Widespread population variability of intron size in evolutionary old genes: implications for gene expression variability

Intronic CNVs and gene regulation

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 were originally thought to be “junk DNA” without function but accumulating evidence has shown that they can have important functions in the regulation of gene expression. In humans and other mammals, introns can be extraordinarily large and together they account for the majority of the sequence in human protein-coding loci. However, little is known about their structural variation in human populations and the potential functional impact of this genomic variation. To address this, we have studied how copy number variants (CNVs) differentially affect exonic and intronic sequences of protein-coding genes. Using five different CNV maps, we found that CNV gains and losses are consistently underrepresented in coding regions. However, we found purely intronic losses in protein-coding genes more frequently than expected by chance, even in essential genes. Following a phylogenetic approach, we dissected how CNV losses differentially affect genes depending on their evolutionary age. Evolutionarily young genes frequently overlap with deletions that partially or entirely eliminate their coding sequence, while in evolutionary ancient genes the losses of intronic DNA are the most frequent CNV type. A detailed characterisation of these events showed that the loss of intronic sequence can be associated with significant differences in gene length and expression levels in the population. In summary, we show that genomic variation is shaping gene evolution in different ways depending on the age and function of genes. CNVs affecting introns can exert an important role in maintaining the variability of gene expression in human populations, a variability that could be related with human adaptation.

Author summary

Most human genes have introns that have to be removed after a gene is transcribed from DNA to RNA because they not encode information to translate RNA into proteins. As mutations in introns do not affect protein sequences, they are usually ignored when looking for normal or pathogenic genomic variation. However, introns comprise about half of the human genome and they can have important regulatory roles. We show that deletions of intronic regions appear more frequent than previously expected in the healthy population, with a significant proportion of genes with evolutionary ancient and essential functions carrying them. This finding was very surprising, as ancient genes tend to have high conservation of their coding sequence. However, we show that deletions of their non-coding intronic sequence can produce considerable changes in their length, significant drops of GC content that could affect splicing or occur in introns harboring regulatory elements. Finally, we found that a significant number of these intronic deletions are associated with under- or over-expression of the affected genes, showing that intronic deletions can be responsible for gene expression variability in ancient genes with highly conserved protein sequences. Our data suggests that the frequent gene length variation in ancient genes resulting from intronic CNVs might have an important role in the fine-tuning of their regulation in different individuals.

Introduction

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 the potential energetic disadvantage that they represent for the cell might be compensated by a number of acquired functionalities [1–3]. For example, introns make possible the expression of multiple transcription products from a single gene by alternative splicing and facilitate the formation of new genes by exon shuffling [3,4].

Human introns are longer than those of other vertebrates and invertebrates [5,6] and their lengths are very variable, contrarily to exon lengths. This variability 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 [7]. Introns can influence several steps of gene transcription [8,9] and it has been seen that a considerable amount of intronic sequence is important in regulating gene expression [10].

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 [3,11,12]. Indeed, intron size is highly conserved in genes associated with developmental patterning [13], suggesting that genes that require a precise time coordination of their transcription are reliant on a consistent transcript length. Highly expressed genes are enriched in housekeeping essential functions [14] and tend to have shorter introns [15]. It has been suggested that selection could be acting to reduce the costs of transcription by keeping short introns in highly expressed genes [15]. Genes transcribed early in development [16–18] and genes involved in rapid biological responses [19] also conserve intron-poor structures. Shorter introns would allow these genes to be transcribed faster and thus they may be particularly sensitive to changes in the time to be transcribed. Interestingly, Keane and Seoighe [20] recently found that intron lengths of coexpressed genes or genes participating in the same protein complexes tend to coevolve (their intron sizes show a significant correlation across species) possibly because a precise temporal regulation of the expression of these genes is required.

Despite their potential importance in regulating transcription levels, transcription timing and splicing, little attention has been payed to the potential role of introns in human population variability studies. Given that direct associations between intronic mutations and certain diseases have been reported [24–27], we need to characterise the normal genetic variability in introns so we can better distinguish normal from pathogenic variations.

It is increasingly apparent that one of the most important sources of variability is the presence of copy number variants (CNVs). CNVs are defined as imbalanced structural variants that result in the gain or loss of >50 bp of genomic sequence [28] and appear in more than one individual of the population. CNVs 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) or gain/loss CNVs (regions that are found duplicated in some individuals - or alleles - and deleted in others). Microarray and next generation sequencing approaches have shown that CNVs are more important and frequent than originally thought. CNVs may have neutral, advantageous or pathological consequences [29]. Initially, CNVs thought to account for about 1% of the entire human genome [30], current estimates range from 4.8 to 9.5% [31].

Here, we have studied the effect of intronic variants using CNV maps of high resolution. Most of these CNVs have been detected using whole genome sequencing (WGS) data, which allows to determine the exact genomic boundaries of these variants and thus their overlap with exons and/or introns. Despite significant advances in the detection of CNVs, discovery and genotyping of these variants remain challenging [32]. To gain consistency, we have analyzed in parallel 5 recent CNV maps obtained by different groups with different experimental and analysis systems [31,33–36]. We have been able to compare the effect of the CNVs overlapping totally or partially with the genes of different evolutionary ages, studying in depth the effect of the intronic variants. We show how intronic variation results in widespread gene length variability in human populations and the potential impact of this variability in splicing and gene expression.

Results

Most genic CNVs fall within introns

CNVs can 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*, **Fig 1A**). As defined, intronic CNVs do not overlap with exons from any annotated transcript isoforms or with exons from overlapping genes.

To analyze the impact of CNVs on protein coding genes in healthy humans, we used five recently published, high resolution CNV maps [31,33–36]. Each of the maps has been derived from a different number of individuals, from different populations and using different techniques and algorithms for CNV detection (**S1 Fig** and **S1 Table)**. Due to these differences, each dataset provides us with a different set of CNVs (**S1 Fig**), which we analysed independently. We only considered the variants present in at least 2 individuals in a dataset, filtering out the variants mapped in sex chromosomes and the private variants within each map.

The total number of autosomal protein coding genes overlapping with common CNVs varies depending on the filtered map, ranging from 1,694 (according to Handsaker's map [34]) to 5,610 (Sudmant-Nature's map [36]), with a total of 7,267 genes (out of 19,430 autosomal protein-coding genes) affected by CNVs when aggregating all 5 maps. Remarkably, only 402 (5.5%) of all genes affected by CNVs coincide in the 5 maps (**S2 Fig**). However, this overlap is non-random (P < 2.2e-16).

Most of the CNVs overlapping with genes fall within intronic regions (~63% of all CNVs) without any overlap with exons. More surprisingly, of the purely intronic CNVs detected, over the 94% are losses or gain/loss CNVs. This is in stark contrast with whole gene CNVs, which tend to be exclusively gains (55% of the cases) or gain/loss CNVs (25% of the cases) (**Fig 1B-F**). 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, in protein-coding genes, there were 13-70% fewer CNV deletions overlapping with exons than would be expected by chance, depending on the map (*P* < 0.05 in all maps).

Given the potential regulatory role of introns and the high frequency of purely intronic deletions in the population, we focused on the impact of CNVs on introns. For all the subsequent analyses we restricted our set of CNVs to loss and gain/loss CNVs, as they together represent sequences that are lost in some individuals. It is important to note that three maps (Sudmant-Nature's [36], Zarrei's [31] and Abyzov's [33]) together represent the 86% of all intronic deletions from our datasets. The methods used to generate the other two maps (Handsaker's [34] and Sudmant-Science's [35]) tend to detect less losses and larger CNV regions, that result in maps with fewer purely intronic deletions (**S1 Fig**).

For the above reasons, we focused our analyses of intronic deletions on the three maps with more intronic deletions (Sudmant-Nature's [36], Zarrei's [31] and Abyzov's [33]). We checked that this enrichment of intronic losses was still significant when controlling for different intron sizes (**S3**

Fig) and DNA replication timing (*P* < 5e-04; permutation testing).

Finally, we tested whether intronic losses were distributed equally between essential and non-essential genes. We separated the protein-coding genes into two groups: those that have been reported to be essential after CRISPR-based genomic targeting [37,38] or gene-trap insertional mutagenesis methodology [39], and those which were not found to be essential. Strikingly, we observed that the proportion of intronic deletions is higher than expected by chance in both essential and non-essential genes **(S2 Table)**. The fact that these intronic deletions can appear in essential genes suggests that they might be an unexpected source of genetic variation that could potentially influence the regulation of functionally relevant genes in human populations.

Intronic losses accumulate in evolutionarily old genes, while losses in coding regions are more frequent in young genes

Intrigued by the overrepresentation of intronic deletions in human protein-coding genes, we next investigated in more detail the quantitative and functional impact of these deletions. We have previously reported that the evolutionarily younger a gene is, the more likely it is to carry whole gene CNVs in the population [40]. Here we confirm that most ancient genes are depleted of CNVs that affect their coding regions, while primate-specific genes are enriched in CNVs (**S4 Fig**). This pattern was also observed when CNV gains were excluded (**Fig 2A**). 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) (**Fig 2A**).

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 (**Fig 2B**). This finding was confirmed with additional analyses considering only genes with introns and adjusting by the different size distributions of introns (**S5 Fig**). The introns of essential genes tend to be shorter [21,22] and essential genes also tend to be ancient [41]. Therefore, we compared the intron size of genes within the same age groups and found that the introns of essential genes are shorter than those of non-essential genes of the same age (**S6A and S6B Fig**).

Even if introns are shorter in essential genes, we found a significant proportion of them present intronic deletions in the Sudmant-Nature's map [36] (**S6C Fig**) suggesting that intronic size variation in the population might be more important than we originally thought.

Intronic deletions result in population variation of gene lengths

The percentage of each intron that can be lost due to CNV losses is highly variable, from 0.03% to 96.8%, representing a loss of the 0.01% to 77.5% of the total genic size. (**Fig 3A-C**). Some examples of genes with a notable change in size after a single intronic deletion are the neuronal glutamate transporter SLC1A1, with a loss of the 37% of its genic size, TCTN3 (tectonic family member 3), which loses 45% of its gene size 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 (**Fig 3D**). 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 [42]. This result shows that genes with a very conserved coding sequence with a general depletion of deletions [36] can have important losses of intronic regions, which might affect their regulation without affecting their protein structure.

Intronic deletions can impact regulatory regions such as enhancers or CTCF binding sites, which are enriched (P < 1e-04) and impoverished (P < 1e-04) in introns, respectively. Indeed, we find very frequently deletions in introns with these regulatory features (P < 2.2e-16). However, the direct overlap of the deletions with both enhancers and CTCF binding sites is significantly lower than expected by chance (P < 1e-04, **S3 Table**). This suggests intronic losses can occur close to regulatory features within introns but that deleting part of a regulatory feature might often have a deleterious impact.

Besides altering the size of the introns or disrupting regulatory regions, intronic deletions could also affect splicing, which is required to avoid the translation of introns. In genes with long introns, the recognition of introns and exons by splicing machinery is based on their differential GC content (Amit et al. 2012, Gelfman et al 2013) as the lower GC content in introns facilitates their recognition. Presumably, this recognition mechanism has contributed to the expansion of introns in higher eukaryotes (Hollander et al 2016). We analyzed the GC content of introns with deletions and found that the deleted sequences had a significantly higher GC content to that of the introns where they are located (*P* = 1.8e-28). Moreover, we observed that the loss of these fragments decreased significantly the overall GC content of the remaining introns (*P* = 2.23e-16). Our results suggest that the deletion of GC rich regions within introns could lower the overall GC content of the intron, increasing the difference of GC content between introns and their flanking exons, what could facilitate exon definition during splicing (**S8** and **S9 Fig**).

We have shown that deletions within introns are widespread in introns of varying sizes and can produce important changes on the sequence composition and the regulatory architecture of many protein-coding genes, which might be relevant for transcription and splicing of those genes through different mechanisms. Intronic deletions constitute a previously unexpected source of variation in gene and transcript length across individuals and can subtly affect ancient genes with important functions that don’t tolerate more drastic alterations.

Effect of intronic deletions on gene expression

Multiallelic CNVs affecting whole genes have been shown to correlate with gene expression: generally, the higher the number of copies of the gene, the higher its expression levels [34,36]. Our data suggests that the intronic size variation could also impact the expression of the affected genes. 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. We used available RNA-seq data from Geuvadis [43] that was derived from lymphoblastoid cell lines for 445 individuals for whom we have the matching CNV data (Sudmant Nature’s map [36]). 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 (**Fig 4A** and **S10 Fig**).

We first studied the effect of intronic deletions on gene expression and we observed significant differences in gene expression in 52 out of the 1,474 genes with intronic deletions (3.5%) in lymphoblastoid cell lines. This percentage is higher than expected by chance (P = 1e-4) **(Fig 4)**, being the expected values the total of differentially expressed genes (DEGs) when randomizing the individuals carrying the mutation. Of the DEGs, 62% were downregulated and the other 38% upregulated, suggesting that intronic deletions might result both in enhancing or repressing gene expression.

We investigated if deletions in introns of genes showing differential expression tended to overlap with regulatory features, but we did not observe any significant enrichment (P = 1). Even though first introns are known to be particularly important for gene regulation [3,44], there was no significant enrichment of DEGs with their first intron affected (P = 0.86). These results suggest that other mechanisms independent of intronic regulatory regions might be responsible for these changes in gene expression. It is also possible that a combination of multiple different mechanisms may be necessary to explain the observed effects. In addition, we cannot rule out that the lack of association between intronic regulatory features and gene expression changes is due to the small number of DEGs in this cell type and/or lack of detailed enough epigenomic annotations.

We wondered how the impact of non-coding intronic deletions in gene expression compared to those of non-coding intergenic deletions. We focused on intergenic regions that show long-range interactions with promoters of protein-coding genes - what it is generally assumed to reflect a regulatory function for these intergenic regions [45]. The impact of noncoding intronic versus intergenic deletions on gene expression was therefore studied. We used promoter-capture Hi-C published data for B-lymphocytes [46] to link deletions in intergenic regions with interacting genes. Significant changes in gene expression were seen in 11 out of 872 (1.26%) genes identified to have a deletion in an intergenic contacting region. Contrary to the effect of intronic deletions within the same gene, this percentage of DEGs was not different to that expected by chance (P = 0.08). Therefore, our data suggests that variation within intronic regions may have a more significant impact on gene regulation than intergenic regions.

The effect on gene expression appears to be greater when coding regions were affected, compared to purely intronic sequence losses: 15 out of 51 (29.4%, P < 1e-4) whole gene deletion CNVs resulted in significant downregulation of gene expression and 30 out of 239 genes with partial exonic deletions that were differentially expressed (12.6%, 28 down- and 2 up-regulated (P < 1e-4). However, given the higher frequency of intronic deletions in the population, the absolute number of DEGs with intronic deletions (52 genes) was similar to the total of DEGs with coding deletions (45 genes, **Fig 4B** and **S4 Table**). Moreover, while coding losses mostly associate to gene down-regulation, intronic losses are frequently associated to gene up-regulation. This shows the potential global relevance of intronic deletions on gene expression, especially considering their frequency in ancient genes (27.9 % in genes older than Sarcopterygii) is almost the double than the one for coding deletions (14.6%, **Fig 2**). In summary, these data suggest that intronic variants could have an important regulatory impact on ancient genes.

Discussion

Several studies have explored copy number variation in healthy humans [31,33–36,47–50], many of which have reported common CNVs overlapping with protein coding regions less frequently than expected by chance [29,31,35,36,47]. However, little attention has been paid to introns and, to the best of our knowledge, no previous study has dissected the differential impact of CNVs on exonic and intronic regions. Taking five recently published CNV maps [31,33–36], we observed an enrichment of deletions in introns resulting in gene length differences among individuals.

The different CNV maps used were built using different datasets and CNV calling algorithms, resulting in very different numbers of CNV sizes and types. Still, we think that each of these studies has their own limitations and probably none of them actually reflects all the variability in the genome. Therefore, instead of merging them into one map, we preferred to analyse the maps in parallel. This allowed us to compare the consistency of the results and, at the same time, helped us to better understand the peculiarities of each CNV set. We saw very consistent trends when we analysed the enrichment of intronic deletions or the differential impact of CNVs depending of the evolutionary age of genes, what show the robustness and generality of our results. At the same time our data suggest that using the different maps in parallel can be a useful way to cross-validate biological findings.

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. We have previously shown that genes of different ages are found in different proportions within current human CNV regions [51].

Whole young gene loci, contrarily to ancient gene loci, are very variable in copy number and tend to be located in late replicating genomic regions, which are more error-prone and have less precise DNA repair mechanisms than earlier regions [51]. Fixation of duplications or losses of whole genes in these regions can lead to the birth of new genes or to their disappearance **(Fig 5)**.

Here, we have observed that also gains and losses affecting only part of the coding sequence are also enriched in young genes (**Fig 5**). Such CNVs can disrupt the protein sequence, but they can also eliminate, duplicate or relocate exons or parts of exons, giving to the organism a mechanism to modify young genes.

On the other hand, evolutionarily ancient genes, generally depleted of CNVs overlapping with their coding regions, are especially enriched with intronic losses (**Fig 5**). This phenomenon shows that although the protein sequence is usually unaffected, changes in the intronic sequence can modulate the expression of the gene and promote variability in the population. We found in lymphoblastoid cells more differentially expressed genes associated with intronic losses than expected by chance. This association is expected to be even stronger as for many genes the effect of their intronic losses will be only observed in other cell types or tissues.

Very interestingly, we observed differences in which genes show changes in gene expression when affected by coding (exonic or whole gene) or purely intronic losses. We see that differences in expression in younger genes are mainly associated to full gene dosage changes or partial disruption of their coding sequence. In contrast, ancient genes that generally are less tolerant to any kind of mutations in their coding sequence, are enriched in intronic deletions which that could be modulating their expression (**Fig 5**). The availability of CNV and population-based gene expression data from several tissues will allow to evaluate more accurately what is the impact of coding and non-coding deletions in the whole organism.

CNVs can be directly disrupting a regulatory feature or affect the distance, for example, between promoter and enhancer. We found that the presence of enhancers is significantly enriched in introns, agreeing with previous findings in plants [11,44]. In general, genes with complex regulation patterns require more regulatory DNA [52] and introns tend to be longer in tissue-specific and transcription factor genes compared to housekeeping genes [21,53]. Since many enhancers are tissue-specific [54] intronic CNVs might frequently have effects on particular cell types. Therefore, the loss of intronic sequence might be affecting the expression of such genes in a tissue-specific manner.

Our results also suggest that non-coding intronic deletions might have a wider impact on population gene expression variability than deletions in non-coding intergenic regions that interact with promoters, given that intronic deletions correlate with gene expression changes more often than expected by chance, while promoter-interacting intergenic regions don’t. However, intergenic deletions were associated with genes using promoter-capture HiC data maps derived from a few pooled genomes [46] and we may need to have personal genome interactomes, more tissues and conditions to evaluate more precisely the effect of intergenic deletions on gene expression. Furthermore, with the necessary experimental and analytical future advances, it will be extremely exciting to see how individual copy number variants change the personal landscape of interactions among promoters and other genomic elements.

We speculate that intronic CNVs might have a previously unsuspected role in shaping gene expression variability in populations with potential important consequences in human evolution and adaptation. After uncovering the relevance of gene length variation in the healthy population by frequent intronic deletions, the next open question will be if any of these common non-coding variants may be associated with disease. In fact, these population-based CNV maps could be useful to identify disease relevant and irrelevant intronic regions. It is now well known that most genome-wide association studies (GWAS) associated SNPs tend to be located in intronic and intergenic regions and the pathogenicity of non-coding CNVs, mostly in upstream promoters, is starting to emerge [55]. Thus, future case-control studies including WGS should also pay attention at potentially important role of purely intronic variation. While exons cover around the 2.8% of the genome, introns cover 35.3%, of the genome (based on the gene set used for this study). WGS studies are starting to focus on distal intergenic enhancers, but intronic regions are commonly ignored in the analyses. 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 [27]. 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.

Being intronic deletions so common in the healthy population, it will also be interesting to explore how frequent are purely intronic somatic deletions in cancer and evaluate their potential contribution to the reprogramming of gene regulation of cancer cells. For example, are there somatic deletions of intronic sequences that result in the shortening of oncogenes, favoring their higher expression. The role of cancer somatic variants in distal regulatory regions is just starting to be explored [56–60]. As we have shown that intronic regions are significantly enriched in regulatory regions in the human genome, understanding the functional effect somatic intronic deletions in cancer could be an attractive new field of research with high potential for discovery. It has been previously proposed that high-order chromatin architecture is influencing the landscape of chromosomal alterations in cancer [61]. We hypothesize that the high-order genome organisation in healthy cells is applying constraints on where variability can be high or low, allowing high variability anywhere in young genes but only in introns for ancient genes. Therefore, it will be interesting to understand better how these constraints change comparing data from healthy cells with the frequent aneuploidy in tumors, especially in radical re-structuring events originated by chromothripsis [62] .

In summary, our data shows that intronic CNVs constitute the most abundant form of CNV in protein-coding genes. These intronic length variation possibly means that the actual size of many genes is not yet fixed in human populations. We show that intronic length variation is particularly frequent in evolutionary old genes, with a significant proportion of them showing associated gene expression changes. This suggests that intronic CNVs might be actively contributing to the evolution of gene regulation in many genes with highly conserved protein sequences. Taken together, 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 fine-tuning the regulation of old genes.

Materials and methods

Origin and filtering of CNV maps

Whole genome CNV maps were downloaded from 5 different publications [31,33–36]. 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 **S5 Table**.

Gene structures

Autosomal gene structures and sequences were retrieved from Ensembl [63] (http://www.ensembl.org; version 75) and principal isoforms were determined according to the APPRIS database [64], 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.

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 [65,66] 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 [67]. 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 [43] 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 [36]. 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 [68] in two groups depending if the individual carried or not the intronic deletion and performed Student's t-tests. 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

The authors thank Salvador Capella, Matthew Bashton, Venetia Bigley, Joris Veltman, 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.

References

1. Lynch M. The Origins of Genome Architecture. Sinauer Associates Incorporated; 2007.

2. Chorev M, Carmel L. The function of introns. Front Genet. 2012;3: 55.

3. Jo B-S, Choi SS. Introns: The Functional Benefits of Introns in Genomes. Genomics Inform. 2015;13: 112.

4. Gilbert W. The Exon Theory of Genes. Cold Spring Harb Symp Quant Biol. 1987;52: 901–905.

5. Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. 2008;18: 88–103.

6. Gelfman S, Burstein D, Penn O, Savchenko A, Amit M, Schwartz S, et al. Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. Genome Res. 2012;22: 35–50.

7. Kirkconnell KS, Magnuson B, Paulsen MT, Lu B, Bedi K, Ljungman M. Gene length as a biological timer to establish temporal transcriptional regulation. Cell Cycle. 2017;16: 259–270.

8. Le Hir H, Nott A, Moore MJ. How introns influence and enhance eukaryotic gene expression. Trends Biochem Sci. 2003;28: 215–220.

9. Nott A, Meislin SH, Moore MJ. A quantitative analysis of intron effects on mammalian gene

expression. RNA. 2003;9: 607–617.

10. Gaffney DJ, Keightley PD. Genomic selective constraints in murid noncoding DNA. PLoS Genet. 2006;2: e204.

11. Rose AB. Intron-mediated regulation of gene expression. Curr Top Microbiol Immunol. 2008;326: 277–290.

12. Le Hir H, Nott A, Moore MJ. How introns influence and enhance eukaryotic gene expression. Trends Biochem Sci. 2003;28: 215–220.

13. Seoighe C, Korir PK. Evidence for intron length conservation in a set of mammalian genes associated with embryonic development. BMC Bioinformatics. 2011;12 Suppl 9: S16.

14. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. 2013;29: 569–574.

15. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. Selection for short introns in highly expressed genes. Nat Genet. 2002;31: 415–418.

16. Heyn P, Kircher M, Dahl A, Kelso J, Tomancak P, Kalinka AT, et al. The earliest transcribed zygotic genes are short, newly evolved, and different across species. Cell Rep. 2014;6: 285–292.

17. Artieri CG, Fraser HB. Transcript length mediates developmental timing of gene expression across Drosophila. Mol Biol Evol. 2014;31: 2879–2889.

18. Swinburne IA, Silver PA. Intron delays and transcriptional timing during development. Dev Cell. 2008;14: 324–330.

19. Jeffares DC, Penkett CJ, Bähler J. Rapidly regulated genes are intron poor. Trends Genet. 2008;24: 375–378.

20. Keane PA, Seoighe C. Intron Length Coevolution across Mammalian Genomes. Mol Biol Evol. 2016;33: 2682–2691.

21. Eisenberg E, Levanon EY. Human housekeeping genes are compact. Trends Genet. 2003;19: 362–365.

22. Vinogradov AE. Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet. 2004;20: 248–253.

23. Carmel L, Koonin EV. A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. Genome Biol Evol. 2009;1: 382–390.

24. Agrawal A, Hamvas A, Cole FS, Wambach JA, Wegner D, Coghill C, et al. An intronic ABCA3 mutation that is responsible for respiratory disease. Pediatr Res. 2012;71: 633–637.

25. Lo Y-F, Nozu K, Iijima K, Morishita T, Huang C-C, Yang S-S, et al. Recurrent deep intronic mutations in the SLC12A3 gene responsible for Gitelman's syndrome. Clin J Am Soc Nephrol. 2011;6: 630–639.

26. Nurnberg ST, Zhang H, Hand NJ, Bauer RC, Saleheen D, Reilly MP, et al. From Loci to Biology: Functional Genomics of Genome-Wide Association for Coronary Disease. Circ Res.

NIH Public Access; 2016;118: 586.

27. Vaz-Drago R, Custódio N, Carmo-Fonseca M. Deep intronic mutations and human disease. Hum Genet. 2017; doi:10.1007/s00439-017-1809-4

28. Arlt MF, Wilson TE, Glover TW. Replication stress and mechanisms of CNV formation. Curr Opin Genet Dev. 2012;22: 204–210.

29. Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, Lee C, et al. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nat Genet. 2010;42: 385–391.

30. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 2010;11: R52.

31. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16: 172–183.

32. Pirooznia M, Goes FS, Zandi PP. Whole-genome CNV analysis: advances in computational approaches. Front Genet. 2015;6: 138.

33. Abyzov A, Li S, Kim DR, Mohiyuddin M, Stütz AM, Parrish NF, et al. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. Nat Commun. 2015;6: 7256.

34. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large multiallelic copy number variations in humans. Nat Genet. 2015;47: 296–303.

35. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015;349: aab3761.

36. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526: 75–81.

37. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, et al. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell. 2015;163: 1515–1526.

38. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and characterization of essential genes in the human genome. Science. 2015;350: 1096–1101.

39. Blomen VA, Májek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, et al. Gene essentiality and synthetic lethality in haploid human cells. Science. 2015;350: 1092–1096.

40. Juan D, Rico D, Marques-Bonet T, Fernández-Capetillo O, Valencia A. Late-replicating CNVs as a source of new genes. Biol Open. 2014;3. doi:10.1242/bio.20147815

41. Chen W-H, Trachana K, Lercher MJ, 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. 2012;29: 1703–1706.

42. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013;9: e1003709.

43. Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501: 506–511.

44. Majewski J, Ott J. Distribution and characterization of regulatory elements in the human genome. Genome Res. 2002;12: 1827–1836.

45. Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre B-M, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. 2015;25: 582–597.

46. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016;167: 1369–1384.e19.

47. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007;318: 420–426.

48. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. Nature Publishing Group; 2008;40: 1166–1174.

49. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464: 704–712.

50. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of human copy number variation and multicopy genes. Science. 2010;330: 641–646.

51. Juan D, Rico D, Marques-Bonet T, Fernandez-Capetillo O, Valencia A. Late-replicating CNVs as a source of new genes. Biol Open. 2013;3: 231–231.

52. Gaffney DJ, Keightley PD. Genomic selective constraints in murid noncoding DNA. PLoS Genet. 2006;2: e204.

53. Kirkconnell KS, Magnuson B, Paulsen MT, Lu B, Bedi K, Ljungman M. Gene length as a biological timer to establish temporal transcriptional regulation. Cell Cycle. 2017;16: 259–270.

54. Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I. Predicting tissue-specific enhancers in the human genome. Genome Res. 2007;17: 201–211.

55. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24: R102–10.

56. Northcott PA, Lee C, Zichner T, Stütz AM, Erkek S, Kawauchi D, et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature. 2014;511: 428–434.

57. Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature. 2015;526: 519–524.

58. Zhang X, Choi PS, Francis JM, Imielinski M, Watanabe H, Cherniack AD, et al. Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. Nat Genet. 2016;48: 176–182.

59. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M. Role of non-coding sequence variants in cancer. Nat Rev Genet. 2016;17: 93–108.

60. Weischenfeldt J, Dubash T, Drainas AP, Mardin BR, Chen Y, Stütz AM, et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat Genet. 2017;49: 65–74.

61. Fudenberg G, Getz G, Meyerson M, Mirny LA. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. Nat Biotechnol. 2011;29: 1109–1113.

62. Zhang C-Z, Leibowitz ML, Pellman D. Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. Genes Dev. 2013;27: 2513–2530.

63. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016;44: D710–6.

64. Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink J-J, Lopez G, et al. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res. 2013;41: D110–7.

65. Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA. Identification of higher-order functional domains in the human ENCODE regions. Genome Res. 2007;17: 917–927.

66. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci U S A. 2010;107: 139–144.

67. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build.

Genome Biol. 2015;16: 56.

68. 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. 2010;6: e1000770.

Figure captions

Fig 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. (E-F) Venn diagrams showing the number of genes with whole gene, exonic and intronic losses (E) or gains (F) in the three maps with the higher number of deletions reported (Zarrei, Sudmant-Nature and Abyzov). Circle sizes are proportional to the number of genes affected.

Fig 2. 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 S7 Fig.

Fig 3. 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.

Fig 4. Differential expression. (A) Number of genes with whole gene, exonic or intronic deletions or with intergenic deletions in a region in long-range contact with it. (B) Differentially expressed genes (DEGs) for each type of deletion. The colored bars show the observed number of DEGs per group. The white bars represent the median random number of DEGs when randomizing 10,000 times the individuals with and without the deletion. Significant differences ($P < 0.05$) are marked with * ($P < 0.0001$ in all cases) and error bars show median absolute deviation.

Fig 5. Impact of CNVs on genes and their evolution. 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 are sometimes 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.

Supporting information

S1 Fig. Comparison of datasets. Only variants in autosomes are considered and private events are excluded. (A) Number and type of CNVs per dataset. (B) Autosomal Mb that are CNV. Gray part of the bars corresponds to the CNV Mbs that are shared among maps. Colored parts of the bars are map-specific CNV regions. (C) Width distribution of gains and losses in each map bean lines and overall line are means). (D) Number of subjects and number of populations of origin used for building of each filtered map.

S2 Fig. Overlap of genes affected among datasets. Number of genes overlapping with CNVs (A) or found completely within a CNV (B) by number of maps in which the overlap is observed.

S3 Fig. Enrichment of deletions in introns of different sizes. Number of deletions in each size bin. The light blue line represents the expected value, calculated as the mean of the affected introns in the 10,000 random permutations. Red asterisks mark the bins significantly enriched with intronic deletions (P < 0.05).

S4 Fig: Differential effect of CNVs on protein coding genes of different evolutionary ages. The proportion of genes with CNVs (gains, losses and gain and loss CNVs) affecting their coding region in each gene age class is shwon. Red stars show a significantly higher percentage of genes with coding compared to 10,000 randomizations of the CNVs over the genome (P < 0.05). Black stars show in which age groups there is a depletion of coding CNVs (P < 0.05).

S5 Fig. Differential effect of intronic deletions in introns and protein coding genes of different evolutive ages. Figure (A) shows the percentage of introns over 1.5kb that contain intronic deletions. Introns shared among genes of different gene ages were assigned the youngest age. Figure (B) represents the percentage of genes with introns over 1.5kb that have intronic deletions. In all plots, red stars show a significantly higher percentage of introns or genes with intronic deletions compared to 10,000 randomizations of the deletions over the genome (P < 0.05). Black stars show in which age groups there is a depletion of intronic deletions (P < 0.05).

S6 Fig. Essential genes. (A) Percentage of essential genes per evolutionary age. (B) Intron sizes of non-essential and essential genes. (C) Percentage of non-essential (solid bars) and essential (empty bars) genes of different evolutionary ages that have intronic deletions. Red stars show a significantly higher percentage of genes with intronic deletions compared to 10,000 randomizations of the deletions over the genome (P < 0.05).

S7 Fig. Effect of the different types of deletions on all evolutionary ages. Proportion of genes with deletions that have the whole locus deleted, only part of their exons (exonic) affected by deletions or intronic deletions only. (Equivalent to Fig. 2C, but here separated by CNV map.)

S8 Fig. GC content in introns and intronic deletions. (A) Bean-plots showing the different GC distribution between the flanking exons of introns with or without deletions, separated by intron size bins (with equal number of introns per bin). (B) GC content distributions in introns with or without deletions, separated by intron size bins. Significance is considered for p-values < 0.05. Beans show the estimated density of each distribution; horizontal lines show the mean values of each side of the bean and the dashed horizontal line line represents the overall average of all values.

S9 Fig. Examples of introns with a drop of GC content. X-axis represents the coordinates of the intron with its flanking exons (black boxes). Y-axis shows the GC content, calculated with sliding 200 bp windows. The deleted region is highlighted in grey.

S10 Fig. Overlap of genes with different types of deletions. Venn diagram showing the overlaps for genes carrying different types of mutations (A) and for differentially expressed genes (DEGs) (B).

S1 Table. Number of individuals in each map, project the variants belong to and methods used for CNV detection.

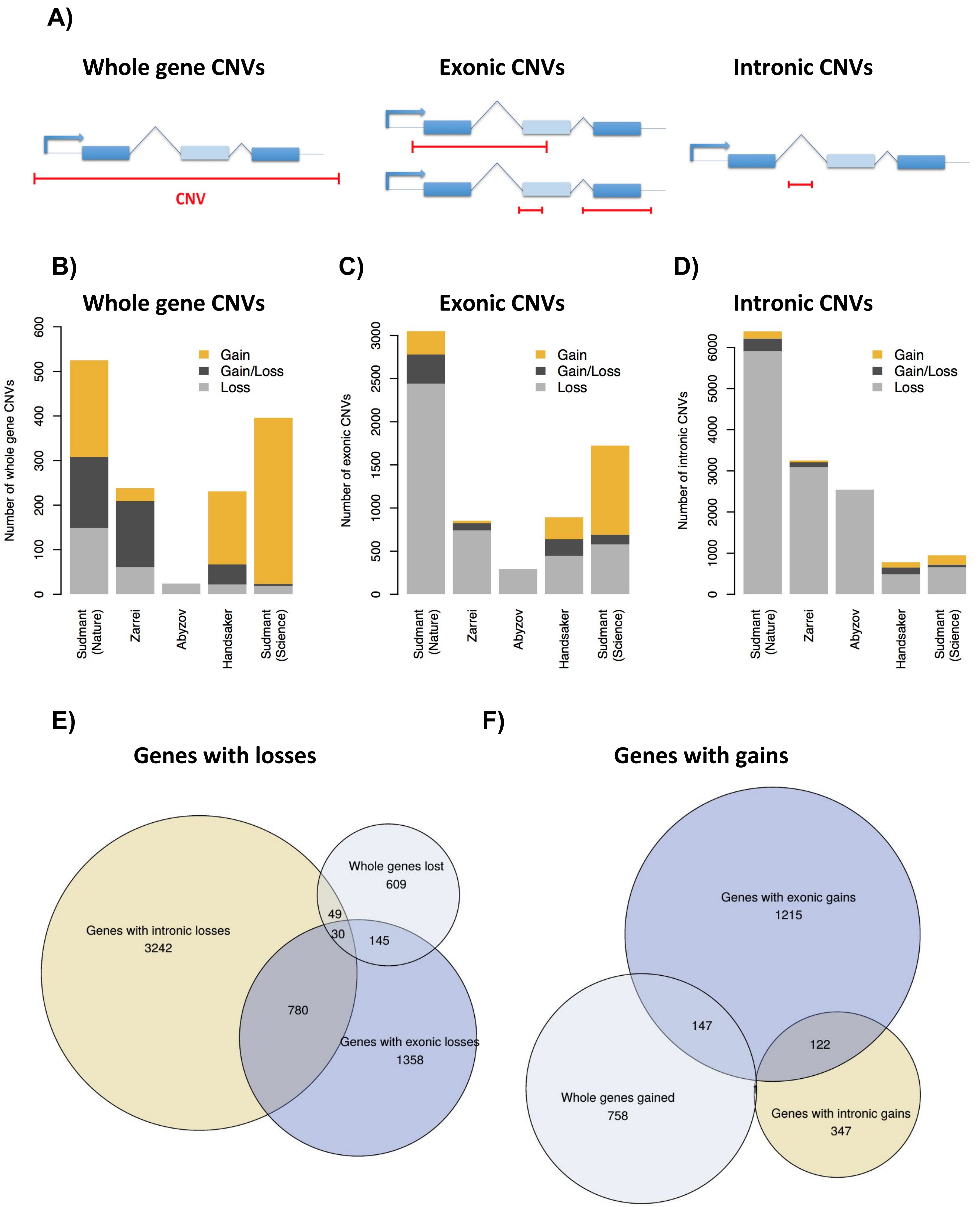
S2 Table. Enrichment of intronic deletions in non-essential and essential genes.

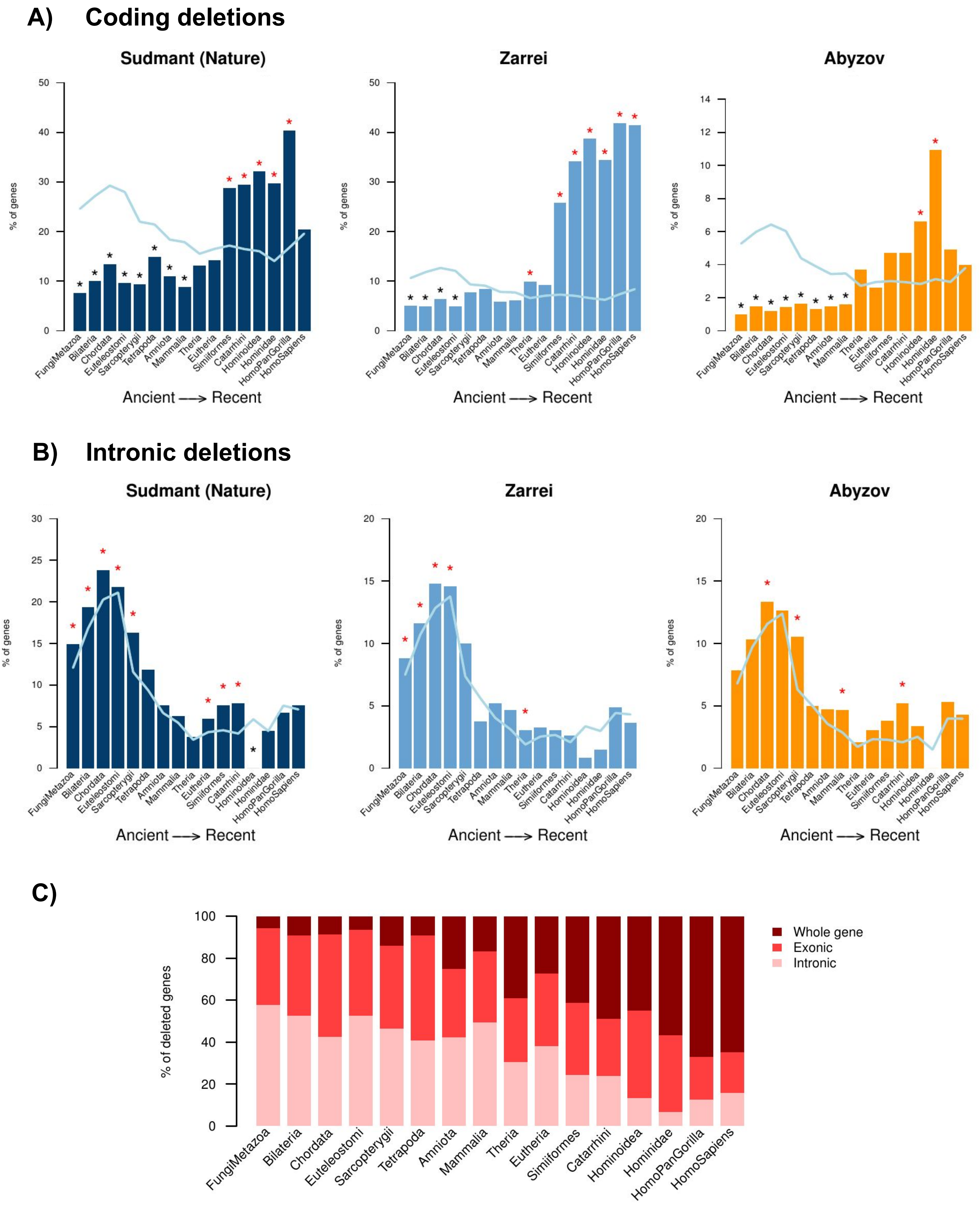
S3 Table. Overlap of intronic deletions with regulatory features.

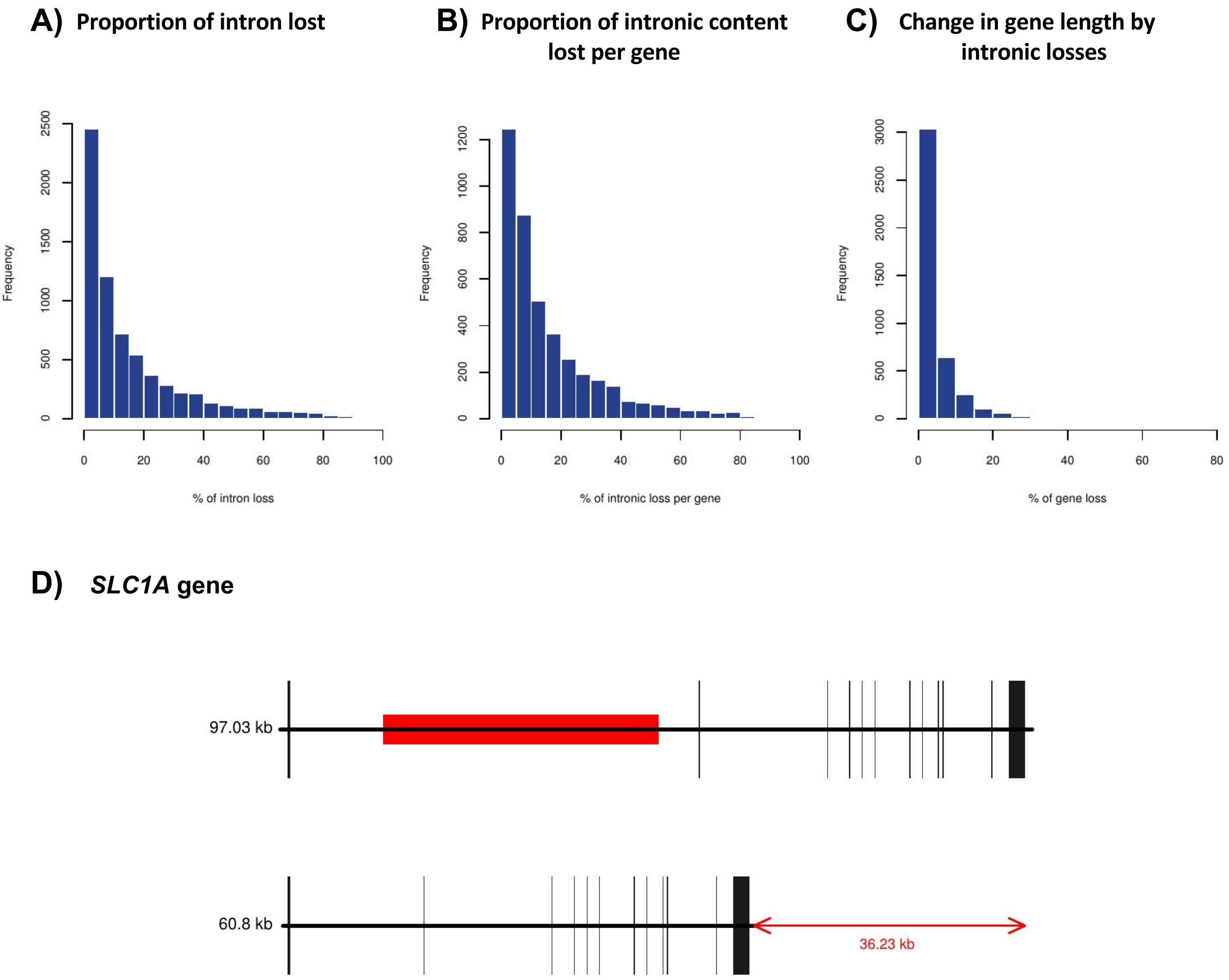
S4 Table. Differentially expressed genes.

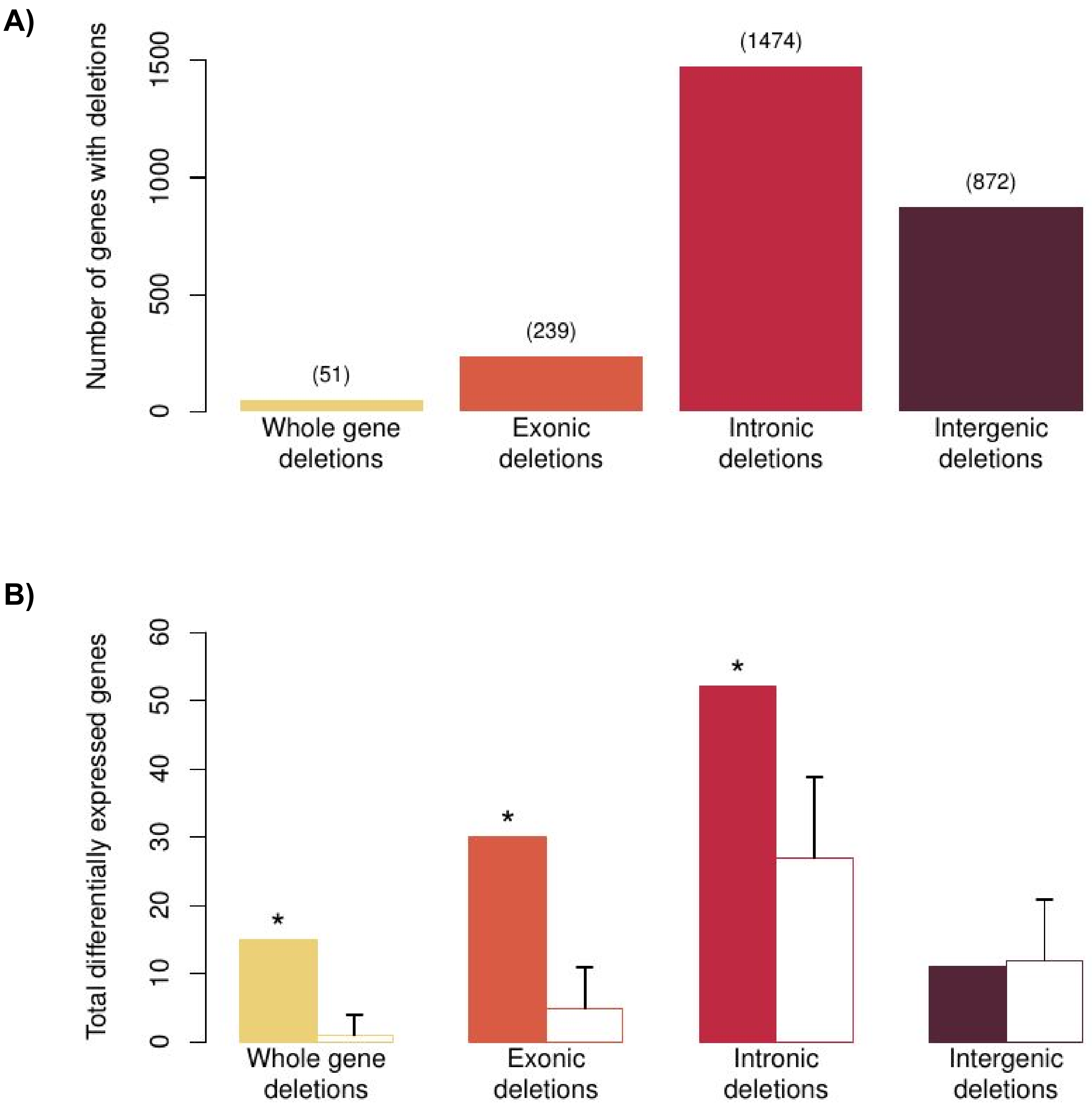
S5 Table. Filtered CNV maps used in this study.

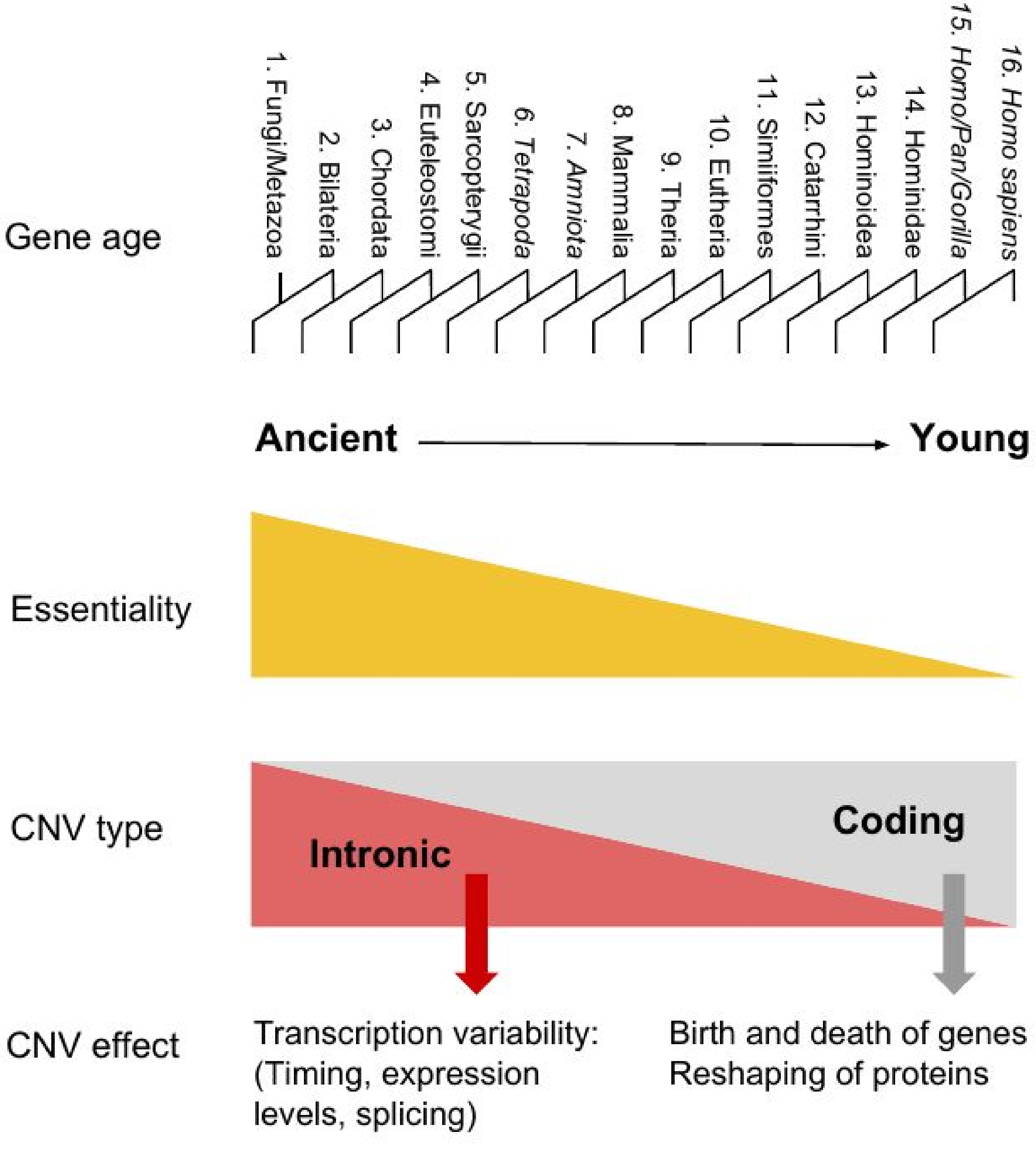
S6 Table. Genes affected by CNVs, including their ages.


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)
E)
Genes with losses
Whole genes lost
609
49
30
145
Genes with intronic losses
3242
780
Genes with exonic losses
1358
F)
Genes with gains
Genes with exonic gains
1215
147
122
1
Whole genes gained
758
Genes with intronic gains
347

A) Coding deletions
Sudmant (Nature)
Zarrei
Abyzov
% of genes
0
10
20
30
40
50
2
4
6
8
12
14
FungiMetazoa
Bilateria
Chordata
Euteleostomi
Sarcopterygii
Tetrapoda
Amniota
Mammalia
Theria
Eutheria
Simiiformes
Catarrhini
Hominoidea
Hominidae
HomoPanGorilla
HomoSapiens
Ancient ⟶ Recent
*
B) Intronic deletions
Sudmant (Nature)
Zarrei
Abyzov
% of genes
0
5
10
15
20
25
30
Ancient ⟶ Recent
*
C)
% of deleted genes
0
20
40
60
80
100
Whole gene
Exonic
Intronic


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


A)
Number of genes with deletions
0
500
1000
1500
(51)
(239)
(1474)
(872)
Whole gene deletions
Exonic deletions
Intronic deletions
Intergenic deletions
B)
Total differentially expressed genes
0
10
20
30
40
50
60
*
*
*
Whole gene deletions
Exonic deletions
Intronic deletions
Intergenic deletions


Gene age
1. Fungi/Metazoa
2. Bilateria
3. Chordata
4. Euteleostomi
5. Sarcopterygii
6. *Tetrapoda*
7. *Amniota*
8. Mammalia
9. Theria
10. Eutheria
11. Simiiformes
12. Catarrhini
13. Hominoidea
14. Hominidae
15. *Homo/Pan/Gorilla*
16. *Homo sapiens*
Ancient
Young
Essentiality
CNV type
Coding
Intronic
CNV effect
Transcription variability:
(Timing, expression levels, splicing)
Birth and death of genes
Reshaping of proteins