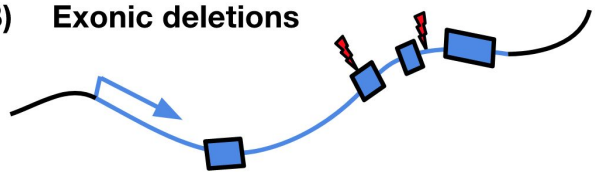

D)
SLC1A gene
97.03 kb
60.8 kb
36.23 kb


E)
Number of different allele sizes per gene
Total gene sizes found in the population
2
5
10
20
50
100
200
CSMD1
CNTNAP2
PARK2
CTNNA3
IMMP2L
PTPRD
FHIT
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
24
27
29
30
38
48
66
Number of intronic deletions in a gene

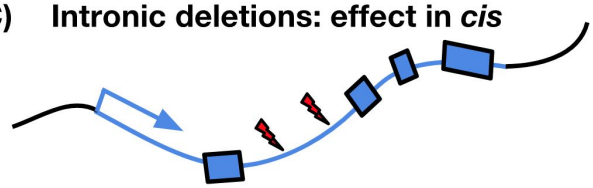
A) Whole gene deletions



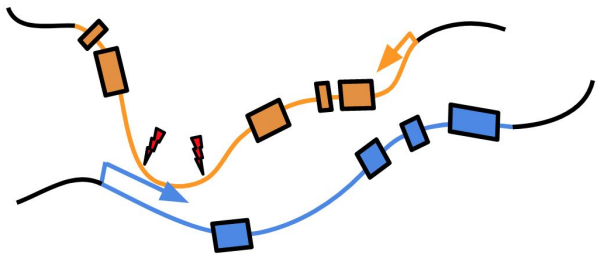
B) Exonic deletions



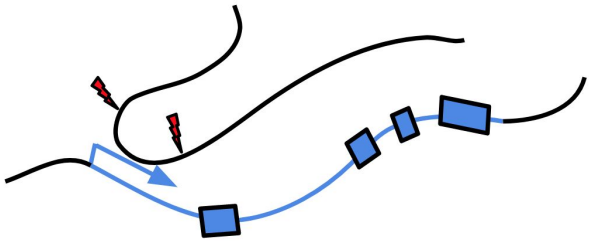
C) Intronic deletions: effect in *cis*



D) Intronic deletions: effect in *trans*



E) Intergenic deletions

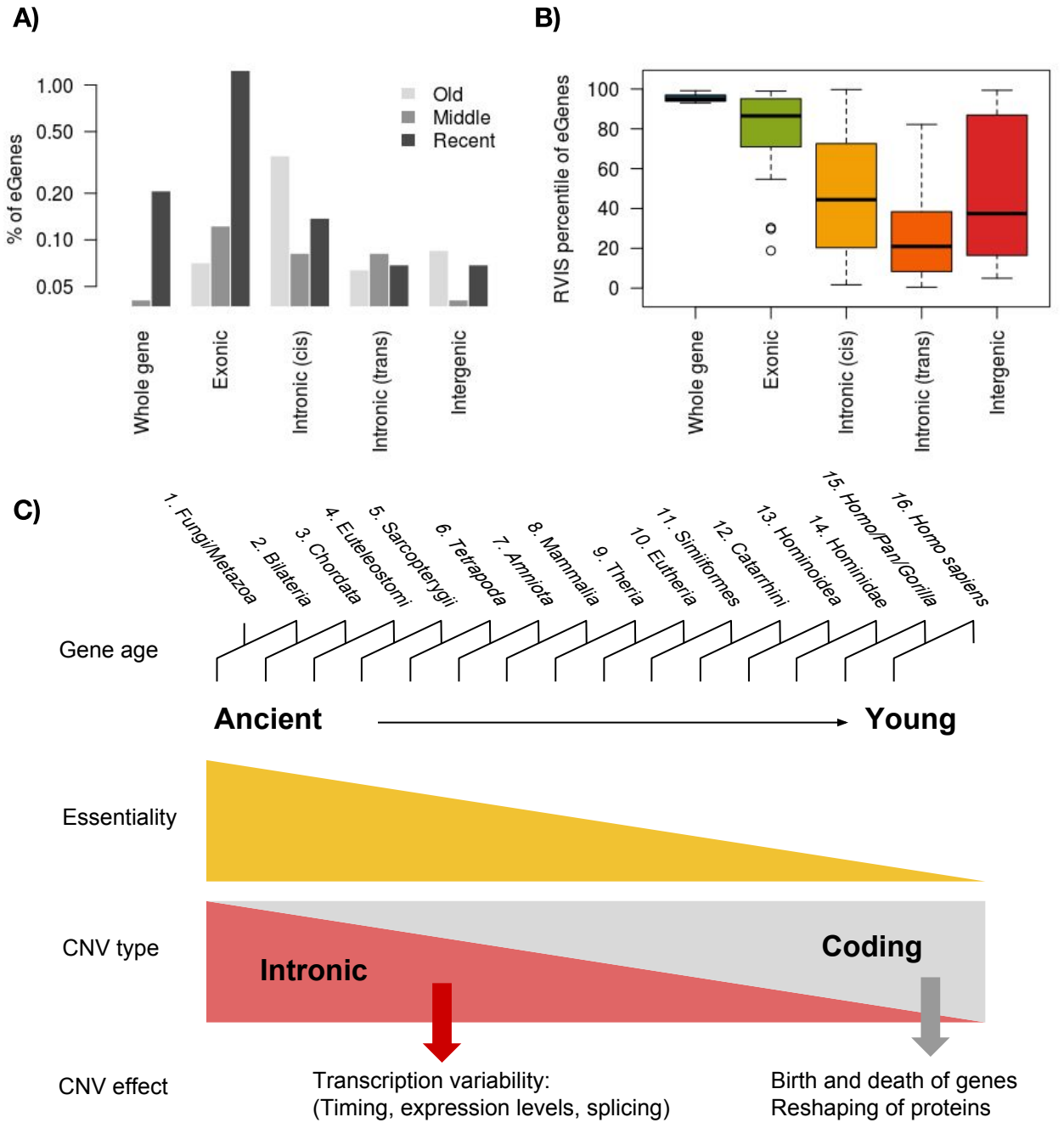


F) Differentially expressed genes

	Whole gene	Exonic	Intronic - cis	Intronic - trans	Intergenic
Observed umber of eGenes (number of eDeletions)	7*** (8)	35*** (36)	53*** (56)	12 (12)	16 (18)
Expected number of eGenes (median ± median absolute dev.)	1 ± 1.48	8 ± 2.97	27 ± 5.93	9 ± 2.97	14 ± 4.45
Proportion of downregulated eGenes	100%	92.5%	68%	58%	83%
Total genes tested (total deletions tested)	50 (45)	437 (472)	1505 (2046)	672 (322)	1011 (545)

G) Differentially expressed transcripts

	Whole gene	Exonic	Intronic - cis	Intronic - trans	Intergenic
Number of eTranscripts (number eDeletions)	22*** (11)	135*** (92)	217* (199)	81 (54)	123 (96)
Expected number of eTranscripts (median ± median absolute dev.)	4 ± 1.48	67 ± 10.38	173 ± 19.27	75 ± 10.38	109 ± 13.34
Number of genes ≥ 1 eTranscript	11 **	87 ***	185 **	65	104
Expected number of genes with ≥1 eTranscript (median ± median absolute dev.)	4 ± 1.48	53 + 7.41	143 ± 14.83	64 ± 8.90	94 ± 10.38
Proportion of downregulated eTranscripts	100%	91%	79%	81%	89%
Total genes tested for transcript differential expression (total deletions tested)	47 (43)	403 (440)	1401 (1886)	653 (319)	972 (529)


A)
% of eGenes
1.00
0.50
0.20
0.10
0.05
Old
Middle
Recent
Whole gene
Exonic
Intronic (cis)
Intronic (trans)
Intergenic
B)
RVIS percentile of eGenes
100
80
60
40
20
0
Whole gene
Exonic
Intronic (cis)
Intronic (trans)
Intergenic
C)
1. Fungi/Metazoa
2. Bilateria
3. Chordata
4. Euteleostomi
5. Sarcopterygii
6. Tetrapoda
7. Amniota
8. Mammalia
9. Theria
10. Eutheria
11. Similiformes
12. Catarrhini
13. Hominoidea
14. Hominidae
15. Homo/Pan/Gorilla
16. Homo sapiens
Gene age
Ancient
Young
Essentiality
CNV type
Intronic
Coding
CNV effect
Transcription variability:
(Timing, expression levels, splicing)
Birth and death of genes
Reshaping of proteins