nature ecology & evolution

Review Article

https://doi.org/10.1038/s41559-023-02014-y

# Evolution and implications of de novo genes in humans

Received: 24 July 2022

Accepted: 6 February 2023

Published online: 16 March 2023

Check for updates

Luuk A. Broeils[1], Jorge Ruiz-Orera[2], Berend Snel[3], Norbert Hubner[2,4,5] & Sebastiaan van Heesch[1]

Genes and translated open reading frames (ORFs) that emerged de novo from previously non-coding sequences provide species with opportunities for adaptation. When aberrantly activated, some human-specific de novo genes and ORFs have disease-promoting properties—for instance, driving tumour growth. Thousands of putative de novo coding sequences have been described in humans, but we still do not know what fraction of those ORFs has readily acquired a function. Here, we discuss the challenges and controversies surrounding the detection, mechanisms of origin, annotation, validation and characterization of de novo genes and ORFs. Through manual curation of literature and databases, we provide a thorough table with most de novo genes reported for humans to date. We re-evaluate each locus by tracing the enabling mutations and list proposed disease associations, protein characteristics and supporting evidence for translation and protein detection. This work will support future explorations of de novo genes and ORFs in humans.

Gene and protein discovery have historically relied on a combination of genomic, transcriptomic and proteomic approaches, most prominently driven by computational assessment of sequence conservation[1–3]. The ribosome profiling technique, which visualizes messenger RNA (mRNA) translation at nucleotide resolution, recently revealed that translation is widespread at many annotated long non-coding RNA (lncRNA) transcripts and presumed untranslated regions of mRNAs[4–6]. Although this would suggest that the human genome harbours many more protein-coding regions than previously imagined, most newly discovered translated open reading frames (ORFs) are evolutionarily young, restricted to humans or primates[7] and hence do not meet the traditional requirements for being annotated as protein-coding[8]. Consequently, the roles of these ORFs, or the proteins they may produce, have not been systematically investigated.

To be classified as a canonical protein-coding region, a new protein-coding gene should ideally conform to certain criteria (for example, a reliable transcript model, an intact ORF, signatures of evolutionary conservation at the coding region and protein-level evidence). Despite not ticking all these boxes, 7,264 mostly evolutionarily young human ORFs have now been nominated by GENCODE/Ensembl and will be systematically evaluated by other gene and protein reference annotation projects[7]. Centralized and standardized annotation of translated ORFs enhances their visibility and accessibility within the scientific community, which we anticipate being a crucial first step towards further in-depth investigations into their putative and possibly human- or primate-specific roles. Decades after their first discovery[9], upstream ORFs will find their way into reference annotations, and many lncRNAs will be accompanied by translated short ORFs more recently detected across human cell types and tissues.

Many of these ORFs, and the genes that encode them, have originated de novo, which means they newly emerged from DNA that was previously non-genic or non-coding[7,10]. While the possibility of genes originating from non-genic DNA was initially disregarded[11], genes and ORFs do frequently emerge de novo in many species, including humans[12–17]. Here, we focus on the potential implications of de novo gene and ORF birth for human health and disease. We summarize the

[1]Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands. [2]Cardiovascular and Metabolic Sciences, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany. [3]Theoretical Biology and Bioinformatics, Department of Biology, Faculty of Science, Utrecht University, Utrecht, The Netherlands. [4]Charité-Universitätsmedizin, Berlin, Germany. [5]DZHK (German Centre for Cardiovascular Research), Partner Site Berlin, Berlin, Germany. ✉e-mail: s.vanheesch@prinsesmaximacentrum.nl

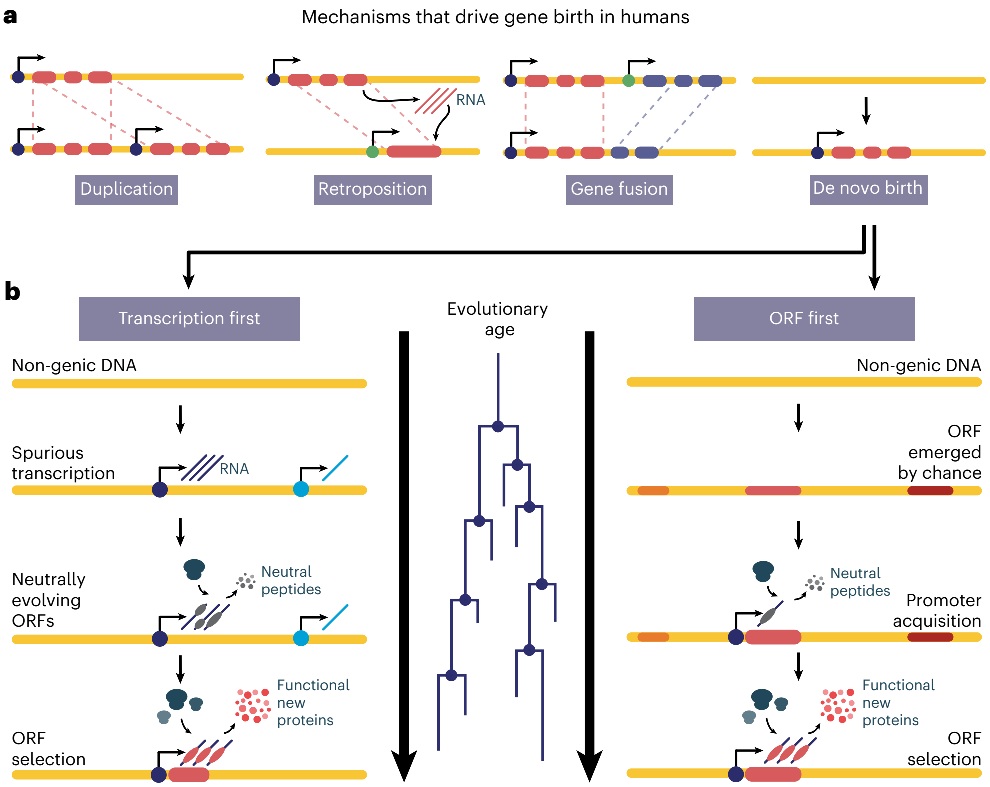


**Fig. 1 | Mechanisms that drive gene birth in humans. a**, New genes can arise through various mechanisms: duplication of genetic material, retroposition of genes, gene fusion or through de novo gene birth. Black arrows indicate transcription start site, coloured blocks indicate gene exons. **b**, Models of de novo gene birth. The 'transcription first' route suggests that spurious transcription precedes the formation of ORFs. Spurious transcription of genes that contain translated ORFs, coined proto-genes, act as a pool of neutrally evolving genes/peptides, out of which new genes and proteins can evolve. The 'ORF first' route proposes that first an ORF is formed in non-genic DNA, but becomes transcribed only after acquiring the necessary regulatory sequences. Through positive selection, the transcribed ORF can transform from a string of amino acids into a biologically relevant protein.

experimental and computational approaches through which de novo genes and protein-coding signatures can be identified. We provide an overview of human genes reported to have evolved de novo in the primate lineage and describe several examples that have been characterized in more detail. Through literature review and reanalysis of published data, we compiled a table with putative human de novo protein-coding genes as previously reported in 13 different studies[18–29]. We have re-evaluated their sequence evolution and expression characteristics and have listed ribosome profiling (also known as ribosome sequencing (Ribo-seq)) and mass spectrometry evidence of protein translation. Last, we reflect on recent annotation efforts that have increased or will increase the visibility and awareness of these genes and ORFs within the broader scientific public. We anticipate such efforts will lead to a better understanding of the fraction of de novo genes that might have roles unique to our species, and will help to visualize the potential risks and benefits of de novo gene birth for human health and disease.

## Genes emerging from scratch

Genes can evolve through processes that do not require the duplication or reorganization of ancestral genes; a process termed de novo gene evolution—the new acquisition of a (protein-coding) gene from previously non-genic DNA[10,30–33] (Fig. 1a). It was initially thought to be extremely unlikely for genes to originate through mechanisms other than the reorganization of existing genetic material, which includes processes such as gene duplication, fusion, fission, exon shuffling or retrotransposition[11]. However, several studies have shown that homology was lacking for certain genes found in eukaryotes, indicating that these genes were restricted to specific lineages or species[1,34,35]. These genes of unknown origin were first coined as orphan genes in yeast[35]. Yet, as more genomic sequencing data became available, orthologues of some of these orphan genes were detected in (closely) related species[34,36]. Therefore, some of the orphan genes were renamed into taxonomically restricted genes if they were detectable only in a particular species or clade. There are two possible explanations for an orphan gene to be considered as such. First, the gene may have duplicated and diverged rapidly, resulting in a failure to detect orthologues through phylostratigraphy[37]. Second, the gene may have evolved de novo, meaning there is no ancestral gene. All genes that evolved newly from previously non-genic sequences by having followed this pattern of molecular evolution would be considered de novo. This makes de novo gene categorization more inclusive and less ambiguous[30,31].

Research into human de novo genes has gained momentum in recent years[32,38]. Using genomic synteny combined with strict sequence similarity searches, two landmark studies found 15 primate-specific genes and 3 human-specific genes transcribed from genomic regions for which no matching protein-coding gene sequences could be found in their evolutionary predecessors, confirming their non-genic and non-coding origin through sequence homology[28,39]. The shift from non-coding to protein-coding was shown to be instigated by random mutations that occurred after the human–chimpanzee split 6.5 to 7.5 million years ago[40], and proteomics data confirmed protein expression[28]. While these putative de novo genes and their protein products were expressed in humans, other primate species showed multiple disabling mutations or small insertions or deletions (indels), preventing translation from these genes[28]. Similar enabling mutations have been found to contribute to the emergence of putative de novo genes in other species[14,17,41,42].

## Steps required for the evolution of de novo protein-coding genes

Multiple consecutive steps are required to turn a non-genic sequence into an expressed protein-coding gene (Fig. 1b). A gene needs to attain the ability to get transcribed and it needs to acquire an ORF that can be translated[31]. These events can take place in either order, as there is evidence for both a non-coding RNA intermediate step ('transcription first'), as well as for ORFs that subsequently attain the ability to be transcribed and translated ('ORF first')[12,14,16,17,31,43].

The capacity of a gene to be expressed is directly dependent on the capability of the transcription machinery to bind near the gene and transcribe it. While only 1–2% of the human genome encodes for protein-coding sequences, transcription across the entire genome is pervasive[44]. Resulting transcripts are often non-coding and lack sequence homology or typical protein-coding RNA properties, frequently resulting in their classification as lncRNA genes[45]. However, in the process of de novo protein-coding gene evolution, expressed lncRNAs could serve as RNA intermediates to new protein-coding genes[12,29,46] (Fig. 1b). In support of this, a large proportion of lncRNAs can localize to the cytosol and physically associate with ribosomes, in a manner very similar to known protein-coding mRNAs[46–51].

The majority of ORFs found in human lncRNAs did not match any homologous sequences in the human genome (paralogues) or related species (orthologues), as revealed by BLAST sequence similarity searches[46]. Additionally, comparisons between macaque, chimpanzee and human tissue indicated that the transcripts of 24 human de novo genes were expressed in other primates and annotated as lncRNAs, but were not translated in these species[29]. These 24 genes had readily acquired transcriptional regulatory sequences in the other primate species, supporting an 'ORF first' route for these genes[29]. Building further on these findings, a combination of homology- and synteny-based genomic analyses on human and chimpanzee transcripts revealed that over 5,000 transcriptional events were unique to each of these species[19]. Expression of species-specific transcripts could be associated with the recent acquisition of upstream transcription start sites and downstream splice sites. Although the sequences of most of these transcripts did not show any signatures of purifying selection, the study demonstrated the importance of gaining transcriptional regulatory sequences as a step in de novo gene evolution[19].

Once a non-coding transcript is being transcribed in a specific species, the accumulation of mutations can result in the formation of a translatable ORF in these transcripts. While ORFs continuously emerge in non-coding transcripts over time, only some acquire the ability to be translated. The exact steps that allow a new ORF to be translated are unclear. Some properties, such as the translation initiation context, codon usage and the relative position of the initiation codon, have been associated with a higher ORF translatability in mammals[13,52–54]. Furthermore, changes in the composition of scanning complexes with different abilities to unwind RNA secondary structures can selectively activate ORFs with suboptimal initiation contexts under specific conditions[52]. In rice, frameshift indel mutations were the most common type of enabling mutation, being observed in over 150 out of the 175 investigated de novo emerged coding sequences (CDSs)[17]. Even though nucleotide substitutions occur over ten times more frequently in rice genomes, indels were found as the enabling mutations almost ten times more often. This is surprising, assuming the null hypothesis of neutral evolution, in which the rates of mutation and evolution would be equal. This suggests that indel-enabling mutations are crucial in the emergence of new de novo genes[17]. A similar model of emergence enabled by indel mutations has been detected for the human putative de novo genes *CLLU1*, *C22orf45* and *DNAH10OS*[28]. However, the mouse- and oviduct-specific de novo gene *Gm13030* emerged due to enabling changes that removed disrupting stop codons (TGA > TGC) and were thereby selected as the most probable node of emergence, indicating that specific substitutions can also help a de novo gene to acquire a CDS[55].

De novo genes may also emerge via mechanisms where the intact ORF is present prior to transcript emergence (Fig. 1b). Eukaryotic genomes contain a large number of intergenic and intronic ORFs[12,16]. Although such ORF structures are frequent, these ORFs are not under selection and are frequently gained and lost[15,23,56]. In the *Drosophila melanogaster* genome, there are several reported de novo ORFs that emerged before the host RNA transcript acquired the means to become transcribed. In closely related *Drosophila* species, orthologous ORFs are genomically identical but the genomic locus is not transcribed[16]. In rice, genomic comparisons of transcriptomes assembled by RNA-seq of ten different species revealed that about 10% of all de novo genes had originated in an 'ORF first' fashion[17]. However, some of these identified genes might exhibit species-specific patterns of tissue-specific, condition-specific or spatiotemporal expression, and their transcription could have hence remained undetected in matched samples of other species. Inspecting the evolution of candidate loci can give further insights into the emergence of a transcript within a specific species or lineage. A primary example is the antifreeze glycoproteins (afgps) that originated de novo in certain types of codfish[14,41]. A translocation event resulted in the regulation of the *afgp* gene by a promoter region that probably originated de novo as well, but only after the formation of the *afgp* ORF[41]. An *afgp*-like sequence can be found in closely related codfish, but it is not expressed there[14]. These examples illustrate how 'ORF first' modes of gene evolution raise the possibility that future coding regions with out-of-the-box roles might be readily present in our genomes and could have a cellular role once transcribed and translated.

## The preservation, selection and features of de novo genes

If the criteria for de novo gene evolution are met, namely the ability to get transcribed and the presence of a translated ORF, the next question is how translated ORFs are established as new protein-coding genes, and to what extent this process occurs randomly. The most prominent model is that the spurious expression of ORFs results in a pool of proto-genes that form the basis from which beneficial ORFs are selected through evolutionary pressure[12,13,15,46,57,58]. Proto-genes are expressed RNA transcripts from non-genic sequences that contain a translated ORF. The model of proto-genes as an intermediate of de novo gene evolution was first investigated in *Saccharomyces cerevisiae*, where over 1,000 proto-genes were found to be unique to *S. cerevisiae* and absent from closely related yeast strains[12]. While most of these proto-genes did not show signs of purifying selection or translation, a small fraction appeared to have evolved from the proto-gene reservoir into genes, transitioning from a neutral proto-gene state into a translated gene under selection[12]. A similar mechanism was observed in humans, mice and *Drosophila*[13,24,42,56,59].

The formation of proto-genes seems common in humans and primates, yet so is their eventual loss[15,56]. The stepwise transition from a non-genic region to a transcribed locus with a translated coding region is known as the continuum model[12]. Another model of de novo gene evolution is known as the preadaptation model[60]. This model includes genes that undergo an 'all-or-nothing' transition. Pre-existing gene characteristics have to be present for a de novo gene to emerge, as any non-functional intermediate would only act as a toxic byproduct for the cell[60]. The initial publication describing the preadaptation model demonstrated that intrinsic structural disorder (ISD) of the CDSs of young genes was higher compared with those of older genes over a longer timescale, in concordance with preadaptation[60]. However, others have suggested that the ISD observed in favour of the preadaptation model was due to the regular occurrence of de novo genes overlapping older, conserved protein-coding sequences, and that de novo genes emerging in intergenic areas do not have a substantially higher ISD[15,61]. Interestingly, several studies demonstrated preadaptation potential for individual genes whose mechanisms of origin were studied in detail[14,62,63], suggesting that the continuum and preadaptation models are both likely to contribute to the emergence of de novo genes. The predicted properties of the proteins translated from these genes, including ISD and aggregation propensity, do not seem to correlate with evolutionary age, especially on a more recent evolutionary timescale[15]. Additionally, random sequences modelled in silico to match the length and amino acid composition of proto-genes possess predicted structural properties that are very similar to those of proto-genes[64]. However, proto-gene amino acid sequences have been found to be more soluble in vitro than synthetic random proteins with matched amino acid frequencies and length distributions[64], indicating that certain intrinsic properties of proto-genes may allow them to be better tolerated, despite their apparent similarities to random peptides.

Genes and ORFs that have originated de novo generally display a relatively short length and span a lower number of exons, with the length of the gene as well as its ORF increasing with age (that is, older genes and ORFs are longer)[15,65,66]. Furthermore, expression levels of young genes are often relatively low and restricted to specific tissues, compared with older genes[12,24,29,65,66]. Approximately 67.9% of all proto-genes in humans and primates contain at least one intron[42]. GC content has been found to be positively correlated with ISD, and de novo genes are more often found in GC-rich regions[56]. Contrary to what was found in other species, ISD is not positively associated with de novo gene birth in *S. cerevisiae*, but the frequency of thymine-rich transmembrane domains is[57,63]. In the human genome, the MYEOV protein was found to have evolved de novo and contains a putative transmembrane domain as well[24,67]. However, in a more systematic approach that investigated microproteins translated from evolutionarily young human de novo genes, transmembrane domains did not seem to be prevalent[43]. Although having a more ordered protein sequence could make it easier to integrate into existing molecular pathways[57], it could have a higher propensity for misfolding and being toxic to the cell at the same time. Genes that require relatively limited cellular resources to be transcribed and translated could possibly be compensated for, as the gene products are present in such low quantities that any RNA- or peptide-induced toxicity is probably not detrimental for the cell[68,69].

The amino acid composition of a protein's carboxy terminus (C terminus) seems to be the important determinant for degradation[70]. Hydrophobicity of the C-terminal domain was strongly correlated with degradation and established protein-coding genes have evolved to remove hydrophobic C-terminal tails[70]. When expressing random peptide sequences in *Escherichia coli*, growth advantages were correlated with higher ISD, indicating that intrinsically disordered peptides can be tolerated or even result in a growth advantage[71]. Another study found similar results, demonstrating that higher ISD of de novo proteins resulted in easier expression in *E. coli*[72]. The acquisition of (rudimentary) protein domains by elongation of a de novo protein could be another mechanism by which a de novo gene can integrate into existing molecular machinery[73,74]. This 'constructive neutral evolution' could explain how neutrally evolving peptides exist despite purifying selection[75]. These peptides are allowed to undergo rapid changes, which can subsequently be selected for if the new product is beneficial to the organism[13,15,46].

How can we determine whether a recently emerged de novo gene has acquired a level of biological activity that could qualify the gene for being considered 'functional' in humans[76]? ORFs detected by ribosome profiling data outside canonical (annotated) protein-coding sequences (Supplementary Information), as reported by Chen et al.[77], were recently re-evaluated by Vakirlis et al. to pinpoint the fraction of these ORFs that emerged de novo[43]. This showed that 155 translated ORFs included by Chen et al. in a CRISPR–Cas9-based genetic perturbation screen[77] had originated de novo from the mammalian lineage onwards. This revealed that 44 out of 155 de novo ORFs could be connected to cellular growth[43,78]. Of these, 14 had originated after the split of the higher primates and 6 after the human and chimp split. Rescue experiments with translationally disabled versions of these genes showed that it was not the non-coding RNA, but the protein-coding sequence, that gave this growth advantage[43,77]. While these data would indicate that the fraction of ORFs that can fulfil a role within a cellular context 'out of the box'—that is, quickly after emergence—is larger than previously anticipated, it is important to keep in mind that 155 de novo ORFs is still a limited dataset to accurately draw these conclusions. Although high-throughput CRISPR screens can give valuable new insights into the importance of a gene for cell viability, higher-resolution in-depth single-gene studies are necessary to pinpoint the precise molecular roles of these young proteins, as well as the networks, protein complexes and processes they act in, as has been done for other species[41,79,80].

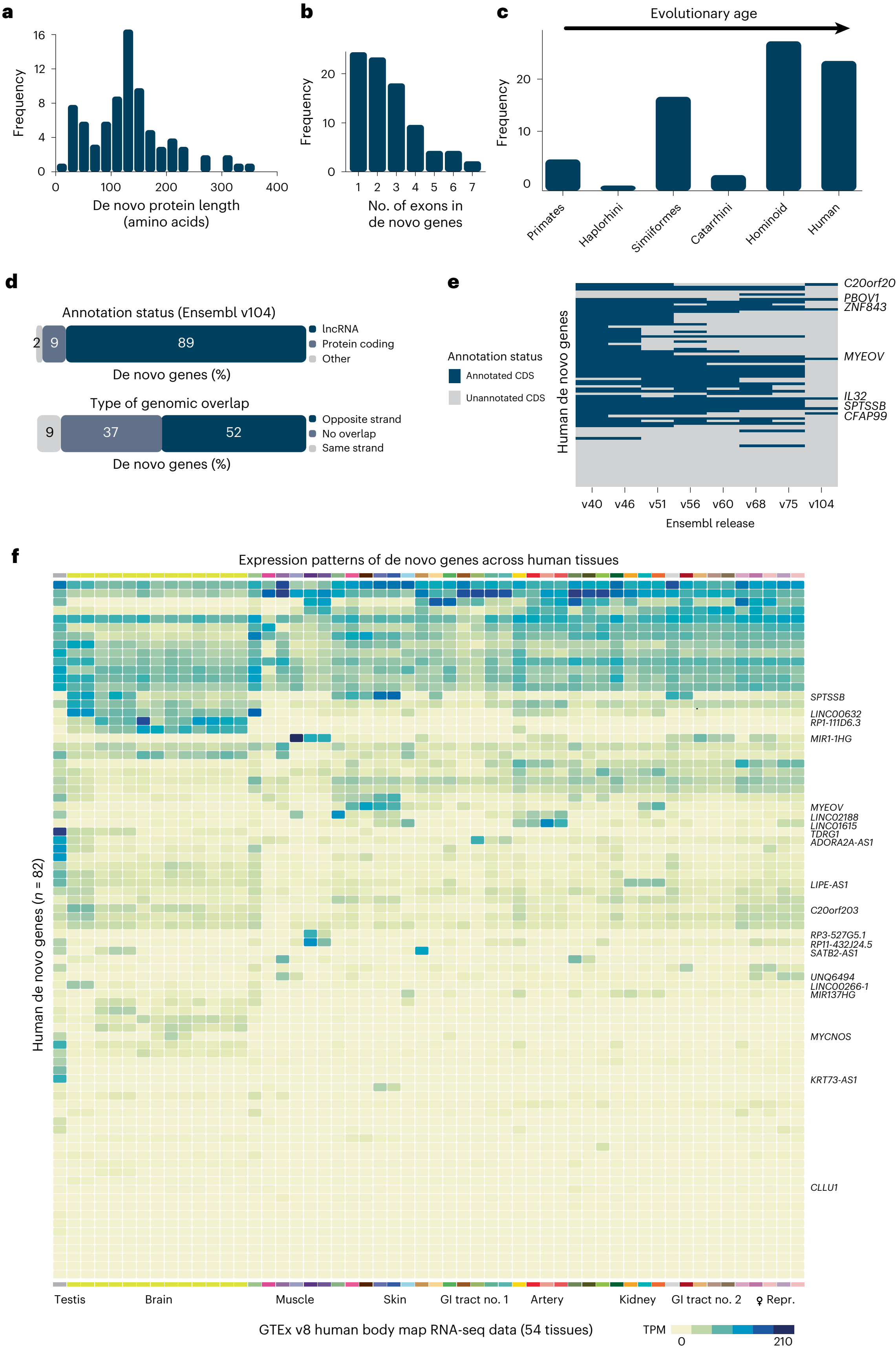
## A manually curated list of de novo protein-coding genes in humans

Although de novo protein-coding gene birth is more prevalent throughout evolution than previously thought, only a few human de novo genes have been investigated in more detail, to date[20,25,26,67,81]. For this Review, we have curated a list of most (to our knowledge) human protein-coding genes that are currently known or reported to have originated de novo in the primate lineage, that is, protein-coding genes discovered and subsequently annotated as putatively de novo in humans by a collection of individual studies[18–20,22–29,43,82] (Supplementary Table 1). We performed our search according to various criteria and using several methods (Supplementary Information) and ended up with 82 de novo genes. For these genes, we validated their mode of evolution, and evaluated evidence on whether, and how, protein expression from these genes has been validated, in which tissues or cell types these genes are expressed, and in which closely related species they may also be present (Supplementary Table 1).

We also evaluated shared characteristics and reported layers of evidence in support of the 82 de novo genes. The median number of exons was 2 and the median protein length was 129 amino acids (Fig. 2a,b). This median amino acid length is three times greater than that of the average sequence length of the 7,264 Ribo-seq ORFs as now catalogued by GENCODE (44 amino acids)[7]. In agreement with the recent study by Vakirlis et al.[43], this probably indicates that more and mostly smaller de novo proteins are yet to be nominated. Of 82 putative de novo genes, 73 (89%) are currently annotated as lncRNA (GENCODE/Ensembl v104), with 7 being annotated as a protein-coding gene (Fig. 2d,e). The remaining 2 out of 82 de novo genes are currently annotated as transcribed unitary pseudogenes.

**Fig. 2 | Characteristics of a newly compiled list of human de novo genes.** **a**, Histogram of the amino acid lengths of all compiled de novo proteins. **b**, Histogram displaying the number of exons found in each de novo gene. **c**, Histogram depicting the species in which the de novo gene originated. **d**, Biotype annotation (GENCODE/Ensembl v104) of compiled human de novo genes and type of genomic overlap of compiled de novo genes with other genes. **e**, Boxes depict the presence (dark blue) or absence (grey) of the reported CDS in the annotations of eight Ensembl releases. The selected Ensembl releases were used by the main studies included in our compiled list of de novo genes and we additionally included the Ensembl version used in our review (v104). Genes are sorted by year of study, considering the oldest study in which the gene was reported. The gene names of seven de novo genes annotated as protein-coding in Ensembl v104 are highlighted. **f**, Expression profiles of human de novo genes across various tissue types. A darker colour denotes a higher expression. A subset of genes has been highlighted on the right for being mentioned as examples in the main text, or for displaying strong tissue-restricted expression patterns. TPM, transcripts per million; Repr., organs of the female reproductive system (vagina, uterus, ovary, fallopian tube and cervix). Figure prepared using the GTEx v8 portal. See also Supplementary Table 1.

A comparison of the compiled list of 82 putative human de novo genes whose transcripts are annotated in Ensembl v104 (Supplementary Table 1), with Ribo-seq datasets of various human cell lines and tissues[19,21,77,83–88], revealed that 35 out of the 73 putative de novo genes that are currently still annotated as lncRNA can in fact be translated by ribosomes. Because the human genome contains at least 2,000



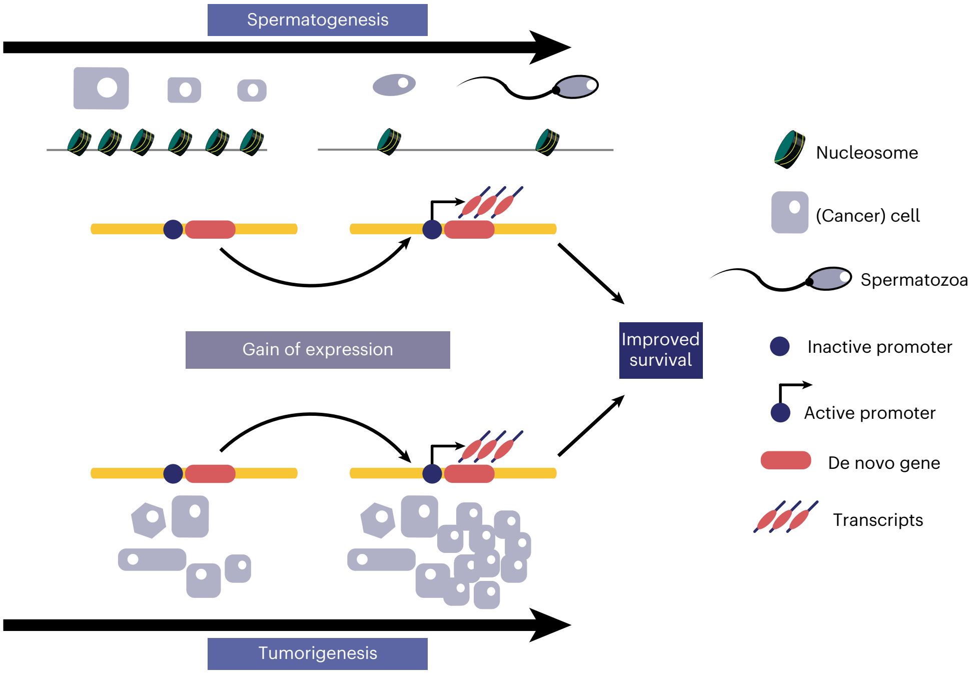


**Fig. 3 | De novo genes expressed in the testis could be positively selected for, despite their tumour-promoting role.** During spermatogenesis (top panel), the chromatin opens, which could allow de novo genes without endogenous promoters to be expressed (out-of-testis hypothesis). When this de novo gene has an implication in preventing apoptosis and improving cell survival in spermatocytes, it can be positively selected for. However, when this gene is aberrantly expressed during tumorigenesis (bottom panel), it can have a cancer-promoting role. Despite this negative effect on overall fitness by being disease-promoting, it can still be under positive selection due to its positive effect on spermatocyte survival.

translated ORFs encoded within genes currently annotated as lncRNAs[7], some of these might encode yet-to-be-characterized de novo proteins. For several putative de novo genes, studies have reported evidence of protein synthesis through western blots with custom antibodies[20,25–27], or from public mass spectrometry databases[89–91]. We would like to stress that we did not re-evaluate the spectra identified to corroborate the unique peptide evidence. We are aware of frequent false-positive spectrum assignments to new ORFs and know that not all reported protein-level evidence meets the Human Proteome Organization's criteria[92].

Looking at the genomic location, over 50% of the putative de novo genes were found to have overlap with another gene, either on the opposite strand (53%) or on the same strand (10%; Fig. 2d). This frequent overlap could be the result of promoter hijacking, in which the de novo gene is transcribed through the usage of a nearby promoter of a conserved gene, to facilitate the acquisition of transcriptional capabilities. Over time, such sense–antisense transcript overlaps might be undesired and selected against, as observed for metazoan genomes[93]. Most of the genes we compiled (53 out of 82) had evolved de novo in the past 25 million years and are specific to the hominoid lineage (Fig. 2c). More than a quarter were found to be human-specific (25 out of 82). Similar to de novo gene evolution in various rice species[17], we found the majority of de novo genes (46 out of 82) to be 'enabled' by small deletions or insertions, often extending the ORF sufficiently to be embraced as a de novo gene.

## De novo genes in human physiology and cancer

As compared with other human tissues, a large number of putative human de novo genes were found to be primarily expressed in the brain (24 out of 82). This is consistent with previous observations[18,94], and one de novo gene has been associated with Alzheimer's disease, suggesting a possible brain-related role[26]. There are several factors that could explain why the brain expresses such a high number of de novo genes. First of all, the brain is an immune privileged tissue, which prevents putative new proteins from being recognized by the immune system as foreign antigens[95]. Additionally, the brain is a complex organ with different regions defined by specific chromatin accessibility profiles that are highly variable during neurogenesis[96–98]. Young genes may acquire critical roles in human brain development, as was previously shown for new genes that arose through duplication events (Supplementary Information). Putative de novo genes expressed in the brain are often absent from the genomes of other primate species (Supplementary Table 1). Two recent studies, published after this Review was submitted[99,100], have now identified two putative de novo genes (*LINC00632* (ENSG00000203930) and *LINC00634*/*SMIM45* (ENSG00000205704)) that might have roles in cortical expansion during human brain development. A recent study that profiled the translatomes of developing, paediatric and adult human brains identified several novel human-specific short ORFs unique to developing brain tissues[101]. Like duplicated genes, de novo genes with brain-restricted expression patterns might facilitate evolutionarily recent morphological changes to our cerebral neocortex and, consequently, help to explain our enhanced cognitive abilities.

In addition to the brain, a relatively large group of human de novo protein-coding genes are primarily expressed in the testes (19 out of 82; Supplementary Table 1). The observation that many de novo genes were selectively expressed in testicular tissue was initially investigated in *D. melanogaster*[102], but also held true for evolutionarily young genes in humans[19,103,104]. This led to the 'out-of-testis' hypothesis, a model in which the testes can act as a catalyst for the evolution of new genes[10] (Fig. 3). Spermatocytes and spermatids might facilitate new gene formation due to their open chromatin state, which allows pervasive transcription of DNA that might otherwise not be transcribed[10,105]. In further *D. melanogaster* studies, it was seen that young genes with a putative de novo origin had different expression patterns in the testes than genes that had formed through recent duplications[106,107]. Additionally, two genes with a putative de novo origin in *D. melanogaster* are crucial for sperm production and function, and their knockdown causes infertility[79,80,108]. Their molecular and evolutionary origin has

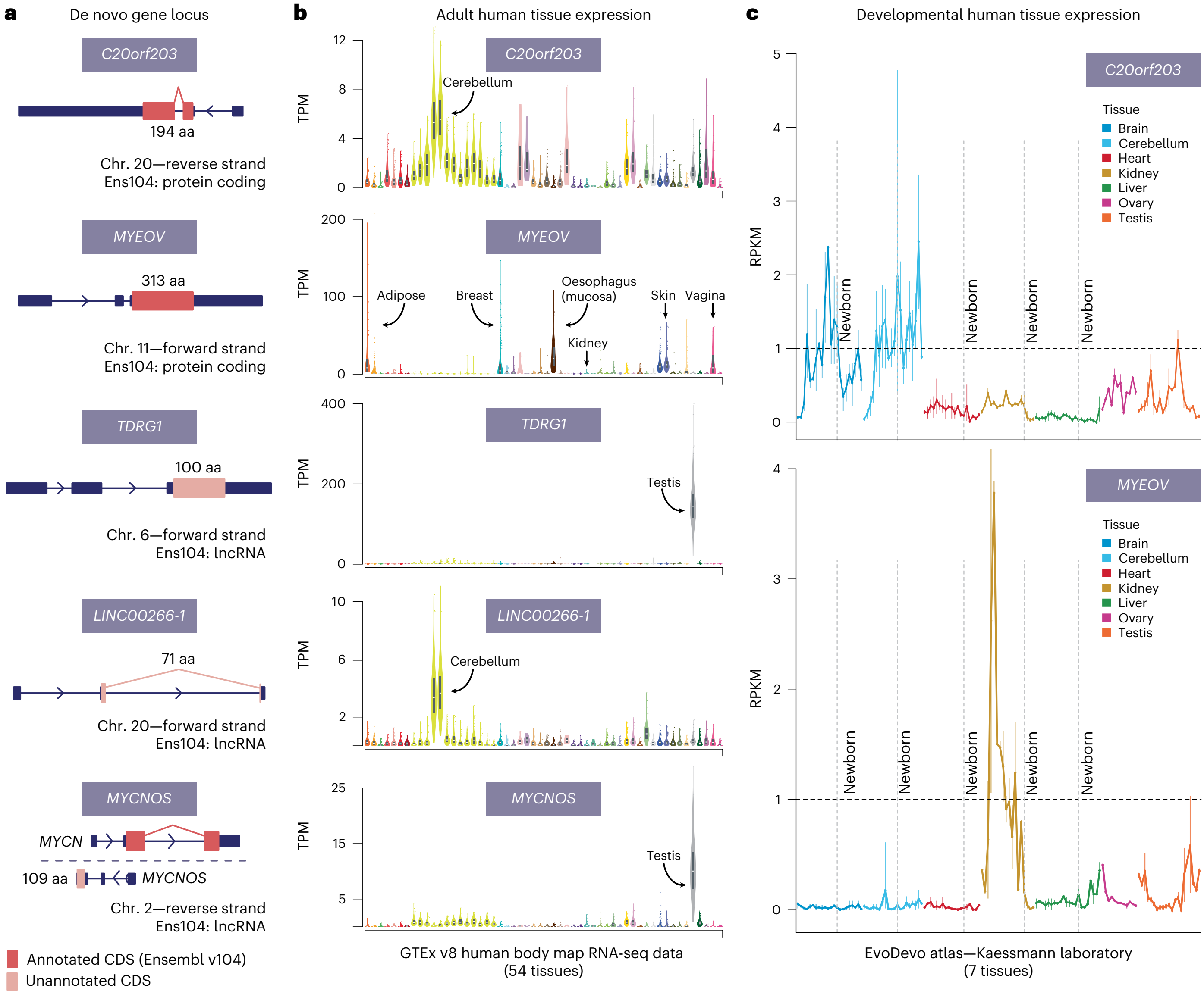


**Fig. 4 | Evolutionary trajectories and expression characteristics of five human de novo genes associated with disease. a**, Schematic depicting the gene locus with CDS highlighted. For two of the genes (*C20orf203* and *MYEOV*), the CDS is annotated in Ensembl v104 (Ens104), while the CDS of the other three genes (*TDRG1*, *LINC00266-1* and *MYCNOS*) is currently not annotated. aa, amino acids; Chr., chromosome. **b**, Expression pattern in adult human tissue. Figure prepared using the GTEx v8 portal. **c**, Expression patterns in various tissues during development. Vertical dashed lines indicate timepoints that denote newborn tissues. RPKM, reads per kilobase of transcript per million reads mapped. Not all genes listed in **a** and **b** could be retrieved from the EvoDevo atlas[135]. Figure prepared using the EvoDevo atlas by the Kaessmann laboratory. See also Supplementary Information.

been extensively studied, although definitive proof of their de novo status could not be given due to difficulties with syntenic alignments in non-*Drosophila* species, making it impossible to identify the non-coding regions in these species[108]. These putative *D. melanogaster* de novo genes could be excellent examples of how de novo genes evolve through testis-related expression. De novo genes are also expressed at higher levels than expected in post-meiotic testicular tissue in *D. melanogaster*[106,109], suggesting that their ability to escape the expected RNA degeneration in these cells could benefit their potential for rapid natural selection. Similar results were found in mice testes, where genes expressed in post-meiotic tissue were more often lineage-specific and those with *cis*-regulatory elements showed higher expression changes between species, both important characteristics of de novo genes[110]. This could lead to the establishment of proto-genes, of which some could then evolve further into new biologically relevant de novo protein-coding genes. Concurrently, new genes that have a positive influence on traits like sperm competitiveness in post-meiotic cells can quickly undergo positive selection and get fixed in the population[10,109].

Although the physiological and cellular roles of most human de novo proteins are unknown, many de novo genes have been associated with various types of cancer (Supplementary Table 1). Genes related to tumour suppression and apoptosis were found to be under positive selection in humans and other primates, just as genes related to spermatogenesis[103]. It was theorized that genes associated with cancer can undergo positive selection due to their capability to repress apoptosis in spermatocytes[32] (Fig. 3). De novo protein-coding genes could confer an advantage in the survival of germ cells and thereby be under positive selection. However, when these same genes become active in other cell types, it could lead to the promotion of tumorigenesis (Fig. 3). For example, the NCYM protein (encoded by the *MYCNOS* gene)—located in the direct vicinity of the *n-myc* oncogene—is normally specifically expressed in the testes, but is associated with various types of cancer[78,111–113]. MYC (*c-myc*) and MYCN (*n-myc*) have also been found to promote spermatogonial stem cell self-renewal in mice[114]. It could be that NCYM facilitates a survival advantage for spermatocytes through the stabilization of MYCN[27], allowing it to become selected for and thereby kept in the population despite its role in tumorigenesis.

For *MYCNOS* and four other putative de novo genes with reported disease-associated roles (*C20orf203*, *MYEOV*, *TDRG1* and *LINC00226-1*; Fig. 4), we provide detailed descriptions of their origin, putative roles and association with disease, including multiple alignment views that detail the enabling mutations underlying their de novo origin, in the Supplementary Information.

## Considerations for future studies of de novo genes in humans

The complete and correct annotation of de novo protein-coding genes remains challenging and has been a topic of discussion in the field (Supplementary Information). Additionally, major gene databases routinely annotated these genes as non-coding for not having conserved coding regions, unless manual curation and literature evidence suggested otherwise. Protein annotation is traditionally a conservative and manual process that generally relies on strong signals of sequence conservation to demonstrate selection for the CDS. The widespread translation of ORFs has challenged this conservative way of annotating CDSs[77,83,84], as current annotation strategies penalize evolutionarily young ORFs, including the ones translated in de novo genes. Displaying evidence of translation by ribosomes, but mostly no detectable signs of evolutionary constraints, 7,264 mostly young human ORFs have now been compiled and catalogued by GENCODE/Ensembl[7]. With support of other main reference gene and protein annotation projects, these ORFs will be evaluated for their protein-coding potential[7]. Reference annotation makes these ORFs visible and accessible to the scientific community in a centralized manner for the first time, to our knowledge, so that their regulatory or coding aspects can be better investigated. In addition to the new annotation of short ORFs in presumed non-coding RNAs, Ribo-seq can help to identify recent changes in translation initiation sites, estimate the frequency of near-cognate start codon usage, and pinpoint human- or primate-specific amino- and C-terminal extensions to CDSs, as well as recent out-of-frame translations that overlap, and possibly overprint, existing CDSs. This illustrates that standardization of the use of Ribo-seq data for gene and protein annotation can provide a powerful data-driven complementary means to existing computational strategies for the identification of ORFs. Ribo-seq signals are additionally capable of pinpointing young variations of, or alterations to, existing (that is, conserved) protein sequences. We would like to stress that capturing most translated ORFs for a given species requires good-quality data generated of a broad range of tissues, cell types, conditions and developmental stages.

For many young genes or transitory proto-genes with evidence of translation into potential proteins, it remains challenging to define the most reasonable gene biotype: protein-coding or non-coding? Several well-studied examples of de novo protein-coding genes (described in more detail in the Supplementary Information) could all have a role in the promotion of cancer in the form of a protein or as RNA molecules. The fact that RNA can have a role in cancer has become evident and many lncRNAs have been found to be dysregulated in cancer[115], though this may hold true for any gene investigated in this context[116]. This suggests that the RNA transcripts of some de novo genes can fulfil a dual role[117–119]. It is possible that genes that have recently evolved from non-coding sequences of DNA by acquiring an ORF can retain their lncRNA functionality. Therefore, it is crucial to investigate and dissect both the coding and non-coding capabilities of a gene.

As of yet, only 9% (7 out of 82) of the putative human de novo genes we compiled were annotated as protein-coding in the Ensembl database, despite additional evidence suggesting the presence of a translated CDS (for example, via previously reported mass spectrometry data, western blots, or Ribo-seq) for 80 out of 82 de novo genes (Supplementary Table 1 and Fig. 2d). Recognition and awareness of the possible translational capacity of these evolutionarily young genes would stimulate more studies to focus on the implications of both coding and non-coding gene functions that might turn out to be specific to humans or primates, having roles in developmental conditions or diseases that rarely or never occur outside our species. There are multiple databases that have aimed to enhance the annotation of these genes and ORFs[120–122], and an ongoing effort has recently joined these and other scientists worldwide to establish standardized and centralized ORF annotation[7]. This effort aims to ensure ORFs in lncRNAs and additional ORFs in protein-coding genes no longer go unnoticed[7]. It advocates for more comprehensive annotation of 'non-canonical' ORFs when evidence of translation by Ribo-seq and/or mass spectrometry exists. We anticipate and hope this will result in a broader and better understanding of the coding and non-coding capabilities of evolutionarily young genes.

While studying newly discovered putative de novo genes in more detail, scientists should ideally adhere to a consensus definition of 'functionality'—an often difficult term or claim to assign to genes that emerged recently and lack clear signs of evolutionary constraint[32,76,123]. Because sequence conservation can be a consequence of functionality, (parts of) a gene and (domains in) a protein are often labelled as important when they are conserved across multiple species[123]. However, when investigating genes that have originated de novo, definitions of functionality are notoriously hard to prove through interspecies conservation, because conservational proof of

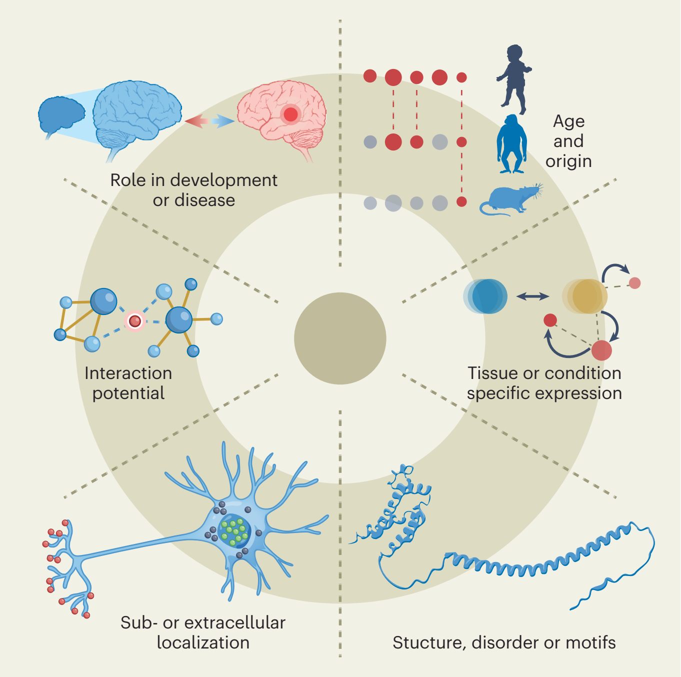

Role in development or disease
Age and origin
Interaction potential
Tissue or condition specific expression
Sub- or extracellular localization
Stucture, disorder or motifs

**Fig. 5 | Open questions for de novo genes in humans.** How and when did de novo genes evolve? When, where and under what circumstances are they expressed? Do the proteins translated from de novo genes possess structure or motifs, and with that inherit biological activity, readily at birth? Are proteins translated from de novo genes stably expressed, and to what subcellular compartments do they localize? Do young proteins have the capacity to interact with their much older and more conserved environment, such as important pathways and cellular processes? Do these genes have roles in recent evolutionary adaptations specific to humans, such as dedicated developmental processes (for example, neocortex formation), and can they contribute to disease? The protein structure is an AlphaFold[136] prediction of the protein encoded by the de novo gene *IL32*.

functionality invalidates the de novo definition. It is possible to look at genetic variation across human populations to determine selection and thereby functionality[124]. However, because of the relatively short length of de novo genes, it could be hard to achieve significant results. Evidence of translation of the gene is not enough as proof for functionality either, as neutrally evolving peptides have been found to be translated as well[13]. Therefore, experimental assays might be the most conclusive way to determine de novo gene functionality[32], and validating each gene separately will yield the best insights. This can be approached from various angles (Fig. 5): interactome profiling (for instance, through immunoprecipitation coupled with mass spectrometry)[83], localization mapping (through immunofluorescence or proximity labelling)[83], protein structure investigations[73], cell viability assays in human-tissue-derived organoids[77,101,125], transgenic studies in mice using de novo gene knock-ins[126–128], and investigations focused on development and disease. Especially organoids are a unique in vitro model for studying the roles of de novo genes and the proteins they encode in a close-to-natural environment, as they mimic healthy and diseased tissues[129]. Organoids can be expanded while maintaining the (epi)genetic architecture and cellular composition of the tumour or tissue sample they were derived from. They can be genetically engineered to create lines with endogenously tagged de novo proteins or deletion mutants[130,131]. Studies investigating the role of human-specific genes that originated through gene duplication have previously used organoids[132–134] and transgenic mice[126–128], illustrating the strength and importance of advanced in vitro model systems to properly dissect the roles of young human genes and proteins. With these ideas in mind, future studies should be performed to answer the remaining open questions in the field of de novo gene evolution and determine which fraction of de novo protein-coding genes is functional in humans, and in what way.

## References

1. Casari, G., De Daruvar, A., Sander, C. & Schneider, R. Bioinformatics and the discovery of gene function. *Trends Genet.* **12**, 244–245 (1996).
2. Boguski, M. S., Tolstoshev, C. M. & Bassett, D. E. Gene discovery in dbEST. *Science* **265**, 1993–1994 (1994).
3. Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).
4. Kong, S., Tao, M., Shen, X. & Ju, S. Translatable circRNAs and lncRNAs: driving mechanisms and functions of their translation products. *Cancer Lett.* **483**, 59–65 (2020).
5. Lu, S. et al. A hidden human proteome encoded by 'non-coding' genes. *Nucleic Acids Res.* **47**, 8111–8125 (2019).
6. Ruiz-Orera, J., Villanueva-Cañas, J. L. & Albà, M. M. Evolution of new proteins from translated sORFs in long non-coding RNAs. *Exp. Cell. Res.* **391**, 111940 (2020).
7. Mudge, J. M. et al. Standardized annotation of translated open reading frames. *Nat. Biotechnol.* **40**, 994–999 (2022).
8. Frankish, A. et al. GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
9. Kozak, M. Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J. Biol. Chem.* **266**, 19867–19870 (1991).
10. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326 (2010).
11. Jacob, F. Evolution and tinkering. *Science* **196**, 1161–1166 (1977).
12. Carvunis, A.-R. et al. Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
13. Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J. L., Messeguer, X. & Albà, M. M. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat. Ecol. Evol.* **2**, 890–896 (2018).
14. Baalsrud, H. T. et al. De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol. Biol. Evol.* **35**, 593–606 (2018).
15. Schmitz, J. F., Ullrich, K. K. & Bornberg-Bauer, E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat. Ecol. Evol.* **2**, 1626–1632 (2018).
16. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* **343**, 769–772 (2014).
17. Zhang, L. et al. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nat. Ecol. Evol.* **3**, 679–690 (2019).
18. Wu, D.-D., Irwin, D. M. & Zhang, Y.-P. De novo origin of human protein-coding genes. *PLoS Genet.* **7**, e1002379 (2011).
19. Ruiz-Orera, J. et al. Origins of de novo genes in human and chimpanzee. *PLoS Genet.* **11**, e1005721 (2015).
20. Zhu, S. et al. An oncopeptide regulates m$^6$A recognition by the m$^6$A reader IGF2BP1 and tumorigenesis. *Nat. Commun.* **11**, 1685 (2020).
21. Guo, Z.-W. et al. Translated long non-coding ribonucleic acid ZFAS1 promotes cancer cell migration by elevating reactive oxygen species production in hepatocellular carcinoma. *Front. Genet.* **10**, 1111 (2019).
22. Shao, Y. et al. GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. *Genome Res.* **29**, 682–696 (2019).
23. Guerzoni, D. & McLysaght, A. De novo genes arise at a slow but steady rate along the primate lineage and have been subject to incomplete lineage sorting. *Genome Biol. Evol.* **8**, 1222–1232 (2016).
24. Chen, J.-Y. et al. Emergence, retention and selection: a trilogy of origination for functional de novo proteins from ancestral lncRNAs in primates. *PLoS Genet.* **11**, e1005391 (2015).
25. Samusik, N., Krukovskaya, L., Meln, I., Shilov, E. & Kozlov, A. P. *PBOV1* is a human de novo gene with tumor-specific expression that is associated with a positive clinical outcome of cancer. *PLoS ONE* **8**, e56162 (2013).
26. Li, C.-Y. et al. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput. Biol.* **6**, e1000734 (2010).
27. Suenaga, Y. et al. *NCYM*, a *cis*-antisense gene of *MYCN*, encodes a de novo evolved protein that inhibits GSK3β resulting in the stabilization of MYCN in human neuroblastomas. *PLoS Genet.* **10**, e1003996 (2014).
28. Knowles, D. G. & McLysaght, A. Recent de novo origin of human protein-coding genes. *Genome Res.* **19**, 1752–1759 (2009).
29. Xie, C. et al. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* **8**, e1002942 (2012).
30. Van Oss, S. B. & Carvunis, A.-R. De novo gene birth. *PLoS Genet.* **15**, e1008160 (2019).
31. Schlötterer, C. Genes from scratch – the evolutionary fate of de novo genes. *Trends Genet.* **31**, 215–219 (2015).
32. McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* **17**, 567–578 (2016).
33. Weisman, C. M. The origins and functions of de novo genes: against all odds? *J. Mol. Evol.* **90**, 244–257 (2022).
34. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
35. Dujon, B. The yeast genome project: what did we learn? *Trends Genet.* **12**, 263–270 (1996).
36. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T. C. G. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* **25**, 404–413 (2009).

37. Weisman, C. M., Murray, A. W. & Eddy, S. R. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol.* **18**, e3000862 (2020).
38. Levy, A. How evolution builds genes from scratch. *Nature* **574**, 314–316 (2019).
39. Toll-Riera, M. et al. Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* **26**, 603–612 (2009).
40. Suntsova, M. V. & Buzdin, A. A. Differences between human and chimpanzee genomes and their implications in gene expression, protein functions and biochemical properties of the two species. *BMC Genom.* **21**, 535 (2020).
41. Zhuang, X., Yang, C., Murphy, K. R., Christina Cheng, C. H. & Cheng, C.-H. C. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc. Natl Acad. Sci. USA* **116**, 4400–4405 (2019).
42. Grandchamp, A., Berk, K., Dohmen, E. & Bornberg-bauer, E. New genomic signals underlying the emergence of human proto-genes. *Genes* **13**, 284 (2022).
43. Vakirlis, N., Vance, Z., Duggan, K. M. & McLysaght, A. De novo birth of functional microproteins in the human lineage. *Cell Rep.* **41**, 111808 (2022).
44. Clark, M. B. et al. The reality of pervasive transcription. *PLoS Biol.* **9**, 5–10 (2011).
45. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26–46 (2013).
46. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *eLife* **3**, e03523 (2014).
47. Wilson, B. A. & Masel, J. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol. Evol.* **3**, 1245–1252 (2011).
48. Aspden, J. L. et al. Extensive translation of small open reading frames revealed by poly-Ribo-seq. *eLife* **3**, e03528 (2014).
49. Van Heesch, S. et al. Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol.* **15**, R6 (2014).
50. Cabili, M. N. et al. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* **16**, 20 (2015).
51. Brar, G. A. et al. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335**, 552–557 (2012).
52. Andreev, D. E. et al. Non-AUG translation initiation in mammals. *Genome Biol.* **23**, 111 (2022).
53. Kozak, M. Pushing the limits of the scanning mechanism for initiation of translation. *Gene* **299**, 1–34 (2002).
54. Ruiz-Orera, J. & Albà, M. M. Conserved regions in long non-coding RNAs contain abundant translation and protein-RNA interaction signatures. *NAR Genom. Bioinform.* **1**, e2 (2019).
55. Xie, C. et al. A de novo evolved gene in the house mouse regulates female pregnancy cycles. *eLife* **8**, e44392 (2019).
56. Dowling, D., Schmitz, J. F. & Bornberg-Bauer, E. Stochastic gain and loss of novel transcribed open reading frames in the human lineage. *Genome Biol. Evol.* **12**, 2183–2195 (2020).
57. Vakirlis, N. et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat. Commun.* **11**, 781 (2020).
58. Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* **1**, 0127 (2017).
59. Palmieri, N., Kosiol, C. & Schlötterer, C. The life cycle of Drosophila orphan genes. *eLife* **3**, e01311 (2014).
60. Wilson, B. A., Foy, S. G., Neme, R. & Masel, J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* **1**, 0146 (2017).
61. Casola, C. From de novo to "de nono": the majority of novel protein-coding genes identified with phylostratigraphy are old genes or recent duplicates. *Genome Biol. Evol.* **10**, 2906–2918 (2018).
62. Durand, É. et al. Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res.* **29**, 932–943 (2019).
63. Vakirlis, N. et al. A molecular portrait of de novo genes in yeasts. *Mol. Biol. Evol.* **35**, 631–645 (2018).
64. Heames, B. et al. Experimental characterisation of de novo proteins and their unevolved random-sequence counterparts. Preprint at https://doi.org/10.1101/2022.01.14.476368 (2022).
65. Albà, M. M. & Castresana, J. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol. Biol. Evol.* **22**, 598–606 (2005).
66. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genom.* **14**, 117 (2013).
67. Janssen, J. W. G. et al. Concurrent activation of a novel putative transforming gene, *myeov*, and *cyclin D1* in a subset of multiple myeloma cell lines with t(11;14)(q13;q32). *Blood* **95**, 2691–2698 (2000).
68. Lynch, M. & Marinov, G. K. The bioenergetic costs of a gene. *Proc. Natl Acad. Sci. USA* **112**, 15690–15695 (2015).
69. Ángyán, A. F., Perczel, A. & Gáspári, Z. Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? *FEBS Lett.* **586**, 2468–2472 (2012).
70. Kesner, J. S., Chen, Z., Aparicio, A. A. & Wu, X. A unified model for the surveillance of translation in diverse noncoding sequences. Preprint at https://doi.org/10.1101/2022.07.20.500724 (2022).
71. Castro, J. F. & Tautz, D. The effects of sequence length and composition of random sequence peptides on the growth of *E. Coli* cells. *Genes* **12**, 1913 (2021).
72. Eicholt, L. A., Aubel, M., Berk, K., Bornberg-Bauer, E. & Lange, A. Heterologous expression of naturally evolved putative de novo proteins with chaperones. *Protein Sci.* **31**, e4371 (2022).
73. Papadopoulos, C. et al. Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. *Genome Res.* **31**, 2303–2315 (2021).
74. Bornberg-Bauer, E., Hlouchova, K. & Lange, A. Structure and function of naturally evolved de novo proteins. *Curr. Opin. Struct. Biol.* **68**, 175–183 (2021).
75. Brunet, T. D. P. & Doolittle, W. F. The generality of constructive neutral evolution. *Biol. Philos.* **33**, 2 (2018).
76. Keeling, D. M. et al. The meanings of 'function' in biology and the problematic case of de novo gene emergence. *eLife* **8**, e47014 (2019).
77. Chen, J. et al. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).
78. Yu, J. et al. lncRNA *MYCNOS* facilitates proliferation and invasion in hepatocellular carcinoma by regulating *miR-340*. *Hum. Cell* **33**, 148–158 (2020).
79. Lange, A. et al. Structural and functional characterization of a putative de novo gene in *Drosophila*. *Nat. Commun.* **12**, 1667 (2021).
80. Rivard, E. L. et al. A putative de novo evolved gene required for spermatid chromatin condensation in *Drosophila melanogaster*. *PLoS Genet.* **17**, e1009787 (2021).
81. Jiang, X. et al. Characterization of a novel human testis-specific gene: testis developmental related gene 1 (*TDRG1*). *Tohoku J. Exp. Med.* **225**, 311–318 (2011).
82. Florio, M. et al. Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *eLife* **7**, e32332 (2018).

83. van Heesch, S. et al. The translational landscape of the human heart. *Cell* **178**, 242–260.e29 (2019).
84. Martinez, T. F. et al. Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* **16**, 458–468 (2020).
85. Raj, A. et al. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife* **5**, e13328 (2016).
86. Gaertner, B. et al. A human ESC-based screen identifies a role for the translated lncRNA LINC00261 in pancreatic endocrine differentiation. *eLife* **9**, e58659 (2020).
87. Calviello, L. et al. Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13**, 165–170 (2016).
88. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890 (2015).
89. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).
90. Craig, R., Cortens, J. P. & Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **3**, 1234–1242 (2004).
91. Deutsch, E. W. et al. State of the human proteome in 2014/2015 as viewed through PeptideAtlas: enhancing accuracy and coverage through the AtlasProphet. *J. Proteome Res.* **14**, 3461–3473 (2015).
92. Deutsch, E. W. et al. Human Proteome Project mass spectrometry data interpretation guidelines 3.0. *J. Proteome Res.* **18**, 4108–4116 (2019).
93. Wright, B. W., Molloy, M. P. & Jaschke, P. R. Overlapping genes in natural and engineered genomes. *Nat. Rev. Genet.* **23**, 154–168 (2022).
94. Zhang, Y. E., Landback, P., Vibranovski, M. D. & Long, M. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol.* **9**, e1001179 (2011).
95. Bekpen, C., Xie, C. & Tautz, D. Dealing with the adaptive immune system during de novo evolution of genes from intergenic sequences. *BMC Evol. Biol.* **18**, 121 (2018).
96. Deng, Y. et al. Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* **609**, 375–383 (2022).
97. Majic, P. & Payne, J. L. Enhancers facilitate the birth of de novo genes and gene integration into regulatory networks. *Mol. Biol. Evol.* **37**, 1165–1178 (2020).
98. Zhang, S. et al. Open chromatin dynamics reveals stage-specific transcriptional networks in hiPSC-based neurodevelopmental model. *Stem Cell Res.* **29**, 88–98 (2018).
99. An, N. A. et al. De novo genes with an lncRNA origin encode unique human brain developmental functionality. *Nat. Ecol. Evol.* **7**, 264–278 (2023).
100. Qi, J. et al. A human-specific de novo gene promotes cortical expansion and folding. *Adv. Sci.* **10**, e2204140 (2023).
101. Duffy, E. E. et al. Developmental dynamics of RNA translation in the human brain. *Nat. Neurosci.* **25**, 1353–1365 (2022).
102. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl Acad. Sci. USA* **103**, 9935–9939 (2006).
103. Nielsen, R. et al. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, 0976–0985 (2005).
104. Vinckenbosch, N., Dupanloup, I. & Kaessmann, H. Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl Acad. Sci. USA* **103**, 3220–3225 (2006).
105. Rödelsperger, C. et al. Spatial transcriptomics of nematodes identifies sperm cells as a source of genomic novelty and rapid evolution. *Mol. Biol. Evol.* **38**, 229–243 (2021).
106. Witt, E., Benjamin, S., Svetec, N. & Zhao, L. Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in Drosophila. *eLife* **8**, e47138 (2019).
107. Kondo, S. et al. New genes often acquire male specific functions but rarely become essential in *Drosophila*. *Genes Dev.* **31**, 1841–1846 (2017).
108. Gubala, A. M. et al. The *goddard* and *saturn* genes are essential for *Drosophila* male fertility and may have arisen de novo. *Mol. Biol. Evol.* **34**, 1066–1082 (2017).
109. Su, Q., He, H. & Zhou, Q. On the origin and evolution of *Drosophila* new genes during spermatogenesis. *Genes* **12**, 1796 (2021).
110. Kopania, E. E. K., Larson, E. L., Callahan, C., Keeble, S. & Good, J. M. Molecular evolution across mouse spermatogenesis. *Mol. Biol. Evol.* **39**, msac023 (2022).
111. Kaneko, Y. et al. Functional interplay between *MYCN*, *NCYM*, and *OCT4* promotes aggressiveness of human neuroblastomas. *Cancer Sci.* **106**, 840–847 (2015).
112. Suenaga, Y., Nakatani, K. & Nakagawara, A. De novo evolved gene product NCYM in the pathogenesis and clinical outcome of human neuroblastomas and other cancers. *Jpn. J. Clin. Oncol.* **50**, 839–846 (2020).
113. Zhao, X. et al. CTCF cooperates with noncoding RNA MYCNOS to promote neuroblastoma progression through facilitating MYCN expression. *Oncogene* **35**, 3565–3576 (2016).
114. Kanatsu-Shinohara, M. et al. *Myc/Mycn*-mediated glycolysis enhances mouse spermatogonial stem cell self-renewal. *Genes Dev.* **30**, 2637–2648 (2016).
115. Zhang, R., Xia, L. Q., Lu, W. W., Zhang, J. & Zhu, J. S. lncRNAs and cancer. *Oncol. Lett.* **12**, 1233–1239 (2016).
116. de Magalhães, J. P. Every gene can (and possibly will) be associated with cancer. *Trends Genet.* **38**, 216–217 (2022).
117. Li, J. & Liu, C. Coding or noncoding, the converging concepts of RNAs. *Front. Genet.* **10**, 496 (2019).
118. Nam, J.-W., Choi, S.-W. & You, B.-H. Incredible RNA: dual functions of coding and noncoding. *Mol. Cells* **39**, 367–374 (2016).
119. Dinger, M. E., Gascoigne, D. K. & Mattick, J. S. The evolution of RNAs with multiple functions. *Biochimie* **93**, 2013–2018 (2011).
120. Brunet, M. A. et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* **47**, D403–D410 (2019).
121. Neville, M. D. C. et al. A platform for curated products from novel open reading frames prompts reinterpretation of disease variants. *Genome Res.* **31**, 327–336 (2021).
122. Olexiouk, V., Van Criekinge, W. & Menschaert, G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* **46**, D497–D502 (2017).
123. Graur, D. et al. On the immortality of television sets: 'function' in the human genome according to the evolution-free gospel of encode. *Genome Biol. Evol.* **5**, 578–590 (2013).
124. Ruiz-Orera, J., Albà, M. M. & Alba, M. M. Translation of small open reading frames: roles in regulation and evolutionary innovation. *Trends Genet.* **35**, 186–198 (2019).
125. Prensner, J. R. et al. Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nat. Biotechnol.* **39**, 697–704 (2021).
126. Xing, L. et al. Expression of human-specific *ARHGAP11B* in mice leads to neocortex expansion and increased memory flexibility. *EMBO J.* **40**, e107093 (2021).
127. Schmidt, E. R. E., Kupferman, J. V., Stackmann, M. & Polleux, F. The human-specific paralogs SRGAP2B and SRGAP2C differentially modulate SRGAP2A-dependent synaptic development. *Sci. Rep.* **9**, 18692 (2019).
128. Suzuki, I. K. et al. Human-specific *NOTCH2NL* genes expand cortical neurogenesis through Delta/Notch regulation. *Cell* **173**, 1370–1384.e16 (2018).

129. Pollen, A. A. et al. Establishing cerebral organoids as models of human-specific brain evolution. *Cell* **176**, 743–756.e17 (2019).
130. Lancaster, M. A. et al. Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373–379 (2013).
131. Sidhaye, J. et al. Integrated transcriptome and proteome analysis in human brain organoids reveals translational regulation of ribosomal proteins. Preprint at https://doi.org/10.1101/2022.10.07.511280 (2022)
132. Fischer, J. et al. Human-specific *ARHGAP11B* ensures human-like basal progenitor levels in hominid cerebral organoids. *EMBO Rep.* **23**, e54728 (2022).
133. Heide, M., Huttner, W. B. & Mora-Bermúdez, F. Brain organoids as models to study human neocortex development and evolution. *Curr. Opin. Cell Biol.* **55**, 8–16 (2018).
134. Fiddes, I. T. et al. Human-specific *NOTCH2NL* genes affect Notch signaling and cortical neurogenesis. *Cell* **173**, 1356–1369.e22 (2018).
135. Cardoso-Moreira, M. et al. Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).
136. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

## Author contributions

L.A.B. and S.v.H. carried out conceptualization, methodology, visualization and writing of the original draft. L.A.B. was responsible for data curation, investigation and software. L.A.B. and J.R.-O. carried out formal analysis. Supervision was carried out by S.v.H. L.A.B., S.v.H., J.R.-O., B.S. and N.H. reviewed and edited the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41559-023-02014-y.

**Correspondence** should be addressed to Sebastiaan van Heesch.

**Peer review information** *Nature Ecology & Evolution* thanks Chuan-Yun Li, Lars Eicholt and Nikolaos Vakirlis for their contribution to the peer review of this work.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.