UNIVERSIDAD AUTÓNOMA DE MADRID

Programa de Doctorado

en Biociencias Moleculares

Analysis of the Impact of Population Copy Number Variation in Introns

Maria Rigau de Llobet

Madrid, 2019

DEPARTMENT OF BIOCHEMISTRY

UNIVERSIDAD AUTÓNOMA DE MADRID

FACULTY OF MEDICINE

UAM
UNIVERSIDAD AUTONOMA
DE MADRID

Analysis of the Impact of Population Copy Number Variation in Introns

Maria Rigau de Llobet

Graduate in Biology

Thesis Directors:
Dr. Alfonso Valencia Herrera and
Dr. Daniel Rico Rodríguez

BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Barcelona Supercomputing Center

Certificado del director o directores

Dr Alfonso Valencia Herrera, Head of the Computational Biology Life Sciences Group (Barcelona Supercomputing Center) and **Dr Daniel Rico Rodríguez**, Head of the Chromatin, Immunity and Bioinformatics Group (Institute of Cellular Medicine, Newcastle, UK)

CERTIFY:

That Ms. **Maria Rigau de Llobet**, Graduate in Biology by the University Pompeu Fabra (Barcelona, Spain), has completed her Doctoral Thesis entitled "**Analysis of the Impact of Population Copy Number Variation in Introns"** under our supervision and meets the necessary requirements to obtain the PhD degree in Molecular Biosciences. To this purpose, she will defend her doctoral thesis at the Universidad Autónoma de Madrid. We hereby authorize its defense in front of the appropriate Thesis evaluation panel.

We issue this certificate in Madrid on April 30th 2019

Alfonso Valencia Herrera	Daniel Rico Rodríguez
Thesis Director	Thesis Director

This thesis was supported by a

“La Caixa” – Severo Ochoa International PhD Programme

Obra Social
Fundación ”la Caixa”

Aknowledgments

Para empezar, me gustaría dar las gracias a mis dos directores de tesis, Alfonso Valencia y Daniel Rico.

Muchas gracias Alfonso por la oportunidad de hacer el doctorado en tu grupo. Puedo decir sin dudar que ha sido el comienzo de una carrera profesional con la que disfruto. Me siento muy afortunada por estos casi cinco años, una etapa de aprender con libertad y con apoyo, de trabajar a gusto y de estar rodeada de gente brillante, tanto en el ámbito laboral como personal. Gracias por crear y mantener un entorno de trabajo tan estimulante y por dejarme formar parte de él.

Dani, a ti te debo, para empezar, mi inesperada entrada al mundo de la bioinformática. Gracias por animarme a hacer el doctorado con vosotros, que sigo convencida de que fue la mejor decisión. Te estoy muy agradecida por todos tus consejos, por tu entusiasmo, por valorar siempre mis comentarios y opiniones y por tratarme como igual desde el principio. También te agradezco la manera cómo has afrontado y revertido con toda naturalidad mis malos momentos y que me hayas ayudado a mantenerme motivada.

A David el pelirrojo también le estoy eternamente agradecida. David, muchísimas gracias por tus lluvias torrenciales de ideas, por enseñarme a descifrar e interpretar resultados, por sacar tiempo de donde no lo había y, a ti también, gracias por el apoyo moral.

Quiero agradecer a Óscar, Fátima y Tomàs, mi comité de tesis, todas sus aportaciones.

De mi etapa en el CNIO, quiero dedicarle un agradecimiento especial a Marta, ya que formó parte del comité de reclutamiento y bienvenida al grupo, y porque sin su compañía en esos primeros desayunos y sin todos los descansos compartidos, mi etapa en el CNIO no habría tenido ni la mitad de las risas. Y a Ángel, a quien también le agradezco la acogida desde el principio y, además, el ayudarme a tener una relación medianamente cordial con mi ordenador. Y a Enrique, por las tardes de tertulia y cacahuetes. Y a Belén, por estar en todo y hacernos la vida más fácil.

Y no me olvido de todos los compañeros de la planta baja - ala norte del CNIO, de mis compañeras puntuales de oficina (Cata, Mónica, Mago e Irene) y de toda la gente de los *Supresores de Telómeros* con quien he compartido o bien muchas horas de trabajo o bien muchas de no-trabajo.

Mi etapa madrileña no habría sido lo mismo sin mis compañeras de piso. Muchas gracias Cristina, Mónica, Miriam, Nerea y Leire por hacer tan familiares, alegres y cómodos mis tres años en Sandoval. Y cómo no, este agradecimiento también aplica a mis vecinos, compañeros de vacaciones, anfitriones y grandes amigos Alejandra y Juan.

And then there are the Chicolas. I couldn't be happier to have shared most of the PhD with you. I miss you and I hope we can meet soon as Doctor Chicolas. For all these years, grazie, Fede; danke, Teresa; eskerrik asko, Leire; ευχαριστώ, Dafni.

Mi doctorado tiene también una etapa Barcelonesa compartida con otro gran equipo muy multidisciplinar: de trabajo, de yoga, de kebabs, de excursiones, de celebraciones, y de karaokes (con y sin lesión). Estos son: Miguel, Vera, Davide, Alba L, Alba J, Eva, Vicky, François, Arnau, Iker, Carlos, Hugo, Mónica, Laure, Eduard, Miguel V, Patricia, Mattia, Victor, los INBs y los mineros. Y cómo no, Juan y Jon, los mejores compañeros de doctorado posibles.

Juancho, mil gracias por estos años. Por el soporte informático y moral, por las discusiones científicas y las reflexiones filosóficas, por obsequiarnos con un peinado diferente cada día y por estar siempre dispuesto a bajar de tu mundo de máxima concentración para ayudar en lo que haga falta.

Y Jon, tu párrafo debería estar en mayúsculas. Muchas gracias por aguantarme del primer al último día, por tu paciencia y por estar allí siempre que lo he necesitado. Creo que nadie me discutirá que, sin ti, nuestro grupo no estaría ni tan unido, ni tan entretenido, ni tan cuidado. Mil gracias por todo.

Gràcies també a tots els del grup d'en David Torrents, perquè amb qui no he compartit CNVs he compartit menjars del pingüí, congressos, aires condicionats o màrfegues de ioga.

També vull donar les gràcies a la colla de la universitat (+2) per tots els moments compartits durant, ara ja, més d'una dècada. Perquè ens hem fet grans junts i mai hem perdut les ganes veure'ns.

I a les amigues de tota, o quasi tota la vida. Ester i Núria, no sé quina de les dues ha aguantat més "quan acabi la tesi", però espero que a partir d'ara poguem fer totes les sortides i viatges que tenim pendents. I Silvieta, merci per estar sempre disposada a àpats improvisats, sempre riallera contagiosa i per ser sempre una font de tranquil·litat.

Vull aprofitar l'ocasió per donar les gràcies a la meva gran família, especialment als meus pares, pel seu exemple i per haver-me donat suport en totes les meves decisions, i als meus altres tres quarts: la Gemma, l'Anna i l'Anton. Us estimo infinit.

I a en Carlos. Per aquests perfectes i inesperats últims mesos de doctorat.

Abstract/Resumen

Abstract

Introns cover most of the DNA sequence in human protein-coding genes and represent approximately half of the non-coding genome. Very little is known about the patterns of structural variation in introns and little attention has been paid to their functional implications, even if several pathogenic intronic mutations have already been characterized. Through the combined analysis of the five most extensive maps of Copy Number Variants (CNVs) in human populations we show that intronic losses are the most frequent type of CNV in protein-coding genes. The lower density of CNVs in introns compared to intergenic regions supports the presence negative selection on intronic CNVs.

We identified many intronic deletions associated with gene expression changes by integrating genotype with RNA-seq and promoter-capture Hi-C data, supporting the implication of many CNVs in genetic regulation. Remarkably, a noteworthy number of these associations are better interpreted by long-range genome interactions. Supporting the possible impact of intronic CNVs on splicing, we have found 185 genes differentially expressed transcripts associated with deletions. Moreover, we have found changes in exon inclusion associated with deletions that alter the GC content of the intron. This finding suggests that the structure of the fragments deleted in introns play a significant role on which exons are included in the mature messenger RNA. Altogether, our findings additionally support the substantial role of intronic CNVs on gene regulation.

Interestingly, we have observed that CNVs are not equally distributed among genes of different evolutionary ages. Ancient genes are, in general, depleted of losses covering their exons, but they carry the majority of intronic deletions, including intronic deletions associated with expression changes. On the other hand, recent primate-specific genes are enriched in CNVs implicating exons. Taken together, our findings suggest that CNVs have a role in shaping gene evolution, possibly acting at different levels at large and short evolutionary times (old and young genes). While in young genes CNVs contribute to directly alter protein sequences, in ancient genes CNVs seem to be preferentially contributing to population variability at the level of regulation with possible adaptive implications.

Resumen

Los intrones cubren la mayor parte de la secuencia de ADN en genes codificantes para proteínas y representan aproximadamente la mitad del genoma no codificante en humanos. Se sabe muy poco acerca de los patrones de variación estructural en los intrones y se ha prestado poca atención a sus implicaciones funcionales, incluso si ya se han caracterizado varias mutaciones intrónicas patógenicas. A través del análisis combinado de los cinco mapas más extensos de las Variantes en Número de Copia (CNVs) en poblaciones humanas, mostramos que las pérdidas intrónicas son el tipo más frecuente de CNV en los genes codificantes para proteínas. La menor densidad de CNVs en intrones en comparación con regiones intergénicas sugiere la presencia de selección negativa sobre las CNVs intrónicas.

Integrando datos de CNVs con datos de RNA-seq y PCHi-C hemos identificado deleciones intrónicas asociadas a cambios en la expresión génica. Parte de estas asociaciones se interpretan mejor por interacciones genómicas entre fragmentos distantes. Apoyando el posible papel de las CNVs intrónicas en el proceso de *splicing*, hemos encontrado 185 genes con tránscritos diferentialmente expresados en los individuos con deleciones. Además, hemos encontrado cambios en la inclusión de exones asociados a CNVs que alteran el contenido GC del intrón. Esto sugiere que la estructura de los fragmentos perdidos en los intrones desempeña un papel importante en la selección de exones en el *splicing*. En conjunto, nuestros hallazgos muestran el importante papel de las CNVs intrónicas en la regulación génica.

Curiosamente, hemos observado que las CNV no están distribuidas equitativamente entre los genes de diferentes edades evolutivas. Los genes antiguos están empobrecidos de pérdidas en sus exones pero tienen la mayoría de deleciones intrónicas, incluidas muchas de las asociadas a cambios de expresión. Por otro lado, los genes recientes están enriquecidos en CNVs exónicas. Nuestros hallazgos sugieren que las CNVs contribuyen a la evolución de los genes, posiblemente actuando a diferentes niveles en genes antiguos y jóvenes. Mientras que en los genes jóvenes las CNVs contribuyen a alterar directamente las secuencias de proteínas, en los antiguos, las CNVs parecen estar contribuyendo de manera preferencial a la variabilidad en la regulación, con posibles implicaciones adaptativas.

Table of contents

Abbreviations

1KGP	1000 Genomes Project
aCGH	Array Comparative Genomic Hybridization
AS	Alternative Splicing
CBS	CTCF Binding Site
CDS	Coding Sequence
CN	Copy Number
CNV	Copy Number Variant
CNVD	Copy Number Variation in Disease database
CNVR	Copy Number Variant Region
CTCF	CCCTC-binding factor
DE	Differentially expressed
DGV	Database of Genomic Variants
DSB	Double Strand Break
ENCODE	ENCyclopedia Of DNA Elements (ENCODE) project
GRC	Genome Reference Consortium
GWAS	Genome Wide Association Study
HGP	Human Genome Project
MAD	Median Absolute Deviation
mCNV	Multiallelic CNV
mRNA	Messenger RNA
NHP	Non-Human Primates
NMD	Nonsense-Mediated Decay
OMIM	Online Mendelian Inheritance in Man
PCHi-C	Promoter-Capture Hi-C
PPI	Protein-protein interaction
RF	Regulatory Feature

RD	Read-depth
RT	Replication Time
RVIS	Residual Variation Intolerance Score
SCNA	Somatic Copy Number Alteration
SD	Standard Deviation
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SR	Split-Read
SV	Structural Variant
TAD	Topologically Associating Domain
TFBS	Transcription Factor Binding Site
WGS	Whole Genome Sequencing

Introduction

Introduction

The evolution of the reference genome

The Human Genome Project (HGP) was initiated in 1990 with the aim of sequencing and mapping of the human genome and that of some model organisms. By that moment, it had been widely anticipated that knowing the complete human DNA sequence would help to better understand the genetic bases of disease, human evolution and the interplay between genes and environment.

The HGP was carried with the DNA of a small number of donors, obtaining a final sequence that was a mosaic of the volunteers' genomes. Since the completion of the HGP in 2003, the reference genome has been constantly improved and updated. The current human reference genome (GRCh38), released by the Genome Reference Consortium (GRC), is the twentieth version of it. This last version, although it has reduced or eliminated more than 100 gaps relative to the previous version (GRCh37, the one used in this thesis) and is considered the best-assembled mammalian genome, still contains 875 gaps (Paten et al., 2017). Long-read sequencing technologies are allowing the resolution of large gaps (>50kb) (Jain et al., 2018), but the reference genome now faces another problem: the variability that is being detected by current techniques, including most previously unidentified Structural Variants (SVs), is too large to be properly referenced by single reference sequences (Paten et al., 2017).

By the moment the HGP started, it was estimated that the 99.9% of the DNA sequence was shared between any two individuals (National Human Genome Research Institute, 1996), and the idea that Single Nucleotide Variants (SNVs) were the main source of genetic variation in humans remained for years after the completion of the first reference genome. Nonetheless, the development of techniques such as Comparative Genomic Hybridization (CGH) arrays led to a burst of population studies that revealed that SVs spanning more than 50 nucleotides contributed to human variation at least as much as SNVs (Escaramís et al., 2015). Current estimates using Next Generation Sequencing (NGS) techniques indicate that a typical genome differs from the reference genome in 3.5-4.3 million SNVs (~0.1%) and harbors a median of 18.4 Mbp of SVs (0.6%) (The 1000 Genomes Project Consortium, 2015; Sudmant et al., 2015a).

This previously unsuspected variability is raising concern about the possible biases that are derived from using a single reference genome to study all other human genomes, to the extent that the GRC has announced that they postpone the next release (GRCh39) indefinitely, while they evaluate new models to provide the best reference(s) (GRC website). Ideally, this improved reference genome should be able to reflect all this structural variability and even the variability within the SVs.

Genome organization and regulation

Besides providing a complete and accurate sequence of the human DNA, the HGP also intended to provide a complete catalogue of all the genes in the human genome. They were surprised to see that the number of protein-coding genes was much lower than initially expected (20,000-25,000, compared to previous estimates as high as 120,000 (Liang et al., 2000)) (International Human Genome Sequencing Consortium, 2004). This finding suggested that the complexity of the human genome is not limited to the number of protein-coding genes, but on how the genome is regulated.

The HGP marked a turning point in the study of the genome by enabling the development of different high-throughput "Omics" technologies on the genomics, transcriptomics, and epigenomics. The HGP also opened the way to other significant biological scientific efforts, such as the ENCyclopedia Of DNA Elements (ENCODE) project (Consortium, 2004), which had the goal to create a complete catalog of different classes of functional elements codified in the human genome. The ENCODE and the later Roadmap Epigenomics Project (Roadmap Epigenomics Consortium et al., 2015) have focused on the annotation of putative regulatory elements by mapping tissue-specific enhancers, based mainly on chromatin modifications and accessibility of the DNA. More recent molecular approaches (Chromosome Conformation Capture techniques) have permitted the analysis of the spatial organization of the chromatin in a cell, showing that the 3-dimensional organization of the genome plays a critical role in gene regulation and that enhancers can regulate gene expression of distal genes, even Mb away through physical interactions (Li et al., 2018).

These enhancers, which are normally found in non-coding regions, have the potential to physically interact with several regions, increasing the regulatory potential of the genome and, together with cis-regulatory elements, drive the identity of any cell type (González-Barrios et al., 2015). It has been suggested that trans factors make larger contributions to gene expression variability among individuals of the same species than cis-regulators (Signor and Nuzhdin, 2018).

Although non-coding regions have been proved to have a relevant regulatory role in many cases, how much of the genome is functional (and even the definition of "functional" itself) remains controversial. A possible classification of the genome based on its functionality and non-functionality, based on the proposal by Graur et al. (Graur et al., 2015) is:

- **Functional DNA**. A sequence that is selected naturally because of its function. It can be classified into two subgroups:
    - **Literal DNA**, if the order of the nucleotides is under selection, as in protein-coding regions.
    - **Indifferent DNA**, when the presence or absence of the fragment is under selection, but not the order of the nucleotides. Such sequences act as spacers, fillers or protectors against frameshift.
- **Non-functional DNA.** Sequences without a positively selected effect function
    - **Neutral non-functional DNA**. DNA that does not contribute not detracts from the fitness of the organism and thus selection does not operate on it. This term appeared for the first time in the 1960s and was formalized by Ohno in 1972 (Ohno, 1972).
    - **Detrimental non-functional DNA.** Negatively selected DNA that decreases the fitness of its carriers. It exists because natural selection is not immediate nor omnipotent.

"Indifferent DNA" can have an essential role in the spatial organization of the genome. These regions should not show selection against base pair substitutions, but SVs are expected to be under selection in these sequences. This means that SVs need to be analysed in a structural context.

Protein coding genes

Since the term "gene" was coined at the beginning of the 20th century, its definition has been evolving with the discoveries in the field of genetics (Portin and Wilkins, 2017). The Ensembl group defines a gene as a "genomic locus where transcription occurs" that may or may not encode or proteins and can have one or more transcripts. Other definitions add to this description that the final product has to be functional (Gerstein et al., 2007).

Currently, according to the annotation of functional elements made by the GENCODE Project, in the human genome, there is a total of 58,381 genes, out of which 19,901 encode for proteins (GENCODE Version 28, November 2017 freeze, GRCh38 - Ensembl 92). In Ensembl version 75, build GRCh37 (the version used in this thesis), there are 22,836 protein-coding genes out of 64,162 genes. The differences between versions exist because the list of genes that encode for proteins is continuously being updated and some genes and transcripts change status between databases and releases. A recent study from our group combining different lines of evidence, including transcript expression, protein expression, and genetic variation, suggests that many protein-coding annotated genes are in fact non-coding and that the number of protein-coding genes is probably close to 19,446 (the number of genes annotated in all Ensembl/GENCODE, RefSeq and UniProtKB reference databases) (Abascal et al., 2018).

In human, protein-coding genes have very variable sizes, ranging from less than 200 bp to more than 2 Mbp (Yates et al., 2016) (**Figure 1**). In most genes, however, the sequence that encodes a protein is discontinuous, distributed in sequences called **exons** that are interrupted by **introns**. An "average" gene[1] contains 11 exons separated by 10 intronic sequences. While the size of exons is quite stable (mean length 309 bp, standard deviation (SD) = 725bp) the size of introns is very variable mean 6355 bp, SD = 20,649 bp). The ratio between intron and exon size in an average gene is about 21:1 (Piovesan et al., 2016).

[1] Unless otherwise specified, when talking about genes we will be referring to protein-coding genes, which are the focus of this thesis.

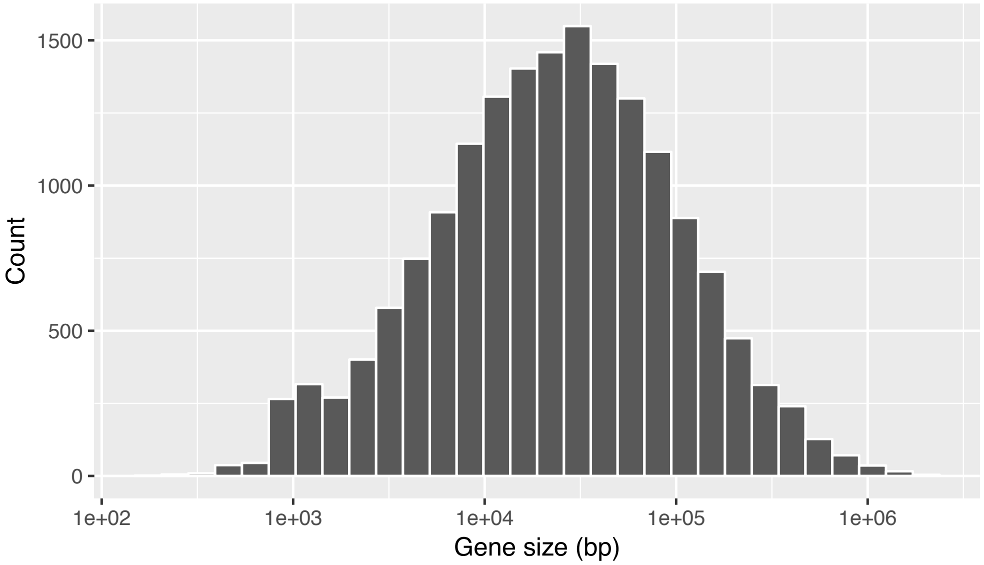


Figure 1 | Gene length distribution

Introns

In large genomes, introns account for most of the genic sequences. In the human genome, they constitute 93% of the protein-coding fraction and about half of the non-coding genome (Francis and Wörheide, 2017). The amount of intronic sequence is, in fact, similar to that of non-coding intergenic DNA, and this happens in most animal species (Francis and Wörheide, 2017).

Every time a gene is transcribed, the intronic content has to be excised at the exact correct positions with complex spliceosomal machinery. Whether (or how) introns compensate for the amount of energy that they cost to the cell is still not fully understood. What is clear, though, is that introns have been key in eukaryotic evolution (Rogozin et al., 2012).

The origin of introns

Extensive research strongly suggests that introns in eukaryotic cells originated after the establishment of the endosymbiosis between an alpha-proteobacterium and an archaeal host. Group II introns from the endosymbiont (typical mobile elements that actively spread around the host genome), would have invaded the genome of the emerging eukaryote (Rogozin et al., 2012). To survive to such an invasion, eukaryotes had to develop mechanisms that allowed the coordination of the (slow) process of splicing and the (faster) process of transcription. According to Martin and Koonin (Martin and Koonin, 2006), the emergence of the nuclear envelope was mandatory to prevent ribosomes from translating unspliced premessengers. Other mechanisms such as nonsense-mediated decay (NMD) — a post-transcriptional surveillance process that ensures the degradation of mRNAs with premature stop codons — also became necessary to ensure that only correctly spliced mRNAs are translated (Lambowitz and Belfort, 2015; Celik et al., 2017).

The roles of introns

Even though introns do not code for protein and need to be removed from the messenger RNA (mRNA) before it is translated into an amino acid sequence, introns can benefit the cell and the organism and participate actively in gene evolution. Some of the principal direct and indirect roles of introns are:

- **Alternative splicing** (AS): Introns break the protein-coding information of a gene. The step of cutting out introns from the pre-mRNA gives the possibility to generate alternative coding messages through the alternative splicing of the introns. In other words, multiple mature mRNAs can be obtained from one single gene thanks to alternative splicing, supposedly resulting in an extended protein repertoire without increasing the number of genes. Approximately 95% of the multi-exon genes in human undergo AS (Pan et al., 2008), although the extent to which AS contributes to proteomic complexity is still largely unknown (Liu et al., 2017; Tress et al., 2017).

- **Trans-splicing**: Although it is a rare process in humans, splicing can also happen *in trans* by combining two pre-mRNA molecules from different genes. The trans-spliced chimeric RNAs potentially can encode for a novel protein or act as regulatory RNAs (Lei et al., 2016).

- **Source of regulatory elements**: Introns (especially first introns) host many regulatory sequences such as enhancers and silencers that regulate the upstream promoter and can modulate transcription (Chorev and Carmel, 2012).

- **Source of non-coding RNAs (ncRNAs)**: Several **ncRNAs** including micro RNAs (miRNAs), short-interfering RNAs (siRNA), piwi-interacting RNAs (piRNAs), long non-coding RNAs (lncRNAs) and small nucleolar RNAs (snoRNAs) are preferentially located within introns. These ncRNAs have a broad spectrum of regulatory functions, and the processing of the ncRNAs itself can modify the expression of the host gene (Rearick et al., 2011; Heyn et al., 2015).

- **mRNA recognition, transport, and stability:** Introns have been suggested to act as identity markers, helping the cell machinery to detect mRNAs among the pool of transcripts (Palazzo and Gregory, 2014). Introns may also be affecting mRNA stability (Bonnet et al., 2017), transport (Valencia et al., 2008) and NMD (Wong et al., 2013).
- **Formation of new genes by exon shuffling.** The intron-mediated recombination of exons from different genes has been an important mechanism to create new genes through evolution (França et al., 2012).

Evidence of the importance of intron size

Gene length influences the time needed to transcribe a gene. Since gene size is primarily determined by intron size, intron length largely determines the expression timing and can provide a mechanism for temporal regulation of gene expression. A number of studies have shown different situations in which the size of the gene is relevant for the function of sets of proteins.

In 2002, Castillo-Davis et al. saw in *Homo sapiens* that introns of highly expressed genes are, on average, 14 times shorter than those of low-expressed genes, suggesting that selection could be acting to reduce the costs of transcription by shortening or keeping short the more highly expressed genes (Castillo-Davis et al., 2002). Similar results were observed for housekeeping genes (genes with a constitutive expression in all tissues), which are enriched in essential functions (Eisenberg and Levanon, 2003).

Genes expressed in rapidly cycling tend to be short and have few or no introns so that they can be efficiently expressed during a short cell cycle. The shortest cycles occur in early embryo development, during which the expressed genes are short and, in many cases, intronless (Heyn et al., 2015). On the opposite extreme, in terminally differentiated cells such as neurons, we find the longest human genes (Heyn et al., 2015) (**Supplementary table 1**). Moreover, long genes are enriched for neuronal functions (Gabel et al., 2015).

Intron length has been shown to affect the dynamics of transcriptionally controlled feedback loops and increase oscillatory periods of gene expression, processes that are essential in numerous contexts such as vertebrate somitogenesis, cell cycle, hormonal signaling and circadian rhythms (Swinburne et al., 2008).

When the transcription of a gene is activated or silenced, the time required to obtain a protein product will depend on the size of the gene. Thus, activation, but also shutting down, will be faster in shorter genes (Heyn et al., 2015). For this reason, long introns can cause delays in dynamic gene expression. In this line, Takashima et al. (2011) found that introns are required for *Hes7* gene oscillations in somite segmentation in mouse (Takashima et al., 2011). Further work by the same group showed that if the number of introns of *Hes7* was reduced, the time delay was shortened, oscillation time increased, and embryos developed more somites and vertebrae than wild-type mice (Harima et al., 2013).

An evolutionary study evidenced high levels of conservation in intron length in genes associated with embryonic development in mammals, suggesting that genes whose transcription requires precise time coordination are sensitive to changes in transcript length (Seoighe and Korir, 2011). Moreover, the comparison of mammalian genomes found that intron lengths of co-expressed genes or genes participating in the same protein complexes tend to coevolve, possibly because a precise temporal regulation of the co-expression of these genes is required (Keane and Seoighe, 2016).

Altogether, these studies suggest that the size of introns in different types of protein-coding genes can impact the proper functioning of a cell or an organism, and that intron length has been regulated, molded and shaped through evolution.

Intron splicing: introns vs. exons recognition theories

Gene structure is largely determined by its location in a region of low GC or high GC content. During the evolution of homeotherms (mammals and birds), a major GC increase happened that was accompanied by changes in gene structure (Bernardi, 2000).

Gil Ast and co-workers found a general negative correlation between exonic GC content and length of the flanking introns in mammalian and avian genomes [2] (homeotherm vertebrates) (Amit et al., 2012) similar to what had also been observed in human and chimpanzee (Gazave et al., 2007). On top of this observation, this work defined two exon-intron architectures that resulted from the evolution from an ancestral state of low GC exons flanked by short introns with even lower GC content (**Figure 2**), which here will be named "exon high – intron high" and "exon low – intron lower":

1) Exon high – intron high: Exons found in regions of high GC content flanked by **short introns** of a similar GC content. This group would have undergone a GC content elevation that abolished the differential GC content between exons and introns.
2) Exon low – intron lower: Exons in low GC content regions, flanked by **long introns** with a significantly lower GC than the exons. This group would have retained the low GC content and the GC drop in introns.

For proper removal of the intron from the pre-mRNA, the splicing machinery needs to recognize the splicing units (exons and introns) within the genic sequence. These two architectures require different mechanisms of splicing that differ in the splicing unit recognition, which can be an intron (intron definition model) or an exon (exon definition model).

[2] This study included genomes from mammals (human, cow, mouse, opossum, and platypus), birds (chicken) other vertebrates (frog, fugu, and zebrafish), invertebrates (*Caenorhabditis elegans*), and plants (*Arabidopsis thaliana*).

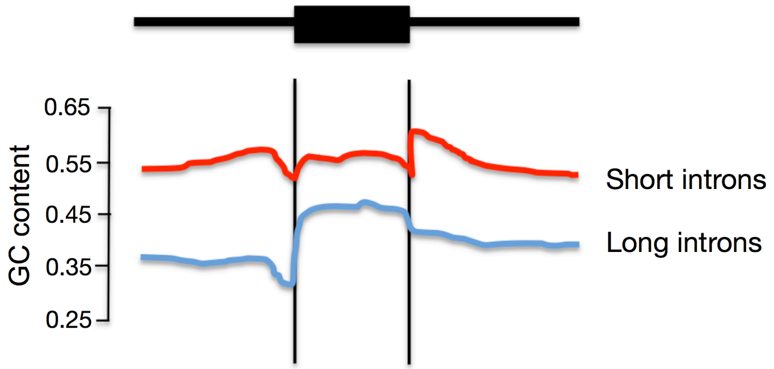


Figure 2 | Relationship between intron size and GC content. Average GC for exons (black box) flanked by long (blue) and short (red) introns (black horizontal line). Adapted from an article by Amit and others (Amit et al., 2012).

In the **intron definition** model, the machinery recognizes introns and places the basal splicing machinery across them. Genes with an "exon high – intron high" structure require this system. Intron definition, which is thought to be the ancestral splicing mechanism and widespread in modern lower eukaryotes, is limited to introns of a certain length. Introns recognized through this mechanism are under evolutionary selection to remain short (Amit et al., 2012; Hollander et al., 2016).

On the other hand, in the **exon definition** model, the splicing machinery recognizes exons among long introns and places the basal splicing machinery across exons instead of introns. This mechanisms is presumably an adaptation to overcome a general lengthening of introns (Hollander et al., 2016), and is used in genes with an "exon low – intron lower" structure. Increasing the GC content differential between exons and introns contributes to better recognition of the exon (Amit et al., 2012).

In higher eukaryotes, where the majority of introns are long, the predominant mode of splicing is probably exon selection (Hollander et al., 2016)

Structural variants

As mentioned above, the emergence of novel technologies uncovered the presence of a previously inconceivable amount of SVs in healthy individuals. The classification of these

SVs has been changing as the resolution of the techniques has increased. As a result, even today there is a lack of consensus on the classification of SVs.

Insertions or deletions under 50bp long are not considered SVs. Instead, they are typically called *short indels* (Lin et al., 2017) or *microindels* (Gonzalez et al., 2007). Notwithstanding, there is no real consensus on the maximum number of base pairs that fall in this category.

SVs are all variants larger than 50bp, and they encompass translocations (change of position of a segment of DNA, without a gain or loss of genetic material), inversions (inverted nucleotide sequence in the same position), insertions, and copy number variants (CNVs) (Escaramís et al., 2015). **CNVs, the focus of this thesis, are fragments of DNA longer than 50bp (Alkan et al., 2011; Zarrei et al., 2015) whose number of copies varies compared to the reference genome.** There is no maximum size for CNVs, although in some cases, such as in the Database of Genomic Variants (DGV) (MacDonald et al., 2014), they keep a record of CNVs up to 3Mb.

More extensive losses or duplications of portions of chromosomes are usually called chromosomal abnormalities or aberrations, or aneuploidies if they involve the loss or gain of a whole chromosome.

Mechanisms of CNV formation

Four major mechanisms generate genomic rearrangements and probably account for the majority of CNVs in humans. These mechanisms are:

- **Non-Allelic Homologous Recombination (NAHR).** NAHR is a recombination error that occurs during mitosis or meiosis (Zhang et al., 2009) when there is a misalignment of regions of extensive sequence similarity. Depending on the orientation and location, NAHR can cause deletions or duplications (Conrad et al., 2010).
- **Non-Homologous End Joining (NHEJ)**. NHEJ is a process of double-strand break (DSB) repair that fuses the ends of the break with little or no sequence homology (<4bp), generating short insertions or deletions at the breakpoint junction.

Breakpoints of NHEJ-mediated rearrangements often fall within DNA repetitive elements such as LTR, LINE, Alu, MIR, and MER2 (Zhang et al., 2009).

- **Fork Stalling and Template Switching** and **Microhomology-Mediated Break-Induced Replication (FoSTeS /MMBIR).** These mechanisms involve erroneous DNA replication and the shift, by microhomology, of the polymerase from the original template to another replication fork. The resulting rearrangements can have sizes ranging from kilobases to several megabases (Lee et al., 2007; Ottaviani et al., 2014).
- **Mobile Element Insertions (MEIs).** Most mobile elements annotated in the human genome are remnants of ancient retrotransposons that are no longer capable of active retrotransposition. However, some are still active, usually belonging to the Alu, L1 and SVA families of retrotransposons (Stewart et al., 2011). MEIs also have a role in the generation of SVs through the previously explained mechanisms, since copies of mobile elements maintain high levels of homology (Escaramís et al., 2015).

Each of these mechanisms leaves a detectable particular molecular signature in and around the breakpoints of the SV (Escaramís et al., 2015).

Distribution of CNVs in the genome

CNVs are distributed unevenly across the genome. To date, a number of studies have identified links between different genomic features and the formation of CNV.

Genomic repeats, both low and high-copy repeats, play an essential role in CNV formation and instability (Chen et al., 2014). A recent study showed that low-mappability regions are five times more likely to harbor CNVs than the remaining 90% of the genome (Monlong et al., 2018). However, because of the scarce coverage in these regions in most of the studies, the structural variation occurring within them is usually missed (Monlong et al., 2018). The temporal order in which DNA replicates (replication time or RT) is associated with different types of CNV mechanisms. While CNVs associated with NAHR are commonly found in early-replicating regions, CNVs caused by non-homologous repair (NH) are enriched in late-replicating DNA (Koren et al., 2012). Replication dynamics also appear to be linked to CNV distribution, and CNV breakpoints are enriched in genomic regions with a slowed replication (which can be a result of fork barriers, less fork initiation or reduced replication speed) (Chen et al., 2015).

SVs in healthy populations

Genetic diversity is essential for adaptation to environmental changes. While SNV variability has been largely studied, the contribution of SVs to traits, disease and gene regulation is still unclear. From the thousands of CNVs that have been detected in healthy populations (a median of 3,145 CNVs per person), some might contribute to susceptibility to diseases (Martin et al., 2015).

In 2015, the most extensive maps[3] of CNVs were published for healthy populations:

- **Abyzov:** Abyzov et al. did a systematic genome-wide study of deletion breakpoints detected from 1,092 individuals sequenced in phase 1 of the 1000 Genomes Project (1KGP) and studied their formation mechanisms (Abyzov et al., 2015).
- **Handsaker:** Handsaker et al. created a CNV map by analysing 849 genomes from phase 1 of the 1KGP. Their study aimed to detect and characterize multiallelic CNVs (mCNVs), defined as variants that appear at high frequency in the population and that vary over widely different numbers of copies (Handsaker et al., 2015).
- **Zarrei:** Zarrei et al. developed a map of CNVs and CNV regions (CNVRs, regions containing at least two CNVs that overlap and that may have different breakpoints) by selecting variants found in 2647 controls from the entire Database of Genomic Variants (DGV) collection. The selected CNVs had been dected using different methods including SNP or CGH array, NGS and Sanger sequencing (Zarrei et al., 2015).
- **Sudmant-Science:** Sudmant et al. sequenced the genome of 236 individuals from 125 distinct human populations from across the globe. They identified new CNVs that could be population-specific (Sudmant et al., 2015b).

[3] For the sake of simplicity, from now on, each study will be referred to using the name is its first author, followed by the journal in the case of the two maps with the same first author.

- **Sudmant-Nature:** This map published by the 1KGP consortium consists of an integrated map of SVs analysing the phase 3 whole-genome sequencing data (Sudmant et al., 2015a) obtained from 2504 individuals and analysed using multiple algorithms for the calling of the SVs.

In addition to the abovementioned studies, which are the ones that will be analysed in this thesis, during the last decade several countries have started national projects to sequence the genome of inhabitants within the country, in order to describe the genetic background of their population groups, and, ultimately, to improve their health care (Dubow and Marjanovic, 2016; An, 2017) (**Table 1**). These projects, however, will only reflect part of the variability in the human genome and underrepresent or miss variants that are specific from other populations.

National Genome Projects
deCODE genetics (*Iceland*)
The Estonian Biobank / Estonian Genome Centre, University of Tartu (EGCUT)
The Singapore Genome Variation Project
Genome of the Netherlands (GoNL)
GenomeDenmark
The Faroe Genome Project (FarGen)
Cymru DNA Wales
The National Centre for Indigenous Genomics (NCIG) (*Australia*)
Kuwait legislation introducing mandatory DNA testing (no project name)
The Precision Medicine Initiative Cohort Program (*US*)
SardiNIA
China Kadoorie Biobank (CKB)
UK Biobank
The Slim Initiative in Genomic Medicine for the Americas (SIGMA) (*Mexico*)
UK10K
The Deciphering Developmental Disorders (DDD) Study (UK)
Genomics England (The 100,000 Genomes Project)
The Saudi National Genome Program
The Belgium Medical Genomics Initiative (BeMGI)
The Initiative on Rare and Undiagnosed Diseases (*Japan*)
The National Centre for Excellence in Research in Parkinson's Disease (*Luxembourg*)

Table 1 | National genome sequencing initiatives. List of national initiatives to sequence the genome of a representative part of their population.

To date, the 1KGP (Sudmant-Nature) is the most comprehensive available study, combining SNV and SV calls and including 26 populations from 5 major population groups (Africa, America, Europe, and South and East Asia) (Sudmant et al., 2015a).

The impact of CNVs on protein-coding genes

CNVs can affect protein-coding genes in different ways:

- **Protein disruption**: CNVs can modify the amino acid sequence if they overlap with exons or splicing signals.
- **Alteration of gene dosage**: CNVs that cover whole genes represent complete loss (homozygous or heterozygous) or gain (of one or more copies) of a gene. The number of copies of a gene correlates in many cases with its expression levels (Handsaker et al., 2015; Rice and McLysaght, 2017).
- **Impact on gene regulation:** CNVs can affect gene expression by either inserting new regulatory elements, by disrupting existing regulatory regions or modifying their distance from the regulated gene. At times, gained copies of a gene can occur in other chromatin environments than the original copy, or be surrounded by new regulatory elements that can produce expression changes (Harewood et al., 2012; Weischenfeldt et al., 2013; Gamazon and Stranger, 2015).

In general, CNVs are more likely to contribute to variation in the expression levels of a gene than SNPs (Bryois et al., 2014; Chiang et al., 2017), and during the last few years, many studies have linked CNVs to changes in gene expression in humans (Gamazon et al., 2011; Chiang et al., 2017; Sudmant et al., 2015a; Glassberg et al., 2019).

How mutations shape evolution

CNVs can produce changes that alter the fitness of an allele and, consequently, selective forces might act upon them. The mechanisms that can alter the fitness include the previously explained gene expression modifications, the changes in the coding sequence, and also the creation of paralogues (Iskow et al., 2012).

Gene duplications arise as CNVs and provide a substrate for evolution. If the original and/or the copied genes mutate, divergence can result in **neofunctionalization** (where the old function of the gene is maintained and a new function evolves in one of the copies) or **subfunctionalization** (where the original function is distributed between the two copies due to mutations partially but complementarily inactivating each copy). This mechanism has been crucial in evolution, as most innovations in gene functions seem to be associated with gene duplication in one way or another (Conant and Wolfe, 2008).

The selective forces acting on CNVs can be purifying (negative) or positive, and they will act on harmful or beneficial CNVs, respectively. Both scenarios will usually lead to fixation (by removing detrimental CNVs or by increasing the frequency of the beneficial ones). An obvious depletion of CNVs overlapping with functional regions has been reported in several studies (Khurana et al., 2013; Sudmant et al., 2015a; Zarrei et al., 2015), suggesting a strong purifying selection on CNVs that disrupt coding sequences. Moreover, big CNVs (of over 500kb) seem to be under stronger purifying selection than smaller CNVs, probably due to the higher probability of overlapping with a functional region (Iskow et al., 2012).

In healthy individuals, however, we find thousands of CNVs. However, most of them are expected to be benign CNVs, with no visible impact on the phenotype or associated with benign polymorphic traits (Zhang et al., 2009). A few CNVs are thought to be positively selected, based on their population distribution (Iskow et al., 2012). This seems to be the case of an mCNVs encompassing the *HPR* gene, which is involved in response to trypanosomes and is present at high copy numbers in the African population (Handsaker et al., 2015; Sudmant et al., 2015b), or the salivary amylase gene (*AMY1*), present at high copies in populations with high-starch diets (Perry et al., 2007).

The association between CNVs and genes and genomic features

Different studies have shown that not all genes or genomic structures are equally affected by CNVs. Zarrei and others observed that genes associated with different types of diseases are less variable in copy number than expected in healthy individuals (Zarrei et al., 2015). Another study showed that most ancient genes, which are enriched in housekeeping and essential functions, have a fixed number of copies (they are not variable in copy number), while young genes, which tend to be more tissue-specific, are more often variable in copy number (Juan et al., 2013). An analysis of regulatory features revealed that regulatory regions are not equally affected by CNVs either: while, in general, promoters are enriched with CNVs, enhancers are depleted (Zarrei et al., 2015).

However, more analyses are needed giving more consideration to the size, boundaries, and activity of these and other functional elements as well as to the location and amount of overlap with the CNVs. Also, compared to variation in exons, much less attention has been paid to the impact of mutations in introns, even if several pathogenic variants have been found deep within introns (Vaz-Drago et al., 2017) and even if many SNVs associated with disease detected through GWAS are located in introns (Hsiao et al., 2016; Xiong et al., 2015).

An interesting feature of the variation in introns is that it can affect regulatory elements such as enhancers, which often act in a tissue-specific manner (Vermunt et al., 2019). A disruption of their function caused by a mutation can show up in a cell-type specific way. A study combining the detection of regulatory regions active in different tissues with GWAS found that disease-associated SNPs are frequently located in enhancers active in a tissue or cell type relevant to the disease (Ernst et al., 2011). In this thesis, we want to study how intronic variants can affect gene regulation and if any groups of genes are more susceptible to carry such variants.

Given that copy number variation started being in the spotlight only recently, the "normal" (healthy) distribution of CNVs in introns is less studied than that of SNVs and small indels. Currently, there is no consensus on whether introns are enriched or depleted of CNVs, or none (Khurana et al., 2013; Mu et al., 2011; Sudmant et al., 2015a). One of the goals of this thesis is to study the distribution of CNVs in the genome, analysing different maps of CNVs in parallel in order to understand the causes of the dissenting results.

Objectives

Objectives

1. Analyse the overlap between copy number variable regions (CNVs) and genomic features in human genomes, focusing on the differential distribution of CNVs in protein-coding genes.

2. Investigate the effect of intronic deletions on gene structure and splicing.

3. Explore the potential effect of intronic deletions on gene regulation and gene expression changes.

4. Study the different impact of intronic CNVs on genes depending on their function, essentiality and evolutionary history.

Materials and methods

Materials and methods

Obtention and filtering of CNV maps

Whole genome CNV maps from healthy populations were downloaded from 5 different publications from 2015 (Abyzov et al., 2015; Handsaker et al., 2015; Sudmant et al., 2015a, 2015b; Zarrei et al., 2015). We selected autosomal and not private CNVs. In Handsaker's map we removed low quality CNVs and all the variants from samples NA07346 and NA11918 because they were missing in the phased map. Abyzov and Handsaker are maps based on all (in Abyzov) or most (in Handsaker) low-coverage alignents from phase 1 of the 1KGP (1000 Genomes Project Consortium et al., 2012). In both cases the samples originate from 14 different populations from Africa, America, East Asia and Europe. Sudmant-Nature is the analysis of the third phase of the 1KGP, which analyses more samples from a total of 26 populations (including samples from South Asia), uses different input sequence data, aligns against an improved version of reference genome GRCh37, and uses different variants callers.

In the case of the Zarrei's map, which is a curated selection of CNVs from the entire Database of Genomic Variants (DGV) collection, we selected the more stringent map that includes CNVs present in at least two individuals and in two studies. It is important to note that this meta analysis includes variants from the pilot and phase 1 of the 1KGP (The 1000 Genomes Project Consortium, 2010, 2015).

Sudmant-Science includes 236 from 125 populations from across the globe (including Siberia and Oceania), with 1 to 3 samples per populations (except for 14 Papuan samples).

Ancestral state

To unravel the ancestral state of the CNVs marked as deletions, we have compared the Final 1000 Genomes Project dataset (Sudmant-Nature) with recent high-quality genomic data of great apes (Kronenberg et al., 2018). In detail, when comparing Sudmant-Nature to Kronenberg's data, an SV was considered identical if there was a reciprocal overlap higher than 80%. Deletions were confirmed when they appeared in a genomic region that can be found in non-human primates (NHP), without any SV in the NHP or with insertions only. Conversely, insertions were confirmed if the fragment is annotated as a deletion in all NHP at an allelic frequency = 1.

Gene structures

Coordinates and sequences of protein-coding gene structures were retrieved from Ensembl (Yates et al., 2016) version 75. Principal and alternative isoforms were retrieved from the APPRIS database (Rodriguez et al., 2013), Ensembl version 74. Intronic regions were defined as the constitutively intronic parts of genes, i.e. parts of introns that don't overlap with any exon from any other gene or isoform. When analysing real introns, for example when we look at the position of the intron, we used only the principal isoform. To avoid duplicate identification of introns, in the cases of more than one principal isoform, we selected the isoform with a higher exonic content.

Genome coordinates and low-mappability regions were obtained from R package "BSgenome.Hsapiens.UCSC.hg19.masked" (The Bioconductor Dev Team, 2014).

Essential genes

The essential genes list is a combination of sets of genes reported as essential in different studies based on CRISPR genomic targeting (Hart et al., 2015; Wang et al., 2015), gene-trap insertional mutagenesis (Blomen et al., 2015) and shRNA (Cheung et al., 2011; Marcotte et al., 2012; Silva et al., 2008).

Statistical assessment of genome-wide distribution of CNVs

To estimate enrichment or impoverishment of CNVs in different genomic functional elements or regions we performed permutation tests in which we compared the number of overlaps of CNVs with the regions to the number of overlaps in a background model. We did these analysis using three types of background models in which we relocated 10,000 times the CNVs in the genome following different criteria. The "global" background model was obtained by relocating all CNVs anywhere in the genome, avoiding low-mappability regions. The “local” background model was obtained by segmenting the genome in 278 of at least 10Mb and afterwards relocating the CNVs within their respective 10Mb window of origin, also avoiding low-mappability regions. Finally, the “RT” or “Replication time” background model consisted of the segmentation of the genome in regions of similar replication time and relocating all CNVs within a region of similar RT. Replication time was obtained from publicly available data from 15 cell lines, downloaded from ENCODE (Hansen et al., 2010; Thurman et al., 2007). Each 1kb window of the genome was assigned the median RT value of all cell lines. Then, the genome was divided in 5 RT intervals with the same number of windows and all CNVs were relocated within windows belonging to the same interval of RT.

Enrichment/Impoverishment ratios and P-values were computed using a function derived from the permTest function from package RegioneR version 1.6.2 (Gel et al., 2016). Code available in https://github.com/orgs/IntronicCNVs.

Comparison of intronic and intergenic regions

The comparison of number and size of deletions in intronic and intergenic regions was done by randomly selecting a subset of 500 intronic regions and finding for each of them the intergenic region with the most similar size possible. Then, we calculated the overall number of deletions and their characteristics in the 500 intronic regions and the 500 intergenic regions. This process of sampling was repeated 10,000 times and the distribution of deletions in the intronic and the intergenic regions was compared using paired Student’s T-test.

Regulatory features

The genome coordinates of regions likely to be involved in gene regulation were downloaded from the Ensembl Regulatory Build (Zerbino et al., 2015), assembled from IHEC epigenomic data (Stunnenberg et al., 2016).

To calculate if such regions are enriched in introns we generated background models similar in the same way as the "global" background model for CNVs.

In order to study if deletions and a regulatory regions that cooccur in the same intron tend to overlap or not, we took each intronic deletion and randomly relocated it 10,000 times within its intron of origin and compared the number of cases in which an intronic deletion overlaps with a RF in the original set and in the randomized sets. P-values are the fraction of random values superior or inferior to the observed values.

In the analysis of the overlap with RFs by the number of cell types in which the RF is active, all intronic deletions from all five datasets (except for exact duplicates) were taken into account.

Gene expression analysis

RNA-seq data for 445 individuals from the 1KGP (Sudmant-Nature) was available from the Geuvadis Consortium (Lappalainen et al., 2013). We analysed the expression of the 763 genes with only one intronic deletion in the population and with at least two of the 445 samples carrying the deletion. For each of these deletions we compared using Student's t-test the PEER-normalized (Stegle et al., 2010) gene expression levels (GD462.GeneQuantRPKM.50FN.samplename.resk10.norm.txt.gz) in the individuals homozygous for the reference genotype and in the individuals with a deletion in one of the alleles. We corrected for multiple testing using the p.adjust R function, using the Benjamini-Hochberg method. In addition to the multiple testing correction, to verify if the number of significant differentially expressed genes is different from expected by chance, we shuffled 10,000 times the genotypes of the individuals and compared in the same way the gene expression levels of the artificial groups of homozygous and heterozygous individuals. Each of the 10,000 times we calculated the number of eGenes and finally we compared the random percentages of eGenes to the percentage observed in the real dataset. P-values were calculated as the fraction of random values superior or inferior to the observed values.

Differential expression at the level of individual transcripts was calculated in a similar way, using data froma file GD462.TrQuantRPKM.50FN.samplename.resk10.txt.gz of the Geuvadis consortium (Lappalainen et al., 2013).

Exon inclusion/exclusion

To study changes in exon inclusion or exclusion we used alternative exon overexpression as a proxy for higher inclusion and underexpression for exon exclusion. The expression levels of alternative exons upstream or downstream of an intronic deletion was compared between the individuals carrying an allele with the deletion and wild-type individuals. Exon expression data was obtained from the Geuvadis Consortium (GD462.ExonQuantCount.45N.50FN.samplename.resk10.txt.gz) (Lappalainen et al., 2013).

CNV mechanisms

We had CNV mechanism information for Abyzov and Sudmant-Nature maps. In Sudmant-Nature, though, the dataset with mechanisms assigned did not correspond exactly to the main CNVs dataset. For this reason, we used the coordinates from the main CNV set and assigned the mechanism of the CNV with the same identifier in the mechanisms dataset.

Population stratification

Population stratification of deletions was estimated using the Vst statistics extracted from Sudmant-Nature (Sudmant et al., 2015a). This Vst statistic (Redon et al., 2006) is a mesure of the variance of a CNV between populations. It is caculated by considering $(V_T-V_S)/ V_T$ where V_T is the variance in copy number genotypes among all unrelated individuals and V_S is the average variance within each population, weighted for population size (Sudmant et al., 2015a). As in the study from which we Vst statistics (Sudmant-Nature, Sudmant et al. 2015a), we selected a cutoff of 0.2 to indicate high population stratification of a locus.

Observed vs expected intronic deletion content score

To rank the genes according to their enrichment of intronic deletions we created a score comparing the observed and expected deletions per gene. For this analysis, a map with all deletions from Sudmant-Nature, Abyzov and Zarrei maps was created (Abyzov et al., 2015; Handsaker et al., 2015; Sudmant et al., 2015a). The expected values were calculated in two different ways: 1) relocating 10,000 times all deletions in the whole genome and 2) relocating 1,000 times all intronic deletions within the intronic regions. In both cases, low-mappability regions were avoided. The enrichment score was calculated after ranking the genes by 1) number of intronic deletions per gene divided by their median expected value, 2) position of the observed divided by the median expected size of the deletions, 3) position of the percentage of intronic content that is lost, 4) the inverse of the expected intronic loss and 5) ranked added minor allele frequencies of deletions per each gene in Sudmant-Nature. Once all rankings were calculated and normalized from 0 to 1, a score was assigned to each gene by averaging their five ranks.

Because this 5 step procedure was done fore the two types ofrandomizations, as a result we obtained two lists of genes from more to less enriched. We then took the top and bottom 500 genes from each list and selected the genes that were in the intersection of the two lists. The intersections resulted in 469 genes with a lowest score and 483 with a highest score (less and more deletions than expected, respectively).

Functional enrichment analysis

Functional enrichment analysis of the genes with a lower scores and higher scores was performed with GSEA (Subramanian et al., 2005) and STRING v11 (Szklarczyk et al., 2015) using default parameters. Enrichment of selected sets of genes within our sets of genes with more and less deletions were done performing Fisher tests. The background in these tests was the list of genes for which we were able to assign an enrichment score.

Dating gene and intron ages

Duplicated and singleton genes were assigned an evolutionary age as described in Juan et al. 2014. Briefly, using the gene family phylogenetic reconstructions of ENSEMBL Compara (Herrero et al., 2016), which uses gene sequences from 52 different species and assigns speciation or duplication events. Using this information, we assigned to each duplicated gene the age of the phylostratum assigned to the last duplication leading to the birth of the extant protein-coding genes. Singleton genes were defined as the ones without a detectable duplication origin and their ages were assigned from the last common ancestor to all the genes in their family.

The resulting gene ages groups and the number of genes per age, from ancient to recent, are the following: 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 analysis, ages were grouped as follows: Ancient genes are all the genes from age groups FungiMetazoa to Sarcopterygii, Middle-aged genes are all genes from Tetrapoda to Eutheria, and Young or Primate genes, all genes from Simiiformes to HomoSapiens.

The ages of intronic regions were given according to the gene they belonged to. When an intronic region was part of more than one gene, the most recent age was assigned.

SCNA data

SCNAs were obtained from 2583 samples from the ICGC/TCGA Pan-Cancer project (Campbell et al., 2017). A filtering of the samples was done to select euploid samples, since the category of gain and loss is difficult to define in very fragmented genomes. Ploidy levels and percentage of diploid genome were calculated for each patient and euploid genomes were defined as all samples with a ploidy (average copy number in the whole genome) between 1.1 and 2.9 (2+- 0.9) and at least the 50% of the genome at copy number = 2. The remaining set consisted of 1068 euploid samples, from which the coordinates of deleted fragments were extracted, considering deletions all fragments with a copy number lower than that of the flanking fragments. The overlap between SCNAs and RFs was calculated as in section “Regulatory features”.

Analysis of differential GC content

Genomic sequences were obtained from the primary GRCh37/hg19 assembly, and were used for calculating the GC content of introns and intronic CNVs. Differences in GC content between a CNV and the intron where it is located were calculated with paired Student’s t-tests taking as statistical unit the CNV. The same was done for changes in intronic GC content before and after a deletion.

Results

Results

Overview and comparison of CNV maps

With the aim of characterizing the impact of CNVs on protein coding genes in healthy humans we used five high resolution CNV maps published in 2015:

- **Handsaker** (Handsaker et al., 2015)
- **Abyzov** (Abyzov et al., 2015)
- **Zarrei** (Zarrei et al., 2015)
- **Sudmant-Nature** (Sudmant et al., 2015a)
- **Sudmant-Science** (Sudmant et al., 2015b)

Each one of the maps has been derived from a different number of individuals from various populations and using different techniques and algorithms for CNV detection (**Supplementary table 2**), representing five differing views of population CNVs.

The datasets contrast notably in number, type and size of CNVs detected, even in cases where the majority of the genomes analysed are the same (Hansaker's and Abyzov's maps, see Materials and methods and **Supplementary table 2**). We decided to analyse each CNV map separately instead of combining them into a single map, avoiding a merging of independent CNVs or the opposite: considering as independent two CNVs that are in fact the same but which have been called differently in two studies. In this thesis, only autosomic CNVs present in at least 2 individuals in the same map are taken into consideration.

We observed that the third phase of the 1KGP (Sudmant-Nature) is, by far, the map that provides more CNVs (**Figure 3A**). More than the half of the genomic regions that are seen to be affected by CNVs (CNV regions or CNVR) in Sudmant-Nature are not reported to be variable in copy number in any of the other maps (**Figure 3B**).

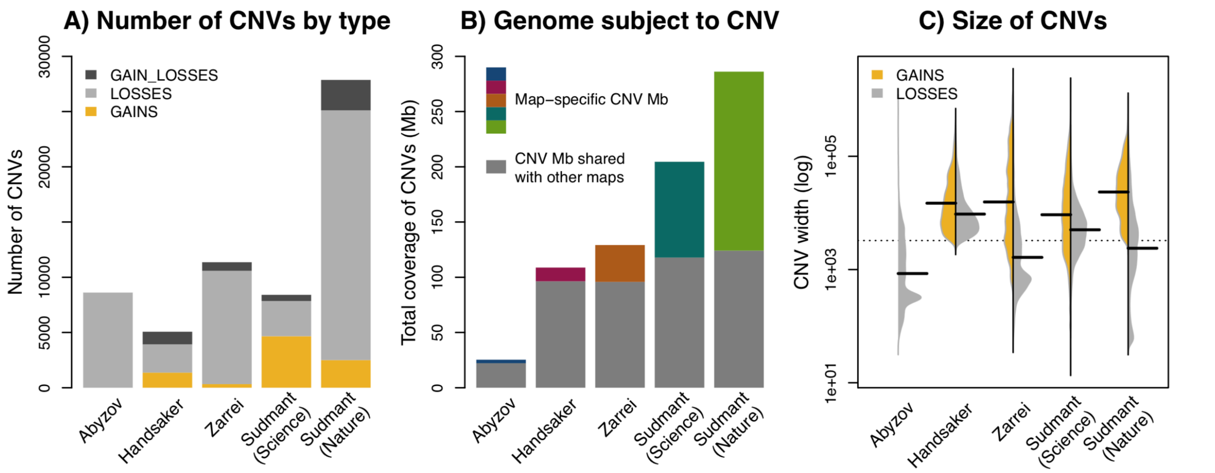


Figure 3 | Comparison of datasets. Comparison of the 5 maps of copy number variability in healthy population used in this study. A) Number of CNVs reported by each map, separated by type of CNV. B) Amount of the genome subject to copy number variation in each map. In gray, regions reported in more than one map; coloured, the amount of CNV genome detected only in one map. C) Size of the CNVs in each map, separated in gains and losses.

Despite the variability among studies, deletions are, on average, consistently smaller than gains (**Figure 3C**). This observation can be due to technical biases in the detection of gains and losses or it can reflect the reality of SV in the genome.

Overlap of CNVs with protein coding genes

The number of autosomal protein coding genes affected by CNVs is very variable, ranging from 1,694 in Handsaker's map to 5,610 in the Sudmant-Nature map. This difference in number is not surprising if we consider that the number and sizes of CNVs are so diverse among maps (**Figure 3**). However, it is striking to see very little overlap among the lists of affected genes: only 402 (5.5%) of all genes affected by CNVs coincide in the five maps.

The impact of a CNV on a gene is likely to be different depending what part of it is affected. A CNV can cover the totality of a gene, affecting gene dosage, or it can delete or duplicate a fraction of it. The high resolution of the CNV maps based on whole genome sequencing (WGS) data allowed us to classify the variants that overlap with protein coding genes in three groups (**Figure 4**):

- **Whole gene CNVs:** CNVs that encompass entire genes
- **Exonic CNVs:** CNVs overlapping with part of the coding sequence

- **Intronic CNVs:** CNVs falling in intronic regions, not overlapping with any exon

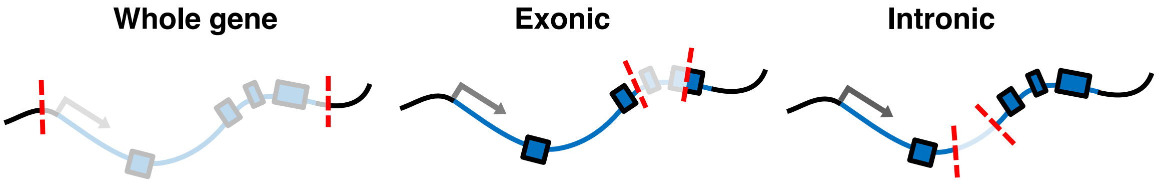


Figure 4 | Protein-coding overlapping CNVs. Schematic representation of the different types of protein-coding overlapping CNVs

For the definition of **intronic CNVs**, we selected CNVs that did not overlap with exons of any annotated transcript isoform or exons from other genes that reside in introns. It is important to bear in mind that some exonic CNVs do overlap with introns, but they are excluded from the intronic CNVs group because they also affect coding regions.

The number of CNVs and their type (gain, loss or gain/loss) changes with the type of overlap with the gene. Intronic CNVs are the most common of gene-overlapping CNVs, while whole-gene CNVs are rarer (**Figure 5**).

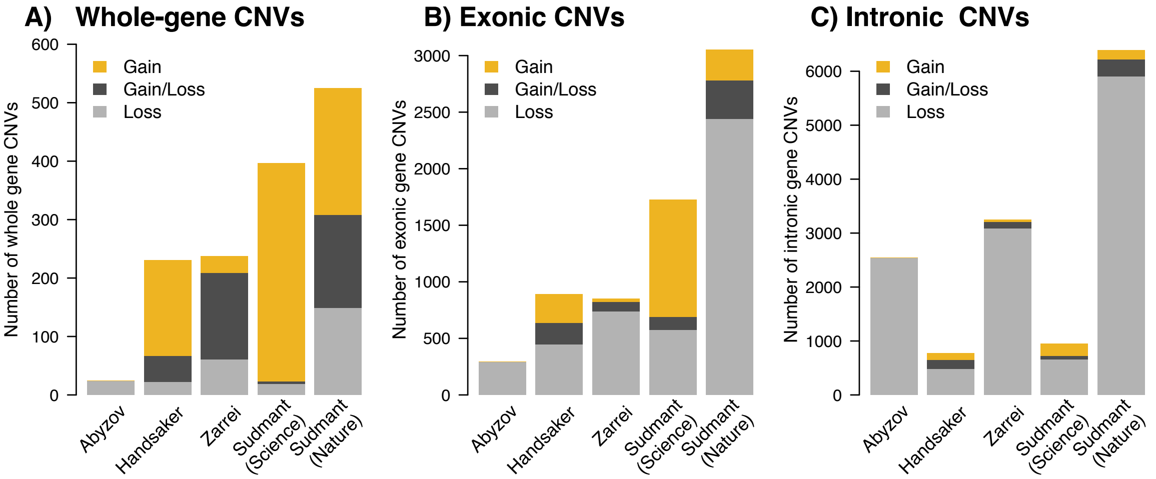


Figure 5 | Number of CNVs per map depending on type and overlap with a gene. Number of CNVs covering A) whole genes, B) exons but not the whole gene or C) falling within introns. Each bar represents a dataset and the types of CNVs (gain, gain/loss or loss) are depicted with different colours.

Whole gene CNVs

All maps include genes that are completely duplicated or deleted. These genes that have different *dosage* in the population are called *CNV-genes*. Whole-gene CNVs are more frequently gains (55% of the cases) or gain-losses (25%) than losses (**Figure 5**).

Considering all maps, we find a total of 1,212 genes entirely overlapping CNVs. However, we only observed 6 genes with this status in all the maps. Interestingly, two of them are associated with disease (**Table 2**).

Gene name	Description	CNV type	AF
LCE3C and LCE3B	Precursors of the cornified envelope of the stratum corneum. Involved in antimicrobial activity. Diseases associated with LCE3C include Psoriatic Arthritis.	Loss (4 maps) Gain-Loss (1 map)	0.5629 (del)
AADAC	Involved in drug metabolism. Diseases associated with AADAC include Chanarin-Dorfman Syndrome and Gilles De La Tourette Syndrome.	Loss (5 maps)	0.0042 (del)
MRGPRG	May regulate nociceptor function and/or development, including the sensation or modulation of pain.	Loss (4 maps) Gain-Loss (1 map)	0.03275 (del) 0.0357 (dup) 0.00599 (trip)
OR52N5	Olfactory receptor	Loss (4 maps) Gain-Loss (1 map)	0.2356 (del)
OR5P2	Olfactory receptor. May be involved in taste perception.	Loss (4 maps) Gain-Loss (1 map)	0.1210 (del) 0.0006 (dup)

Table 2. CNV-genes common in all maps. Description of the six CNV-genes detected in all the maps and allelic frequencies (AF) for each allele in Sudmant-Nature.

We reasoned that frequent CNV-genes will more probably be detected by more maps, while rarer events will be detected in fewer studies. Indeed, if we look at the allelic frequencies (AF) of the CNVs affecting genes in Sudmant-Nature, we see that the genes detectes as variable in copy number in more maps tend to overlap with CNVs with a higher AF (**Figure 6**). This result shows that each map is able to collect a partial subset of the CNV-genes.

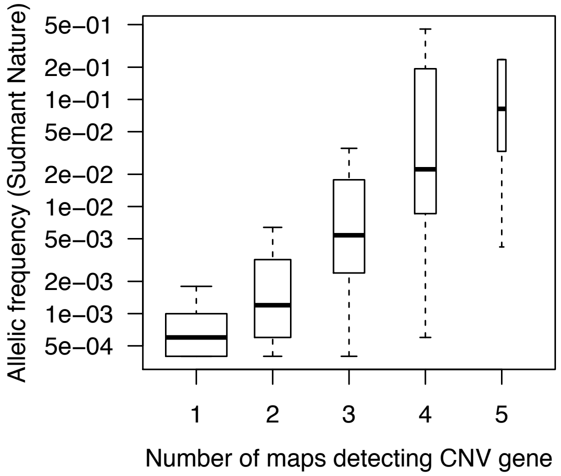


Figure 6 | Frequency of whole gene CNVs. Allelic frequencies of the CNVs enterily with overlapping with genes, group by number of maps where they are detected. The AFs for all alleles (different copy numbers of a CNV) are extracted from the Sudmant-Nature map and all the AF of alleles different than the reference were summed, representing the frequency of the gene having a copy number different than the reference. Because the AF are extracted from Sudmant-Nature, CNV-genes not observed in this map are not included.

Exonic CNVs

Exonic CNVs are CNVs that overlap with exons of a protein-coding gene but do not cover the whole gene. These CNVs represent the 31% of the CNVs that overlap with genes. A substantial proportion of exonic CNVs are losses (66%), while the 23% are gains and the remaining 11% are gain-losses (**Figure 5**).

Losses of exonic sequence will necessarily result in changes or even in the disruption of the protein sequence. On the other hand, it is not possible to predict the impact of gains of exonic sequence with the data provided by the maps in our study. The maps in our study do not give information on where gained regions are inserted. Depending on where the insertion happens, the gain can modify or disrupt the protein sequence, or it can have no impact at all on the protein (for example, if it is inserted in an a distant intergenic region). However, even without knowing where the insertion happens, we can assume that exonic **deletions will more often be deleterious than exonic gains, due to a higher probability to disrupt the coding sequence**.

Exons can belong to principal or to alternative isoforms, or both. High-impact variation has been shown to be substantially lower in alternative exons than in exons belonging to the principal isoform(s) (Tress et al., 2017). This past study on the variation in principal and alternative exons has not included the analysis of CNVs.

Given that losses will more probably have a high impact on the protein and will probably be under stronger negative selective pressure than gains, we assumed that the ratio of gains to losses will differ in principal and alternative exons. To check if our hypothesis was true, we classified all exons into principal, alternative and intersection (exons or parts of exons that belong to both principal or alternative isoforms, as in the abovementioned article (Tress et al., 2017)). For each type of exons, we calculated the ratio of genes with losses versus genes with gains. We observed in all datasets that this ratio is the lowest for the principal exons and the highest for alternative, showing that principal exons have a lower tendency to be lost than alternative exons and in agreement with our assumption that coding losses are probably more deleterious than gains (**Figure 7**).

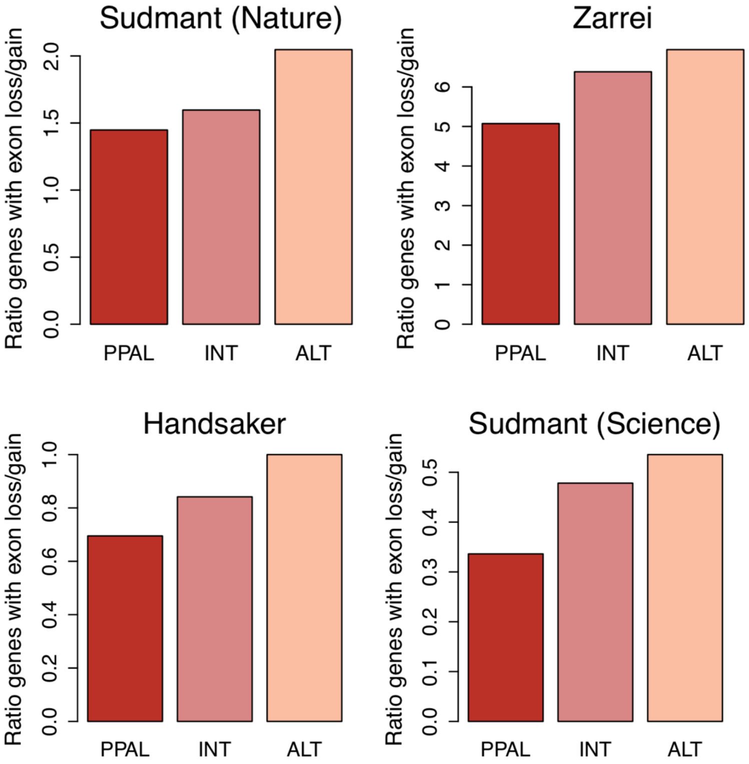


Figure 7 | Impact of CNVs on principal and alternative exons. Ratio of genes with losses versus genes with gains in their exons, by type of exon. PPAL = exons or part of exons that only belong to principal isoforms, INT = exons or part of exons that belong to principal and alternative isoforms. ALT = exons or parts of alternative exons. The ratios were not calculated for Abyzov's map because it only contains losses.

Intronic CNVs

Most CNVs that overlap with genes fall within intronic regions (from 31 to 88.9% of all CNVs, 63% from all CNVs in all 5 maps), without overlapping with any exon.

Almost all (91%) of the intronic CNVs are deletions (**Figure 5 and Table 3**). Thus, because intronic CNVs are the most common type of gene-overlapping CNVs and most of them are losses, **intronic deletions are the most prevalent form of CNVs overlapping protein-coding genes** (57.3%) (**Table 3**).

	Gain	Gain-Loss	Loss
Whole gene	783 (3.5%)	356 (1.6%)	275 (1.2%)
Exonic	1,587 (7.2%)	726 (3.3%)	4,500 (20.3%)
Intronic	575 (2.6%)	654 (3.0%)	**12,680 (57.3%)**

Table 3 | Proportion of each type of CNVs. Absolute and relative number of CNVs by type, taking all CNVs from all datasets together.

Two of the CNV maps, Handsaker and Sudmant-Science have a very limited number of intronic CNVs and they contribute altogether with only the 14% of the purely intronic deletions in our study. This is probably because the tools used in both studies have biases towards the detection of larger CNV regions (mostly gains) so the probability of covering exons is higher.

As mentioned before, the maps that we have analysed do not inform of the position where gains are inserted. While we know that intronic deletions lead to the shortening of an intron, we cannot know if a gain has a consequence or not on the intron. For this reason, and because deletions represent most of the intronic CNVs, we focused our subsequent analyses of the impact of CNVs on introns on deletions.

Ancestral state of the variants

Most deletions reflect losses relative to an ancestral genome

The deletions detected in the five datasets provide us with the fragments of introns that can be absent in part of the population without an obvious deleterious impact, since they are observed in the healthy individuals. In total, intronic deletions cover the 2,95% of the reference "introme".

It is important to note that the status of deletion or gain of a CNV is determined in relation to the reference human genome in all the maps in our study. However, the reference genome, which is a composite derived from the DNA of many individuals, does not necessarily reflect the ancestral genome. If, for example, a **gain** of 50bp relative to the ancestral genome is by chance present in the reference genome (because it was present in sequenced individuals), in the CNV maps based on this reference genome, this fragment will be annotated as a **deletion** in the individuals who lack this fragment but who, in fact, carry the ancestral genotype. For this reason, we cannot assume that the CNVs marked as deletions in these 5 maps correspond to ancestral regions that can be lost without a deleterious effect, as some of them might be recent insertions not fixed in the population.

With the aim of understanding the extent to which the CNVs annotated as deletions really represent deletions relative to the ancestral genome, we took recent high-quality data of great apes (Kronenberg et al., 2018) and checked if the variants from the 1KGP (Sudmant-Nature) were specific of humans. Deletions could be confirmed if they appeared in regions of the human genome that do not present deletions in non-human primates (NHP).

We were able to confirm that at least 72.8% (16,319/22,412) of the deletions are actual deletions compared to the ancestral state, implying that in most of the cases the genotype of the reference genome reflects the ancestral genome. **Regarding the subgroup of intronic deletions, the 79.2% were confirmed to be ancestral deletions**. On the other hand, 0.3% of the deletions in Sudmant-Nature are in fact insertions (0.42% of the intronic deletions). For the remaining 27% it was difficult to unravel the ancestral state, sometimes because more species would be needed and other times because of the presence of other SVs in the same region in some NHP make it difficult to assignment a state (**Supplementary figure 1**).

In summary, **most of the detected deletions reflect deletions relative to an ancestral genome.**

Population variability in intron and gene size

Intronic deletions result in drastic changes of gene length in the population

The part of an intron that is subject to loss is very variable, from 0.03% to 98.01% (51 bp to 293 kb). Taking whole genes as units, the part of the "introme" that is subject to losses represents 0.01% to 77.5% of the total genic size (**Figure 8**).

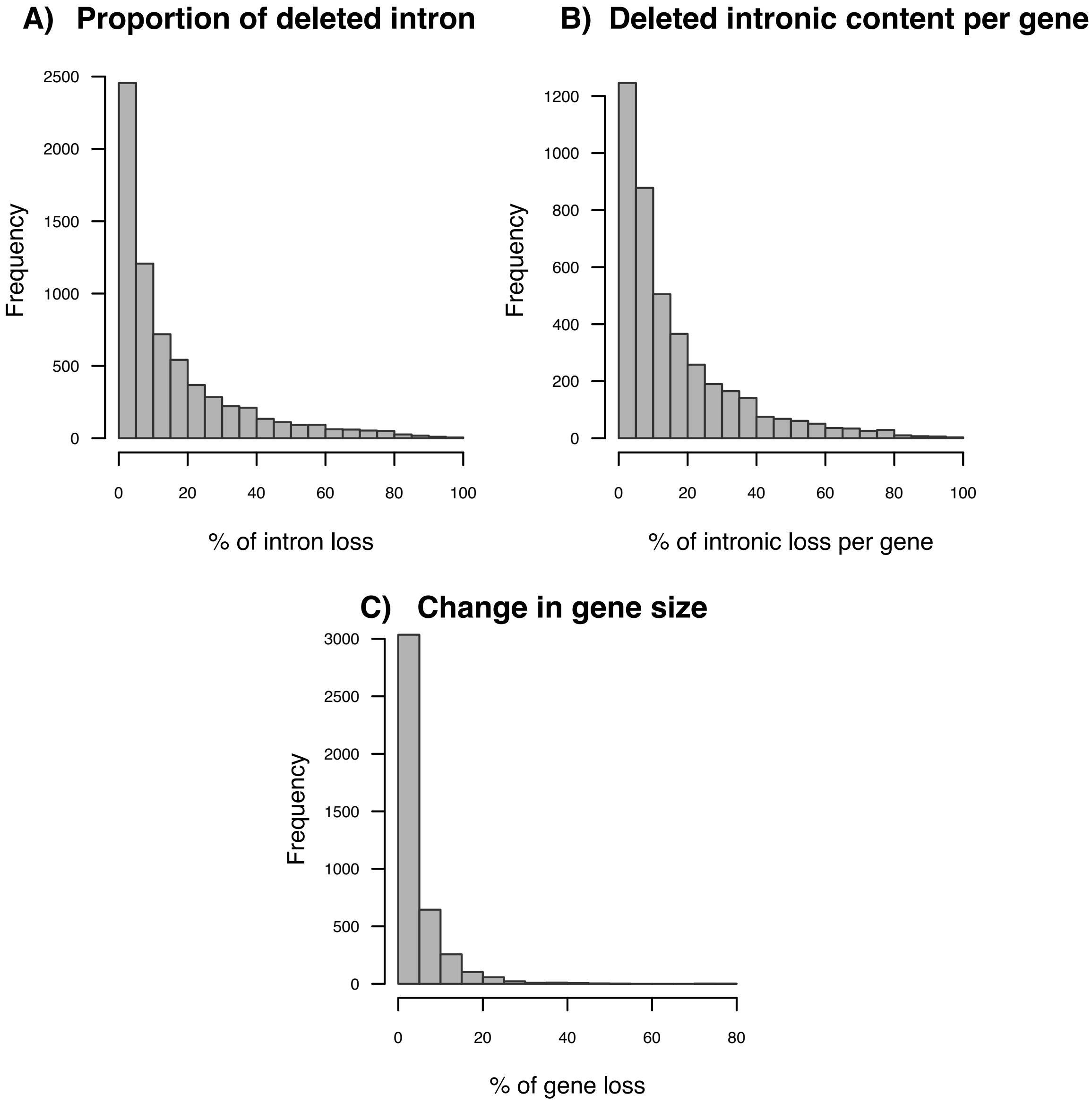


Figure 8 | (A) Proportion of the intron that has deletions in at least one study. (B) Percentage of all the intronic content of a gene that has deletions in at least one study. (C) Observed changes in the size of the gene caused by intronic deletions.

Even single deletions can cause very big changes in intron and gene size. Two examples of genes with a single deletion causing one of the largest changes in gene size are the neuronal glutamate transporter *SLC1A1* (*Solute Carrier Family 1 Member 1*), with a loss of the 37% of its genic size and the *LINGO2* (*Leucine Rich Repeat And Ig Domain Containing 2*) gene with a loss of the 34% of its size. Both genes are highly conserved at the protein level and have variants associated with diseases. In the case of *LINGO2*, a SNP within an intron has been associated with body mass (Rask-Andersen et al., 2015). In total, **we found 1,638 genes associated with disease** (present in the Online Mendelian Inheritance in Man - OMIM - database) **carrying intronic deletions in the healthy population**.

The combination of different intronic deletions affecting the same gene in a same individual can give rise to several alleles of different size in a population. The gene with more alleles in the 1KGP population (Sudmant-Nature) is *CSMD1* (*CUB And Sushi Multiple Domains 1*), with a total of 66 intronic annotated deletions that, combined, produce 150 alleles of different sizes. Notably, this gene is among the most conserved genes in the human genome, with only 0.168% genes more intolerant to variation in the coding sequence, according to the the Residual Variation Intolerance Score (RVIS), which is a scoring system that assesses whether genes have more or less functional (missense, stop and splicing) variants than expected by chance given the amout of neutral variants that they carry (Petrovski et al., 2013).

These results show that **many human genes are subject to losses in the population, generating genes with variable sizes. Even in genes with very high conservation in the coding sequence we observe extreme variation in intron sizes**.

Association of deletions genomic elements

Intronic regions are enriched with deletions

Interested in the possible relevance of high number of CNVs in introns, we asked ourselves if the deletions in our datasets are more or less prevalent in introns than expected by chance. To estimate the amount of deletions expected to fall by chance within introns, we generated three independent background models. The three models differ in the regions of the genome where each deletion can be relocated:

- The **global** model: All the deletions from a map are relocated in random places anywhere in the genome, avoiding low-mappability regions.
- The **local** model. All deletions are relocated within 10Mb windows (each within their window of origin) under the assumption that there is a similar underlying genomic structure within each window. To do this, we segmented all chromosomes in fragments of up to 10Mb as previously done in other studies (Bickel et al., 2010; Mu et al., 2011) and relocated the deletions avoiding low-mappability regions.
- The **replication-timing** (RT) model. Since DNA RT influences the rates of CNV formation (Koren et al., 2012), we created this model consisting in the relocation of each deletion within a region of similar RT, avoiding low-mappability regions.

After generating 10,000 randomizations for each model and each one of the five datasets, we compared the observed distributions the of the deletions overlapping with exons, introns or intergenic regions with the expected distributions obtained with the background models. Using the global background model, we found that there is a general depletion of deletions overlapping with exons and with introns (**Figure 9**). However, if we focus our attention on the deletions that are **purely intronic** (i.e., that do not overlap with an exon) **we see that introns are significantly enriched with deletions in 3 out of 5 maps** (4.14-9.3% more deletions than expected) and depleted in one of the maps (Sudmant-Science) (**Figure 9**). **Intergenic regions are more enriched with deletions than introns in most of the maps**, except in Zarrei's map (**Figure 9**).

Similar results were obtained using the local and the RT background model (**Supplementary figure 2**).

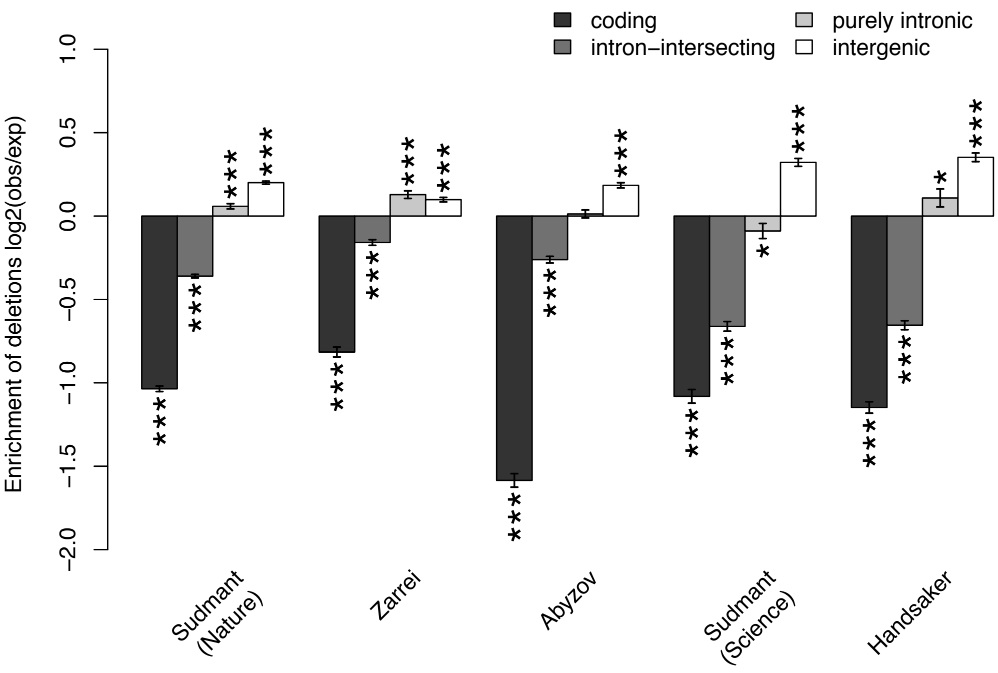


Figure 9 | Enrichment of deletions. Ratios of observed versus expected number of deletions overlapping with exons ("coding"), overlapping with introns ("intron-intersecting"), falling within intronic regions ("purely intronic") or within intergenic regions ("intergenic"). Height of the bar is the median of the ratio between the observed number of overlaps and each of the 10,000 randomized sets. Whiskers show median absolute deviation and asterisks mark significance: * for P<0.05, ** for P<0.005 and *** for P<0.0005.

The enrichment of deletions within introns apparently contradicts previous studies that have described that introns have less CNVs than expected by chance (Khurana et al., 2013; Mu et al., 2011; Sudmant et al., 2015a). We have analysed different explanations that can justify the apparently contrary results:

1. **Distinct definition of genic elements:** our definition of intronic regions includes only constitutive (purely) intronic regions, while in other studies, introns are taken from the principal isoform and thus, can contain alternative exons.
2. **Different classification of what CNVs are "intronic".** In our study, intronic deletions never overlap with an exon. Other studies reporting impoverishment of CNVs in introns included those CNVs intersecting exons. If we look at intron-intersecting deletions without removing those also intersecting exons, then we also find a clear depletion of deletions (**Figure 9**).

3. **Differences in the background model**: Differences in the background model can give different results. We tried two models used in other articles (Mu et al., Khurana et al.) and a novel RT-based model. The results differ slightly, especially in Sudmant-Science (in which we see a depletion with two background models). However, with all background models we see an enrichment of intronic deletions in 3 maps.
4. **Differences in the dataset**: Different datasets show different levels of enrichment within introns, also among the 5 maps in our study. The differences among datasets can be due to real differences among the populations, but most probably will be due to biases in the methods and algorithms used for CNV calling. For example, we did not see an enrichment of intronic deletions with Abyzov's map, but if we classify the deletions by their formation mechanism, then we find that the subset of **deletions generated through NAHR are enriched in introns**, a result consistent with a previous study (Mu et al., 2011) and also with the enrichment observed when we separate by mechansims the deletions from Sudmant-Nature in which the mechanism is annotated (**Figure 10**). The proportion of deletions caused by each of the mechanisms is different in Abyzov and Sudmant-Nature(See legend in **Figure 10**). The percentage of TEI-caused CNVs is over 8 times higher in Abyzov than in Sudmant-Nature. TEI deletions are significantly smaller than NAHR (P= 5.30e-31) and NH deletions (P = 4.02e-21) (**Supplementary figure 3**) and they are not enriched within introns. The overrepresentation of TEI deletions, possibly because Abyzov's map is biased towards smaller deletions, could be an underlying cause of the observed differences in enrichment in the two maps.

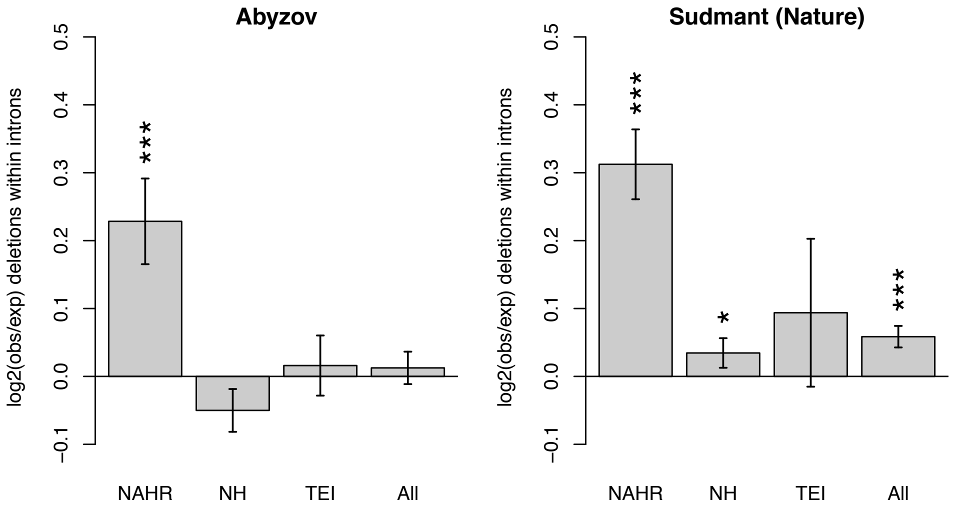


Figure 10 | Enrichment of deletions with created through different mechanisms. Differential enrichment within introns of the deletions generated through different mechanisms. Bar height is the log2 ratio of observed versus expected values, obtained using the global background random model. Asterisks mark the significantly enriched or depleted groups of deletions. Significance: * for P<0.05, ** for P<0.005 and *** for P<0.0005. Proportion of deletions for each mechanism:
Sudmant-Nature: NAHR = 13.7%, NH = 78.4%, TEI = 2.9%, Other/unsure = 5%.
Abyzov: NAHR = 13.2%, NH = 60.4%, TEI = 25.5%, Other/unsure = 0.9%.

Intergenic regions carry more and larger deletions than intronic regions

Intergenic and intronic regions have a different size distribution (**Supplementary figure 4**), with intergenic regions being, on average, larger (median intergenic size = 17.33 kb, 1.47 kb). As mentioned before, we observed that intergenic regions are more enriched with deletions than introns. To better understand if these distinct levels of enrichment are caused by the differences in size or not, we compared the load of deletions of intronic regions with that of intergenic regions of similar size. We found that intergenic regions with sizes comparable to intronic regions have a significantly higher number of deletions than intronic regions (FC = 1.23, P = 2.23e-308). Also, we observed that intergenic deletions are on average bigger than intronic deletions (**Figure 11**). A possible explanation for this difference could be the presence of a stronger purifying selection on the deletions happening in introns than in intergenic regions.

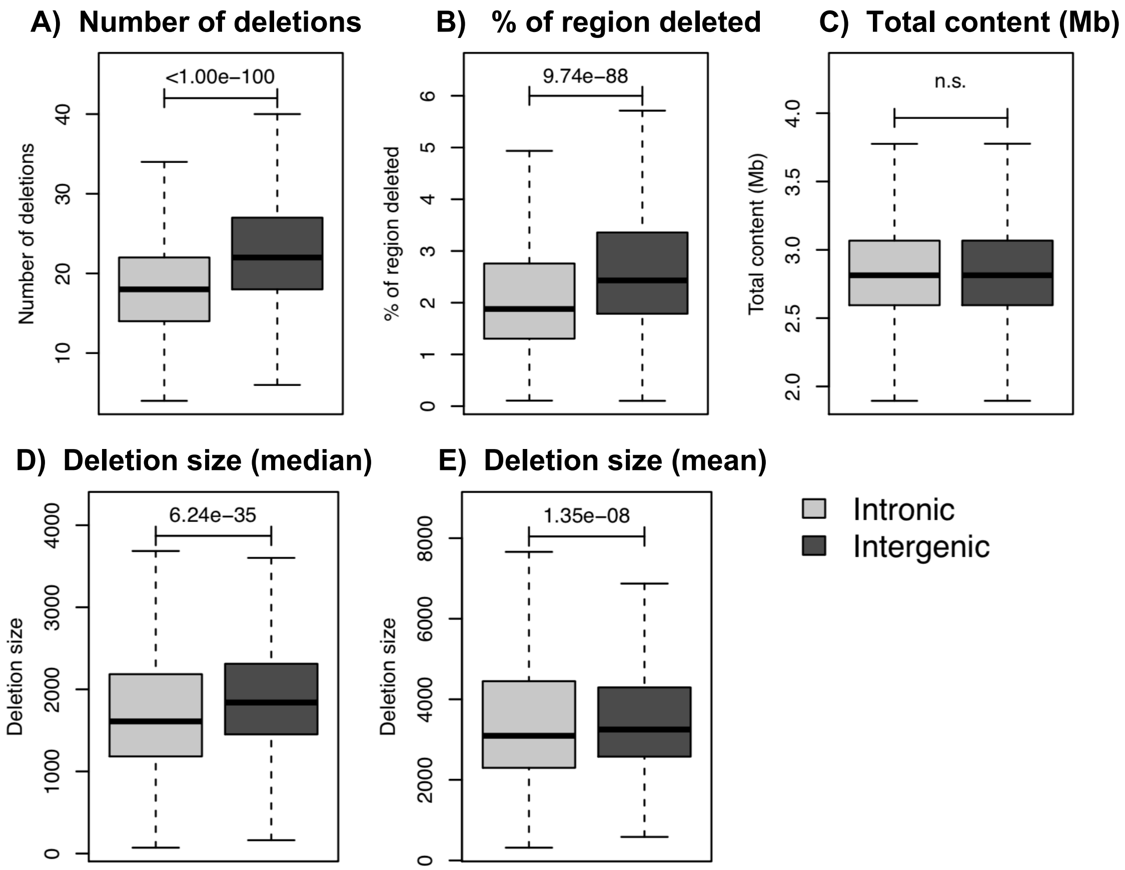


Figure 11 | Comparison of intronic and intergenic regions. For a random sample of 500 introns, we selected 500 intergenic regions of similar sizes and we calculated the deletion content and size for both groups independently, using Sudmant-Nature's dataset. We repeated this procedure 10,000 times. Each permutation is represented as one point in the boxplot. A) Number of deletions in the subset of introns or intergenic regions. B) Proportion of the sampled regions that is lost in the population. C) Total genome selected in each randomization (control to verify that the intronic and the intergenic regions are comparable in size). D) Median and E) mean size of the deletions comprised in each subset of intronic or intergenic regions. P-values were calculated with paired Student's T-tests.

Enrichment of deletions is independent of intron size

We wondered if the enrichment of deletions within introns was explained by a specific size range of introns. For this, we classified all introns in 10 groups by size (deciles, with a similar number of introns) and observed that **most of intronic deletions (90%) are found within introns larger than 1500 bp** (median size of intronic regions is 1,470 bp) and the 61% of the intronic deletions are located in the 10% largest introns (**Figure 12A**).

If we compare the number of deletions in each size group to those expected by chance according to our global background model, we find that the enrichment pattern is not particularly accentuated in any size bin, neither in the largest introns (**Figure 12B**).

Because Sudmant-Science and Handsaker maps have very few intronic deletions, we only calculated the enrichment by intron size using Sudmant-Nature, Zarrei and Abyzov's maps, which represent the 86% of all intronic deletions from our datasets.

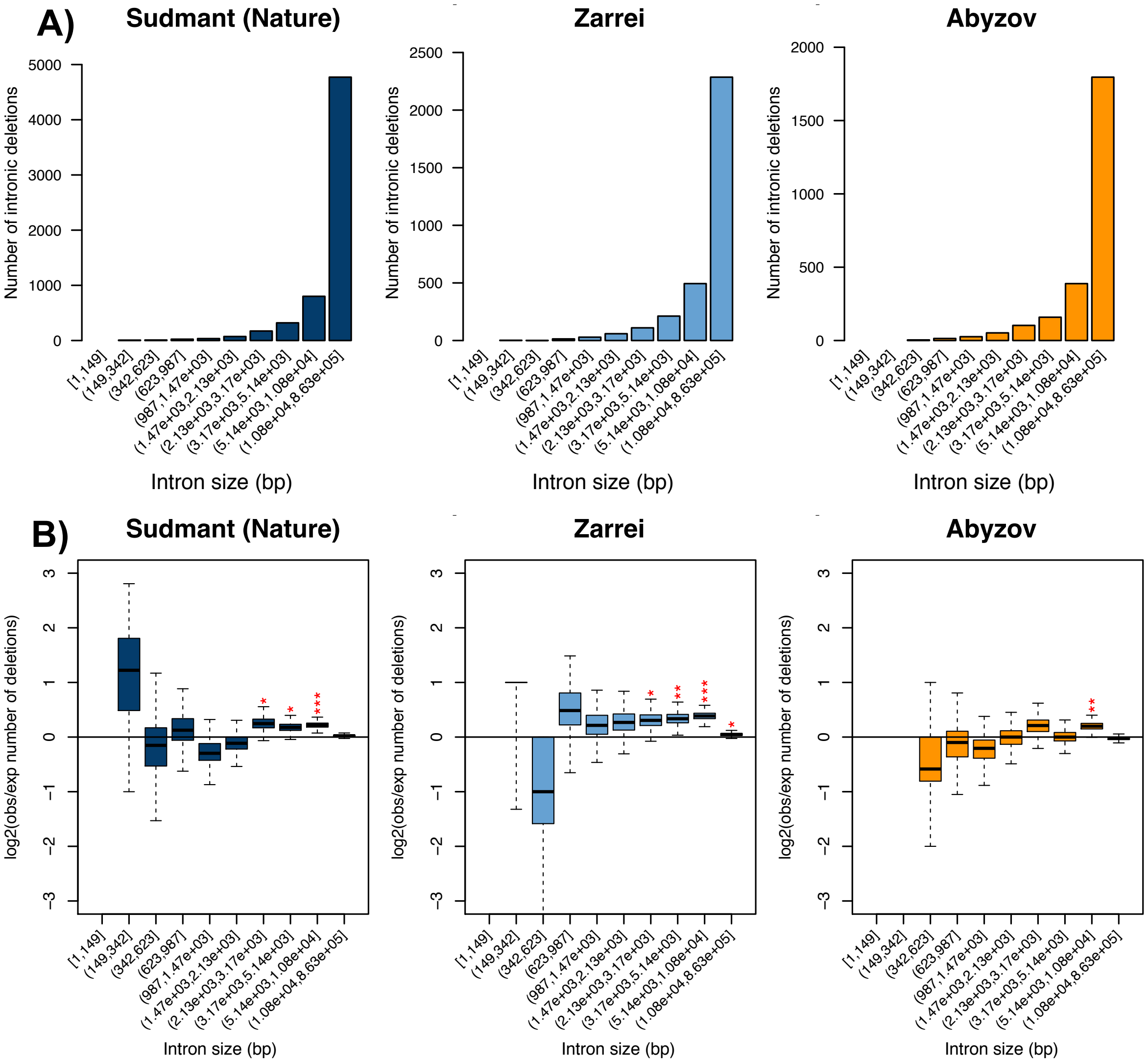


Figure 12 | Distribution and enrichment of deletions by intron size. (A) Number of deletions within introns of different sizes. (B) Enrichment of deletions in introns of different sizes, compared to the global background model. Red asterisks mark the significantly enriched groups of genes. Significance: * for P<0.05, ** for P<0.005 and *** for P<0.0005. All size bins (deciles) contain a similar number of intronic regions, between 18807 and 18913.

These results show that the enrichment of intronic deletions is not a centered in a specific size of introns, and that the number of intronic deletions is close to the expected values in most size bins.

Evolutionary age of genes affected by CNVs

Genes of different evolutionary ages show different patterns of overlap with CNVs

A previous study from our group showed that the percentage of CNV-genes increased as gene age decreased (Juan et al., 2013). Motivated by this finding, which was derived from the analysis of **whole-gene CNVs** specifically, we asked ourselves if partial CNVs, either exonic or intronic, also followed the same trend and how the structure of the genes (their size and exon-intron content) could be influencing these patterns.

Whole gene CNVs

In the article by Juan and others they used data derived from array technologies, which does not achieve the same resolution as the NGS technologies from the CNV maps in our study. Desite the limitation in resolution in the previous study, our results reproduced their findings (Juan et al., 2013) with all of the 5 maps (**Figure 13A, Supplementary figures 4 and 5**). The results were similar when we only considered losses, which are the focus of our study[4] (**Figure 13B**) .

[4] As previously mentioned, losses are the the focus of this study given that in introns, most CNVs are deletions. Thus, to be able to compare the impact of intronic CNVs to that of coding CNVs, we removed gains from most subsetquent analysis.

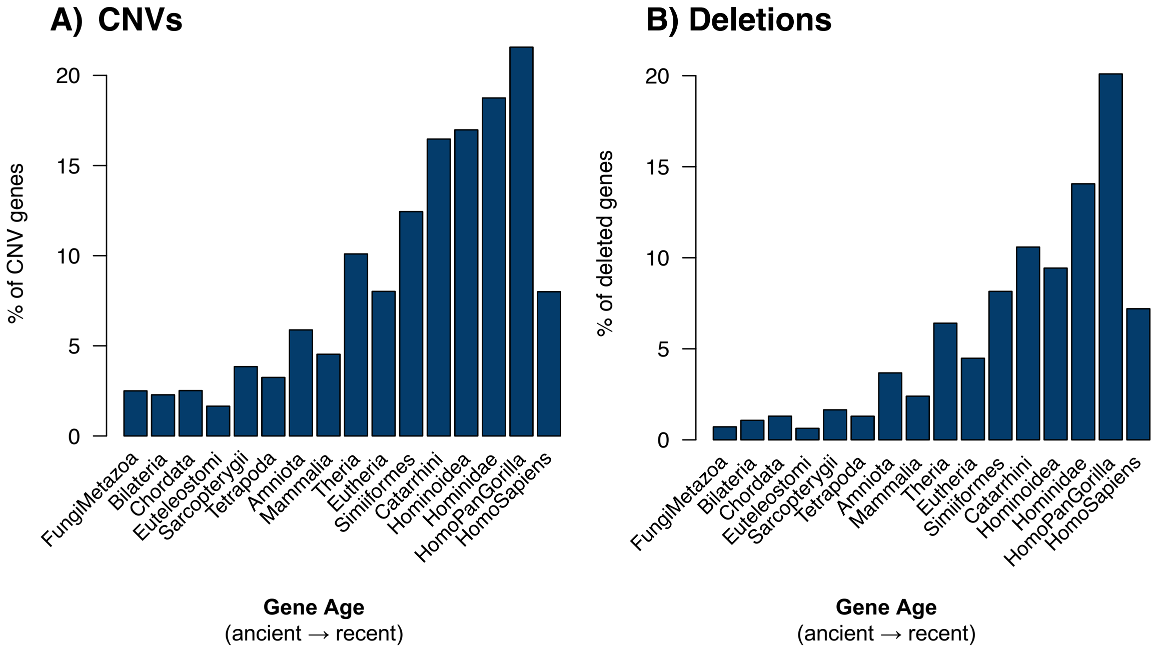


Figure 13 | Proportion of CNV-genes in different evolutionary ages. Percentage of genes from each evolutionary age group that are completely covered by a CNV (A) or a deletion (B). See figures **Supplementary figure 5** and **Supplementary figure 6** for the results in the remaining 4 CNV maps.

Partially overlapping deletions

Regarding partially overlapping deletions (exonic and purely intronic), we observed that exonic deletions show a very similar pattern of overlap to whole-gene CNVs (**Figure 14**), **with younger genes being more likely to have their coding sequence affected by CNVs.**

Strikingly, we find the opposite pattern for intronic deletions, where ancient genes harbor intronic deletions more frequently.

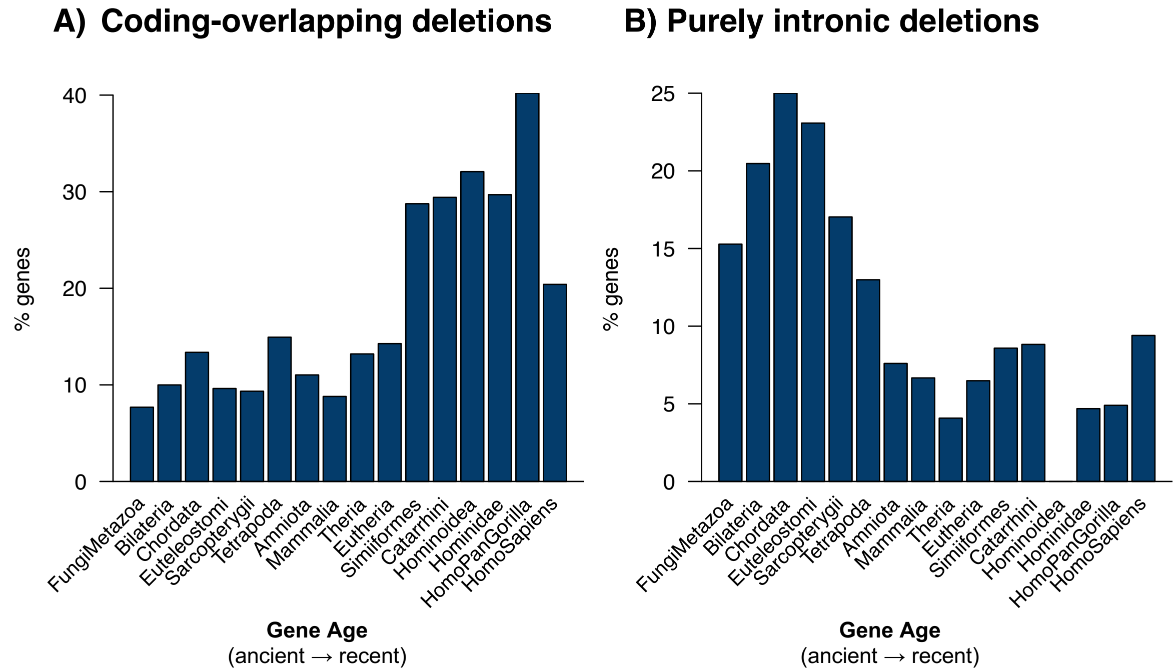


Figure 14 | Impact of deletions on genes of different evolutionary ages. Percentage of genes from each gene evolutionary age that contain exon-overlapping deletions (A) or contain intronic deletions (B). See **Supplementary figure 7** for the results obtained with other maps.

Influence of gene structure on the observed patterns of CNVs

Genes of different evolutionary ages have differences in gene structure. Ancient genes have more exons (and therefore introns) and are, in general, longer than young genes (**Figure 15**). These differences in gene structure could be underlying the different pattern of deletions throughout evolutionary ages. For example, the probability of an ancient (usually long) gene to be fully duplicated or deleted is lower than that of a younger smaller gene, because a much larger CNV is needed to duplicate or delete a bigger gene. Contrarily, the probability for an ancient gene to have an intronic deletion is higher because ancient genes have a higher intronic content (**Figure 15D**). This higher intronic content is as a combination of having more and longer introns (Figure 15 B and C). Also, ancient genes are more rarely intronless (**Supplementary figure 8**). In order to discriminate between the differences caused by the structure of the gene and possible functionally or evolutionary relevant causes, we compared the observed distribution to what would be expected by chance, according to our background models, which will also be affected and thus correct by the gene structure.

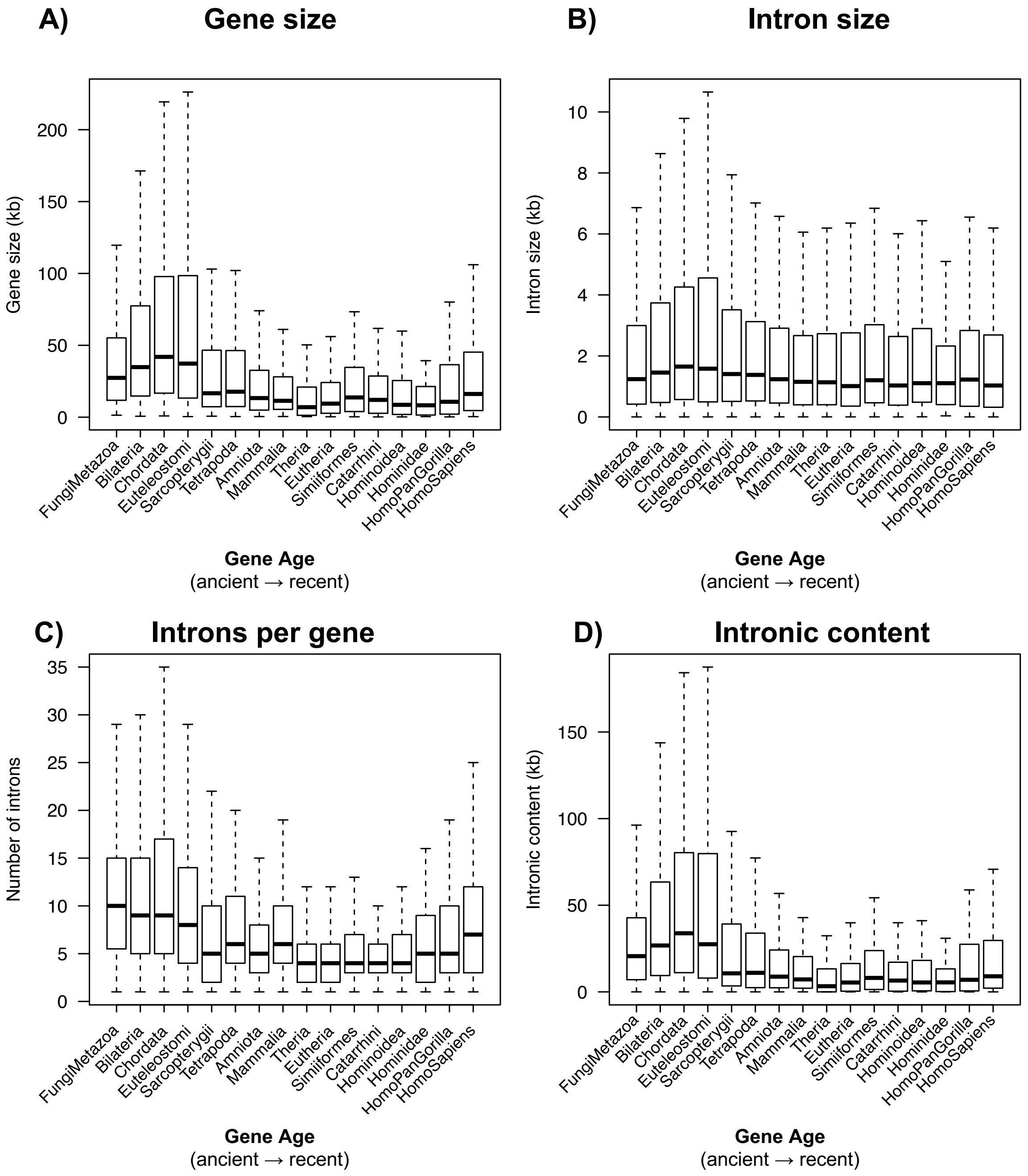


Figure 15 | Gene and intron sizes by evolutionary age. (A) Sizes of genes of different evolutionary ages, (B) size of their introns, (C) number of introns per gene, and (D) total intronic content per gene.

We observed that ancient genes do not just have fewer coding deletions (both whole-gene and exonic) but also, they have fewer coding deletions than expected by chance. This impoverishment is present in all age groups before the appearance of Mammals and increases with age. Contrarily, young genes present only in Primates are enriched with coding deletions (**Figure 16 A and B**). This pattern is similar if we add gains to the analysis (**Supplementary figure 9**). Regarding intronic deletions we observe a quite flat pattern of enrichment, **meaning that genes of different evolutionary ages have similar densities of deletions within their introns** (**Figure 16C**).

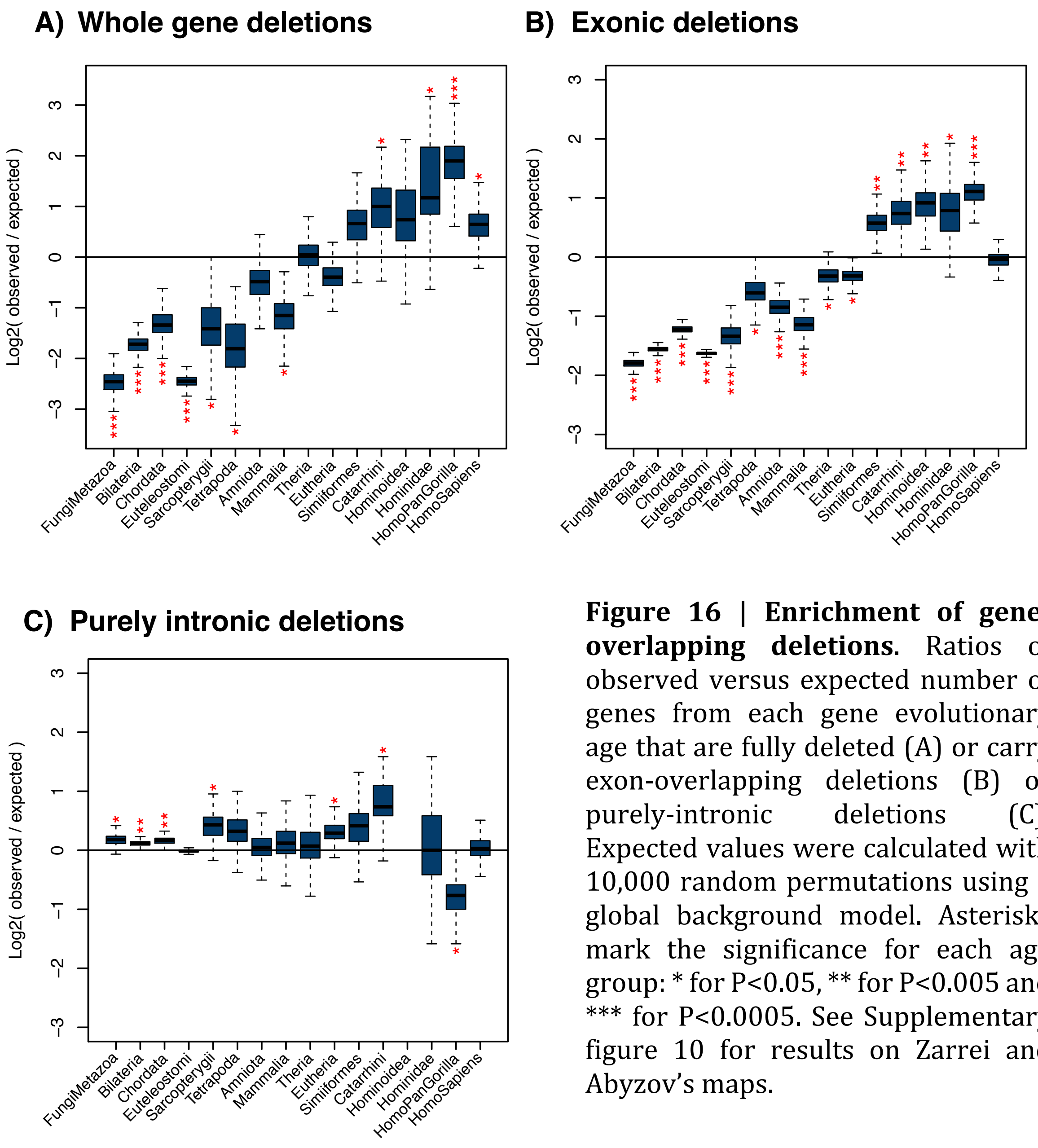


Figure 16 | Enrichment of gene-overlapping deletions. Ratios of observed versus expected number of genes from each gene evolutionary age that are fully deleted (A) or carry exon-overlapping deletions (B) or purely-intronic deletions (C). Expected values were calculated with 10,000 random permutations using a global background model. Asterisks mark the significance for each age group: * for P<0.05, ** for P<0.005 and *** for P<0.0005. See Supplementary figure 10 for results on Zarrei and Abyzov's maps.

Interestingly, **we can see that whole-gene CNVs cover genes that are shorter than the rest of the genes of their same age** (**Figure 17**).

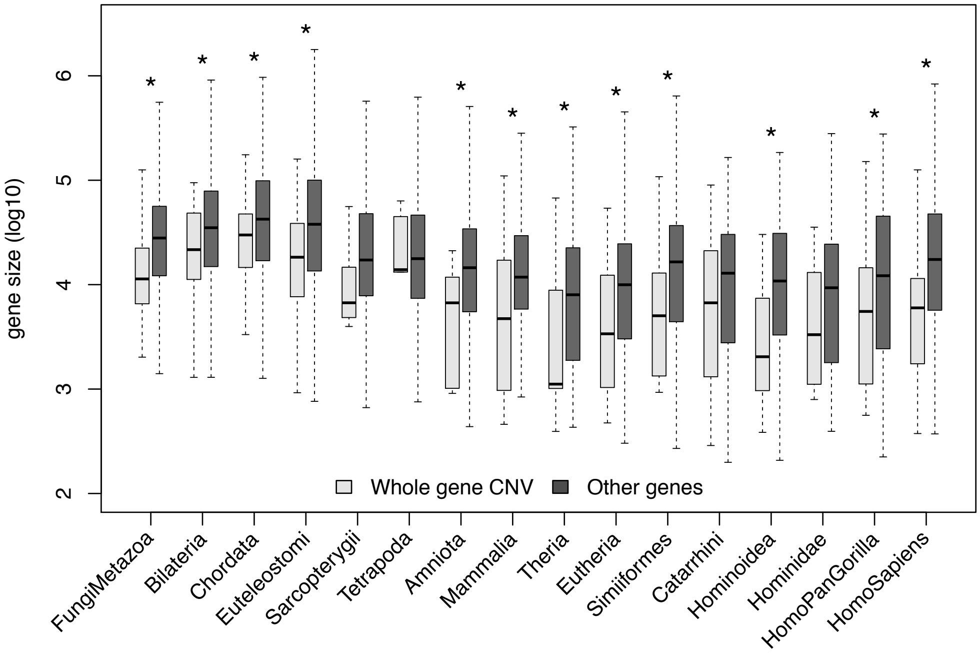


Figure 17 | Size of whole CNV-genes and non-variable genes by age. Sizes of all genes are represented, depending on if they are affected by whole-gene CNVs or have a fix copy number in the population. Differences in gene size are represented by age group and tested with Wilcoxon rank sum tests. Asterisks mark significance at P <0.05.

Intronic deletions in young genes are more frequent in the population

Depletion of CNVs on specific genes or regions suggests the presence of negative selection acting on variants occuring un such regions. We have seen that deletions occurring on an exon of an ancient gene, for example, seem to be under a stronger purifying selection. Thus, it is expectable to find such impoverished variants at lower frequencies in the population.

Indeed, if we look at the AFs of the all deletions overlapping coding sequences and classify them by the age of the gene, we see that deletions affecting ancient genes are less frequent in the population than the ones that overlap with young protein-coding sequences (**Figure 18**).

The AFs of intronic deletions are always significantly higher than those of deletions overlapping with exons of genes with a similar age (**Figure 18**), probably due to a stronger purifying selection acting on exonic deletions.

Surprisingly, intronic deletions, which showed similar patterns of enrichment regardless of gene age (**Figure 16C**), also have lower frequencies in ancient genes compared to young genes (**Figure 18**), suggesting that selection is not acting equally on intronic deletions on genes of different ages.

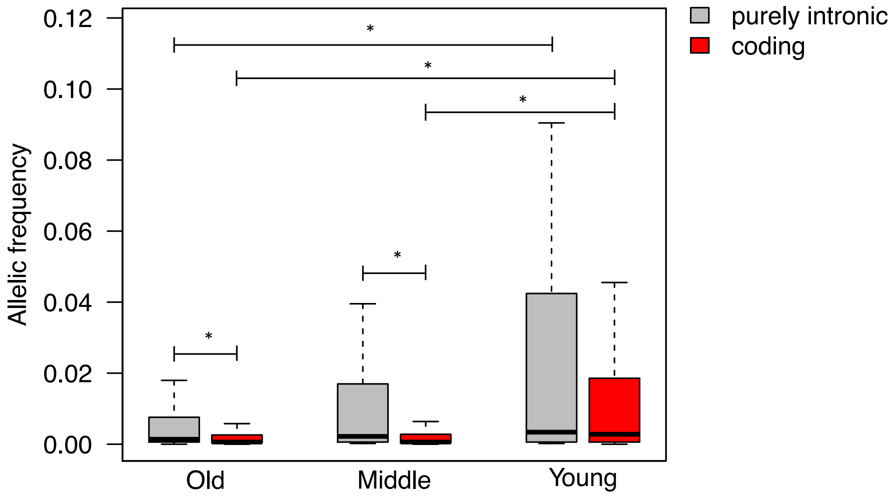


Figure 18 | Allelic frequencies of intronic and coding CNVs. AF of intronic or coding CNVs in genes of different evolutionary age groups. Gene ages are grouped as follows: “Old” genes (FungiMetazoa to Sarcopterygii), “Middle” (Tetrapoda to Eutheria) and “Young” (Simiiformes to HomoSapiens). Significant differnces are marked for P < 0.05.

Essential genes also show variable intron size

Variability of intron size in ancient genes was suprising because older genes are more frequently essential at the cellular or organismal level (**Figure 19A**). We wondered if the intronic variabilitiy of ancient genes was restricted to non-essential genes.

Essential tend to be more compact (to have shorter introns) than non-essential genes of a similar evolutionary age (**Figure 19B**).

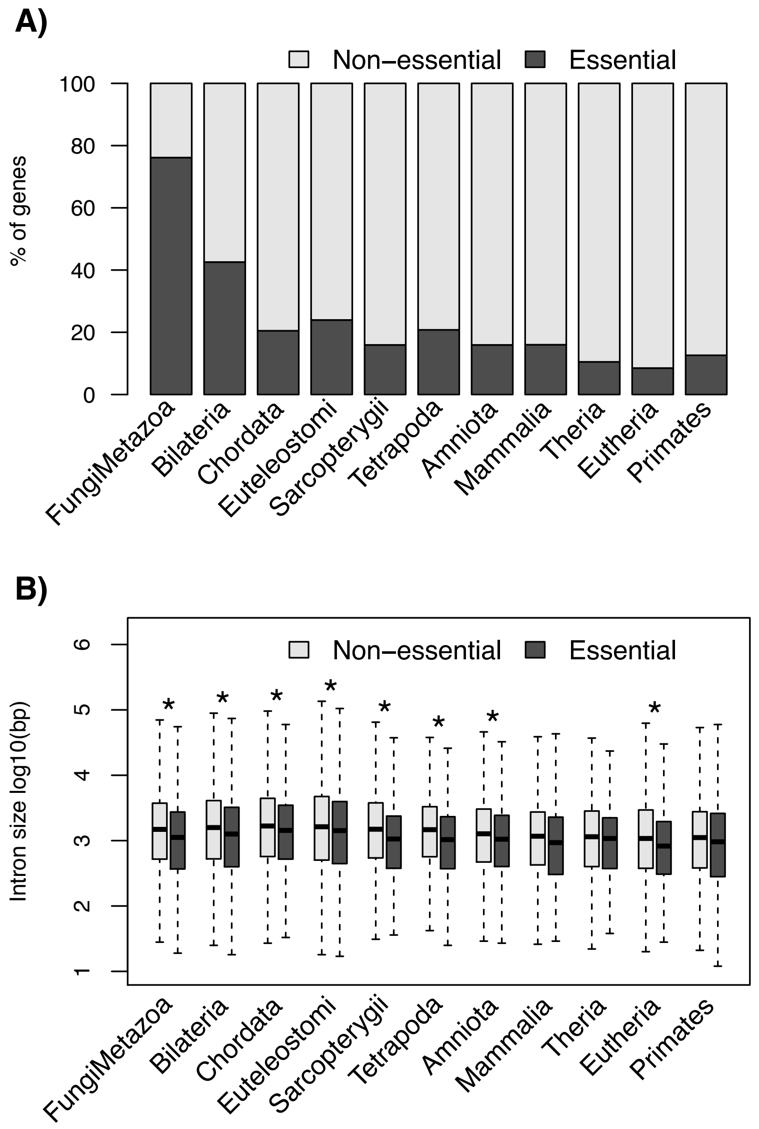


Figure 19 | Essential genes. (A) Percentage of essential genes in each age group. (B) Size of introns in non-essential and essential genes. Significant differences between the two types of genes in each age group were calculate using Wilcoxon tests and significance is marked with asterisks at P < 0.05.

As expected, essential genes are more depleted of whole-gene losses than the rest of the genes, and perhaps more surprisingly, essential genes are also more depleted of whole-gene gains than other genes (**Figure 20**).

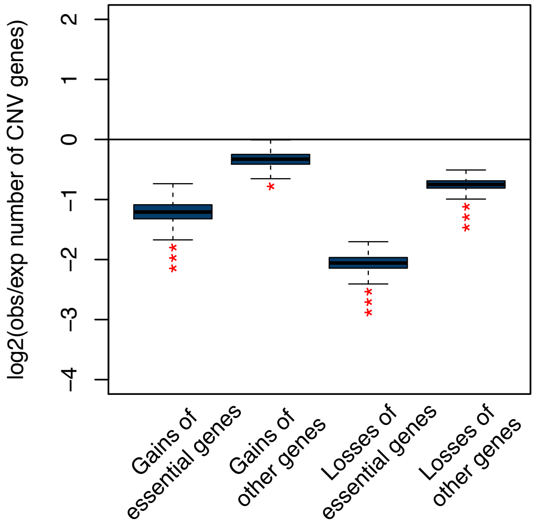


Figure 20 | Enrichment of whole-gene CNVs affecting essential or non-essential genes. Ratio of observed/expected total genes affected by whole-gene CNVs. Each box represents the number of observed CNV-genes divided by each of the expected number of CNV-genes(x10,000 randomizations). Asterisks mark significant depletions: * for P<0.05, ** for P<0.005 and *** for P<0.0005. Observed values per group: 48, 239, 43, and 296. Source of CNVs: Sudmant-Nature.

However, if we look at their introns we found that the number of essential genes with intronic deletions (907 genes in Sudmant-Nature) is also higher than expected by chance (*P* = 0.034) (See **Supplementary table 3** for results by map). Moreover, the AF of intronic deletions are slightly higher in essential genes, as opposed to coding deletions, which have lower AF when they affect essential genes (**Figure 21**).

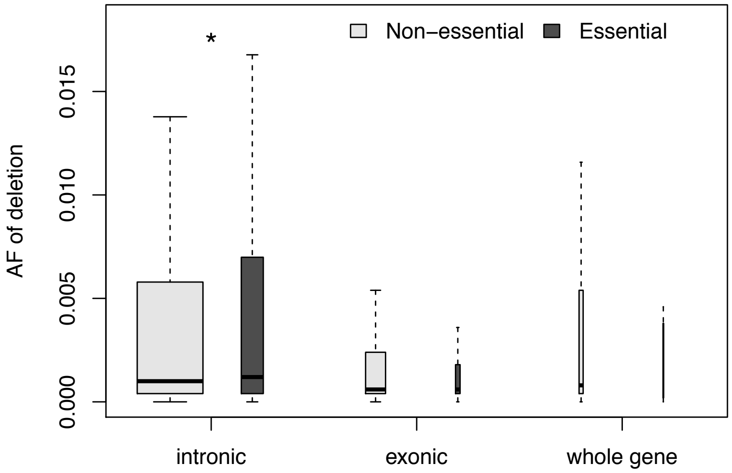


Figure 21 | Frequencies of deletions affecting essential genes. Allelic frequency (AF) of the deletions from Sudmant-Nature that overlap with non-essential and essential genes.

These results altogether show that, contrarily to what we expected, introns of essential genes do not seem to be under stronger purifying selection than introns of non-essential genes.

Characteristics of genes that do not show CNV variability in introns

Even if the introns of essential genes do not seem to be have less intronic deletions than other genes, we expect that there might be sets of genes intolerant to CNVs within the introns, for example if the size of the intron or of the gene is important for the proper functioning of the cell.

Intron length has been shown to be evolutionarily conserved or coevolving in some sets of genes. For example, genes related to embryonic development intron size seems to be especially conserved (Seoighe and Korir, 2011). To see if these sets of intron-conserved or intron-coevolving genes are also depleted of CNVs in the actual human populations, we ranked all genes according to their enrichment of intronic deletions, from more enriched to the most impoverished genes (see Methods for details and **Supplementary tables 4 and 5** for the lists of genes). We observed that **genes with a stronger depletion of intronic deletions show significantly more protein-protein interactions (PPI) among them than expected by chance** (P-value < 1.0e-16, calculated with STRING, v11 (Szklarczyk et al., 2015)), while the genes with more intronic deletions do not show this enrichment (P-value = 0.207) (**Table 4**).

	Genes with less deletions than expected	Genes with more deletions than expected
Number of nodes	469	480
Number of edges	889	542
Expected number of edges	626	523
Average node degree	3.83	2.26
PPI enrichment P-value	< 1.0e-16	0.207

Table 4 | Protein-protein interaction networks of genes with less or more deletions than expected. Statistics calculated using STRING. Low P-values (< 0.05) indicate that the network has more connections than expected for a random set of proteins of similar size.

Genes with more or less intronic deletions than expected show different levels of intolerance to functional mutations in their coding sequence. **Genes with less deletions have significantly lower RVIS scores** (P < 2-16)**, meaning that they are significantly more intolerant to functional mutations affecting their coding sequence**. According to these results, it seems that genes with a more conserved intronic sequence tend to also have less coding variability in the population, while genes with more variability within their introns also are more tolerant to coding variants (**Figure 22**).

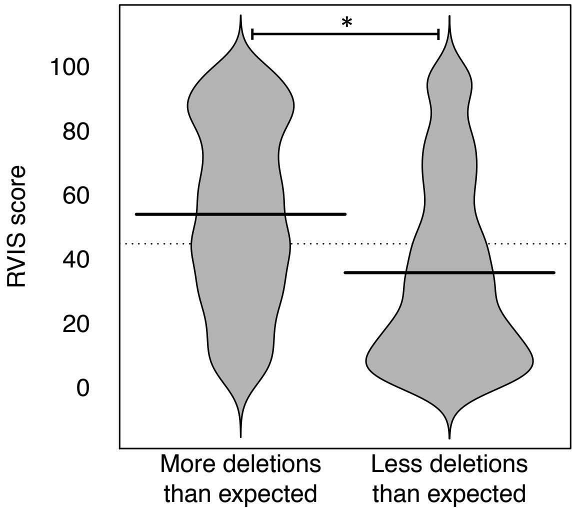


Figure 22 | Relationship between exon and intron conservation in the population. Comparison of the conservation level of the exons in genes with less conserved (more deletions than expected) or more conserved introns (less deletions than expected). Conservation is estimated using RVIS scores, which are inversely proportional to coding sequence conservation. Statistically significance was assessed using Wilcoxon tests (P-value = 1.28e-20).

However, it is surprising to see that among the genes with more deletions than expected there are genes with a very conserved coding sequence, some of which are associated with diseases (**Table 5**).

Gene name	Description	Diseases associated	RVIS percentile
CNOT1	Deadenylation-dependent mRNA decay and Gene expression	Iritis	0.46%
SETD1A	Histone methyltransferase involved in chromatin organization	Schizophrenia and cerebritis	0.74%
SCN3A	Generation and propagation of action potentials in neurons and muscle	Epilepsy	0.83%
NUP205	Active transport of proteins, RNAs and ribonucleoproteins between nucleus and cytoplasm.	Steroid-resistant nephrotic syndrome	1.02%
SCAP	Binds and mediates the transport of sterol regulatory element binding proteins.	Familial hyper-cholesterolemia	1.56%

Table 5 | Genes with highly variable introns and very conserved coding sequence. Genes with the lowest RVIS scores (high coding-sequence conservation) found among the list of genes with more intronic deletions than expected.

A possible explanation for having very conserved introns is to have a higher concentration of regulatory elements in them. Indeed, we find that **genes with less intronic deletions than expected have a significantly higher proportion of their introns occupied by regulatory features** (RFs) (**Figure 23**).

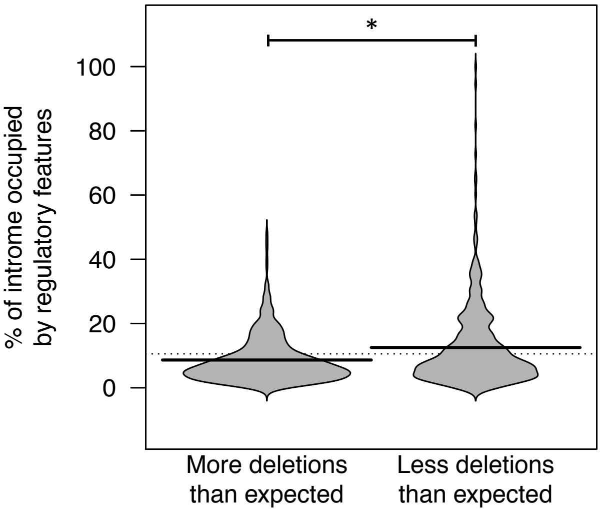


Figure 23 | Introme of a gene occupied by regulatory features. Percentage of the intronic regions of a gene that are covered by a RF. Statistical significance was calculated with Wilcoxon test (P-value = 1.04e-5)

We checked how these two gene groups are related to gene sets whose intron size has been shown to be important. Taking the sets of genes showing intron coevolution from the study by Keane and Seoighe (Keane and Seoighe, 2016), we found that 5 of 9 sets are significantly enriched with genes with less deletions than expected and significantly depleted of genes with more deletions than expected (**Figure 24**).

In genes that are activated in response to serum, the size of the gene has been suggested to regulate the order in which genes that are simultaneously induced finish transcription (Kirkconnell et al., 2017). We observe that genes with less deletions than expected were enriched in the set of serum-induced genes, and genes with more deletions impoverished, although none of the tendencies was significant (**Figure 24**).

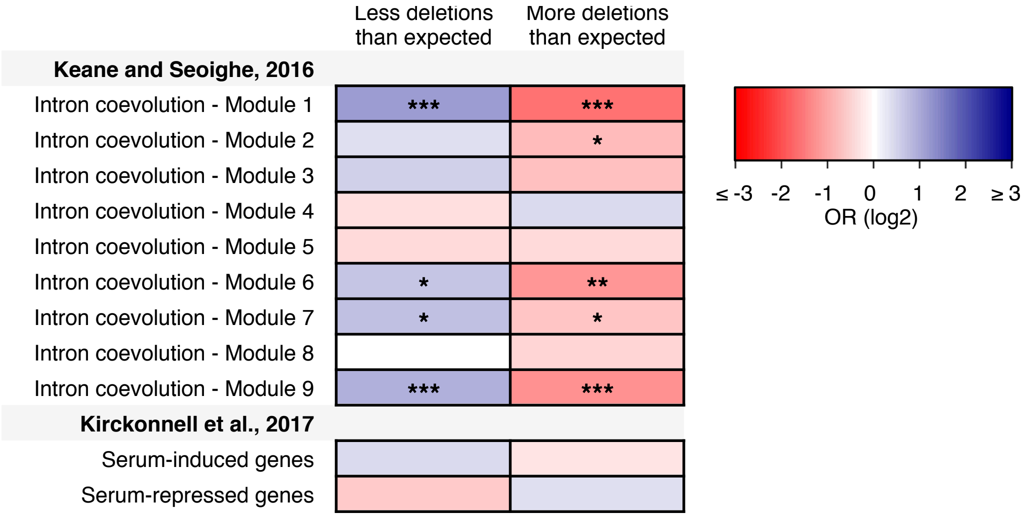


Figure 24 | Relationship between genes with introns depleted or enriched with deletions and sets of gene-length sensitive genes. Enrichment of the sets of genes with more or less intronic deletions than expected in sets of genes whose size has been claimed in previous studies to be under evolutionary pressure. Statistical significance was calculated using chi-squared tests. Color scale represents odds ratio and significance is marked for each set of genes with asterisks: * for P<0.05, ** for P<0.005 and *** for P<0.0005.

Genes with conserved introns are enriched in brain and developmental processes

Further gene set analysis using GO biological processes on the ranked list of genes according to their score of enrichment/depletion of intronic deletions, we found that the genes with less intronic deletions than expected are enriched in neuron recognition and somitogenesis, with FDR<5% (results consistent in the two randomizations). The enrichment in somitogenesis is in agreement with previous research showing the importance of gene length in processes involving oscillations in gene expression (Swinburne et al., 2008). Other biological functions significant at an FDR<25% included in **Table 6** are related to the detection of mechanical stimulus and to segmentation.

Genes with less intronic deletions than expected (lowest scores)		
	FDR q-val	
	Whole genome randomization	Within intron randomization
Neuron recognition	**0.0151**	**0.0239**
Somitogenesis	**0.0499**	**0.0094**
Nerve development	0.0611	**0.0000**
Forebrain cell migration	0.0591	0.1016
Positive regulation of axon extension	0.0541	0.0950
Adherens junction organization	0.0781	0.0561
Cerebral cortex cell migration	0.0834	0.2164
Detection of mechanical stimulus	0.0782	0.2103
Semaphorin plexin signaling pathway	0.0882	0.0871
Axon extension	0.1060	**0.0463**
Segmentation	0.1142	0.0861
Clathrin mediated endocytosis	0.1231	0.2074
Regulation of syaptic transmission – glutamatergic	0.2092	0.0540

Genes with more intronic deletions than expected (highest scores)		
	FDR q-val	
	Whole genome randomization	Within intron randomization
Ribonucleoprotein complex subunit organization	0.1964	0.1508
DNA templated transcription termination	0.2029	0.1292
Ribonucleoprotein complex biogenesis	0.1575	0.1297

Table 6 | Enriched processes in genes with more and less deletions than expected. Biological processes significant at FDR < 25% in both randomizations. Enrichment was calculated using GSEA on a preranked list of genes with the highest scores corresponding to the genes with more intronic deletions than expected and the lowest scores to genes with less intronic deletions than expected.

Genes expressed in the brain are among the longest genes in our genome. Some **pathogenic intronic CNVs** have been found associated with neurological or psychiatric disorders, annotated in the Copy Number Variation in Disease database (CNVD) (Qiu et al., 2012). We checked if these genes have their introns in general depleted of intronic deletions. The overlap with the genes with a more conserved intronic sequence was not significant (OR = 1.72, P = 0.26), but we observed that none of the 49 genes with pathogenic intronic CNVs carries deletions in the healthy population, even if most of them (68.4%) are found among the top 10% largest genes in the human genome.

In summary, by ranking all genes by their observed compared to expected content of intronic deletions, we conclude that **most gene sets whose intron sizes have been previously reported to be more conserved or to coevolve seem to have less deletions than expected in the actual human population**. We have found that, in addition to these previously described gene sets, brain-specific genes also seem to have more conserved introns.

Relationship between intronic deletions and regulatory features

Introns are enriched with regulatory features

Several studies have identified regulatory elements hosted in introns (Chorev and Carmel, 2012) in different genes. We retrieved the data from the Ensembl Regulatory Build (Zerbino et al., 2015), which provides a genome-wide set of regions that are likely to be involved in gene regulation, to better understand the association of the different types of RFs with introns. We found that introns are enriched with RFs such as promoters, enhancers, promoter flanking regions or transcription factor binding sites, compared to our (global) random background model (**Figure 25**).

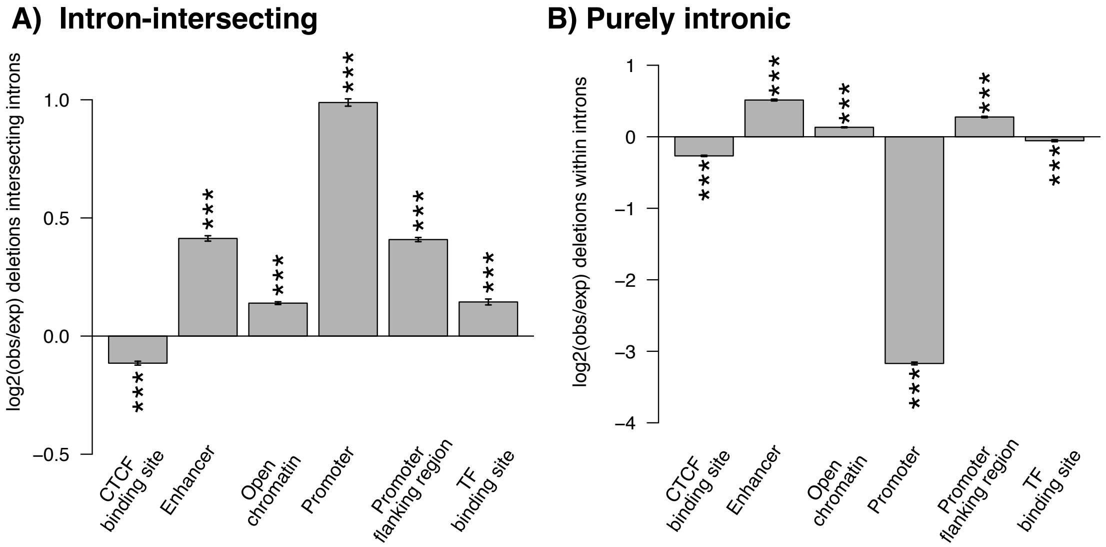


Figure 25 | Enrichment of regulatory regions in introns. Ratio of observed versus the median of expected number of regulatory regions overlapping (A) or completely falling within an intron (B). Error bars denote median absolute deviation and asterisks mark significance: * for P<0.05, ** for P<0.005 and *** for P<0.0005.

Deletions and RFs tend to be found in the same intron but overlap less than expected

Since deletions and of most types of RF are all enriched in introns, we wondered how frequently these two elements coocurred within a same intron. We observed that intronic deletions tend to be found in introns that also contain RFs (**Table 7**). This greater than expected coocurrence can probably be explained by fact that most deletions occur within longer introns, which are the ones harboring most RFs (**Supplementary figure 11**).

	Sudmant (Nature)	**Zarrei**	**Abyzov**	**Handsaker**	**Sudmant (Science)**
Enhancer	7.44 (<1e-100)	6.1 (<1e-100)	5.93 (<1e-100)	8.44 (2.57e-83)	9.16 (<1e-100)
Promoter	1.95 (1.41e-51)	1.7 (5.23e-19)	1.75 (5.06e-18)	2.11 (4.33e-10)	2.38 (4.66e-15)
Promoter Flank. Reg.	4.64 (<1e-100)	4.54 (<1e-100)	4.07 (<1e-100)	5.23 (1.05e-69)	5.67 (3.07e-86)
TFBS	5.02 (<1e-100)	5.37 (<1e-100)	4.69 (<1e-100)	6.81 (2.05e-63)	6.65 (1.31e-67)
Open chromatin	7.59 (<1e-100)	7.1 (<1e-100)	6.96 (<1e-100)	12.03 (<1e-100)	13.5 (<1e-100)
CBS	4.58 (<1e-100)	4.73 (<1e-100)	4.34 (<1e-100)	6.72 (1.88e-91)	6.31 (3.78e-93)
All RFs	5.58 (<1e-100)	5.55 (<1e-100)	4.85 (<1e-100)	9.41 (<1e-100)	11.45 (<1e-100)

Table 7 | Coocurrence of deletions and regulatory features in introns. Each cell shows the Odds ratio and, in brackets, the P-value, calculated using Fisher's test.

But, do the deletions and RFs coocurring within a same intron overlap? To calculate this, we randomly relocated 10,000 times all intronic deletions within their host introns. Then, we compared the number of observed overlaps between deletions and RFs with the number of overlaps obtained after randomization. The results showed that the overlap between deletions and **enhancers** is significantly lower than expected overlap with deletions in 4 out of 5 maps (**Table 9**). For **promoters**, **promoter-flanking regions** and **CTCF Binding Sites (CBS)**, this tendency to not overlap was also detected but in less maps (2 out of 5). The overlaps between deletions and **Transcription Factor Binding Sites (TFBS)** and **open chromatin** were similar to those expected by chance in all maps..

These results show that, despite the high prevalence of deletions within introns and their tendency to occur in introns that also contain RFs, **deletions tend to be found elsewhere in the intron, without overlapping with RF, especially with enhancers**.

	Sudmant (Nature)	**Zarrei**	**Abyzov**	**Handsaker**	**Sudmant (Science)**
Enhancer	**-0.14 (0.03)**	**-0.45 (0.001)**	**-0.48 (0.004)**	-0.05 (0.406)	**-0.34 (0.021)**
Promoter	**-0.74 (0.001)**	**-0.68 (0.011)**	-0.50 (0.113)	-1.32 (0.066)	-0.19 (0.378)
Promoter-Flanking Region	**-0.12 (0.019)**	-0.11 (0.105)	**-0.34 (0.002)**	0.04 (0.375)	0.06 (0.363)
TF Binding Site	-0.09 (0.191)	-0.05 (0.391)	-0.05 (0.42)	-0.21 (0.192)	-0.26 (0.132)
Open chromatin	-0.02 (0.33)	-0.12 (0.052)	-0.12 (0.103)	-0.08 (0.186)	0.06 (0.25)
CTCF binding site	**-0.18 (0.007)**	-0.14 (0.11)	**-0.81 ($<10^{-4}$)**	-0.16 (0.189)	-0.18 (0.146)

Table 9 | Overlap with regulatory regions. Relative enrichment or depletion of overlaps between deletions and each type of regulatory features, calculated by comparing the number of observed overlaps to a background model. Values show log2(Observed/Expected deletions overlapping with a RF) and the p-value in brackets. The background randomizations were performed by relocating each intronic deletion within its intron 10,000 times.

The significant depletion of overlaps between deletions and RFs suggests a negative selection on the losses of intronic RFs. This finding implies that **intronic losses occur more often in the regions of the intron that do no have regulatory function.**

In cancer, regulatory elements are not depleted of intronic deletions

We wondered if somatic copy number alterations (SCNAs) in cancer show the same patterns of overlap with RFs. We took all deletions from 2583 patients **from the Pan-Cancer project** (Campbell et al., 2017) **and, in this case, we found that the overlap between deletions and regulatory regions is similar to what is expected by chance** (**Table 9**), in contrast with the results obtained with germline deletions from healthy individuals.

	Pan-Cancer SCNAs
Enhancer	0.01 (0. 368)
Promoter	-0.044 (0.38)
Promoter Flanking Region	-0.023 (0.123)
TF Binding Site	-0.016 (0.309)
Open chromatin	-0.014 (0.132)
CTCF binding site	-0.004 (0.439)

Table 9 | Overlap of SCNAs with regulatory regions. Relative enrichment or depletion of overlaps between somatic deletions and each type of RF, calculated by comparing the number of observed overlaps to a background model. Values show $\log_2$(Observed/Expected deletions overlapping with a RF) and the p-value in brackets. The background randomizations were performed by relocating each intronic deletion within its intron 10,000 times.

Intronic TFBSs active in more tissues are more depleted of deletions

We hypothesized that RFs active in more cell types would show lower overlaps with deletions than tissue-specific RFs, as we expected that the disruption of a widely used regulatory element would have an impact on more tissues. We classified each type of RF by the number of tissues in which they are active.

Surprisingly, only the **overlaps between TFBS or CTCF and deletions are more strongly depleted when the number of tissues in which the RF is active** (**Figure 26**). However, given the number of deletions overlapping with TFBS is very limited and thus, the results should be interpreted carefully.

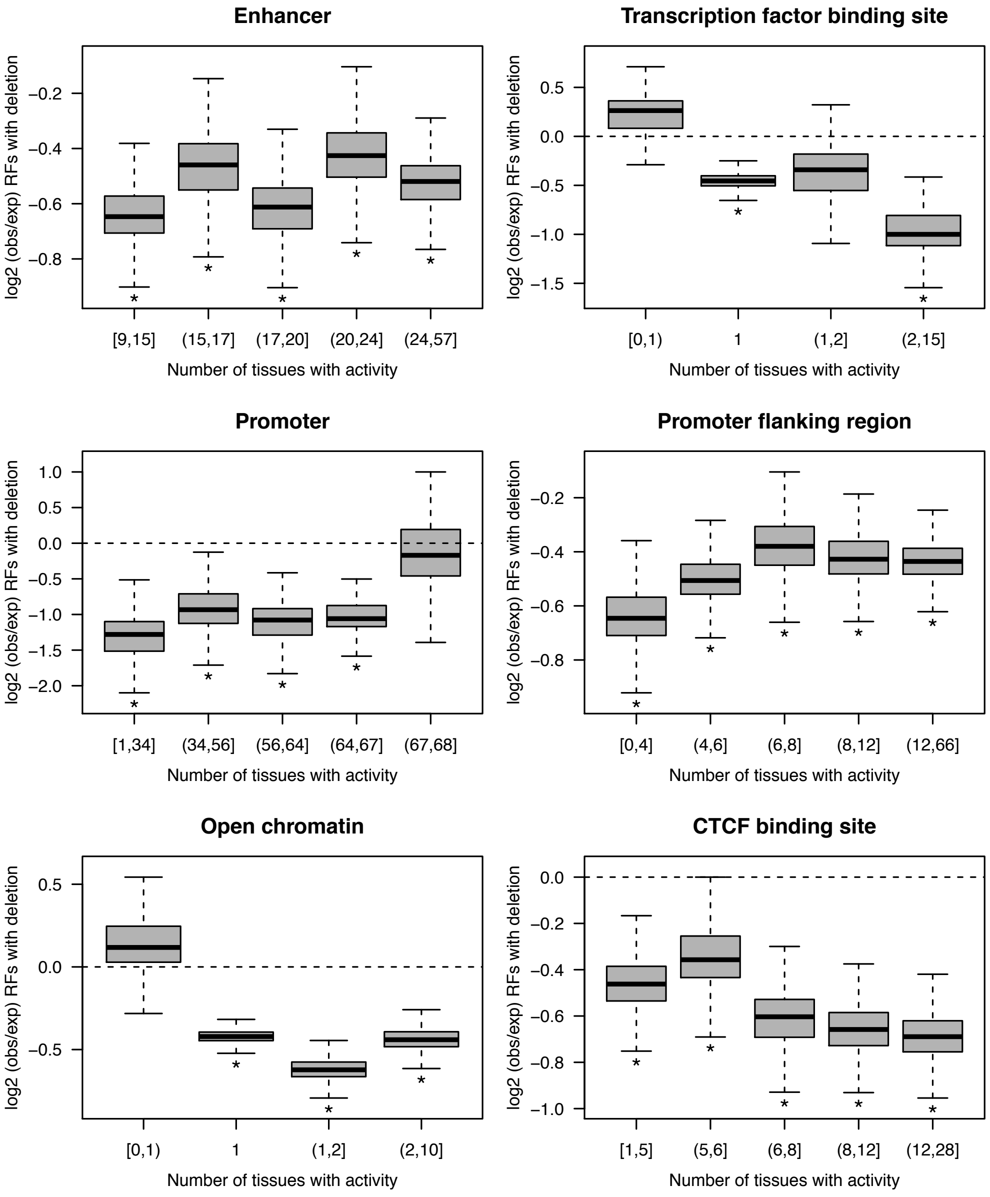


Figure 26 | Overlap between deletions and RFs active in different number of tissues. Each box shows the log2 ratio between the observed and the expected number of RFs overlapping with an intronic deletion. Expected values were calculated by relocating each deletion within the host intron 10,000 times. Asterisks mark significanct differences between observed and random values, at $P < 0.05$. The median number of regions per box in each RF type and, in brackets, the number of median number of RFs with an overlapping deletion are: Enhancer = 2672 (70), Transcription factor binding site = 1034 (16.5), Promoter = 2226 (9), Promoter flanking region = 4568 (116), Open chromatin = 8397 (213), CTCF binding site = 3237 (64).

Association between CNVs and gene expression changes

Intronic deletions are associated with changes in gene expression

CNVs have been shown to be associated with gene expression levels: the higher the number of copies of the gene, the more it is expressed (Handsaker et al., 2015; Sudmant et al., 2015a). We explored if intronic CNVs could also affect the expression of genes without altering the dosage of their coding sequence. To do this, we took RNA-seq data from the Geuvadis project (Lappalainen et al., 2013) including 445 lymphoblastoid cell lines from individuals from the 1KGP with CNV data (Sudmant-Nature's map).

We compared the gene expression levels of wild-type individuals (copy number = 2) with that of individuals with a deletion in one of the alleles (copy number = 1) to identify deletions associated with gene expression changes. We selected the deletions from Sudmant-Nature's dataset that were present in at least two wild-type (diploid) and two heterozygous individuals and classified them in different groups according to the impact on the gene: whole gene, exonic and intronic. From now on, we will refer to the deletions associated with gene expression changes as "eDeletions" and to the differentially expressed genes as "eGenes".

To compare the impact of intronic deletions with that of coding deletions, we first tested 45 whole gene deletions overlapping with 50 genes and 472 exonic deletions that affected a total of 437 genes. Seven out of the fifty whole-gene deletions (14%) were associated with a significantly lower gene expression in the individuals carrying the deletion. The 7.6% of exonic deletions (36 out of 472) were associated with gene expression changes, most of them also downregulations (91.4%) (**Table 10**).

In relation to the 2046 intronic deletions found in 1505 genes, we detected that the 2.7% were associated with a differential gene expression in the heterozygous individual. Interestingly, in intronic eDeletions, the proportion of downregulated genes was lower: 68% of the eGenes had lower expression in the group that carried the eDeletion while the remaining 32% had a higher expression. These results suggest that, while coding losses mostly associate to gene down-regulation**, intronic deletions might result in both gene expression repression and enhancement**.

	Whole gene	Exonic	Intronic
Number of eGenes	7***	35***	53***
Number of eDeletions	8	36	56
Expected number of eGenes (median ± MAD)	1 ± 1.48	8 ± 2.97	27 ± 5.93
% of downregulated genes	100%	92.5%	68%
Total genes tested	50	437	1505
Total deletions tested	45	472	2046

Table 10 | Differentially expressed genes. Number of DEGs in association with whole gene, exonic or intronic deletions. The expected number of DEGs was calculated after randomly shuffling the genotype of the subjects for whom we had gene expression data. MAD = Median Absolute Deviation. A list of all eGenes can be found in supplementary material (**Supplementary table 6**).

No significant differences exist among the effect size of the different types of eDeletions. The median effect size is higher in whole-gene eDeletions, but intronic eDels present more variable effect sizes, with some cases showing very strong effect sizes (**Figure 27**).

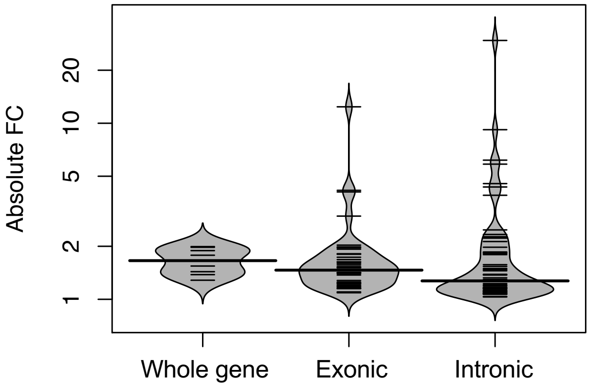


Figure 27 | Effect size of different types of eDeletions. Absolute log2 ratio between the median gene expression of wild-type versus heterozygous individuals. No significant differences were detected using Wilcoxon tests.

Taken together, these results show that **most deletions associated with gene expression changes are intronic.** While coding deletions are almost always associated with downregulation, **intronic eDeletions are associated with both up and downregulation**.

eDeletions frequently overlap with regulatory features

We previously showed that intronic deletions are preferentially located in non-regulatory regions of the intron. We wondered if the eDeletions overlap more or less with enhancers than the deletions not associated with changes in gene expression. We found that 15 intronic eDeletions overlap with enhancers active in B-lymphocytes (the cell-type used to obtain lymphoblastoid cell lines, for which we have gene expression data). This number of eDeletions overlapping with active enhancers is higher than expected (P = 0.023, odds ratio = 2.04, Fisher's test) and corresponds to the 24% of deletions overlapping with active enhancers in B-lymphocites. However, there are 422 other intronic deletions overlapping with enhancers active in other tissues that may be eDeletions in these other tissues but not in the cell-type of our study.

Among the deletions that do not overlap with enhancers, we found that eDeletions tended to be closer (in linear distance) to an enhancer than other deletions not associated to changes in expression (P = 9.2e-04, Student's t-test). **This suggests that disrupting sequences proximal to enhancers could be affecting regulatory interactions without removing the enhancer region itself**.

It is known that regulatory regions are preferentially located in first introns (Chorev and Carmel, 2012). We checked if the intronic eDeletions are also preferentially found in the first introns. We found that 17 (30.4%) of the eDeletions are found in first introns, but this percentage is not significantly higher than that of the remaining (non-significant) intronic deletions (26%, P = 0.54, Fisher's test).

Intronic deletions are associated with changes in the expression of distant genes

Gene expression regulation can happen through the interaction of distant fragments of DNA, which are brought close to each other by chromatin looping (Vermunt et al., 2019). The interacting fragments can be separated by up to over a megabase, and a fragment of DNA can have contacts with different distant fragments. Thus, genes can be regulated by multiple enhancers, and different genes can be under the control of the same enhancer looping (Vermunt et al., 2019).

We wondered if deletions of one of the two fragments in contact could have an impact on the expression of the target gene. To assess whether deletions could have this *trans* effect, we used promoter-capture Hi-C (PCHi-C) published data (Javierre et al., 2016) to link deleted regions with promoters of other genes. The Hi-C data has been derived from B-lymphocytes, the cell-type used to obtain lymphoblastoid cell lines (the cell-type for which we had expression data), which we assumed have a similar chromatin conformation.

We identified 867 deletions in regions that interact with gene promoters of other genes. We analysed separately intronic and intergenic deleted regions (**Figure 28**).

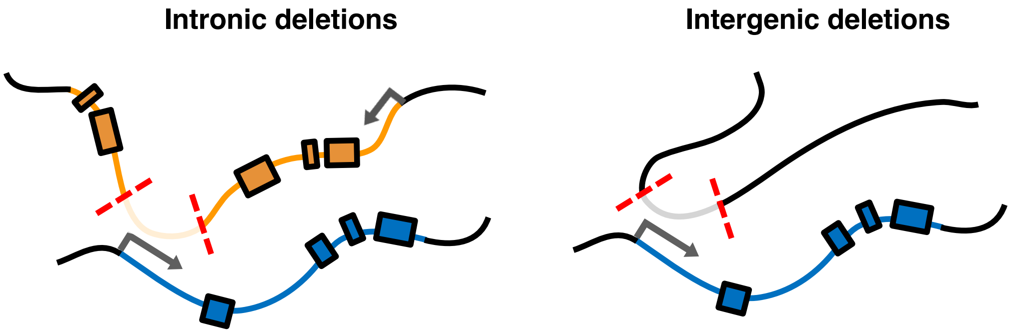


Figure 28 | Deletions with a potential impact in trans. Schematic representation of intronic (A) or intergenic (B) deletions of fragments interacting with promoters from another gene.

The analysis of all possible combinations between 322 intronic deletions and the 672 genes they were in contact with (a total of 758 deletion-promoter interactions) revealed 12 genes that were significantly differentially expressed in the individuals presenting an intronic deletion in another gene in contact. 16 additional eGenes were found in association to 18 intergenic eDeletions, out of the 545 intergenic deletions of fragments in contact with a promoter. The proportion of deletion-gene pairs that are associated with significant changes in expression is similar to that of intronic *cis* deletions (3.7 % of intronic in *trans*, 2.9 of intergenic and 2.7% of intronic in *cis*) (**Table 11**).

In both types of eDeletions (*trans*-intronic and *trans*-intergenic) we found cases of higher and lower expression (**Table 11**), in agreement with previous research that shows that chromatin architectural changes are coupled to both activation and repression (Vermunt et al., 2019).

	***trans*-intronic**	***trans*-intergenic**
Number of eGenes	12	16
Number of eDeletions	12	18
Expected number of eGenes (median ± MAD)	9 ± 2.97	14 ± 4.45
% of downregulated genes	58%	83%
Total genes tested (total deletions tested)	672	1011
Total deletions tested	322	545

Table 11 | Differentially expressed genes in *trans*. Number of eGenes in association with deletions in trans, located in an intron of another gene or in an intergenic region. The expected number of eGenes was calculated after randomly shuffling the genotype of the subjects for whom we had gene expression data. A list of all DEGs can be found in supplementary material (**Supplementary table 6**).

For example, we found *PRSS36* (Protease, Serine 36) to be downregulated in individuals with an intronic eDeletion in *SETD1A* (SET Domain Containing 1A) gene (P = 1.98e-02), while *LIAS* (Lipoic Acid Synthetase) gene is upregulated in individuals with a intronic eDeletion in *PDS5A* (PDS5 Cohesin Associated Factor A) (P = 1.53e-06).

Two interesting cases of intergenic eDeletions are the CDO1 gene (associated with higher expression) and two components of the major histocompatibility complex (MHC), HLA-DPA1 and HLA-DQA1 (associated with lower expression).

The *CDO1* (Cysteine Dioxygenase Type 1) gene has an important role in regulation of cellular cysteine concentrations and it initiates many metabolic pathways. Hypermethilation in the promoter of this gene (typically a repression mark) is a molecular diagnostic and a prognostic indicator in various human cancers (Nakamoto et al., 2018). We found that a deleted fragment located at 179.1kb from the *CDO1* gene is associated with higher expression, suggesting a repressor role of the deleted fragment in 3D contact.

HLA-DQA1 and *HLA-DPA1* are two of the 6 main MHC class II genes. These genes play a central role in the immune system and variation in these genes have been associated with several disorders including type 1 diabetes, oral cancer and celiac disease. We found that both genes have an intergenic deletion of a fragment in contact with their promoter associated with a lower expression of the gene. In both cases the deleted fragment is upstream of the gene (10.1kb and 3.7kb, respectively). In this case, the contacting fragment seems to be an enhancer and by removing this contact the gene will be downregulated (**Figure 29**).

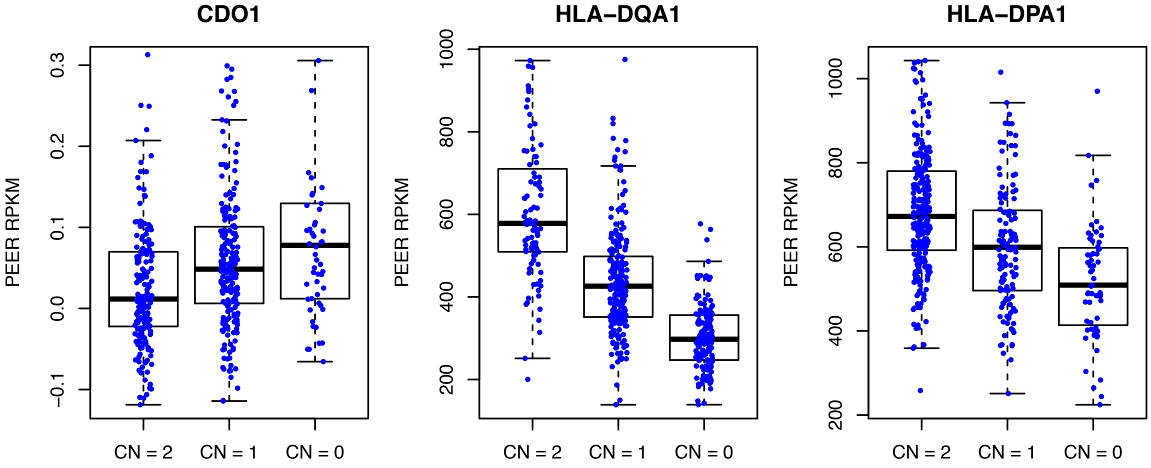


Figure 29 | Gene expression of genes associated with intergenic deletions. Example of a gene with increasing expression in the individuals with an intronic deletion in one or both alleles (CDO1 gene) and of two genes with lower expression in the individuals with an intronic loss (*HLA-DQA1* and *HLA-DPA1*).

In general, deletions of fragments in contact with promoters seem to have a low impact on the expression of the gene in contact, lower than that of intronic deletions in cis (**Supplementary figure 12**). Nevertheless, we have shown that the expression of some genes is strongly associated with very distant eDeletions.

Most non-coding eDeletions are found in ancient genes

We looked at the ages of the of the eGenes and we found that ancient genes are more frequently associated with non-coding eDeletions than to coding eDeletions, and the opposite happens in young genes (**Figure 30**).

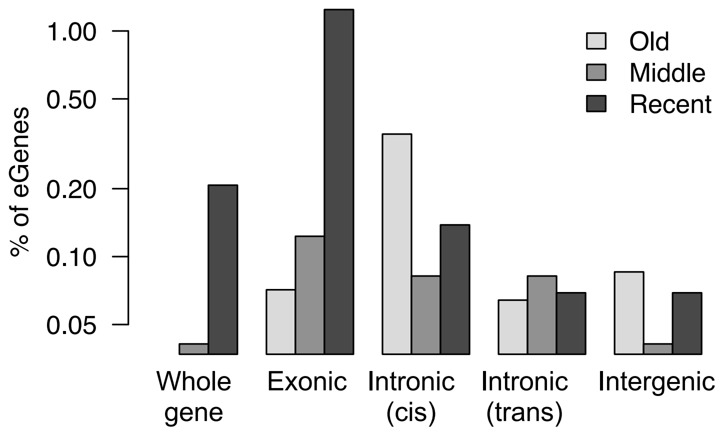


Figure 30 | Differentially expressed genes by gene age and deletion type. Each bar represents the percentage of genes in old, middle-aged or recent genes whose expression is associated with whole-gene, exonic, intronic (in cis or in trans) or intergenic deletions.

If we look at the tolerance to coding mutations of the eGenes associated with whole-gene eDeletions, these eGenes have very high tolerance to mutations (based on their high RVIS scores), while *cis* and *trans* intronic and intergenic eDeletions are associated with eGenes that have very variable tolerance score (**Figure 31**), including some genes with very low RVIS scores (low tolerance). Strikingly, eGenes in association with *trans*-eDeletions are the ones with a lowest RVIS score, maybe pointing at genes whose coding and non-coding sequence is very conserved but whose gene expression can be altered, for example, through changes in the chromatin structure.

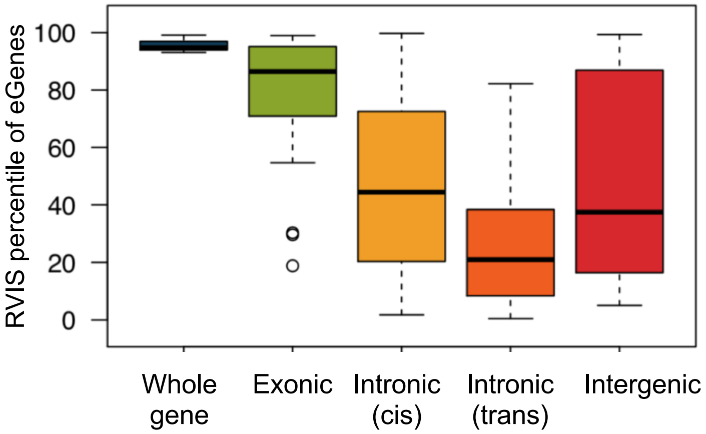


Figure 31 | Coding-sequence conservation of DEGs. RVIS score for coding sequence conservation in eGenes associated with whole-gene, exonic, intronic (in cis or in trans) or intergenic eDeletions. Numbers in the y-axis correspond to the percentile where a gene is located after a ranking based on their protein-coding sequence conservation, with lower values showing more conservation than high values.

Population stratification of deletions associated with DE

Population stratification of CNVs can indicate that a locus is under adaptive selection (Sudmant et al., 2015a, 2015b). To explore population differentiation we used the statistic Vst, a measure that estimates the proportion of variance that is attributable to variation between populations and not within populations (Redon et al., 2006) (See Materials and methods for details). 352 gene-overlapping deletions from Sudmant-Nature appear to be highly stratified (Vst > 0.2). 282 of them are intronic, 53 exonic and 17 whole gene. Surprisingly, the percentage of highly stratified deletions in each of the three types is similar, but not uniformly distributed across gene ages. Young genes have a higher proportion of highly stratified deletions than ancient genes, even for intronic deletions (**Figure 32**).

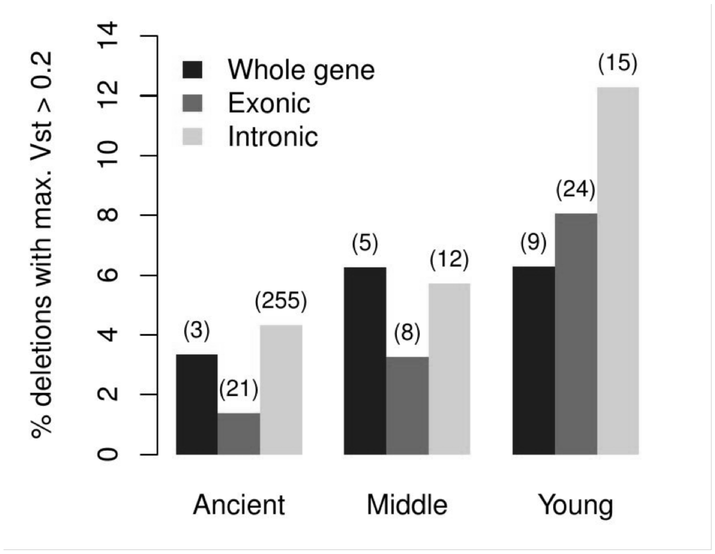


Figure 32 | Population stratification of deletions associated with differential gene expression. Percentage of highly stratified variants (maximum Vst > 0.2) in each age group and by type of overlap with the gene. The absolute number of deletions is indicated above each bar.

We found that four of the intronic deletions associated with gene expression changes in cis are highly stratified and located in four ancient genes (Sarcopterygii or older): *EXOC2*, *SKAP2*, *PTGR1* and *PHYHD1*. *EXOC2* appears among the 5% more conserved genes (RVIS = 3.34). It is possible that we have detected intronic deletions that cause a variability in the gene's expression that contributes to human adaptation and that, in some cases, the variant ends up being positively selected in some populations.

Impact of deletion size and GC content on exon inclusion and transcript differential expression

Intronic deletions are associated with transcript differential expression

The size of the intron can have an impact on the inclusion or exclusion of exons during the process of splicing (Roy et al., 2008). This differential inclusion of exons will result in differences in isoform expression.

We hypothesized that differences in intron size in the population would result in the presence of differentially expressed transcripts (eTranscripts) in the individuals with different intron sizes. By comparing the expression levels of each transcript individually in wild-type and homozygous individuals, we identified 185 genes with at least one eTranscript (**Table 12**), in addition to the previously identified eGenes. Most of the eTranscripts that we detected corresponded to alternative isoforms (174 out of 217).

	Whole gene	Exonic	Intronic (cis)	Intronic (trans)	Intergenic
Number of eTranscripts	22***	135***	217*	81	123
Number of eDeletions	11	92	199	54	96
Expected number of eTranscripts (median ± MAD)	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 genes with ≥1 eTranscript (median ± MAD)	4 ± 1.48	53 + 7.41	143 ± 14.83	64 ± 8.90	94 ± 10.38
Downregulated eTranscripts	100%	91%	79%	81%	89%
Total genes tested	47	403	1,401	653	972
Total deletions	43	440	1,886	319	529

Table 12 | Genes with differentially expressed transcripts. Observed and expected numbers of DETs and of genes with at least one DET in association with deletions whole-gene, exonic or intronic deletions in cis or with deletions in trans(located in an intron of another gene or in an intergenic region). The expected number of DETs was calculated after randomly shuffling the genotype of the subjects for whom gene expression data was available.

These results suggest that intronic deletions can alter the the expression of whole genes or unbalance the expression of one or more isoforms of a gene.

Intronic deleted regions are GC rich and happen in introns with a high exon-intron differential GC content

Different studies have shown that that exons flanked by larger introns are more likely to be alternative spliced than exons flanked by short introns (Fox-Walsh et al., 2005; Kim et al., 2007; Roy et al., 2008). Further bioinformatics and experimental analyses have proved that manipulating the size of the intron can affect the patterns of exon inclusion or exclusion (Amit et al., 2012).

Based on these previous findings, we hypothesized that altering the length of introns could change the level of exon inclusion. Specifically, we expected that shortening of an introns would increase the inclusion of alternative exons located upstream or downstream. However, the impact of changes in intron size on exon inclusion/exclusion is different depending on the GC content structure of the gene. Amit and others showed that introns with a higher GC differential between exon and intron tolerate better changes in intron length and the levels of inclusion or exclusion are less affected (Amit et al., 2012).

A first analysis on the GC content of the intronic deletions showed that, in general, **deleted intronic fragments (intronic deletions) show higher GC content than the rest of the intron** (P = 2.54e-18, paired Student's t-test). Moreover, the removal of these sequences causes a significant drop in the relative GC content of the introns (P = 2.23e-16, paired Student's test).

In agreement with the findings by Amit and others, we see that the GC content differential between exons and introns is higher in the introns carrying deletions (**Figure 33**). This consistency with the previous study is only observed in the largest introns. However, as mentioned before, the top 10% largest introns harbor the 61% of all intronic deletions .

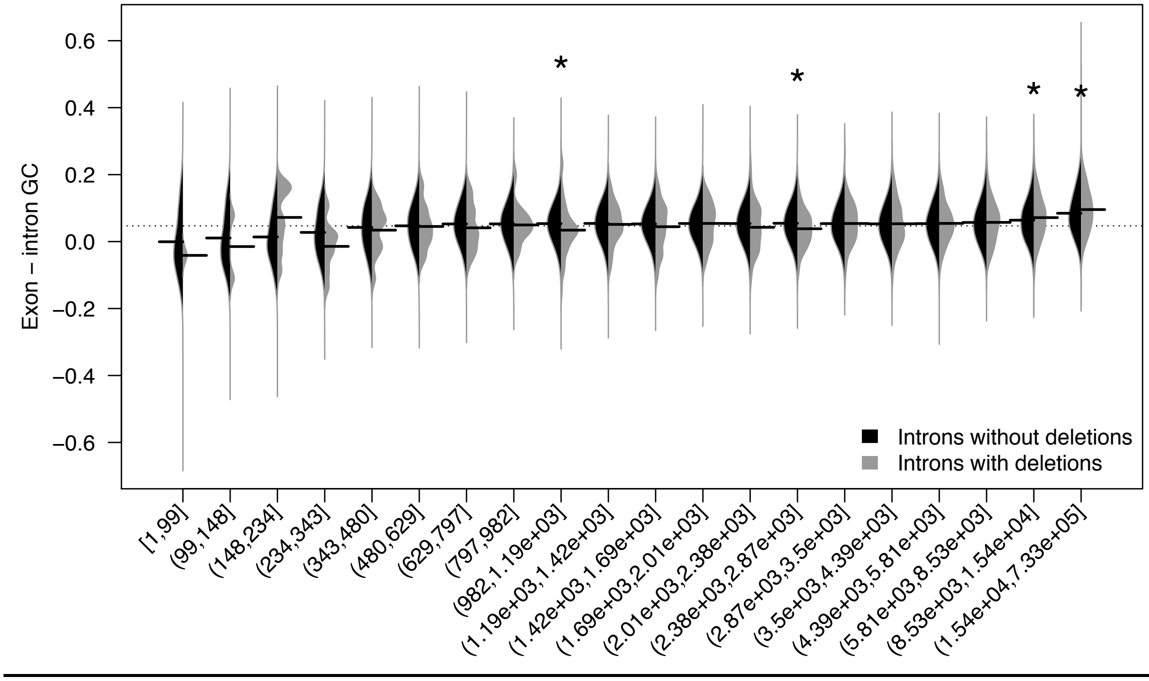


Figure 33 | Exon-intron differential GC content. Difference between the flanking exons of introns with or without deletions. Bean lines show the mean values of each side of the bean and overall line represents the average of all values. Significance calculated with Wilcoxon tests and marked with asterisk at P < 0.05. Note: The relative GC content of the introns with deletions does not take into account the deleted bases.

These results show that **most intronic deletions happen in introns genes a higher exon-intron GC content differential than that of other introns of similar size**, possibly because This results can be interpreted in agreement with Amit and others' findings that suggest that changes in intron size affect less the patterns of exon inclusion/exclusion when the GC differential is higher. Also we hypothesize that the presence or absence of regions with a high GC content within introns is less troublesome for the splicing machinery if the exons are more well defined by a high GC content.

Losses of large or GC-rich intronic fragments are associated with changes in exon inclusion and exclusion

We presumed that the effect of changes in intron length or GC content on exon inclusion or exclusion can be tested by checking if there are any changes in the expression of the exons flanking the intron that contains a deletion.

We looked at the differential expression of the alternative exons flanking intronic deletions and found 49 eDeletions (2.12 % of all tested) associated with changes in expression of the downstream exon (**downstream-eExon**) and 28 eDeletions (1.56% of all tested) associated with differential expression of the upstream exon (**upstream-eExon**) (**Table 13**).

Most (63.3%) of the **downstream-eExons** had lower expression (were more often excluded in the in the individuals with the deletion), while most (75%) **upstream-eExon** were more expressed (were more often included in the individuals with the deletion). From all eDeletions, 9 were linked to both upstream and downstream eExons. In 7 of these 9 cases, the eDeletion was associated with higher inclusion of both exons; in one case, with exclusion and in the remaining one with an one excluded and one included exon.

We tested how all these eDeletions related with intron size and GC content. Because it is not fully understood if the changes in size of the upstream or of the downstream introns is more important in exon inclusion/exclusion (Roy et al., 2008), we analysed separately the eDeletions associated with upstream and downstream-eExons (**Table 13**).

The results showed that **eDeletions are bigger** (P = 1.21e-4 and P = 1e-3 for down and upstream-eExons, respectively) **and they represent, on average, a 2 to 3 bigger proportion of the intron** than deletions not linked to exon inclusion/exclusion (P = 8.88e-5 and P = 9.43e-3 for down and upstream, respectively). Also, the distance from a eDeletion to the downstream eExon is 20% smaller, although this difference is not significant (P = 0.054). On the other hand, the size of the introns with eDeletion was not significantly different (**Table 13**).

A) Number of eDeletions and genes with eExons

	UPSTREAM	**DOWNSTREAM**
Total deletions associated with exon DE	**28 (1.56% of tested deletions)**	**49 (2.12% of tested deletions)**
Total genes affected	18 genes (1.79%)	29 genes (2.72%)

B) Significantly vs. not significantly DE exons

	UPSTREAM	**DOWNSTREAM**
Deletion size	***2.56 (P*** **= 1.00e-3)**	**1.92 (P = 1.21e-4)**
Intron size	*1.59 (P* = 0.36)	0.81 (P = 0.23)
Position of intron with deletion	1 (P = 0.75)	1 (*P* = 0.45)
% of intron that is deleted	**1.91 (*P* = 9.43e-3)**	***2.99 (P*** **= 8.88e-5)**
Distance from deletion to the DE exon	1.05 (P = 0.85)	0.8 (P = 0.054)
Relative GC content of deletion	1.02 (*P* = 0.76)	1.02 (*P* = 0.37)
Relative GC content of intron	1 (*P* = 0.83)	1 (*P* = 0.11)
Relative GC content of intron (after deletion)	***0.75 (P*** **= 4.30e-3)**	0.9 (*P* = 0.43)
Deletion – intron differential GC content	***2.58 (P*** **= 4.58e-3)**	1.44 (*P = 0.26*)
GC change in intron (when deletion occurs)	4.66 (*P* = 0.06)	1.82 (*P* = 0.23)
% of downregulation	25%	63.3%

Table 13. Continues in the next page.

C) Differences between included and excluded exons

	UPSTREAM	DOWNSTREAM
Deletion size	1.97 (*P* = 0.87)	1.17 (*P* = 0.99)
Intron size	**0.67 *(P* = 4.37e-3)**	1.04 (*P* = 0.73)
Position of intron with deletion	1 (*P* = 0.26)	1.67 (*P* = 0.37)
% of intron that is deleted	**2.54 *(P* = 2.71e-3)**	0.99 (*P* = 1)
Distance from deletion to the DE exon	***0.24* (*P* = 1.26e-3)**	1.47 (*P* = 0.25)
Relative GC content of deletion	0.96 (*P* = 0.60)	1.04 (*P* = 0.91)
Relative GC content of intron	0.99 (*P* = 0.87)	1 (*P* = 0.84)
Relative GC content of intron (after deletion)	0.99 (*P* = 0.87)	1.09 (*P* = 0.26)
GC change	1.24 (*P* = 0.96)	0.68 (*P* = 0.26)

Table 13 | Intronic deletions associated with differential exon expression. (A) Number of deletions associated exons exclusion or inclusion and number of genes affected. (B) Comparison of significant and non-significant associated cases of intronic deletions and exon inclusion/exclusion. (C) Comparison of exon inclusion versus exon exclusion. All differences were tested with Wilcoxon tests except for the GC change, which is the difference in relative GC content of an intron after the deletion and was calculated with a paired Student's T-test.

Given the differential GC content between introns and exons is important for the splicing machinery to recognize exons among long introns, we checked if the eDeletions had particularities in terms of GC content. While eDeletions did not show a particularly different GC content than other intronic deletions, **the differential GC content between the eDeletion and the rest of the intron was larger** (P = 4.58e-3, with the deleted fragment having, on average, higher GC). Also, the introns that hosted these deletions showed a lower GC content than the rest (P = 4.30e-3). These significant differences were detected for upstream-eExons. For downstream-eExons no significant differences were detected between the GC content of the eDeletion and the GC content of the intron. However, we can observe that a subset of deletions are peaks of GC content and that have (**Figure 34**), suggesting an impact of these peaks on the inclusion/exclusion of both upstream and downstream exons.

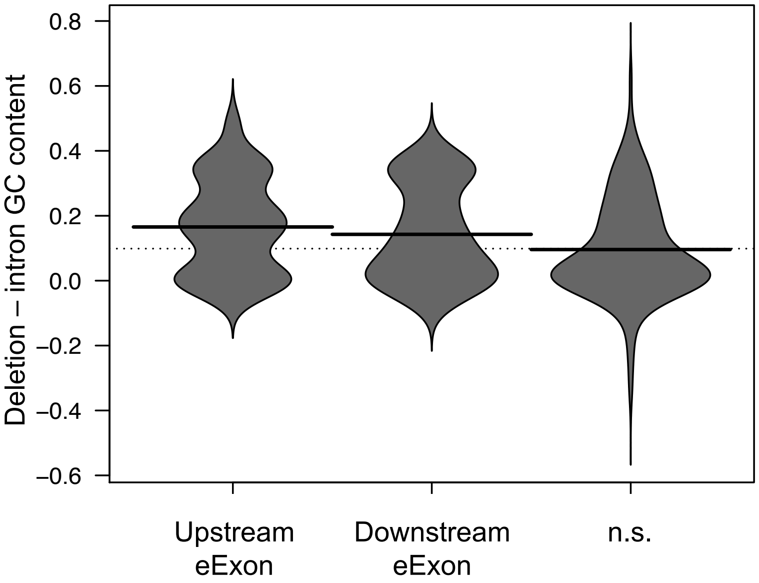


Figure 34 | Difference in relative GC content of an intronic deletion and the rest of the host intron.

Through the comparison of the included versus the excluded eExons we found that, in upstream-eExons, higher inclusion was linked to smaller introns (P = 4.37e-3), deletions removing a larger fraction of the intron (P = 2.71e-3) and to deletions located closer to the exon (P = 1,26e-3).

Discussion

Discussion

CNVs are an important source of genetic variation that might have a previously unsuspected role in evolution and disease. We have stablished that most CNVs overlapping with protein-coding genes fall within introns and we have studied their distribution, functional impact and contribution to the evolution of gene regulation.

A necessary consequence of an intronic loss is a reduction of the size of the gene, which seems to be intolerable for a set of genes that are apparently highly sensitive to changes in gene size. Intron size is also important for the recognition of exon/intron boundaries by the splicing machinery, that is determined by a combination of intron size and differential GC content (Amit et al., 2012). Here, we have observed that CNVs tend to have a GC content higher than the rest of the intron, which means that those losses represent a change in both the size and the overall GC content, with a potential high impact on splicing. Indeed, we have observed cases of differential expression of the alternative exons flanking the CNVs that could contribute to the selection of exons to be included in the processed mRNA.

The different size and position of the intronic CNVs will have different effects on the regulatory sequences contained in the introns, with downstream consequences for the regulation of gene expression. We have observed that CNVs found in populations tend to be located in the regions of introns with less charge of regulatory signals. Moreover, we have confirmed the relation between changes in gene expression and the accumulation of deletions in the regulatory regions of introns. These observations put in value the importance of analysing the specific position of intronic CNVs for the prediction of their potential pathogenicity in disease studies.

Comparison of datasets

In order to understand the distribution of intronic deletions in the genome of healthy individuals, we have analysed five maps of CNVs in large cohorts of individuals from different populations, all of them published in 2015. These maps are, to date, the most extensive in terms of populations represented and the number of individuals sequenced. All of them were obtained using short-read sequencing technologies, except for Zarrei's map, which combines data from different studies present in the DGV, some of which use short-read sequencing but others use SNP arrays or aCGH. We found that these datasets considerably differed in the number, sizes, and types of CNVs, presumably due to a combination of factors such as the methods used for detection of the CNVs, the number of samples and the populations from which they originate.

Regarding the influence of the methods on the size of the CNVs, some characteristics of the maps fit with the limitations or biases of the algorithms that were used for CNV detection. For example, Abyzov et al. used different algorithms to detect deletions at high-resolution (Abyzov et al., 2015). Half of the algorithms that they used are based on split-reads (SR), which can detect exact breakpoints but perform worse with larger SVs (Pirooznia et al., 2015). This is probably the reason why, in this map, deletions are on average smaller than in the rest of the studies. On the other side, Handsaker and Sudmant-Science, the two studies in which the algorithms are based on read-depth (RD), have the higher proportion of gains, probably because these approaches are biased towards the detection of larger CNVs.

The geographic distribution of the samples seems to have a significant impact on the number of observed CNVs. Sudmant-Science has almost as many CNVs as Abyzov, even if the number of sequenced individuals in Abyzov is four times larger. Although Sudmant-Science has many fewer samples, these were selected from very diverse populations from all over the planet. This difference suggests that the sequencing of multiple and geographically distant populations provides a map with many new, population-specific CNVs. However, this hypothesis should be tested by analysing the two groups of samples with the same algorithms. Nevertheless, the fact that the 42.4% of Sudmant-Science CNVRs are specific to the map, against the 11.4% of CNVRs in Handsaker, which uses similar methods to those in Sudmant-Science (based on read-depth), goes in the direction of this hypothesis.

The map from the 1KGP (Sudmant-Nature) is the one with more CNVs, with more than the double CNVs than each of the other four maps, possibly due to the combination of having a large cohort, selecting samples from different populations and using different algorithms for CNV calling.

In general, considering all CNVs in the five maps, we find more deletions than duplications. This balance is different depending on the size of the CNVs, and part of it can probably be explained biologically, but probably a significant part is a consequence of the biases in the detection using short-read sequencing methods, since these methods work very well for detecting SNPs throughout most of the genome but they have substantial limitations for CNV calling. In general, we find that large CNVs are more frequently gains than losses. Probably, large deletions are more likely to be deleterious than large gains, a biological explanation supported by the significant deficit of gains in the OMIM database, compared to deletions. On the other hand, we find that small CNVs are more often losses than gains. Small gains are more difficult to detect than gains of similar sizes using short-read sequencing technologies (Xi et al., 2011). This higher proportion of losses in small CNVs is not observed in a recent study using long-read sequencing (Huddleston et al., 2017). In this study, they suggest that the real landscape of types and sizes of CNVs is far from all the represented in these maps (Huddleston et al., 2017). For CNVs smaller than 1kb, the ratio of gains and losses is relatively close to 1, while almost no gains in this size range were detected in the 1KGP. Regarding larger CNVs, long-read methods reveal more novel deletions than gains, but large gains still seem more frequent than large losses (**Figure 35**).

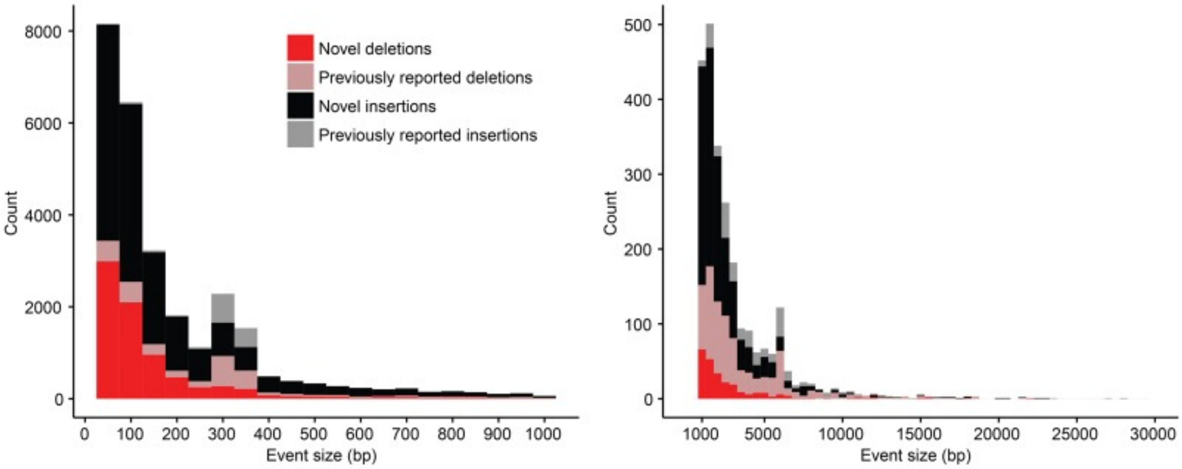


Figure 35 | Structural variant discovery with long-read sequencing. Figure extracted from Huddleston et al., 2017. Deletions (red) and gains (black) identified by long-read sequencing a theoretical diploid human (CHM1 and CHM13) and classified as novel (83%) or previously reported (17%), based on their presence in previously published CNV maps, including Sudmant-Nature and Sudmant-Science.

It is important to note, though, that this study using long-reads shows the CNVs occurring in only two samples (Huddleston et al., 2017), which will not be representative of all the population. Still, the results from long-read sequencing seem to uncover a more balanced proportion of gains and losses than what is observed with short-read based methods.

Assigning CNVs to gains and losses of ancestral genetic material

As explained in the Results section, the classification of CNVs into "gains" or "losses" is done in relation to the reference genome, which does not necessarily reflect the gain or loss of regions from the ancestral genome. For example, if the reference genome assembly includes an inserted region, a loss will be called in this region in those individuals without this insertion (who actually carry the ancestral allele).

Through the comparison of the human genome with the genomes from other non-human primates, we discovered that at least 72.8% of the deletions in Sudmant-Nature reflect deletions relative to an ancestral genome. Still, we were able to detect that at least 0.42% of the deletions are insertions, but we were not able to determine the ancestral state for the remaining 27%).

The reference genome is possibly enriched with insertions given that, when it was assembled, clone selection was biased towards the largest insert clones in order to construct a minimal tiling set (Lander et al., 2001; Nguyen et al., 2006).

To complicate things further, **the allele that is represented in the reference genome also affects the probability of a CNV to be detected**, due to the biases in the methods towards losses or gains. For example, because small gains are more easily missed than small losses using short-read sequencing methods (Xi et al., 2011), if a small CNV sequence is present in the reference genome it will be easier to detect the variant than if the CNV fragment is absent in the reference.

Taken together, our results show that, for a number of technical reasons, **current maps are still far from capturing all the existing structural variability,** implying that the results presented in our work might represent an underestimation of importance of CNVs. Moreover, with the biases in the current maps and the current reference genome, it is difficult (or perhaps impossible) to understand if gains occur more often than losses or the other way around. With an improved reference genome that includes different alleles against which reads can be aligned, together with better methods that allow the detection of all losses and gains, it will be less difficult to study how selection acts on both types of mutations and, also, if there are differences in their occurrence due to repair mechanisms more prone to cause one of the two types of CNV.

Relationship between CNVs and protein-coding genes: differential types and frequency of overlaps with exons and introns

The different ratio of gains and losses in coding and intronic CNVs is partly explained by their different sizes

The different ratios of gains and losses in CNV regions of different sizes are reflected when we look at the different types of overlaps with genes. Intronic CNVs are smaller due to their size limitations (they have to fit within an intron) and thus tend to be losses, while whole-gene CNVs are larger and more frequently gains.

The number of CNVs of each type (intronic, exonic or whole gene) is very variable depending on the map, presumably also due to the biases in the sizes of CNV detected. For example, Handsaker and Sudmant-Science have very few intronic CNVs, probably due to their use of algorithms based on read-depth, which are biased towards larger CNVs. This bias caused the two maps to report very few small variants and thus, very few intronic variants. For this reason, these maps could not be considered in all the analyses of intronic deletions.

The distribution of exonic CNV types is more similar to the one in introns than that whole-gene CNVs. One could expect an exon-disrupting CNV to have an equally deleterious impact than a whole-gene CNV, and consequently, to be biased towards gains to the same extent as whole-gene CNVs. However, many exonic CNVs are very small and thus, are affected by the biases of the algorithms that miss most small gains. Also, in the case of exonic CNVs, many will probably have a milder impact because they can be affecting only alternative isoforms, many of which seem to not be translated into proteins (Tress et al., 2017).

Are deletions enriched in introns?

To characterize the levels of enrichment or depletion of the CNVs overlapping with whole genes, exons or introns, we compared the observed distributions to those seen in randomized sets of CNVs. In three out of five maps, we observed a small but significant general enrichment of deletions in introns and, consistent in all maps, a stronger depletion of deletions in exons.

In the background models used in this analysis, a CNV can be relocated with equal probability anywhere in the genome (global model), within a 10Mb window (local model) or in a region of similar RT (RT model), avoiding in all three cases low-mappability regions. The results obtained with the different backgrounds were similar. The global and local background models were generated in a similar way than other studies (Khurana et al., 2013; Mu et al., 2011; Sudmant et al., 2015a). However, the results regarding the enrichment of CNVs within introns differ from previous studies where they saw a depletion (Khurana et al., 2013; Sudmant et al., 2015a) or no significant differences with the background (Mu et al., 2011). There are different possible explanations for such differences, such as the definition of introns or intronic CNVs (see Results for details). In addition, the biases in size of CNVs in the different studies, caused by the different algorithms that are used, can influence the variable results among maps. Deletions generated by different mechanisms differ in size and seem to be differentially enriched. For example, we and others have observed that NAHR deletions are enriched in introns (Mu et al., 2011) and that TEI deletions (much more represented in Abyzov) do not seem enriched.

However, even if all these factors could explain the apparent contradiction with other studies, these enrichment results need to be interpreted carefully. Strong selective pressures on specific genomic structures can affect the enrichment or depletion observed on another genomic structure. This effect can be explained with an example of a hypothetical genome where CNVs are deleterious in exons (and never found in the healthy population) and neutral anywhere else. After randomizing all CNVs, these will be homogeneously distributed in the hypothetical genome, where some CNVs will overlap with exons and thus, the number of CNVs elsewhere in the genome will be reduced. This will lead to a false enrichment of CNVs in the other places of the genome as a result of the real depletion of CNVs in exons (**Figure 36**).

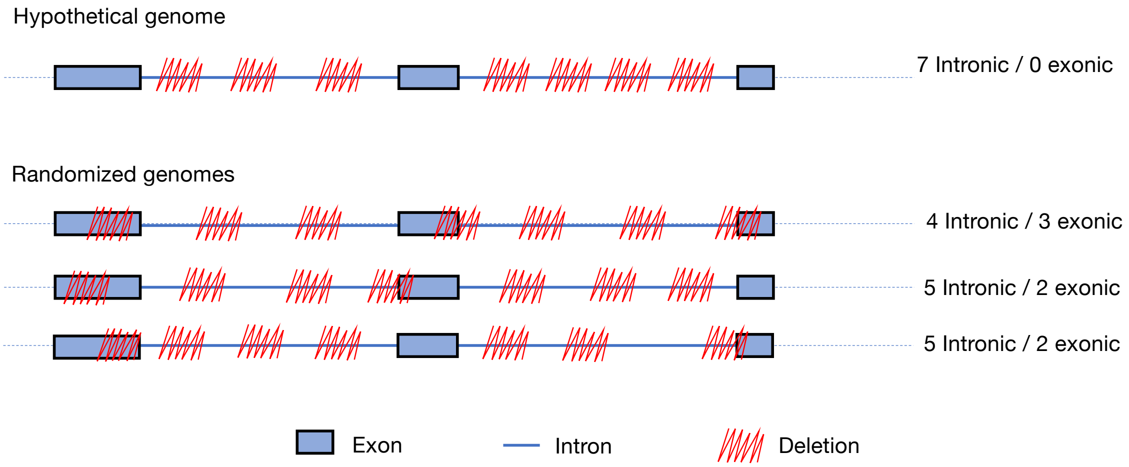


Figure 36 | Effect of negative selection on a genomic feature on the enrichment observed in other regions. Example of a hypothetical genome with exons depleted of deletions and introns with randomly distributed deletions that have no impact on cell fitness. After randomly relocating the deletions over the genome, some of them will overlap with exons by chance. This higher number of exonic deletions in the background models than in the original genome will show that exons are depleted of deletions. However, at the same time, introns will have less deletions in the background models and consequently an enrichment of intronic deletions in the original genome will be detected. This false enrichment (because in this hypothetical genome intronic deletions occur randomly) can lead to wrong assumptions such as the presence of positive selection of intronic deletions.

In this thesis and in previous studies (Khurana et al., 2013; Mu et al., 2011; Sudmant et al., 2015a), the observed values are not very different from the random values, even if in the cases where the difference is significant. Given the limitations in our background models and the CNV maps, we cannot determine if the enrichment we observe in introns is real or not. Instead, larger and more consistent differences with the background models can probably be (and have been in several cases) trusted, such as the depletion of CNVs overlapping with exons (Khurana et al., 2013; Mu et al., 2011; Sudmant et al., 2015a; Zarrei et al., 2015).

However, even if our background models are not sufficient to determine if deletions represent significant enrichments within introns, the results can be interpreted relatively, and that these background models are useful for comparing enrichment levels among groups of elements of the same type, such as introns of different sizes or introns belonging to specific groups of genes. This way, we can conclude that introns carry more deletions than exons but less than intergenic regions. The free-of-background comparison between introns and intergenic regions also shows that introns carry fewer deletions (and smaller) than intergenic regions, possibly due to purifying selection acting on this type of SVs in (at least some) introns. This, at the same time, hints that intronic deletions have a higher probability to produce an impact on the organism than intergenic deletions. It would be interesting to study more in depth if the lower load of deletions in introns is due to purifying selection acting against deletions that disrupt functional sequences or to what extent it is preserving the separation between exons or gene size. Information on sequence conservation can give a hint and help distinguish between "literal" (the order of the nucleotides in under selection) or "indifferent" DNA (presence or absence is under selection).

CNVs overlapping with the protein-coding sequence

In this study, we classified the CNVs that overlap with the protein-coding sequence in “whole-gene CNVs” and partial, “exonic CNVs”.

Whole-gene CNVs and gene dosage

Despite the consistent depletion of coding deletions, taking all CNVs from the five maps together, we find more than 1,200 CNV-genes occurring in healthy individuals. Whole-gene CNVs very often lead to changes in gene expression that cause disease (Wellcome Trust Case Control Consortium et al., 2010; Zhang et al., 2009).

Although there are common multiallelic CNVs where gene expression scales linearly with the number of copies (Handsaker et al., 2015), in many cases, the expression of common extra copies of a gene is lower to the expression of the original copies (Glassberg et al., 2019). It seems that there is a strong constraint on variants that substantially affect gene expression and, often, the expression of the extra copies of genes is buffered to avoid an increase in gene expression proportional to the number of copies (Glassberg et al., 2019; Qian et al., 2010). Moreover, CNVs causing large changes in expression are generally seen at very low frequencies in the population (Glassberg et al., 2019). In cancer, it has been noticed that post-transcriptional mechanisms exist to attenuate amplifications, likely via protein degradation (Gonçalves et al., 2017).

All these mechanisms can exist to buffer the expression of dosage-sensitive genes. However, some studies also claim that expressing a protein is a costly process for the cell, and duplication of highly expressed genes will lead to a depletion of cellular resources and impact the expression of other genes, exerting, indirectly, a deleterious effect (reviewed in Rice and McLysaght, 2017). In agreement with this theory, we have seen that essential genes, which tend to be highly expressed (Wang et al., 2015), are more depleted of gains than other genes.

Essential genes are more depleted of whole gene deletions than the rest, which is not surprising given that, by definition (according to the studies where we have obtained the lists), essential genes are those that, when downregulated or when carrying deleterious mutations, fitness is reduced (Blomen et al., 2015; Hart et al., 2015; Silva et al., 2008). Still, we find 43 "essential" genes lost in the 1KGP population, 6 of them homozygously, suggesting that the lists of essentiality should be revised.

Indeed, in the maps from this study, we see that CNV-genes tend to be smaller than genes that are not variable in the population. Maybe, as it has been observed for highly expressed genes (Rice and McLysaght, 2017), expression of very long genes also ends up exhausting the resources of the cell, causing downregulation of other genes and being this is the reason why we see a difference in gene size of duplicated and not duplicated genes.

Exonic CNVs

The maps that we have analysed in this thesis do not provide with information on where duplications are inserted. However, this information is very relevant to predict the impact of CNVs. For example, a duplication in tandem can disrupt the coding sequence and be as deleterious as a deletion. On the other hand, an *interspersed* insertion of an exonic gain (an insertion of the extra copy elsewhere in the genome) is less likely to affect the fitness of the cell negatively.

To roughly estimate how deleterious gains can be, we compared their impact on principal and alternative exons, as principal exons accumulate less high impact variants than alternative exons (Tress et al., 2017). We see that principal exons have a lower ratio of deletions to gains than alternative exons, suggesting that gains of exons or part of them have a less deleterious effect than losses of part of an exon. It is possible that most gains are either inserted elsewhere in the genome, in tandem but without modifying the coding sequence, or modifying the coding sequence but not disrupting the function of the protein.

Intronic variants in exon-conserved genes

Despite the mild enrichment of deletions within introns observed in most of the maps, we found that not all genes have the same of intronic deletions. By ranking genes by their observed and expected content of intronic deletions, we found that genes with fewer deletions than expected have, on average, a more conserved coding sequence than the genes with more deletions than expected. However, there are exceptions to this tendency, and we found genes among the ones more enriched intronic deletions that belong to the top 2% of genes more conserved genes at a protein level. These results show that it is possible to have a very conserved coding sequence and very variable introns. If some of these intronic deletions affect gene expression, this means that there are genes with a very conserved protein-coding sequence that can have variable gene expression levels.

It would be interesting to study if variability in the introns can affect the performance of DNA damage repair mechanisms that are based on homology and how this could make the gene more prone to have new mutations in the exons. Maybe the variability within the introns is located far enough from the exons to allow repair based on homology if replication errors occur in the coding sequence.

The importance of gene size

We observed that some genes have surprisingly variable gene lengths in the population caused by losses of intronic sequences. In some cases, what is remarkable is the size difference between pairs of individuals, such as the reduction of the 37% of the size in the neuronal glutamate transporter *SLC1A1*; in other cases, what is notable is the number of possible sizes a gene can have in the population. Some of the very variable genes have a very conserved coding sequence. This is the case of the *CSMD1* gene, in which we observe 150 different alleles of different sizes.

Our ranking of more to less intron-conserved genes fits in with the results presented in several studies that advocate the importance of gene size in specific gene sets. We also detected that genes with fewer deletions than expected are enriched in other pathways (at an FDR < 25%) related to brain development, endocytosis and detection of mechanical stimuli. It is possible that similarly to what was hypothesized for the genes activated in response to serum (Kirkconnell et al., 2017), gene length is acting as a biological timer in the genes activated in response to mechanical stimulus. This would mean genes would be induced **simultaneously** by a mechanical stimulus but, because the time taken to obtain the protein correlates the size of the gene, the protein products obtained would be obtained **in a sequential order**, starting with proteins encoded by the shortest and ending with the largest genes.

Regarding brain development, we uncovered additional evidence for the importance of intron conservation in these genes: no intronic deletions were detected in healthy populations in brain genes in which pathogenic intronic deletions causing neurodevelopmental, neurological or psychiatric disorders have been identified. It seems, then, that there are some genes in the brain, that despite being very long, have little structural variation in their introns.

We found that genes with fewer deletions than expected have a higher proportion of their "introme" occupied by regulatory sequences. Thus, our set of intron-conserved genes is probably a combination of genes that do not tolerate changes in their gene size and genes with a high density of essential regulatory regions.

The relationship between intronic deletions and regulatory regions

The enrichment analysis of the genes by their enrichment of intronic deletions suggests certain selection to preserve the intron size. However, we also observed that in introns with regulatory regions and deletions, there is a general tendency for these two elements not to overlap, suggesting a conservation of the regulatory regions, especially of enhancers. When we look at the number of tissues an RF is active in, we find a general depletion, in some cases stronger when the number of tissues increases in CBS and TFBS (although the number of tested TFBS is insufficient to extract conclusions). CTCF is a versatile nuclear factor that can act as a transcriptional activator or repressor, as well as an insulator and a regulator of genomic imprinting. Given the intragenic location of the CBS that we are testing, it is more likely that these are transcriptional regulators rather than insulators. CBS active in many tissues have high affinity and have been previously shown to be highly conservative (Liu et al., 2018).

Interestingly, in cancer samples, we found that SCNAs overlap with RFs as much as expected by chance. This is probably the result of a much lower selective pressure in tumoral cells that allows mutations to accumulate. It could also happen that in some case these deletions that disrupt RFs affect gene expression and contribute to the tumorigenicity of the cell. However, a higher number of samples would be needed to test this hypothesis.

Impact of intronic deletions on gene expression

By combining genotype and gene expression data from a group of 426 individuals, we found some intronic deletions associated with gene expression changes (eDeletions). Intronic eDeletions are associated with higher and lower expression, unlike coding eDeletions, which are all associated with lower expression levels. Although we cannot assume that the eDeletions are the cause of the gene expression changes, we did see that eDeletions overlap significantly more with active enhancers than the rest of intronic deletions, suggesting that the removal of part of the enhancer has an impact on regulation. Also interestingly, we found that eDeletions that do not overlap with enhancers are significantly closer to active enhancers. Since the interaction between an enhancer and a promoter occurs by 3D interaction of the two sequences (Vermunt et al., 2019) it could be that deletions close to these regions affect the formation of loops necessary for these interactions, affecting expression through what is called a "position effect" (Kleinjan and van Heyningen, 2005; Spielmann et al., 2018).

Trans-eDeletions: moving towards a 3D approach

In this study, we proposed a new way of studying the impact of deletions on gene expression in *trans* that requires the combination of genotype, gene expression, and genome organization data. Typically, search for *cis*-eQTLs is restricted to variants within 1Mb from the TSS of a gene. However, there is less consensus in the definition of *trans*-eQTLs. In some cases, *trans*-eQTLs are inter-chromosomal (GTEx consortium, 2017), while in others they are variants beyond 1Mb from the TSS of a gene. The definition of *trans*-eQTL (Gong et al., 2018). A study by the GTEx consortium showed that most *trans*-eQTLs are also *cis*-eQTLs and suggested that the regulation of *trans*-eGenes is via the *cis*-eGenes (GTEx consortium, 2017). However, they also observed that *trans*-eQTLs are enriched in cell-type matched enhancers, suggesting that other regulatory mechanisms may also be involved (GTEx consortium, 2017). In our study, we restricted our search of *trans*-eQTLs to variants in 3D contact with the promoter of a gene. This way, we expect the proportion of eVariants directly linked to the expression change to increase due to a lower detection of eVariants that are indirectly affecting the expression of a gene in *trans* through the modulation of the expression of a gene in *cis*.

Most contacts between enhancer and promoter occur within **Topologically Associating Domains** (TADs), which are self-interacting regions that, in mammalian cells, range in size from hundreds of kilobases to 5 Mb, with an average size of 1Mb (Rocha et al., 2015; Vermunt et al., 2019). We think that genome spatial organization should be taken more into account when looking for associations between variants and gene expression data. It would be interesting to redefine *cis*-variants as those within the same TAD of the tested gene, instead of variants within a window of arbitrary size from the TSS. Correspondingly, *trans*-variants would be inter-TAD and could be limited to regions in 3D contact with the tested gene.

Linking CNV data to genotype and expression data can also be used to see which 3D contacts detected in Hi-C experiments have a regulatory role, and also to determine if its activating or repressing the gene.

Our results are likely to underestimate the effect of losses in introns on gene expression, on the one hand, because we only analysed gene expression in one cell line, limiting the analysis to the genes expressed in these cells. On the other hand, many contacts between enhancer and promoter are tissue-specific (Vermunt et al., 2019), making it necessary to have cell-type interaction maps to search for trans-eQTLs. Besides, interactions in which one of the fragments is deleted in some individuals in the population might be underrepresented in the interaction maps, since this interaction cannot be detected if one of the fragments is missing in the genome. Moreover, the probability of being missed will positively correlate with the frequency of the mutation because the interacting fragment(s) will be less likely to be present in the samples used for the experiment. Another reason why we are probably underestimating the effect of CNVs on gene expression is that the samples from the Geuvadis project belong to 5 of the 26 populations from the 1KGP (four European and one African). Thus, it was not possible to test the variants specific to other populations.

We cannot ignore the fact that other unexplored variants could cause a proportion of the statistical associations that we found between deletions and gene expression, probably in linkage disequilibrium with the CNV, even if CNVs are more likely to be eQTLs than SNPs (Bryois et al., 2014).

Effect of intronic deletions in transcript differential expression

In our study of gene expression, we detected genes with differentially expressed transcripts in individuals with an intronic deletion. We hypothesized that some of these eDeletions could be producing changes in the structure of the intron the individuals with the deletion, causing an imbalance of expressed isoforms by affecting the inclusion or exclusion of upstream or downstream exons. We studied under the assumption that lower expression of an exon reflects exon exclusion or skipping (the processed mRNA does not include this exon), while higher expression reflects higher levels of exon inclusion. We are aware, though, that the differences in exon expression could be in fact due to other factors, such as changes in the regulation or the stability of the different transcripts.

In our analysis, we found a higher proportion of intronic eDeletions associated with the changes in the downstream rather than the upstream exon (2.12% vs. 1.56%), suggesting that maybe, downstream exons are more affected by intronic deletions. Although the number of significant cases is minimal and further analysis with bigger samples should be carried to extract meaningful conclusions, we observed that **intronic deletions seem to be mainly associated with the inclusion of the upstream exon and with the exclusion of the downstream exon**.

According to our analysis, the relative change in the size of the intron seems to be relevant, with proportionally bigger losses being more associated with intron inclusion or exclusion. Moreover, overexpressed upstream exons are associated with **relatively bigger deletions** than underexpressed upstream exons. The size of the deletion, thus, could be not only affecting splicing but determining if the alternative exon is included or excluded.

Given the importance of the differential GC content between exons and introns to allow the splicing machinery to recognize exons flanked by long introns (Amit et al., 2012; Gelfman et al., 2012), we analysed the GC content of the tested deletions and genes. Interestingly, we found that deletions associated with exon inclusion or exclusion were located in introns with a lower GC content, but **the deleted fragment represented a peak of GC within the intron**. These results suggest that peaks of GC within the intron might affect splicing.

The higher GC content of the deleted fragment, however, is not limited to deletions associated with expression changes. Looking at the whole of intronic deletions in the 1KGP, we found a general tendency for the deleted fragments to have a higher GC content than the rest of the intron. Further analyses remain to be done to understand better why the deleted segments show this higher GC. The differential GC content of the deleted fragment could be causing epigenetic differences, for example in **DNA methylation**, which tends to occur in CpG islands and is higher in exons than in introns (Gelfman et al., 2012; Moore et al., 2013). Since DNA methylation is known to influence gene expression and splicing (Shayevitch et al., 2018), deletions could be causing variability in isoform expression by altering the density of epigenetic marks within the intron.

Several questions remain to be answered regarding the impact on splicing and gene expression of the changes in intron structure caused by a deletion. Hopefully, some of these questions will be answered using larger datasets that also combine sequence and gene expression data from the same individuals, or maybe with further experimental studies modifying intron sequence (Amit et al., 2012) or epigenetics (Shayevitch et al., 2018).

Genes of different evolutionary ages have different patterns of CNVs

We have found that human genes of born at different times during evolution accumulate different types of CNVs. While ancient genes accumulate most intronic CNVs, young genes are enriched with coding CNVs. These two types of CNVs can have a different impact on the gene, suggesting that CNVs are shaping the evolution of genes differently, depending on their age. For example, new genes born via whole gene duplications, which are an important substrate for functional innovation, will more probably arise from young genes. At the same time, modifications in the protein sequence caused by partial losses or duplications of the coding sequence will also tend to happen in young genes. In the article previously published in our group by Juan, Rico and others (Juan et al., 2013), they described that young genes replicate later in the S-phase and they suggested that duplications of young genes tend to be inserted in late-replicating regions. These late-replicating regions are enriched in CNVs and thus make the duplicated genes more susceptible to be further duplicated. Contrarily, they observed that ancient genes tend to have a fix copy number. However, they did not look at the variability in introns or intergenic regions in these genes.

We observed that ancient genes, despite being impoverished with coding CNVs, carry most of the intronic deletions in the genome and, more interestingly, most of the intronic deletions associated with gene expression changes. These genes are in general larger, with longer introns that probably carry more regulatory elements. Based on our results, we suggest that intronic CNVs cause gene expression variability in the population, likely through the direct overlap with RFs or interfering with the regulation by contacts in 3D. This effect on gene regulation will happen mainly in ancient genes and might provide with the capacity to adapt to new environments.

In fact, previous studies have suggested that changes in gene expression have been more prevalent in human adaptation than changes at the protein level (King and Wilson, 1975; Fraser, 2013), suggesting a strong evolutionary potential of intronic CNVs. Although we expect the functional effects of the coding CNVs to be stronger than intronic CNVs, we found a similar proportion of coding and intronic eDeletions showing signatures of potential positive selection.

Interestingly, structural variation within introns can affect regulation in a tissue-specific way. We found that there is a negative correlation between the number of tissues in which some regulatory regions are active and their probability to be disrupted by an intronic deletion, suggesting that mutations in ancient genes may have a tissue-specific impact. Indeed, coding mutations can also have a tissue-specific impact. For example, exonic CNVs modifying the protein-coding sequence can have a cell-type specific impact if they the gene is only expressed in some cell-types, but we expect that the impact will be similar anywhere where the protein is produced. Whole gene duplications, as discussed by Juan and others, can also have a tissue-specific effect if they are inserted in a part of the chromatin that is only active in specific tissues (Juan et al., 2013). However, we think that intronic variation will more frequently cause variability in a cell-type specific manner than coding CNVs.

Besides the impact intronic deletions may have on gene expression, these can also affect the time taken to transcribe a gene and their splicing. Modifications in these two processes can also have tissue-specific consequences. For example, an increase in gene length could be deleterious in a tissue where cells are rapidly replicate, if this increase makes it impossible for the protein to be produced before the next cycle starts (Seoighe and Korir, 2011).

We propose a model in which CNVs are shaping the evolution of genes differently, depending on the age of the gene. In ancient genes, CNVs are currently modifying the expression, splicing, and the time taken to transcribe the gene; while in young genes, CNVs impact on the coding sequence, modifying proteins and providing the substrate for the birth of new genes (**Figure 37**).

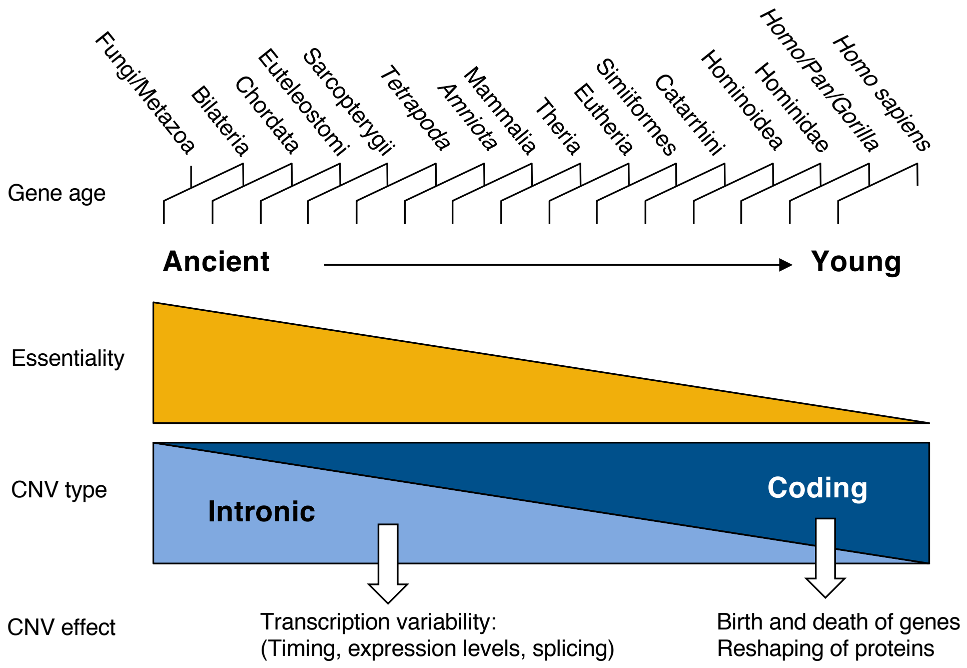


Figure 37 | Impact of CNVs on genes and their evolution. Evolutionarily ancient and young genes accumulate different types of CNVs. While young genes are enriched in coding deletions (which alter gene dosage or modify or disrupt the protein, sometimes affecting gene expression), ancient genes normally have a conserved coding sequence but a high load of intronic deletions, which are sometimes associated with gene expression under or overexpression.

Future perspectives

One of our main findings is that genes that have essential functions and a conserved protein-coding sequence can accumulate SVs in their introns, providing a substrate for adaptation through changes in gene regulation. Although we found some characteristics in the intronic deletions that could be causing an impact on gene expression or splicing, in most of the analyses we had minimal numbers of intronic deletions or to test hour hypothesis or not enough individuals for whom we had expression data. Having more intronic deletions to test (more expression data or more sensitive CNV calls) would help to predict better the impact of the structure of intronic variation on gene expression and splicing. However, even with the variability of CNVs among maps and the low number of intronic CNVs in some of them, most of our results were consistent across maps, such as the distribution of CNVs on the genes of different ages, or the lower-than-expected overlap of CNVs with RFs.

The variability of types of CNVs among the maps, even when the sequenced individuals are essentially the same, highlights the limitations of the current methods. Scientists checking if a gene of interest contains CNVs should consider looking at as many maps of SV as possible.

In our case, we think that our analyses should be repeated with more complete maps of population CNVs. We expect that long-read sequencing methods, which have been shown to improve the sensitivity of SV detection, will provide the scientific community with more accurate maps. In a very recent study, Audano and others have long-read sequenced 15 genomes, and they have detected 99,604 SVs, the 40.8% of which are novel relative to a combination of CNV maps that includes Sudmant-Nature, Sudmant-Science and the long-read sequencing data from Huddleston et al. 2017 (Audano et al., 2019). In this study, besides providing a much more extensive catalog of SVs, they developed an algorithm to improve genotyping SVs from short-read data and generated and released a reference genome containing SVs as alternative loci. In their final map of SVs, insertions outnumber deletions (57,994 insertions, 41,388 deletions).

A future analysis to complement our result could be the analysis of the CNVs provided by Audano and others. In this case, given that the authors inform of the position where the insertions occur(Audano et al., 2019), we could check which genes accumulate insertions (and thus, grow in size), complementing our analysis on which genes tolerate or not size reductions.

In summary, we have studied in depth the distribution of CNVs in the human genome and their possible functional implication, but our study has been based on current maps of SV that show several limitations. We expect that further analyses with more comprehensive maps will have more power to validate our findings. A better understanding of the functional impact of CNVs will help facilitate the functional prediction of new CNVs, for example after checking if they overlap or are close to an RFs, in which tissues the RF is active, if any contacts are disrupted or if the change in gene size is considerable.

Conclusions/Conclusiones

Conclusions

1. In the maps analysed in this study, intronic CNVs are mostly deletions and they are more frequent than those that affect exons.

2. Introns accumulate more CNV losses than expected by chance, although less than intergenic regions of similar sizes. These CNVs in introns are smaller than intergenic ones, suggesting that introns are more sensitive to losses than intergenic regions.

3. Intronic deletions are impoverished in genes related to development or required in stimulus-activated reactions, possibly because the time required for transcription is important in these groups of genes.

4. Intronic deletions are also depleted in neuronal genes in which pathogenic intronic CNVs have been found, highlighting the importance of considering the introns of such genes in future genetic tests.

5. Intronic deletions can be associated with changes in the expression of the host gene or in other genes that show long-range interactions with the intronic CNV region.

6. Intronic deletions associated with changes in gene expression tend to overlap with enhancers or are linearly close to them, suggesting that CNVs in introns can contribute to gene expression variability in the populations by interfering the three-dimensional interactions of promoters and intronic enhancers.

7. Intronic losses tend to occur in genes with a high differential GC content between the exon and the introns that flank it. The regions lost tend to be GC-rich and their disappearance leads to higher exon-intron GC differences that could influence exon recognition during splicing.

8. Intronic deletions appear to affect splicing processes by altering the inclusion or exclusion of the alternative exons that flank them, altering the balance of the isoforms expressed in the cell. This effect seems to be dependent on the size and GC content of the deletion.

9. Genes of different evolutionary ages show different patterns of overlap with CNVs: young genes are enriched in CNV that overlap coding regions, with possible functional impact at the protein level, while old genes are impoverished in coding CNVs and enriched in intronic CNVs, possibly with a weaker functional impact on the proteins but influencing their regulation.

10. According to our model, CNVs are shape the evolution of genes differently depending on the age of the gene. CNVs are modifying the expression, splicing, and the time taken to transcribe of ancient genes while they alter the coding sequence or gene dosage of new genes.

Conclusiones

1. En los mapas analizaos en este estudio, las CNVs intrónicas son en su mayoría deleciones y son más frecuentes que las que afectan a los exones.

2. Los intrones acumulan más deleciones de lo esperado por azar, aunque menos que las regiones intergénicas de tamaños similares. Las CNVs en intrones son más pequeñas que las intergénicas, lo que sugiere que los intrones son más sensibles a las pérdidas que las regiones intergénicas.

3. Las deleciones intrónicas están empobrecidas en los genes relacionados con el desarrollo o activados en la reacción a estímulos, posiblemente porque el tiempo requerido para la transcripción es importante en estos grupos de genes.

4. Las deleciones intrónicas también están empobrecidas en genes neuronales en los que se han encontrado CNVs intrónicas patógenicas, lo que destaca la importancia de considerar los intrones de dichos genes en futuras pruebas genéticas.

5. Las deleciones intrónicas pueden asociarse a cambios en la expresión del gen que las contiene o de otros genes ubicados lejos en la secuencia cuando se pierden contactos con el promotor del otro gen en la estructura tridimendional de la cromatina.

6. Las deleciones intrónicas asociadas con cambios en la expresión génica tienden a solapar con los enhancers o están linealmente cerca de ellas, lo que sugiere que las CNVs intrónicas pueden contribuir a la variabilidad de expresión al interferir sobre las interacciones tridimensionales de los promotores y los *enhancers* intrónicos.

7. Las pérdidas intrónicas tienden a ocurrir en genes con una pronuncidad diferencia de contenido GC entre el exón y los intrones que lo flanquean. Las regiones perdidas tienden a ser ricas en GC y su desaparición conduce a mayores

diferencias de contenido GC que podrían influir en el reconocimiento del exón durante el proceso de *splicing*.

8. Las supresiones intrónicas parecen afectar los procesos de empalme al alterar la inclusión o exclusión de los exones alternativos que los flanquean, alterando el equilibrio de las isoformas expresadas en la célula. Este efecto parece depender del tamaño y el contenido de GC de la eliminación.

9. Los genes de diferentes edades evolutivas muestran diferentes patrones de solapamiento con CNVs: los genes jóvenes están enriquecidos en CNVs que se afectan las regiones codificantes, con un posible impacto funcional a nivel de la proteína, mientras que los genes antiguos están empobrecidos de CNVs codificantes y enriquecidos con CNVs intrónicas, posiblemente con un impacto funcional más débil en las proteínas pero influyendo en su regulación.

10. Según nuestro modelo, las CNVs modelan la evolución de los genes de forma diferente según la edad del gen, modificando la expresión, el *splicing* y el tiempo necesario para transcribir los genes antiguos mientras que alteran la secuencia codificante o la dosis génica en genes evolutivamente recientes.

References

References

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Abascal, F., Juan, D., Jungreis, I., Martinez, L., Rigau, M., Rodriguez, J.M., Vazquez, J., and Tress, M.L. (2018). Loose ends: almost one in five human genes still have unresolved coding status. Nucleic Acids Res. *46*, 7070–7084.

Abyzov, A., Li, S., Kim, D.R., Mohiyuddin, M., Stütz, A.M., Parrish, N.F., Mu, X.J., Clark, W., Chen, K., Hurles, M., et al. (2015). Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. Nat. Commun. *6*, 7256.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., et al. (2012). Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. Cell Rep. *1*, 543–556.

An, J.Y. (2017). National human genome projects: an update and an agenda. Epidemiol. Health *39*.

Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., Cantsilieris, S., Sorensen, M., Welch, A.E., Dougherty, M.L., Nelson, B.J., Shah, A., Dutcher, S.K., et al. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. Cell *176*, 663-675.e19.

Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. Gene *241*, 3–17.

Bickel, P.J., Boley, N., Brown, J.B., Huang, H., and Zhang, N.R. (2010). Subsampling methods for genomic inference. Ann. Appl. Stat. *4*, 1660–1697.

Blomen, V.A., Májek, P., Jae, L.T., Bigenzahn, J.W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F.R., Olk, N., Stukalov, A., et al. (2015). Gene essentiality and synthetic lethality in haploid human cells. Science *350*, 1092–1096.

Bonnet, A., Grosso, A.R., Elkaoutari, A., Coleno, E., Presle, A., Sridhara, S.C., Janbon, G., Géli, V., Almeida, S.F. de, and Palancade, B. (2017). Introns Protect Eukaryotic Genomes from Transcription-Associated Genetic Instability. Mol. Cell *67*, 608-621.e6.

Bryois, J., Buil, A., Evans, D.M., Kemp, J.P., Montgomery, S.B., Conrad, D.F., Ho, K.M., Ring, S., Hurles, M., Deloukas, P., et al. (2014). Cis and Trans Effects of Human Genomic Variants on Gene Expression. PLoS Genet. *10*.

Campbell, P.J., Getz, G., Stuart, J.M., Korbel, J.O., and Stein, L.D. (2017). Pan-cancer analysis of whole genomes. BioRxiv 162784.

Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V., and Kondrashov, F.A. (2002). Selection for short introns in highly expressed genes. Nat. Genet. *31*, 415–418.

Celik, A., Baker, R., He, F., and Jacobson, A. (2017). High-resolution profiling of NMD targets in yeast reveals translational fidelity as a basis for substrate selection. RNA *23*, 735–748.

Chen, L., Zhou, W., Zhang, L., and Zhang, F. (2014). Genome architecture and its roles in human copy number variation. Genomics Inform. *12*, 136–144.

Chen, L., Zhou, W., Zhang, C., Lupski, J.R., Jin, L., and Zhang, F. (2015). CNV instability associated with DNA replication dynamics: evidence for replicative mechanisms in CNV mutagenesis. Hum. Mol. Genet. *24*, 1574–1583.

Cheung, H.W., Cowley, G.S., Weir, B.A., Boehm, J.S., Rusin, S., Scott, J.A., East, A., Ali, L.D., Lizotte, P.H., Wong, T.C., et al. (2011). Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proc. Natl. Acad. Sci. U. S. A. *108*, 12372–12377.

Chiang, C., Scott, A.J., Davis, J.R., Tsang, E.K., Li, X., Kim, Y., Hadzic, T., Damani, F.N., Ganel, L., Montgomery, S.B., et al. (2017). The impact of structural variation on human gene expression. Nat. Genet. *49*, 692–699.

Chorev, M., and Carmel, L. (2012). The function of introns. Front. Genet. *3*, 55.

Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. *9*, 938–950.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704–712.

Dubow, T., and Marjanovic, S. (2016). Population-scale sequencing and the future of genomic medicine.

Eisenberg, E., and Levanon, E.Y. (2003). Human housekeeping genes are compact. Trends Genet. *19*, 362–365.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Escaramís, G., Docampo, E., and Rabionet, R. (2015). A decade of structural variants: description, history and methods to detect structural variation. Brief. Funct. Genomics *14*, 305–314.

Fox-Walsh, K.L., Dou, Y., Lam, B.J., Hung, S., Baldi, P.F., and Hertel, K.J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc. Natl. Acad. Sci. *102*, 16176–16181.

França, G.S., Cancherini, D.V., and Souza, S.J. de (2012). Evolutionary history of exon shuffling. Genetica *140*, 249–257.

Francis, W.R., and Wörheide, G. (2017). Similar Ratios of Introns to Intergenic Sequence across Animal Genomes. Genome Biol. Evol. *9*, 1582–1598.

Fraser, H.B. (2013). Gene expression drives local adaptation in humans. Genome Res. *23*, 1089–1096.

Gabel, H.W., Kinde, B.Z., Stroud, H., Gilbert, C.S., Harmin, D.A., Kastan, N.R., Hemberg, M., Ebert, D.H., and Greenberg, M.E. (2015). Disruption of DNA methylation-dependent long gene repression in Rett syndrome. Nature *522*, 89–93.

Gamazon, E.R., and Stranger, B.E. (2015). The impact of human copy number variation on gene expression. Brief. Funct. Genomics *14*, 352–357.

Gamazon, E.R., Nicolae, D.L., and Cox, N.J. (2011). A Study of CNVs As Trait-Associated Polymorphisms and As Expression Quantitative Trait Loci. PLoS Genet. *7*.

Gazave, E., Marqués-Bonet, T., Fernando, O., Charlesworth, B., and Navarro, A. (2007). Patterns and rates of intron divergence between humans and chimpanzees. Genome Biol. *8*, R21.

Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M.A., and Malinverni, R. (2016). regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinforma. Oxf. Engl. *32*, 289–291.

Gelfman, S., Burstein, D., Penn, O., Savchenko, A., Amit, M., Schwartz, S., Pupko, T., and Ast, G. (2012). Changes in exon–intron structure during vertebrate evolution affect the splicing pattern of exons. Genome Res. *22*, 35–50.

Genome Reference Consortium Human Genome Overview - GRC website. https://www.ncbi.nlm.nih.gov/grc/data. Accessed: 21/03/2019.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. Genome Res. *17*, 669–681.

Glassberg, E.C., Gao, Z., Harpak, A., Lan, X., and Pritchard, J.K. (2019). Evidence for Weak Selective Constraint on Human Gene Expression. Genetics *211*, 757–772.

Gonçalves, E., Fragoulis, A., Garcia-Alonso, L., Cramer, T., Saez-Rodriguez, J., and Beltrao, P. (2017). Widespread Post-transcriptional Attenuation of Genomic Copy-Number Variation in Cancer. Cell Syst. *5*, 386-398.e4.

Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., Feng, J., Liu, R., Diao, L., Guo, A.-Y., et al. (2018). PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. Nucleic Acids Res. *46*, D971–D976.

Gonzalez, K.D., Hill, K.A., Li, K., Li, W., Scaringe, W.A., Wang, J.-C., Gu, D., and Sommer, S.S. (2007). Somatic microindels: analysis in mouse soma and comparison with the human germline. Hum. Mutat. *28*, 69–80.

González-Barrios, M., Fierro-González, J.C., Krpelanova, E., Mora-Lorca, J.A., Pedrajas, J.R., Peñate, X., Chavez, S., Swoboda, P., Jansen, G., and Miranda-Vizuete, A. (2015). Cis- and Trans-Regulatory Mechanisms of Gene Expression in the ASJ Sensory Neuron of Caenorhabditis elegans. Genetics *200*, 123–134.

Graur, D., Zheng, Y., and Azevedo, R.B.R. (2015). An Evolutionary Classification of Genomic Function. Genome Biol. Evol. *7*, 642–645.

GTEx consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Handsaker, R.E., Van Doren, V., Berman, J.R., Genovese, G., Kashin, S., Boettger, L.M., and McCarroll, S.A. (2015). Large multiallelic copy number variations in humans. Nat. Genet. *47*, 296–303.

Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc. Natl. Acad. Sci. U. S. A. *107*, 139–144.

Harewood, L., Chaignat, E., and Reymond, A. (2012). Structural variation and its effect on expression. Methods Mol. Biol. Clifton NJ *838*, 173–186.

Harima, Y., Takashima, Y., Ueda, Y., Ohtsuka, T., and Kageyama, R. (2013). Accelerating the Tempo of the Segmentation Clock by Reducing the Number of Introns in the Hes7 Gene.

Cell Rep. *3*, 1–7.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*, 1515–1526.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources. Database *2016*.

Heyn, P., Kalinka, A.T., Tomancak, P., and Neugebauer, K.M. (2015). Introns and gene expression: Cellular constraints, transcriptional regulation, and evolutionary consequences. BioEssays *37*, 148–154.

Hollander, D., Naftelberg, S., Lev-Maor, G., Kornblihtt, A.R., and Ast, G. (2016). How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing? Trends Genet. *32*, 596–606.

Hsiao, Y.-H.E., Bahn, J.H., Lin, X., Chan, T.-M., Wang, R., and Xiao, X. (2016). Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins. Genome Res. *26*, 440–450.

Huddleston, J., Chaisson, M.J.P., Steinberg, K.M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T.A., Munson, K.M., Kronenberg, Z.N., Vives, L., et al. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. *27*, 677–685.

Iskow, R.C., Gokcumen, O., and Lee, C. (2012). Exploring the role of copy number variants in human adaptation. Trends Genet. TIG *28*, 245–257.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*, 338–345.

Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369-1384.e19.

Juan, D., Rico, D., Marques-Bonet, T., Fernández-Capetillo, O., and Valencia, A. (2013). Late-replicating CNVs as a source of new genes. Biol. Open *2*, 1402–1411.

Keane, P.A., and Seoighe, C. (2016). Intron Length Coevolution across Mammalian

Genomes. Mol. Biol. Evol. *33*, 2682–2691.

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., et al. (2013). Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science *342*, 1235587.

Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. *35*, 125–131.

King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107–116.

Kirkconnell, K.S., Magnuson, B., Paulsen, M.T., Lu, B., Bedi, K., and Ljungman, M. (2017). Gene length as a biological timer to establish temporal transcriptional regulation. Cell Cycle Georget. Tex *16*, 259–270.

Kleinjan, D.A., and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. Am. J. Hum. Genet. *76*, 8–32.

Koren, A., Polak, P., Nemesh, J., Michaelson, J.J., Sebat, J., Sunyaev, S.R., and McCarroll, S.A. (2012). Differential relationship of DNA replication timing to different forms of human mutation and variation. Am. J. Hum. Genet. *91*, 1033–1040.

Kronenberg, Z.N., Fiddes, I.T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O.S., Underwood, J.G., Nelson, B.J., Chaisson, M.J.P., Dougherty, M.L., et al. (2018). High-resolution comparative analysis of great ape genomes. Science *360*.

Lambowitz, A.M., and Belfort, M. (2015). Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. Microbiol. Spectr. *3*.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lappalainen, T., Sammeth, M., Friedländer, M.R., ’t Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Lee, J.A., Carvalho, C.M.B., and Lupski, J.R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell *131*, 1235–1247.

Lei, Q., Li, C., Zuo, Z., Huang, C., Cheng, H., and Zhou, R. (2016). Evolutionary Insights into RNA trans-Splicing in Vertebrates. Genome Biol. Evol. *8*, 562–577.

Li, Y., Hu, M., and Shen, Y. (2018). Gene regulation in the 3D genome. Hum. Mol. Genet. *27*, R228–R233.

Lin, M., Whitmire, S., Chen, J., Farrel, A., Shi, X., and Guo, J. (2017). Effects of short indels on protein structure and function in human genomes. Sci. Rep. *7*, 9313.

Liu, F., Wu, D., and Wang, X. (2018). Roles of CTCF in conformation and functions of chromosome. Semin. Cell Dev. Biol.

Liu, Y., Gonzàlez-Porta, M., Santos, S., Brazma, A., Marioni, J.C., Aebersold, R., Venkitaraman, A.R., and Wickramasinghe, V.O. (2017). Impact of Alternative Splicing on the Human Proteome. Cell Rep. *20*, 1229–1241.

MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. *42*, D986-992.

Marcotte, R., Brown, K.R., Suarez, F., Sayad, A., Karamboulas, K., Krzyzanowski, P.M., Sircoulomb, F., Medrano, M., Fedyshyn, Y., Koh, J.L.Y., et al. (2012). Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov. *2*, 172–189.

Martin, W., and Koonin, E.V. (2006). Introns and the origin of nucleus–cytosol compartmentalization. Nature *440*, 41–45.

Martin, C.L., Kirkpatrick, B.E., and Ledbetter, D.H. (2015). CNVs, Aneuploidies and Human Disease. Clin. Perinatol. *42*, 227–242.

Monlong, J., Cossette, P., Meloche, C., Rouleau, G., Girard, S.L., and Bourque, G. (2018). Human copy number variants are enriched in regions of low mappability. Nucleic Acids Res. *46*, 7236–7249.

Moore, L.D., Le, T., and Fan, G. (2013). DNA Methylation and Its Basic Function. Neuropsychopharmacology *38*, 23–38.

Mu, X.J., Lu, Z.J., Kong, Y., Lam, H.Y.K., and Gerstein, M.B. (2011). Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res. *39*, 7058–7076.

Nakamoto, S., Kumamoto, Y., Igarashi, K., Fujiyama, Y., Nishizawa, N., Ei, S., Tajima, H., Kaizu, T., Watanabe, M., and Yamashita, K. (2018). Methylated promoter DNA of CDO1 gene and preoperative serum CA19-9 are prognostic biomarkers in primary extrahepatic cholangiocarcinoma. PLOS ONE *13*, e0205864.

National Human Genome Research Institute (1996). NHGRI-DOE Guidance on Human Subjects Issues in Large-Scale DNA Sequencing.

Nguyen, D.-Q., Webber, C., and Ponting, C.P. (2006). Bias of Selection on Human Copy-Number Variants. PLOS Genet. *2*, e20.

Ohno, S. (1972). So much “junk” DNA in our genome. Brookhaven Symp. Biol. *23*, 366–370.

Ottaviani, D., LeCain, M., and Sheer, D. (2014). The role of microhomology in genomic structural variation. Trends Genet. *30*, 85–94.

Palazzo, A.F., and Gregory, T.R. (2014). The Case for Junk DNA. PLOS Genet. *10*, e1004351.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Paten, B., Novak, A.M., Eizenga, J.M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. Genome Res. *27*, 665–676.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nat. Genet. *39*, 1256–1260.

Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. *9*, e1003709.

Piovesan, A., Caracausi, M., Antonaros, F., Pelleri, M.C., and Vitale, L. (2016). GeneBase 1.1: a tool to summarize data from NCBI Gene datasets and its application to an update of human gene statistics. Database *2016*.

Pirooznia, M., Goes, F.S., and Zandi, P.P. (2015). Whole-genome CNV analysis: advances in computational approaches. Front. Genet. *6*.

Portin, P., and Wilkins, A. (2017). The Evolving Definition of the Term “Gene.” Genetics *205*, 1353–1364.

Qian, W., Liao, B.-Y., Chang, A.Y.-F., and Zhang, J. (2010). Maintenance of duplicate genes and their functional redundancy by reduced expression. Trends Genet. TIG *26*, 425–430.

Qiu, F., Xu, Y., Li, K., Li, Z., Liu, Y., DuanMu, H., Zhang, S., Li, Z., Chang, Z., Zhou, Y., et al. (2012). CNVD: text mining-based copy number variation in disease database. Hum. Mutat. *33*, E2375-2381.

Rask-Andersen, M., Almén, M.S., Lind, L., and Schiöth, H.B. (2015). Association of the

LINGO2-related SNP rs10968576 with body mass in a cohort of elderly Swedes. Mol. Genet. Genomics MGG *290*, 1485–1491.

Rearick, D., Prakash, A., McSweeny, A., Shepard, S.S., Fedorova, L., and Fedorov, A. (2011). Critical association of ncRNA with introns. Nucleic Acids Res. *39*, 2357–2366.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

Rice, A.M., and McLysaght, A. (2017). Dosage-sensitive genes in evolution and disease. BMC Biol. *15*, 78.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Rocha, P.P., Raviram, R., Bonneau, R., and Skok, J.A. (2015). Breaking TADs: insights into hierarchical genome organization. Epigenomics *7*, 523–526.

Rodriguez, J.M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J., Lopez, G., Valencia, A., and Tress, M.L. (2013). APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res. *41*, D110-117.

Rogozin, I.B., Carmel, L., Csuros, M., and Koonin, E.V. (2012). Origin and evolution of spliceosomal introns. Biol. Direct *7*, 11.

Roy, M., Kim, N., Xing, Y., and Lee, C. (2008). The effect of intron length on exon creation ratios during the evolution of mammalian genomes. RNA *14*, 2261–2273.

Seoighe, C., and Korir, P.K. (2011). Evidence for intron length conservation in a set of mammalian genes associated with embryonic development. BMC Bioinformatics *12 Suppl 9*, S16.

Shayevitch, R., Askayo, D., Keydar, I., and Ast, G. (2018). The importance of DNA methylation of exons on alternative splicing. RNA N. Y. N *24*, 1351–1362.

Signor, S.A., and Nuzhdin, S.V. (2018). The Evolution of Gene Expression in cis and trans. Trends Genet. *34*, 532–544.

Silva, J.M., Marran, K., Parker, J.S., Silva, J., Golding, M., Schlabach, M.R., Elledge, S.J., Hannon, G.J., and Chang, K. (2008). Profiling essential genes in human mammary cells by multiplex RNAi screening. Science *319*, 617–620.

Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nat. Rev. Genet. *19*, 453–467.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). 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.

Stewart, C., Kural, D., Strömberg, M.P., Walker, J.A., Konkel, M.K., Stütz, A.M., Urban, A.E., Grubert, F., Lam, H.Y.K., Lee, W.-P., et al. (2011). A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. PLOS Genet. *7*, e1002236.

Stunnenberg, H.G., International Human Epigenome Consortium, and Hirst, M. (2016). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell *167*, 1145–1149.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. *102*, 15545–15550.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.-Y., et al. (2015a). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., et al. (2015b). Global diversity, population stratification, and selection of human copy-number variation. Science *349*, aab3761.

Swinburne, I.A., Miguez, D.G., Landgraf, D., and Silver, P.A. (2008). Intron length increases oscillatory periods of gene expression in animal cells. Genes Dev. *22*, 2342–2346.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447-452.

Takashima, Y., Ohtsuka, T., González, A., Miyachi, H., and Kageyama, R. (2011). Intronic delay is essential for oscillatory expression in the segmentation clock. Proc. Natl. Acad. Sci. *108*, 3300–3305.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

The Bioconductor Dev Team (2014). BSgenome.Hsapiens.UCSC.hg19.masked: Full masked genome sequences for Homo sapiens (UCSC version hg19).

Thurman, R.E., Day, N., Noble, W.S., and Stamatoyannopoulos, J.A. (2007). Identification of higher-order functional domains in the human ENCODE regions. Genome Res. *17*, 917–927.

Tress, M.L., Abascal, F., and Valencia, A. (2017). Alternative Splicing May Not Be the Key to Proteome Complexity. Trends Biochem. Sci. *42*, 98–110.

Valencia, P., Dias, A.P., and Reed, R. (2008). Splicing promotes rapid and efficient mRNA export in mammalian cells. Proc. Natl. Acad. Sci. U. S. A. *105*, 3386–3391.

Vaz-Drago, R., Custódio, N., and Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. Hum. Genet. *136*, 1093–1111.

Vermunt, M.W., Zhang, D., and Blobel, G.A. (2019). The interdependence of gene-regulatory elements and the 3D genome. J Cell Biol *218*, 12–26.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096–1101.

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. *14*, 125–138.

Wellcome Trust Case Control Consortium, Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature *464*, 713–720.

Wong, J.J.-L., Ritchie, W., Ebner, O.A., Selbach, M., Wong, J.W.H., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K., et al. (2013). Orchestrated intron retention regulates normal granulocyte differentiation. Cell *154*, 583–595.

Xi, R., Hadjipanayis, A.G., Luquette, L.J., Kim, T.-M., Lee, E., Zhang, J., Johnson, M.D., Muzny, D.M., Wheeler, D.A., Gibbs, R.A., et al. (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc. Natl. Acad. Sci. *108*, E1128–E1136.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. Nucleic Acids Res. *44*, D710-716.

Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome. Nat. Rev. Genet. *16*, 172–183.

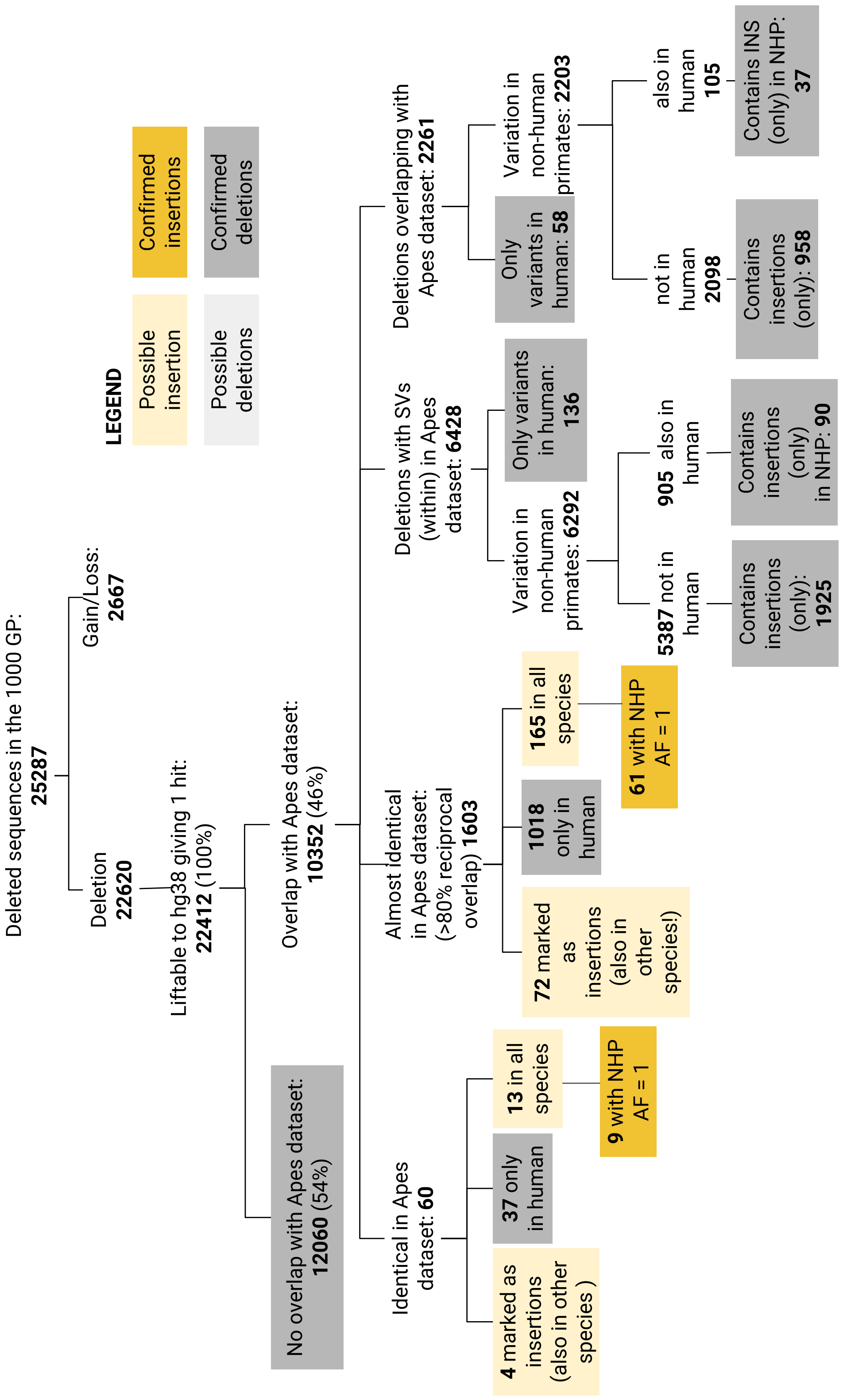
Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T., and Flicek, P.R. (2015). The ensembl regulatory build. Genome Biol. *16*, 56.

Zhang, F., Gu, W., Hurles, M.E., and Lupski, J.R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. Annu. Rev. Genomics Hum. Genet. *10*, 451–481.

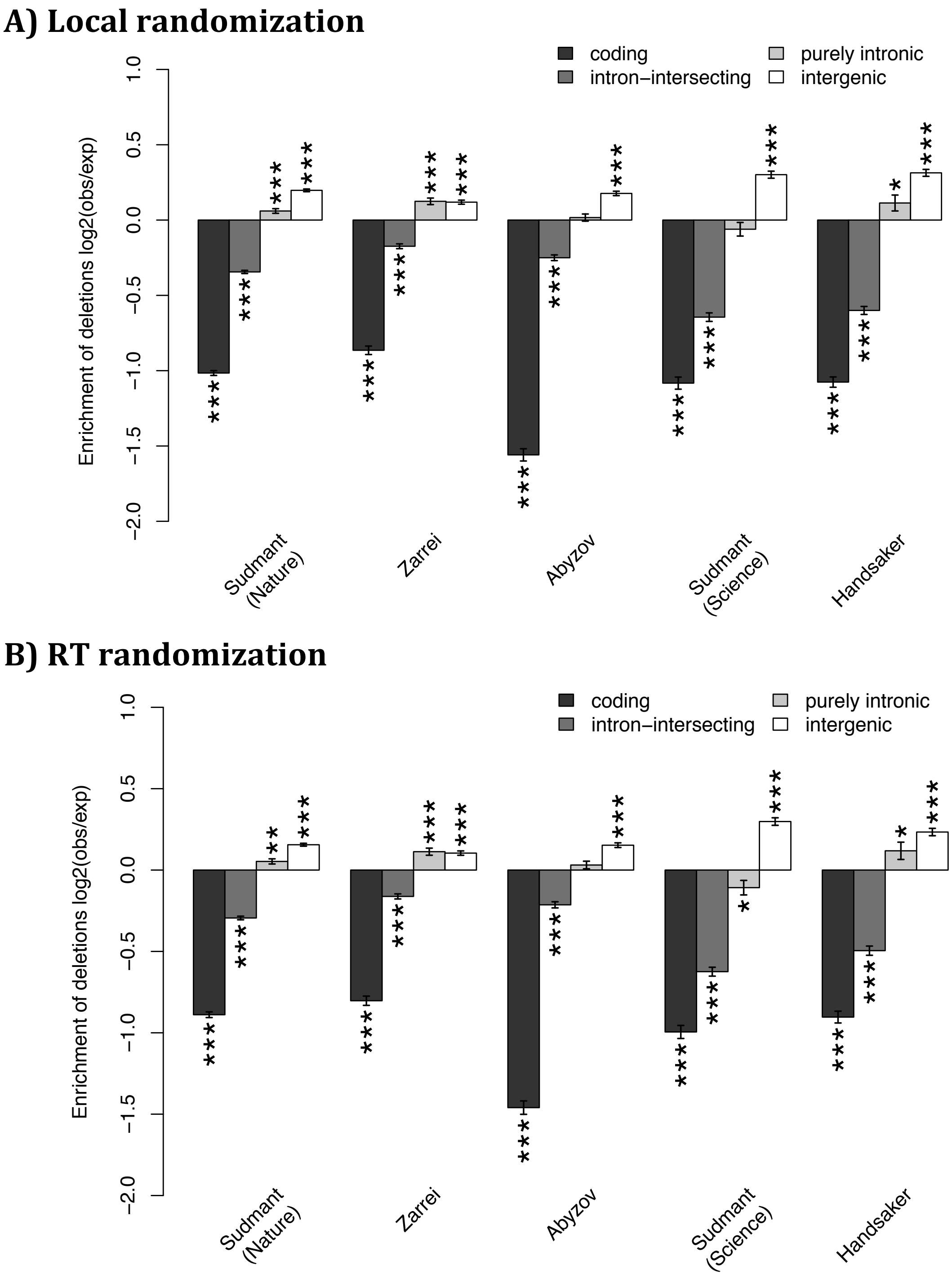
Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J.-Q., and Tian, D. (2009). Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics *10*, 47.

Supplementary material

Supplementary figures


LEGEND
Possible insertion
Confirmed insertions
Possible deletions
Confirmed deletions
Deleted sequences in the 1000 GP: 25287
Deletion 22620
Gain/Loss: 2667
Liftable to hg38 giving 1 hit: 22412 (100%)
No overlap with Apes dataset: 12060 (54%)
Overlap with Apes dataset: 10352 (46%)
Identical in Apes dataset: 60
4 marked as insertions (also in other species)
37 only in human
13 in all species
9 with NHP AF = 1
Almost identical in Apes dataset: (>80% reciprocal overlap) 1603
72 marked as insertions (also in other species!)
1018 only in human
165 in all species
61 with NHP AF = 1
Deletions with SVs (within) in Apes dataset: 6428
Variation in non-human primates: 6292
Only variants in human: 136
5387 not in human
905 also in human
Contains insertions (only): 1925
Contains insertions (only) in NHP: 90
Deletions overlapping with Apes dataset: 2261
Only variants in human: 58
Variation in non-human primates: 2203
not in human 2098
also in human 105
Contains insertions (only): 958
Contains INS (only) in NHP: 37

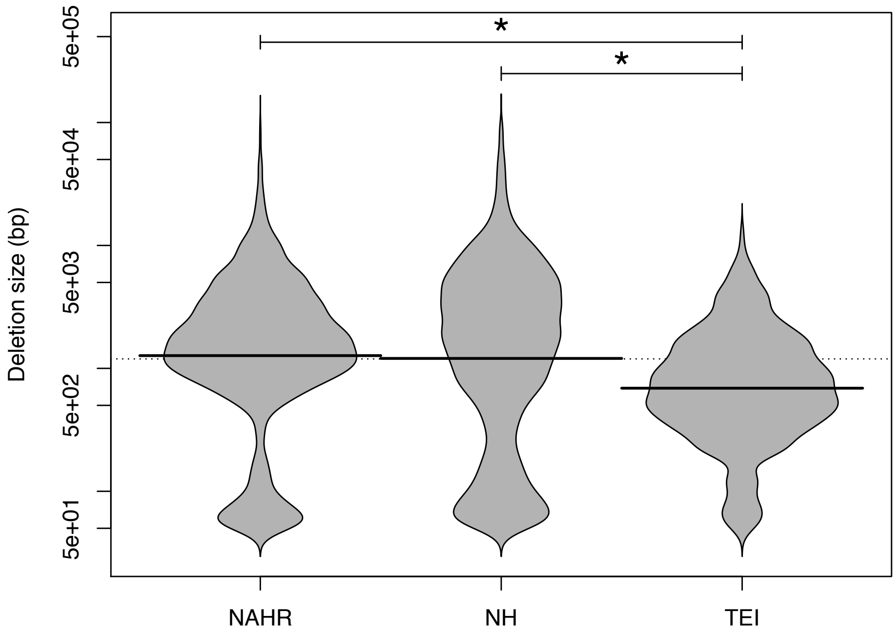
Supplementary figure 1 | Ancestral state of deletions. Decision tree to assign the type of CNV (insertion or deletion) based on the ancestral state.



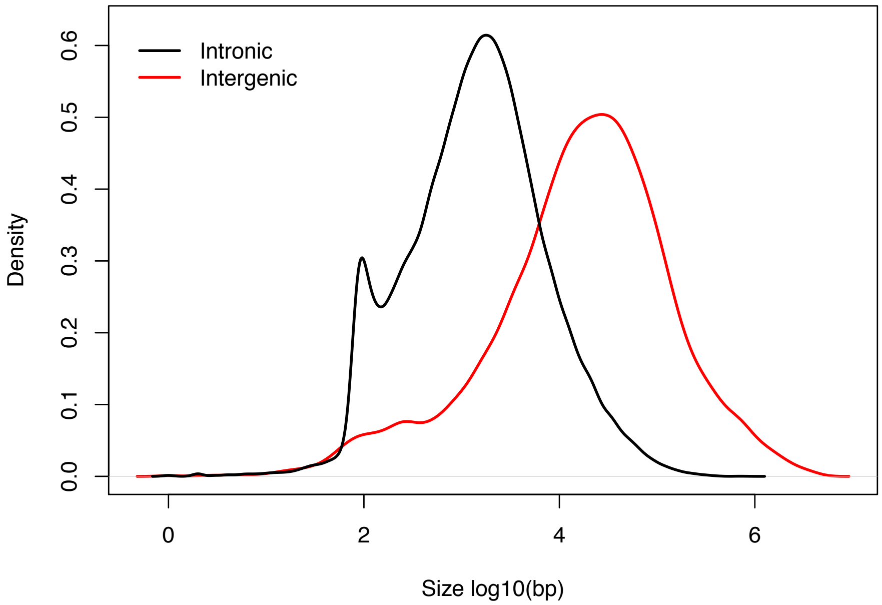
A) Local randomization

B) RT randomization

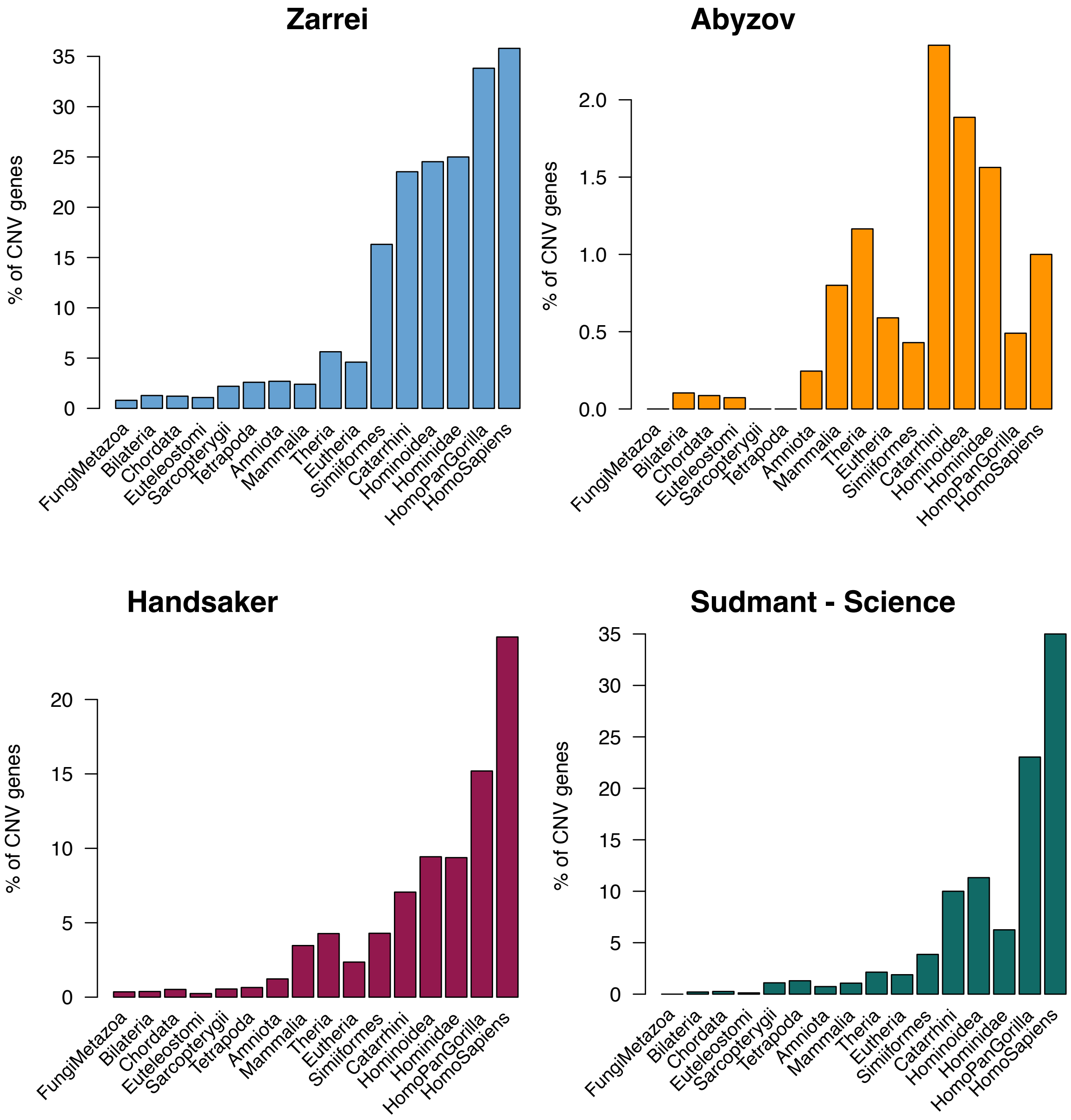
Supplementary figure 2 | Equivalent to **Figure 9** using local model. Enrichment of deletions. Height of the bar is the median of the ratio between the observed number of overlaps and each of the 10,000 randomized sets. Whiskers show median absolute deviation and asterisks mark significance: * for P<0.05, ** for P<0.005 and *** for P<0.0005. The randomized sets were obtained by A) Local or B) RT-controlling randomizations (see section "Statistical assessment of genome-wide distribution of CNVs" from Materials and Methods).



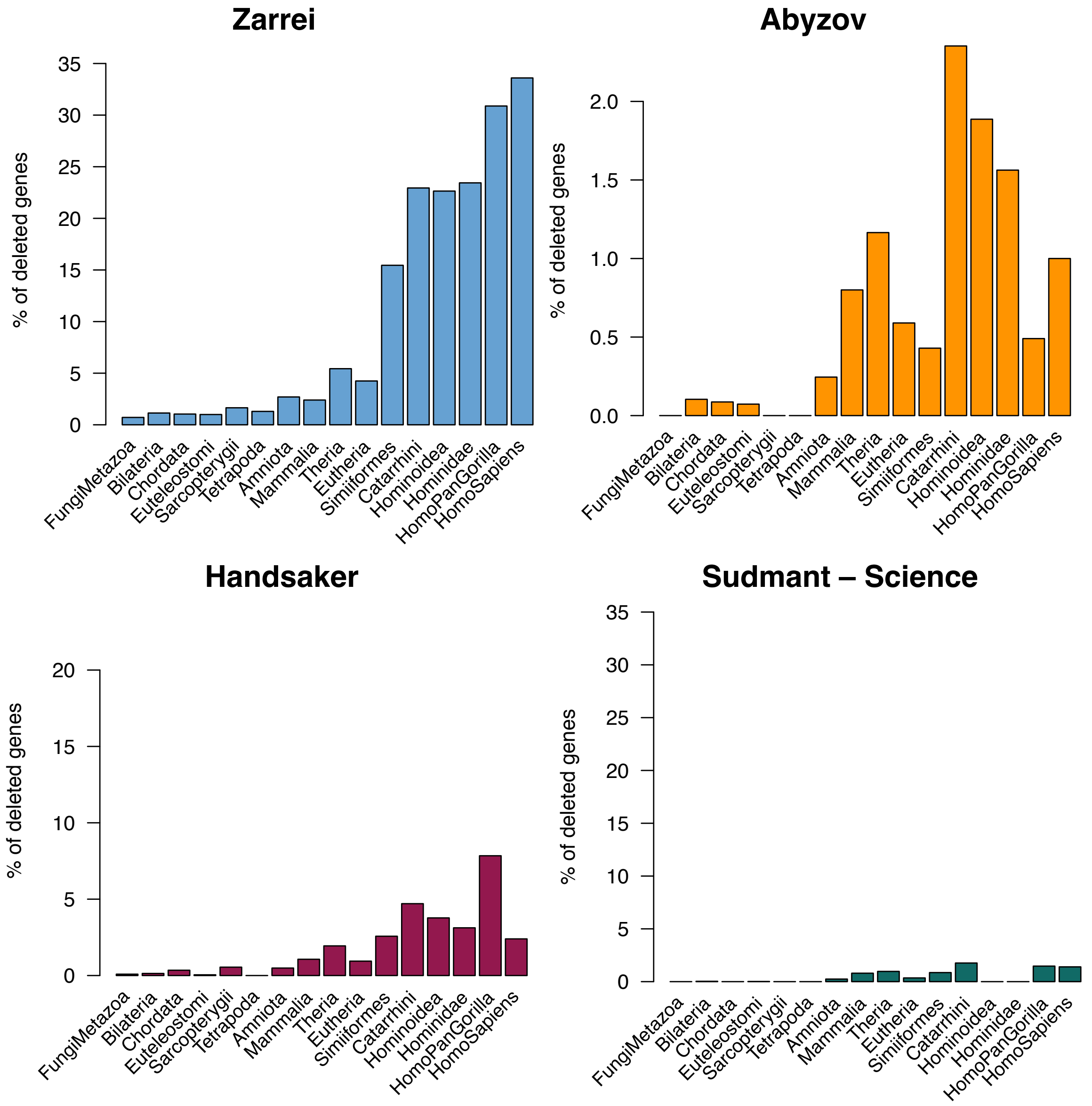
Supplementary figure 3 | Size of the deletions generated through different mechanisms. Differences in the size of deletions from Sudmant-Nature, classified by their generation mechanism. Statistical significance marked with asterisks for P < 0.05, calculated using Wilcoxon tests (NAHR-TEI: P= 5.30e-31, NH-TEI P= 4.02e-21.



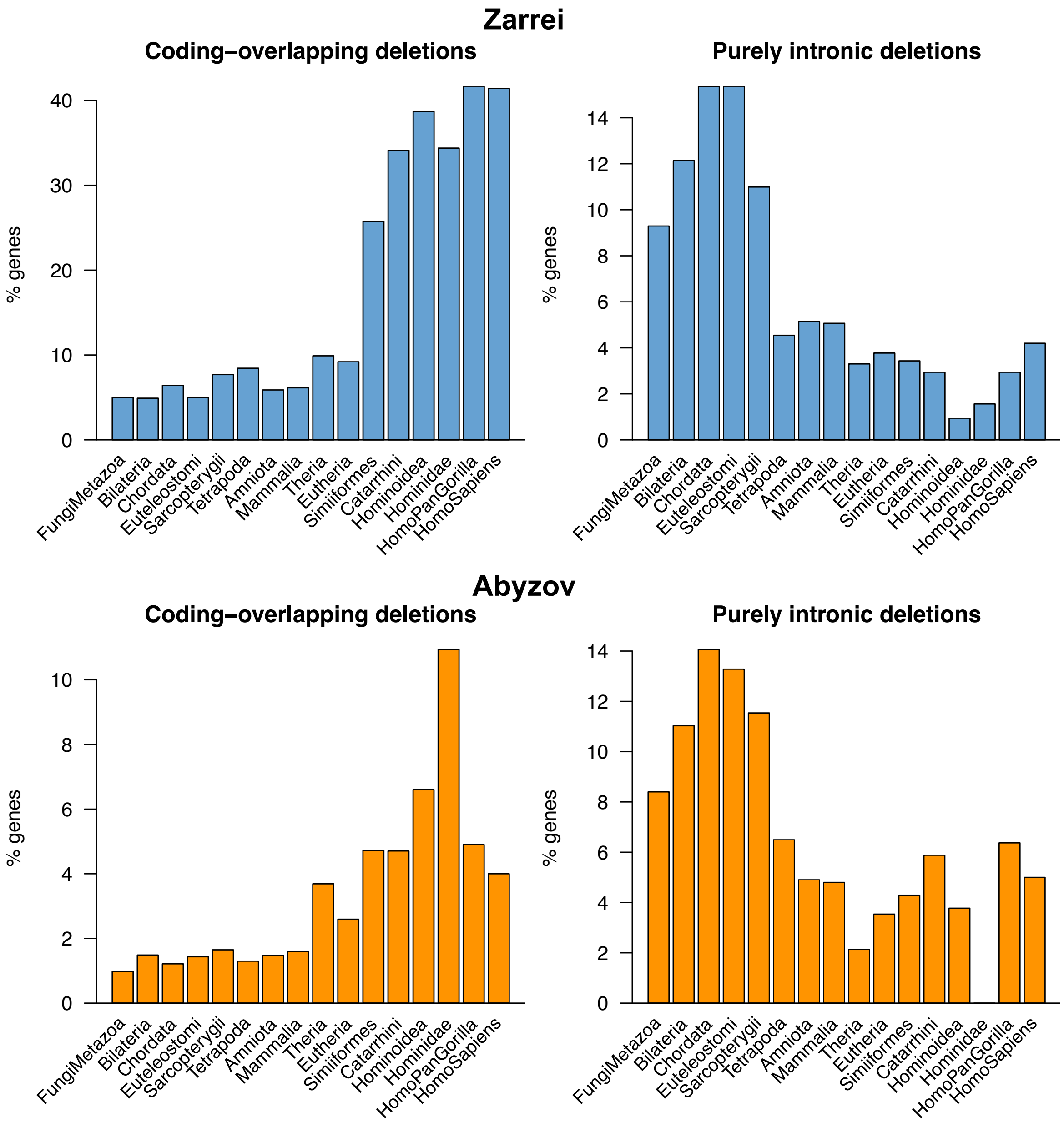
Supplementary figure 4 | Size distribution of intronic and intergenic regions



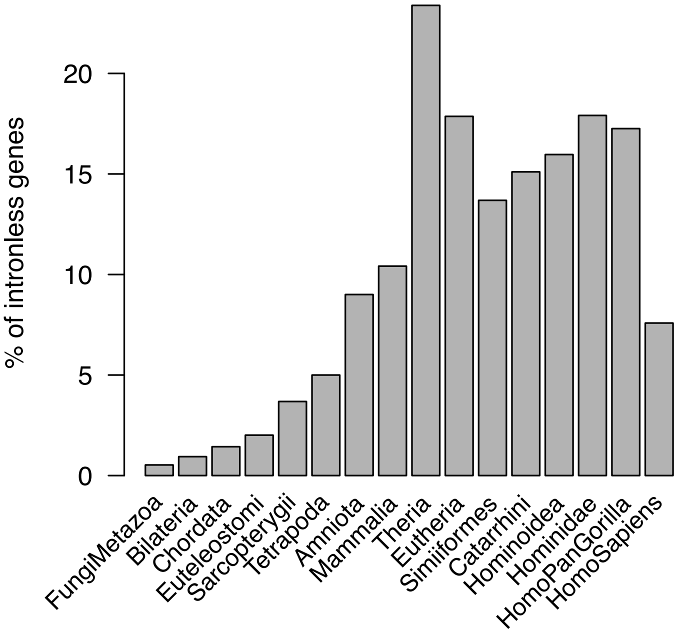
Supplementary figure 5 | Percentage of genes from each evolutionary age group that are completely covered by a CNV in other maps. This figure is equivalent to panel A from Figure 13. Figure 13 represents data from Sudmant-Nature and this figure the results obtained with the other CNV maps. Notice that Abyzov has only deletions.



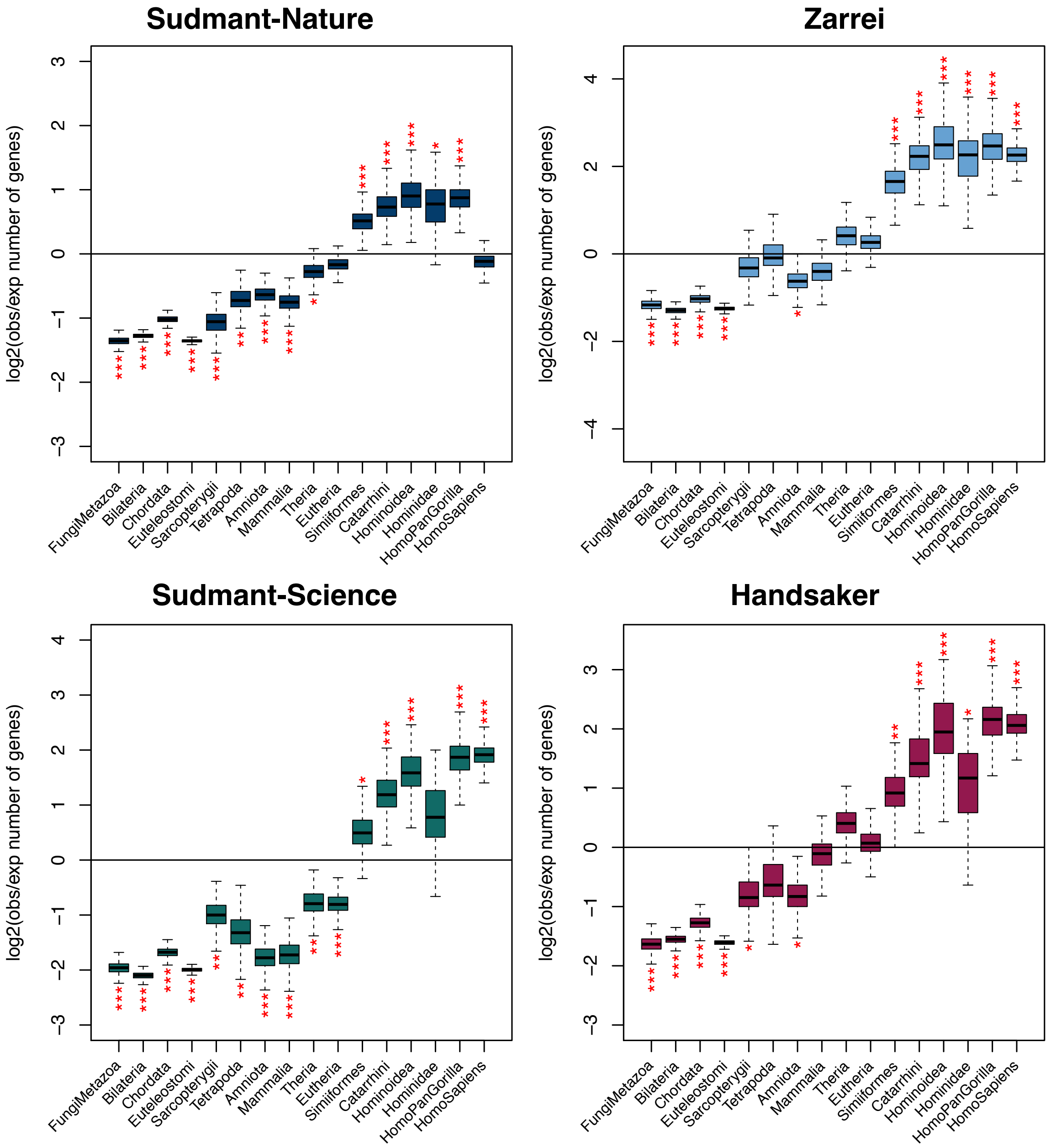
Supplementary figure 6 | Percentage of genes from each evolutionary age group that are completely covered by a deletion in other maps. This figure is equivalent to panel B from Figure 13. Figure 13 represents data from Sudmant-Nature and this figure the results obtained with the other CNV maps. Notice that Abyzov has only deletions.



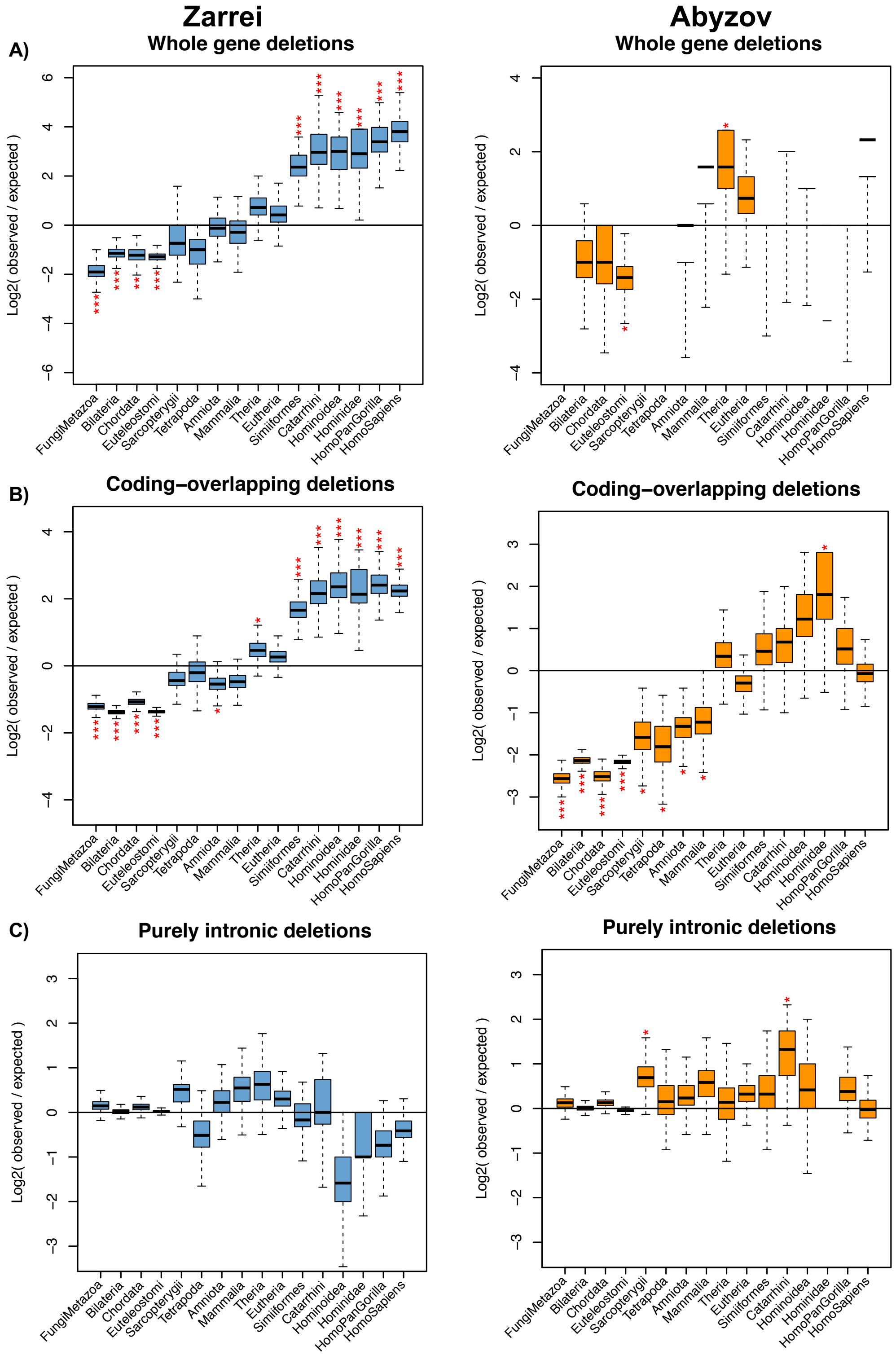
Supplementary figure 7 | Impact of deletions on genes of different evolutionary ages in other maps. Figure equivalent to Figure 14 showing the percentage of genes from each gene evolutionary age that harbor exon-overlapping deletions or that carry intronic deletions (B). Handsaker and Sudmant-Science not shown due to the lower number of intronic deletions in these maps.



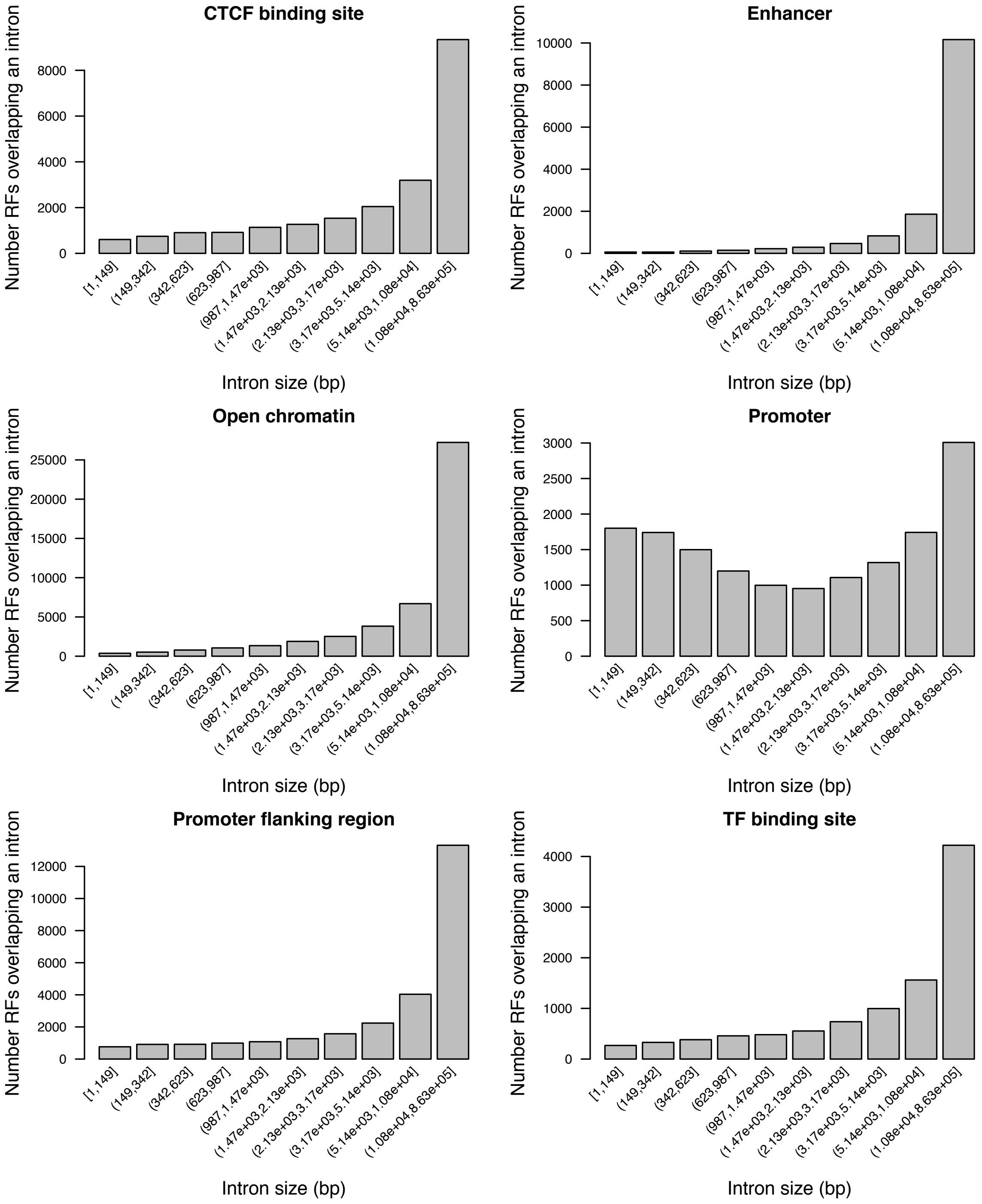
Supplementary figure 8 | Proportion of intronless genes per age. Percentage of genes from each age group that does not have introns.



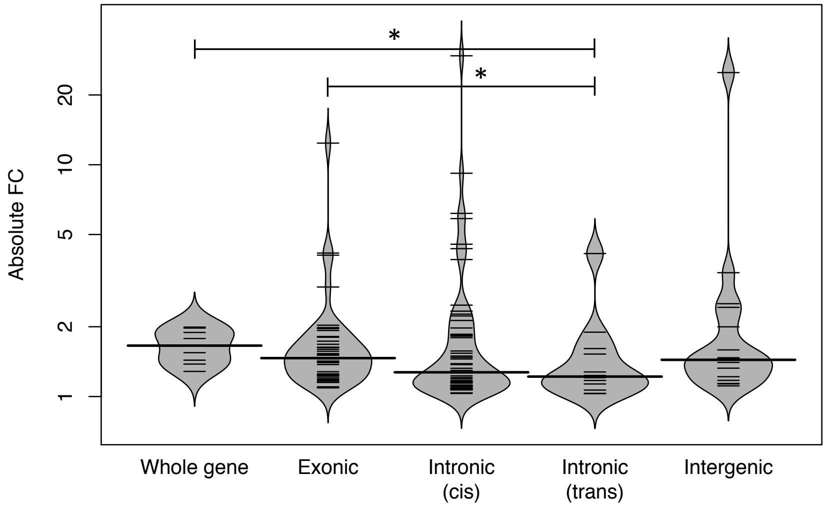
Supplementary figure 9 | Enrichment of CNVs overlapping with coding sequence in different evolutionary ages. Ratios of observed versus expected number of genes from each gene evolutionary age that carry exon-overlapping **deletions or gains.** Expected values were calculated with 10,000 random permutations using a global background model. Asterisks mark the significance for each age group: * for P<0.05, ** for P<0.005 and *** for P<0.0005.



Supplementary figure 10 | Enrichment of gene-overlapping deletions in Abyzov and Zarrei. Ratios of observed vs. expected number of genes per evolutionary age that are fully deleted (A) or carry exon-overlapping deletions (B) or purely-intronic deletions (C). Expected values calculated with 10,000 randomizations using a global background model. Asterisks mark the significance in each group: * for P<0.05, ** for P<0.005 and *** for P<0.0005.



Supplementary figure 11 | Number of regulatory regions overlapping introns of different sizes



Supplementary figure 12 | Effect size of different types of eDeletions. Absolute log2 ratio between the median gene expression of wild-type versus heterozygous individuals. Significant differences using Wilcoxon tests are marked with an asterisk at P < 0.05.

Supplementary tables

Supplementary table 1 | Characteristics of brain-specific genes

	Brain enriched genes (419 genes)			**Brain elevated genes (1460 genes)**		
	Brain genes median	Other genes median	P-value	Brain genes median	Other genes median	P-value
Size of genes (kb)	41.32	25.80	**2.75e-9**	52.78	25.06	**4.59e-56**
Size of introns (kb)	1.89	1.46	**3.37e-25**	2.24	1.42	**1.88e-270**
Size of exons (bp)	144	134	**8.56e-16**	142	134	**3,73e-37**
Exons per gene	9	9	0.52	10	9	**1.14e-4**
Exonic bp per gene	2,376	1,597	**9.23e-21**	3,632	2,561	**1,80e-67**

Differences in structure between brain genes and other genes. The lists of brain-specific genes was retrieved from the The Human Protein Atlas (Uhlén et al., 2015) on April, 2019. Brain-enriched genes include genes at least five-fold higher mRNA levels in brain compared to all other tissues. Brain-elevated genes include all brain-enriched genes plus genes with at least five-fold higher mRNA levels in a group of 2-7 tissues plus genes with at least five-fold higher mRNA levels compared to average levels in all tissues.

Supplementary table 2 | Additional information on the origin of the CNV maps

	Sudmant (Nature)	**Zarrei**	**Abyzov**	**Handsaker**	**Sudmant (Science)**
Individuals	2504	2647	1092	849	236
Project	1KGP, Phase 3	Meta analysis of DGV collection	1KGP, Phase 1	1KGP, Phase 1	
Populations	26*	15	14*	14*	125
Methods	WGS, Multiple algorithms	Multiple genome-wide techniques	WGS, Multiple algorithms	WGS, Read-depth	WGS, Read-depth
CN states	Absolute CN	Gains / Losses	Losses	Absolute CN	Absolute CN

Characteristics of the maps analysed in this thesis. *Abyzov and Handsaker analyse the low-coverage alignments from the phase 1 of the 1KGP. Abyzov analyses all the samples and Handsaker a subset that includes samples from all the populations represented in Abyzov's map. Each of the 26 populations in Sudmant-Nature include 60-100 individuals from Europe, Africa, America (South -not in phase 1- and North) and Asia (South -not in phase 1- and East).

Supplementary table 3 | Enrichment of deletions in essential genes

	Sudmant-Nature	Zarrei	Abyzov
Essential	**0,127 (P = 0,0001)**	**0,1438 (P = 0,002)**	**0,0345 (P = 0,2683)**
Not essential	**0,0358 (P = 0,0299)**	**0,1235 (P = 0,0001)**	**0,0053 (P = 0,4282)**

Log2 ratios of observed versus expected number of intronic deletions in essential and non essential genes. Expected values calculated using the "global" randomization model (see Methods for details).

Supplementary table 4 | List of genes with less intronic deletions than expected

ABI1	ABL2	ADAM23	ADAMTS16	ADAMTS20	ADCY2
ADCY5	ADRA1A	ADRBK2	AGBL1	AGTPBP1	AJAP1
AK5	AKAP7	ALCAM	ALPK2	AMBRA1	ANK3
ANKFN1	ANKH	APBA2	ARFIP1	ARHGAP10	ARHGAP20
ARHGAP22	ARHGEF26	ARHGEF35	ARHGEF4	ARID2	ARSG
ASAP1	ASXL3	ATE1	ATP9A	ATXN10	ATXN7
BBX	BCAT1	BCL11B	BMPER	BRINP1	BRINP2
C10orf76	C12orf56	C4orf45	CACNA1D	CACNA1E	CADM1
CADPS2	CCSER2	CD247	CDC42SE2	CDH11	CDH2
CDH6	CDH9	CDK17	CELF4	CEP85L	CFTR
CGNL1	CHCHD3	CHD6	CHD7	CHRM2	CHRM3
CHST8	CLMP	CLNK	CLPB	CLYBL	CNGB3
CNNM2	CNTNAP3B	COBLL1	COL14A1	COL21A1	COL8A1
COLEC12	CPEB3	CPNE8	CPPED1	CRIM1	CRYL1
CTDSPL	CTNNBL1	CTTNBP2	CYP2C19	CYP7B1	CYSTM1
CYTH3	CYYR1	DAAM1	DAPK2	DCDC1	DEPTOR
DERA	DIAPH3	DKK2	DNAJC6	DNM3	DOK5
DPH6	DPY30	DPYD	DSCAML1	DTD1	DTNBP1
DYNC1I1	E2F3	EBF1	EBF2	EEFSEC	EEPD1
EGFLAM	EHBP1	EIF3H	ELAVL2	ENPP6	ENTHD1
ENTPD1	EPB41L4A	EPB41L5	EPC2	EPHA5	EPHA7
EPHB2	EPM2A	ERP44	EXT2	FAM110B	FAM117B
FAM168A	FAM172A	FAM19A1	FAM53B	FAM78B	FANCC
FANK1	FAR2	FARP1	FAT4	FBXO42	FBXW11
FBXW7	FCHSD2	FER1L6	FIGN	FLI1	FLRT2
FLT1	FNDC3A	FNDC3B	FNIP1	FOXO3	FOXP1
FOXP2	FRK	FRMD6	FSIP1	FTO	FUT10
GAB1	GAB2	GABRA2	GABRB1	GABRB2	GALNT2
GAP43	GBF1	GFOD1	GFRA2	GNAO1	GNG12
GPR176	GRB10	GRB14	GRIN2A	GRM1	GRM3
GTDC1	HCAR1	HCRTR2	HIVEP1	HOMER1	HTR1E
HTR4	HYDIN	IGF1R	IGSF21	IKZF2	INPP4A
INSC	IQCK	ITFG1	ITPKB	ITPR2	JAKMIP2
KALRN	KCNC2	KCND3	KCNH1	KCNH5	KCNH7

Continuation of supplementary table 4

KCNJ6	KCNK13	KCNN2	KCNN3	KCNQ3	KCNU1
KIAA0247	KIAA1199	KIAA1211	KIAA1211L	KIAA1549	KIF21A
KIRREL3	KSR1	KSR2	LAMC1	LARP1B	LCORL
LDB2	LDLRAD4	LEF1	LIN28B	LIN52	LIN7A
LIPC	LMX1B	LNX1	LPAR1	LPHN2	LRCH1
LRP12	LRRC49	LRRC8D	LRRIQ1	LRRIQ3	LYPD6
MAN2A1	MAP2	MAP2K6	MAPK1	MAPK4	MAPK8
MAPKAP1	MARK1	MATN2	MBOAT2	MCF2L2	MCTP1
MCU	MDFIC	MEGF11	MEIS1	MEIS2	MEMO1
METTL8	MICU1	MIPOL1	MKL2	MLLT10	MME
MNAT1	MRPS28	MSRB3	MTSS1	MYLK	MYO18B
MYO1B	MYO1D	MYO1E	MYO3A	MYO5A	NARS2
NAV3	NCOA1	NCOA3	NECAB1	NEDD4	NEK11
NEK7	NFASC	NFAT5	NFATC3	NFIA	NIPBL
NMNAT3	NPSR1	NREP	NRG1	NSF	NTRK3
NXPH2	OCA2	OPRM1	OSBP2	OSBPL3	OTUD7A
PACRG	PARM1	PARP8	PAX3	PAX5	PAX7
PCDH17	PCDHA5	PCDHA6	PCDHA7	PCDHA8	PCDHA9
PCED1B	PCNXL2	PDE10A	PDE1C	PDE3A	PDZRN3
PEBP4	PELI2	PEPD	PGM5	PHF14	PHLDB2
PHTF2	PIBF1	PKIA	PKN2	PKP4	PLA2G4A
PLAGL1	PLCXD2	PLXDC2	PLXNA2	POLN	POU2F1
PPAP2A	PPP1R12B	PREP	PRICKLE2	PRKCQ	PRR5L
PTPN14	PTPN4	PTPRQ	RAI14	RALGPS1	RALGPS2
RANBP17	RARB	RASGRF2	RBMS1	REEP1	REEP3
RERG	RFWD2	RHOBTB1	RIT2	RNF144A	RNF180
RNF217	RORA	RREB1	RSU1	RUNX2	RYR2
SAMD4A	SCAI	SCN2A	SCN8A	SEL1L2	SESN1
SFMBT2	SGPP2	SH3PXD2A	SH3RF2	SH3RF3	SHB
SHC3	SHROOM3	SIL1	SIPA1L1	SIPA1L2	SLC16A10
SLC16A12	SLC1A2	SLC24A3	SLC35F1	SLC41A2	SLC4A10
SLC6A11	SLCO1B3	SLCO1B7	SLX4IP	SMAP1	SMARCC1
SNCAIP	SND1	SNRK	SOBP	SORCS1	SORCS3
SPATA16	SPATA5	SPATS2L	SPECC1	SPECC1L	SPHKAP
SPOCK1	ST18	ST5	ST8SIA1	STAG1	STK3
STON2	STRBP	STX18	STXBP4	SUSD4	SV2C
SYN3	SYNPO2	SYNPR	SYT16	TACR1	TAOK3
TASP1	TBC1D4	TBC1D8	TCF12	TCF4	TCF7L2
TDRD3	TEC	TGFA	TGFB2	TMEM241	TOM1L2
TOX	TOX3	TPD52	TPST1	TRAF3	TRIQK
TSHZ2	TSPAN18	TSPAN5	UBL3	USP12	UVRAG
VEPH1	VKORC1L1	VPS41	VSNL1	VWC2	WARS2
WASF3	WDFY2	WDFY4	WDR49	WDR7	WWC1
XXYLT1	ZBTB16	ZCCHC7	ZEB1	ZEB2	ZFHX4
ZHX2	ZNF608	ZNF618	ZNF644	ZNF704	ZPLD1
ZSWIM6	ZZZ3				

Supplementary table 5 | List of genes with more intronic deletions than expected

A2ML1	ABCB5	ACAA1	ACAP1	ACER1	ACKR4
ACOT11	ACOT12	ACOT6	ACSF3	ACVR1B	ACYP1
ADA	ADAMTS10	ADIPOR2	AK1	AKR1C4	ALKBH3
ALLC	ALPK1	AMN1	ANKRD36C	ANKS3	ANLN
ANO1	APPBP2	AQP8	ARHGAP19	ARHGAP8	ARHGEF10
ARMC7	ASCC2	ASNA1	ATOX1	ATP2A3	ATP5A1
ATP5H	ATP6V0E1	ATP6V1E1	AVPI1	AZU1	B3GALT4
BACE1	BAHD1	BAX	BCL2L11	BCL2L13	BDKRB2
BOK	BOLL	BOP1	C11orf74	C16orf46	C17orf85
C1orf170	C1orf177	C1QBP	C2orf73	C4BPA	C6orf10
C6orf203	CAB39L	CACNA1H	CACNG7	CASS4	CBWD1
CBX1	CCDC114	CCDC169	CCDC50	CCDC66	CCDC77
CCNB2	CCND2	CCR6	CD2	CDA	CDC40
CDCA2	CDCP2	CEACAM4	CENPC	CEP120	CERS5
CHADL	CHMP1A	CHPT1	CLCC1	CLEC17A	CLHC1
CLIC6	CLPTM1	CLSTN1	CLUL1	CNN2	CNOT1
COL10A1	COL18A1	COLEC10	COMMD7	COX20	CRABP2
CSF1R	CSNK1G2	CSNK2A2	CWC25	CWF19L2	CXCL16
CYP4F11	DACT2	DAD1	DAK	DAP3	DAPL1
DBF4B	DCAKD	DCPS	DCST2	DCTN5	DDB2
DEAF1	DEFB107B	DEPDC1B	DHCR24	DIABLO	DMPK
DNAJC8	DOCK5	DOK7	DRAM1	DUSP3	DVL3
DYRK3	EBNA1BP2	EDEM2	EGLN1	EIF2B5	EIF3E
ELMSAN1	ELP6	EMID1	EWSR1	EXD3	F7
FADS6	FAHD1	FAM104A	FAM105A	FAM117A	FAM153A
FAM153B	FAM154B	FAM167A	FAM179A	FAM195B	FAM220A
FBXL12	FBXO28	FBXO41	FBXO6	FGL1	FKBP3
FLYWCH2	FRZB	FSTL1	FUT5	GABRR1	GALNT15
GALNTL5	GAS6	GATA4	GATSL3	GCFC2	GFRA3
GIPC1	GJA3	GLOD4	GMEB2	GMPR	GOLGA8A
GOLGA8B	GPR161	GRK4	GRTP1	GSTA2	GTF2F1
GTF3C5	GUCY1A3	HAT1	HEATR4	HGF	HIF3A
HOPX	HSD17B11	HSF1	IARS2	IDI1	IFT52
IL17REL	IL1A	IL1RL1	IL27RA	IL2RA	IL32
INSIG2	IRAK2	IRS2	JMJD7	KANK2	KBTBD11
KCNE2	KCNJ15	KIAA0101	KIAA0368	KIAA1257	KIAA1467
KIF19	KLB	KMO	L2HGDH	LAIR2	LATS1
LGI1	LHX4	LHX9	LILRA2	LINGO1	LITAF
LMF1	LMNB2	LOXL4	LRRC8E	LRRN4	MAP1LC3B2
MAPK9	MAPKAPK5	MARCO	MAX	MCCC2	MCFD2
MCM3AP	METAP1D	MINPP1	MITF	MOV10	MRPL19
MRPS35	MS4A6A	MSTO1	MTERFD2	MXD1	MYADML2
MYCT1	MZT1	NAA15	NAA20	NAT1	NCMAP
NFE2L3	NINJ1	NIPSNAP1	NLN	NLRP5	NOC4L
NOD1	NOTCH1	NOTCH4	NPAS1	NPAT	NT5M
NTSR1	NUDT4	NUP205	NUP43	NXPE1	ODF1
ODF4	OIP5	OR2W3	OTUD6B	PADI4	PAIP2

Continuation of supplementary table 5

PANK2	PANX1	PARD6B	PBLD	PCCB	PCGF3
PCYOX1	PDE6B	PDLIM3	PDSS1	PFKP	PGAP3
PHAX	PHYHD1	PLA2G2C	PLA2G5	PLBD1	POGZ
POLR3K	POPDC3	POU1F1	PPP1R13L	PPP1R37	PPP2R5A
PPP6C	PPT1	PRKAA1	PRPSAP1	PRPSAP2	PRR11
PRR5-ARHGAP8	PRRG4	PSMA8	PSMD9	PTGR1	PWWP2A
PWWP2B	PYGL	QPRT	QRFPR	QRICH2	RAB19
RAB4A	RAD52	RAP1B	RAP2A	RASA3	RASSF4
RBM25	RD3	REEP5	REV1	RGS13	RHOF
RIBC2	RNF11	RNF168	RNF212	RNF219	RNF38
RPGRIP1	RPH3AL	RPL27	RSPH6A	SAFB	SAMHD1
SBK2	SBNO1	SCAP	SCN3A	SELPLG	SERPINF2
SESN3	SETD1A	SEZ6L2	SGSM3	SH3GLB2	SH3TC1
SH3TC2	SHC2	SIRT4	SLC15A5	SLC16A13	SLC17A5
SLC1A7	SLC22A2	SLC25A18	SLC25A37	SLC27A5	SLC3A2
SLC6A20	SMAD1	SNRNP27	SNRPN	SP3	SPATA24
SPIRE2	SSC5D	STK17A	STMN3	STYXL1	SULT2B1
SUN1	SVOPL	SYNM	SYPL1	TADA2A	TAS1R1
TBC1D2	TBCE	TCEA3	TCTN3	TECR	TEKT5
TERT	TES	TFIP11	TIMMDC1	TIMP4	TJP3
TM4SF19	TMCO1	TMEM11	TMEM121	TMEM165	TMEM192
TMEM39B	TMEM41B	TMEM65	TMEM68	TMEM72	TMIGD2
TNFSF13B	TNN	TOPBP1	TOX4	TPH1	TRAPPC6A
TRAPPC6B	TRIM13	TRIM29	TRIM67	TRMT61B	TSPAN13
TSPAN16	TUBGCP6	TXLNB	UNC119B	USP8	UTP20
UTS2	VANGL1	VPS33A	VSTM5	WDR34	WDR46
WDR47	WDR65	WDR76	WDSUB1	WDYHV1	WFDC8
XCL1	XCR1	XRN2	YIF1B	ZBTB5	ZC3H12D
ZC3H18	ZC3H7A	ZCCHC24	ZDHHC13	ZDHHC19	ZDHHC7
ZER1	ZFAND2A	ZFP14	ZFP2	ZFR2	ZNF106
ZNF14	ZNF142	ZNF143	ZNF207	ZNF264	ZNF268
ZNF429	ZNF43	ZNF454	ZNF483	ZNF487	ZNF490
ZNF492	ZNF554	ZNF566	ZNF57	ZNF570	ZNF665
ZNF675	ZNF700	ZNF701	ZNF730	ZNF763	ZNF814
ZSCAN5A	ZYG11A				

Supplementary table 6 | Table of differentially expressed genes

	Deletion type	Gene (ENSEMBL)	Gene (HUGO)	Adj. p-value	log2(FC)	Gene age
1	whole gene	ENSG00000184674		2.17e-45	-1.89	
2	whole gene	ENSG00000100068	LRP5L	2.11e-03	-1.78	Eutheria
3	whole gene	ENSG00000184923	NUTM2A	2.15e-04	-1.55	HomoSapiens
4	whole gene	ENSG00000187010	RHD	3.10e-11	-1.43	HomoSapiens
5	whole gene	ENSG00000008128		5.93e-04	-1.38	
6	whole gene	ENSG00000008128		8.45e-05	-1.28	
7	whole gene	ENSG00000184022		1.42e-14	-1.97	
8	whole gene	ENSG00000197888	UGT2B17	7.15e-29	-1.99	Catarrhini
9	exonic	ENSG00000100068	LRP5L	9.53e-04	-1.68	Eutheria
10	exonic	ENSG00000136527	TRA2B	6.16e-16	-1.09	Euteleostomi
11	exonic	ENSG00000177335		2.94e-06	-12.4	
12	exonic	ENSG00000214562	NUTM2D	6.07e-05	-1.63	HomoSapiens
13	exonic	ENSG00000215252	GOLGA8B	5.45e-14	-1.52	HomoSapiens
14	exonic	ENSG00000130812	ANGPTL6	2.83e-07	-1.6	Euteleostomi
15	exonic	ENSG00000249679		1.18e-06	-2.03	
16	exonic	ENSG00000179119	SPTY2D1	1.22e-07	-1.17	Bilateria
17	exonic	ENSG00000128383	APOBEC3A	6.31e-16	-1.45	Hominoidea
18	exonic	ENSG00000179750	APOBEC3B	2.85e-32	-1.8	Hominoidea
19	exonic	ENSG00000179750	APOBEC3B	1.11e-24	-1.73	Hominoidea
20	exonic	ENSG00000116791	CRYZ	2.98e-02	-1.57	Bilateria
21	exonic	ENSG00000100197	CYP2D6	1.49e-02	-1.47	Sarcopterygii
22	exonic	ENSG00000117226	GBP3	3.41e-18	-1.37	Simiiformes
23	exonic	ENSG00000160867	FGFR4	5.77e-03	-4.07	Euteleostomi
24	exonic	ENSG00000188677	PARVB	4.24e-03	-1.15	Euteleostomi
25	exonic	ENSG00000105501	SIGLEC5	2.31e-05	1.46	HomoSapiens
26	exonic	ENSG00000157326	DHRS4	2.49e-02	-1.19	Hominoidea
27	exonic	ENSG00000187630	DHRS4L2	7.71e-04	-1.24	Hominoidea
28	exonic	ENSG00000187630	DHRS4L2	3.64e-02	-1.2	Hominoidea
29	exonic	ENSG00000187630	DHRS4L2	7.13e-04	-1.23	Hominoidea
30	exonic	ENSG00000221923	ZNF880	2.78e-04	-1.22	Catarrhini
31	exonic	ENSG00000204267	TAP2	5.89e-10	1.16	HomoSapiens
32	exonic	ENSG00000134184	GSTM1	1.88e-11	-1.93	Simiiformes
33	exonic	ENSG00000134184	GSTM1	4.84e-10	-1.81	Simiiformes
34	exonic	ENSG00000197888	UGT2B17	1.74e-28	-1.96	Catarrhini
35	exonic	ENSG00000197888	UGT2B17	1.52e-28	-1.98	Catarrhini
36	exonic	ENSG00000196620	UGT2B15	1.06e-03	2.97	Catarrhini
37	exonic	ENSG00000188603	CLN3	7.30e-61	-1.53	HomoSapiens
38	exonic	ENSG00000165935	SMCO2	6.74e-04	-1.42	Theria
39	exonic	ENSG00000183486	MX2	2.29e-03	1.4	Eutheria
40	exonic	ENSG00000175265	GOLGA8A	1.71e-11	-1.42	HomoSapiens
41	exonic	ENSG00000112787	FBRSL1	1.17e-02	-1.25	Euteleostomi
42	exonic	ENSG00000134326	CMPK2	3.02e-02	-1.63	Chordata
43	exonic	ENSG00000141569	TRIM65	4.70e-02	-1.1	Euteleostomi
44	exonic	ENSG00000008128		2.13e-02	-1.41	
45	exonic	ENSG00000173272	MZT2A	2.78e-03	-1.15	HomoPanGorilla

Continuation of supplementary table 6

46	exonic	ENSG00000213366		1.57e-02	-1.28	
47	exonic	ENSG00000204449	TRIM49C	1.75e-03	-4.16	HomoSapiens
48	exonic	ENSG00000142794	NBPF3	9.50e-03	-1.51	Hominoidea
49	intronic (cis)	ENSG00000143156	NME7	6.06e-03	1.11	Bilateria
50	intronic (cis)	ENSG00000094975	SUCO	4.24e-05	1.04	FungiMetazoa
51	intronic (cis)	ENSG00000123684	LPGAT1	2.06e-07	-1.08	Bilateria
52	intronic (cis)	ENSG00000143740	SNAP47	6.10e-03	1.08	Euteleostomi
53	intronic (cis)	ENSG00000144451	SPAG16	1.37e-22	-1.3	Bilateria
54	intronic (cis)	ENSG00000163359	COL6A3	1.96e-03	2.13	Chordata
55	intronic (cis)	ENSG00000163686	ABHD6	1.27e-06	-1.5	Euteleostomi
56	intronic (cis)	ENSG00000163754	GYG1	2.89e-08	-1.16	Euteleostomi
57	intronic (cis)	ENSG00000109667	SLC2A9	5.91e-03	1.17	Chordata
58	intronic (cis)	ENSG00000138759	FRAS1	6.79e-03	1.86	Chordata
59	intronic (cis)	ENSG00000151466	SCLT1	6.04e-42	-1.14	Chordata
60	intronic (cis)	ENSG00000112977	DAP	2.45e-07	-1.09	Bilateria
61	intronic (cis)	ENSG00000123213	NLN	3.69e-02	1.11	Euteleostomi
62	intronic (cis)	ENSG00000164176	EDIL3	9.85e-09	2.33	Euteleostomi
63	intronic (cis)	ENSG00000113615	SEC24A	6.58e-04	-1.03	Euteleostomi
64	intronic (cis)	ENSG00000182578	CSF1R	6.22e-04	1.57	Euteleostomi
65	intronic (cis)	ENSG00000170074	FAM153A	2.44e-07	9.19	HomoSapiens
66	intronic (cis)	ENSG00000112685	EXOC2	2.78e-04	-1.11	Bilateria
67	intronic (cis)	ENSG00000112137	PHACTR1	3.02e-02	1.23	Amniota
68	intronic (cis)	ENSG00000112378	PERP	7.78e-04	1.45	Euteleostomi
69	intronic (cis)	ENSG00000005020	SKAP2	9.33e-03	-1.21	Euteleostomi
70	intronic (cis)	ENSG00000164543	STK17A	1.82e-07	-1.26	Euteleostomi
71	intronic (cis)	ENSG00000127952	STYXL1	8.58e-10	1.19	Chordata
72	intronic (cis)	ENSG00000187391	MAGI2	6.23e-06	-4.54	Bilateria
73	intronic (cis)	ENSG00000158528	PPP1R9A	2.43e-04	-4.34	Euteleostomi
74	intronic (cis)	ENSG00000135250	SRPK2	2.30e-03	-1.12	Euteleostomi
75	intronic (cis)	ENSG00000164946	FREM1	2.49e-02	2.23	Chordata
76	intronic (cis)	ENSG00000106853	PTGR1	1.55e-02	-1.32	Bilateria
77	intronic (cis)	ENSG00000136848	DAB2IP	7.78e-07	-3.9	Chordata
78	intronic (cis)	ENSG00000136895	GARNL3	3.31e-02	-1.79	Bilateria
79	intronic (cis)	ENSG00000175287	PHYHD1	2.01e-02	-1.98	Bilateria
80	intronic (cis)	ENSG00000148948	LRRC4C	8.91e-04	-1.54	Euteleostomi
81	intronic (cis)	ENSG00000148948	LRRC4C	2.57e-03	-1.82	Euteleostomi
82	intronic (cis)	ENSG00000118971	CCND2	3.18e-03	-1.16	Euteleostomi
83	intronic (cis)	ENSG00000118971	CCND2	3.18e-03	-1.16	Euteleostomi
84	intronic (cis)	ENSG00000165714	LOH12CR1	1.52e-03	-1.14	Bilateria
85	intronic (cis)	ENSG00000165714	LOH12CR1	1.52e-03	-1.14	Bilateria
86	intronic (cis)	ENSG00000205323	SARNP	3.54e-05	-1.12	Chordata
87	intronic (cis)	ENSG00000196792	STRN3	9.51e-08	1.12	Euteleostomi
88	intronic (cis)	ENSG00000151812	SLC35F4	3.46e-02	-6.17	Euteleostomi
89	intronic (cis)	ENSG00000133985	TTC9	5.88e-03	-1.39	Euteleostomi
90	intronic (cis)	ENSG00000140157	NIPA2	6.84e-25	1.09	Euteleostomi
91	intronic (cis)	ENSG00000066933	MYO9A	2.78e-03	1.84	Euteleostomi
92	intronic (cis)	ENSG00000067225	PKM	3.45e-02	-1.09	Euteleostomi

Continuation of supplementary table 6

93	intronic (cis)	ENSG00000064270	ATP2C2	2.25e-30	-29.52	Euteleostomi
94	intronic (cis)	ENSG00000131469	RPL27	4.82e-02	1.08	FungiMetazoa
95	intronic (cis)	ENSG00000121104	FAM117A	3.05e-02	1.07	Euteleostomi
96	intronic (cis)	ENSG00000141376	BCAS3	9.07e-04	-2.27	Bilateria
97	intronic (cis)	ENSG00000150477	KIAA1328	2.79e-03	-1.37	Euteleostomi
98	intronic (cis)	ENSG00000197256	KANK2	2.47e-03	-1.15	Euteleostomi
99	intronic (cis)	ENSG00000197013	ZNF429	1.02e-13	1.27	Simiiformes
100	intronic (cis)	ENSG00000142065	ZFP14	1.47e-02	-1.07	Eutheria
101	intronic (cis)	ENSG00000149596	JPH2	1.94e-04	1.27	Euteleostomi
102	intronic (cis)	ENSG00000124092	CTCFL	9.33e-03	-2.48	Euteleostomi
103	intronic (cis)	ENSG00000160207	HSF2BP	1.19e-02	-1.47	Euteleostomi
104	intronic (cis)	ENSG00000100154	TTC28	3.31e-02	-5.86	Bilateria
105	intronic (trans)	ENSG00000130939	UBE4B	7.17e-03	-1.28	Bilateria
106	intronic (trans)	ENSG00000121897	LIAS	1.53e-06	1.17	FungiMetazoa
107	intronic (trans)	ENSG00000002745	WNT16	3.74e-02	-1.61	Euteleostomi
108	intronic (trans)	ENSG00000197892	KIF13B	1.75e-27	1.23	Euteleostomi
109	intronic (trans)	ENSG00000154359	LONRF1	9.54e-25	1.21	Euteleostomi
110	intronic (trans)	ENSG00000168092	PAFAH1B2	9.25e-05	-1.13	Euteleostomi
111	intronic (trans)	ENSG00000171471	MAP1LC3B2	6.71e-03	1.03	HomoSapiens
112	intronic (trans)	ENSG00000198146	ZNF770	1.00e-06	1.07	Euteleostomi
113	intronic (trans)	ENSG00000178226	PRSS36	1.98e-02	-4.14	Tetrapoda
114	intronic (trans)	ENSG00000125107	CNOT1	2.81e-11	-1.52	FungiMetazoa
115	intronic (trans)	ENSG00000176401	EID2B	3.21e-02	-1.03	Eutheria
116	intronic (trans)	ENSG00000167619	TMEM145	3.08e-03	-1.89	Bilateria
117	intergenic	ENSG00000155903	RASA2	3.72e-23	-1.44	Euteleostomi
118	intergenic	ENSG00000256825		6.28e-03	2	
119	intergenic	ENSG00000095015	MAP3K1	1.01e-05	-1.33	Chordata
120	intergenic	ENSG00000129596	CDO1	2.30e-03	2.42	Bilateria
121	intergenic	ENSG00000113758	DBN1	1.49e-57	-2.51	Euteleostomi
122	intergenic	ENSG00000196735	HLA-DQA1	6.25e-14	-1.47	Simiiformes
123	intergenic	ENSG00000196735	HLA-DQA1	1.48e-10	-1.4	Simiiformes
124	intergenic	ENSG00000231389	HLA-DPA1	5.85e-04	-1.14	Tetrapoda
125	intergenic	ENSG00000020181	GPR124	2.89e-04	-1.44	Euteleostomi
126	intergenic	ENSG00000177335		9.14e-12	24.97	
127	intergenic	ENSG00000148200	NR6A1	2.26e-18	-1.46	Chordata
128	intergenic	ENSG00000160613	PCSK7	8.71e-03	-1.22	Bilateria
129	intergenic	ENSG00000123297	TSFM	4.71e-02	1.13	Bilateria
130	intergenic	ENSG00000122971	ACADS	1.98e-02	1.17	Bilateria
131	intergenic	ENSG00000139971	C14orf37	3.46e-02	-3.42	Euteleostomi
132	intergenic	ENSG00000007129	CEACAM21	2.37e-07	-1.59	Euteleostomi
133	intergenic	ENSG00000007129	CEACAM21	2.37e-07	-1.59	Euteleostomi
134	intergenic	ENSG00000125772	GPCPD1	7.17e-03	-1.11	FungiMetazoa

Annex

Annex

During this thesis, I have participated in the elaboration of the following manuscripts:

- **Rigau, M.**, Juan, D., Valencia, A., and Rico, D. (2019). Intronic CNVs and gene expression variation in human populations. PLOS Genet. *15*, e1007902.

- Abascal, F., Juan, D., Jungreis, I., Martinez, L., **Rigau, M.**, Rodriguez, J.M., Vazquez, J., and Tress, M.L. (2018). Loose ends: almost one in five human genes still have unresolved coding status. Nucleic Acids Res. *46*, 7070–7084.

- Jodkowska, K.#, Pancaldi, V.#, Almeida, R.*, **Rigau, M.***, Graña-Castro, O., Fernández-Justel, JM., Rodríguez-Acebes, S., Rubio-Camarillo, M., Carrillo-de Santa Pau, E., Pisano, D., Al-Shahrour, F., Valencia, A., Gómez, M., Méndez, J. (2019). Three-dimensional connectivity and chromatin environment mediate the activation efficiency of mammalian DNA replication origins. *In preparation*.