OPINION

Open questions in the study of *de novo* genes: what, how and why

Aoife McLysaght and Laurence D. Hurst

Abstract | The study of *de novo* protein-coding genes is maturing from the *ad hoc* reporting of individual cases to the systematic analysis of extensive genomic data from several species. We identify three key challenges for this emerging field: understanding how best to identify *de novo* genes, how they arise and why they spread. We highlight the intellectual challenges of understanding how a *de novo* gene becomes integrated into pre-existing functions and becomes essential. We suggest that, as with protein sequence evolution, antagonistic co-evolution may be key to *de novo* gene evolution, particularly for new essential genes and new cancer-associated genes.

There are many routes by which new genes are recruited to an organism's genome. Duplication, horizontal transfer, chimaeras and other 'bricolage' origins showcase evolution as a tinkerer, reusing existing functional parts. By contrast, the discovery of *de novo* open reading frames (ORFs) — that is, those that have arisen, at least in part, from previously non-coding sequence (FIG. 1) — testifies to the existence of mechanisms by which functional protein-coding sequences can originate 'from scratch'. The examination of *de novo* gene evolution enables a better understanding of what defines the genetic basis of a species and of how the transition from a functionless to a functional sequence occurs.

The initial studies of *de novo* genes were necessarily 'proof-of-concept'; the primary goal was to identify examples through detailed analyses[1–9]. With the current increase in popularity of broad-brush pangenomic approaches to identifying *de novo* genes — made possible by the upsurge in high-quality, complete genome-level sequencing — now is an opportune time to identify the foundational cornerstones, open questions and potential pitfalls for this emerging field.

Rather than reviewing the current state of knowledge regarding *de novo* genes, which has been well covered elsewhere[10–12], we highlight what we consider to be three key questions. First, how can a putative functional *de novo* gene be confirmed? Second, how do *de novo* genes appear and what are the genomic contexts in which they arise? Third, and perhaps most challenging, why do *de novo* genes become fixed? Here, we highlight two coupled uncertainties. First, why do some new genes take over extant functions (as opposed to specifying new functions during periods of evolutionary innovation). Second, how is it that a *de novo* gene can become not only functional but essential for viability or fertility. We suggest a possible resolution to both uncertainties, one with the potential to explain why many *de novo* genes are involved in cancer.

Identification of *de novo* genes

The defining characteristic of *de novo* protein-coding genes is that they originate, at least in part, from DNA sequence that has not, at least recently, been protein-coding. A new gene may be entirely *de novo* (having arisen entirely from non-coding sequence), mostly *de novo* or partially *de novo*. We suggest that it is most useful to assess them in a hierarchical framework (FIG. 1) and categorize them into one of three logically distinct classes (types I–III). Gene duplicates, chimeric genes and simple horizontal transfer events may be classified as new genes but not as *de novo* genes. Given this definition, two challenges are foremost: being sure that the gene really is *de novo* and being confident that it is functional.

***Is it* de novo?** The first step in *de novo* gene identification is to identify an ancestral node in the phylogenetic tree where the ORF was absent. In practice, identifying this node is challenging, not least because of the possibility of independent lineage sorting[13]. An important distinction needs to be made between methods that infer newness by absence of sequence similarity alone and those that also use synteny data.

The phylostratigraphic approach[14], used for *en masse* analysis of putatively *de novo* genes, typically involves only a search for sequence similarity (for example, using BLAST) between a gene in a focal species and in more distant genomes. A gene is considered to have originated in the common ancestor of all the lineages in which it is detectable. In principle, the sensitivity of this approach in detecting other gene gain events (for example, horizontal gene transfer) depends on its implementation (that is, which genomes are being compared) and parameterization (BLAST search criteria). The approach is tractable for analysis of ancient putative *de novo* genes and can be scaled readily.

Sequence similarity methods are, however, prone to false positives. Likewise, in the reporting of gene loss, close scrutiny has shown that BLAST analysis can indeed fail[15]. This is because BLAST searches are based on the span and level of sequence similarity, and when that overall similarity becomes too low even true homologues are not detected. This loss of signal happens faster for quickly evolving genes and short genes, meaning homologues of such genes are harder to detect, thus creating a bias in the BLAST results. In investigations of both *de novo* gain and gene loss, small and fast evolving proteins are more likely to be missed even when present[16,17]. Importantly, a recent reanalysis[18] challenges the reported high rates of *de novo* sequence creation that were based on BLAST-based methods[19,20]. Of the 16 putative *de novo*

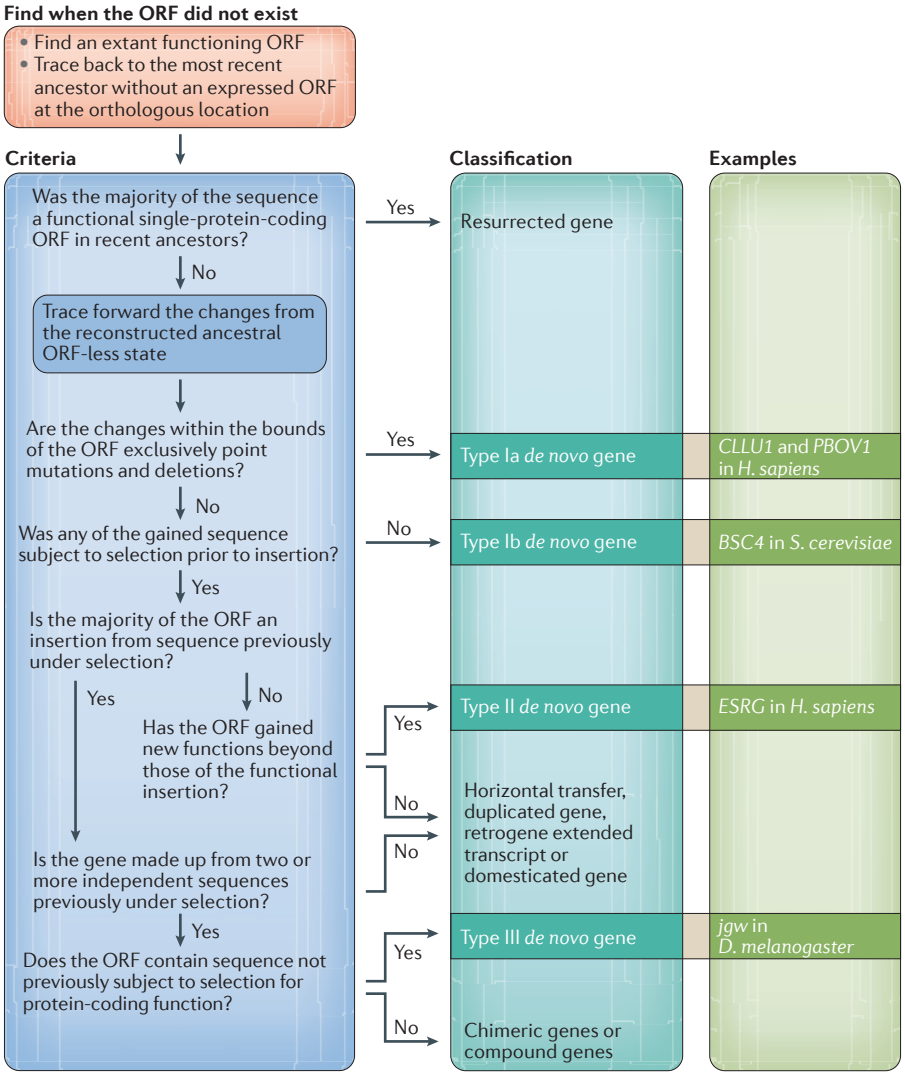


Figure 1 | **A systematic approach to the classification of novel genes.** This classification is based on tracing the evolution of a locus to the most recent common ancestor in which it can be inferred that there was no expressed open reading frame (ORF) at the orthologous location, and then inspecting the evolutionary steps in the origin of the gene. This requires abundant, closely related, high-quality genomes. Evidence of the presence of a translated ORF, even a relatively short one, can be sought through ribosome protection assays[33,121,122]. The defining characteristic of a *de novo* gene is that it evolved from previously non-coding sequence; therefore, it must be possible to identify that sequence otherwise the classification must remain ambiguous. ORFs for which an ORF-less orthologous location cannot be found cannot be defined as *de novo* or new, as we cannot exclude the possibility that they are ancient but fast evolving. The approach using sequence similarity alone can, at best, suggest that the new gene bears no resemblance to extant genes or horizontally transferred genes, thus implicating *de novo* origination by exclusion. If no attempt is made to reconstruct the ancestral state or if it is not practically possible to do so, these are classified as 'putatively *de novo* genes' and are not considered in this taxonomy. We propose a hierarchical classification of *de novo* genes based on whether or not there is newly inserted sequence in the locus and whether or not that sequence had been previously under natural selection: type Ia and type Ib are entirely derived from non-protein-coding sequence; type II contains a minority of sequence that was previously under selection (such as transposable elements or portions of a pre-existing gene), but that does not explain the function of the modern gene; and type III *de novo* genes are chimaeras of sequences previously under selection, but which contain some novel sequence not previously under selection for protein-coding function. Type I genes are the most clean-cut, whereas the gene histories for type II and type III are progressively more convoluted. Real world complexity probably means that type III genes will be contested in some instances. For example, *jingwei* (*jgw*) has not previously been considered as a *de novo* gene, but it is a new gene that includes previously intronic (that is, non-protein-coding) sequence, and classifying it as a type III *de novo* gene acknowledges this mixed history. *CLLU1*, chronic lymphocytic leukaemia upregulated 1; *D. melanogaster*, *Drosophila melanogaster*; *ESRG*, embryonic stem cell related; *H. sapiens*, *Homo sapiens*; *PBOV1*, prostate and breast cancer overexpressed 1; *S. cerevisiae*, *Saccharomyces cerevisiae*.

genes reported to be specific to the focal lineage, *Saccharomyces cerevisiae*[19], 15 were after reanalysis reported to be not species-specific and none have signals of selection or of translation[18].

Although some candidate *de novo* genes are false positives, it can be argued that the BLAST-based method remains valuable as the proportion might be low. One way to address this issue is to simulate gene evolution under the assumption that within the taxa under consideration there are no *de novo* genes (that is, all genes are shared across all lineages). BLAST-based methods can then be applied to the resulting simulated genes to determine how many artefactually appear to be *de novo*. In turn, it can be asked how this number of false discoveries compares with the number claimed to be recent originations in the real data. Recent, albeit conservative, null simulations[18] addressing a prior analysis of *de novo* genes across 14 yeast species[19], report that approximately 11% of simulants artefactually seem to be recent originations, with a few seeming to be species-specific. We estimate that this equates to a false positive rate of approximately 30% (669 false positives[18] out of 2,253 reported *de novo* genes[19]), although this ignores the possibility of simulation-based error, which can introduce false negatives and false positives. In the original analysis, 5% of the putative *de novo* genes were shown to have homologues outside of the study species, a distribution that may be the result of differential gene loss or horizontal transfer events.

It remains unknown how representative this approximate 30% false positive figure might be. More generally, the extent of ascertainment bias is debated, with some[21] noting that spans of slow evolving sequence in a fast evolving gene make the bias less profound[16,22]. The simulations mentioned in (REF. 18), however, included rate heterogeneity. Thus, there can be little doubt that there must be an ascertainment bias of some degree. Given the false positive problem, we refer to candidate *de novo* genes identified by absence of sequence similarity alone as 'putative *de novo* genes'.

Perhaps the more relevant question is whether trends ascribed to putative *de novo* genes might themselves be ascertainment bias artefacts. Simulations show that many, if not most, of the claimed age-related trends[19] can, at least in principle, be qualitatively accounted for by such artefacts[18]. As expected, given the above BLAST bias, genes that are short or have

a high rate of sequence evolution seem to be younger genes, although all genes in the simulation have the same age[17]. Other gene features, such as low expression levels and low codon usage bias are known to correlate with fast rates of sequence evolution and are thus expected to track with the BLAST bias, as indeed they do[18]. Similarly, the claim that *de novo* genes keep expanding[19] is parsimoniously explained as bias towards detection of longer genes in more distant taxa[18]. Less obviously biased trends, such as the claim that human disease-associated genes tend to be old[23], are also observed in null simulations[17] (possibly because disease genes tend to be longer than average[24]). Over-representation of young genes associated with ectoderm in *Drosophila* species[17] is one counter example, as this trend is not observed in simulations.

Although the simulation approach suggests that many of the claimed trends might be due to artefactual signals, they do not show that they actually are. How might we resolve this issue? One option is to exclude from analysis the putative *de novo* genes that are expected, given the simulations, to be strong candidates as false positives (that is, genes that might artefactually seem young) and test for correlations only on the remainder. However, the robustness of this approach is uncertain, not least because the eliminated gene set is likely to be sensitive to simulation assumptions (for example, the degree of intragene and interlineage heterogeneity in rates of evolution).

The other, safer, option is to go beyond sequence similarity in defining the timing of origination. For recent *de novo* gene origination, such problems are circumvented by analysing outgroup genomes to identify positive evidence of the absence of the expressed ORF (FIG. 1). The typical approach is to use conserved synteny to identify the expected location for a gene in the outgroup and then to inspect the DNA sequence to identify the orthologous DNA and determine whether or not it has coding capacity or activity[1,4,6]. The advantage of the synteny approach is that the determination of *de novo* status is near definitive. This strategy, however, is only tractable if there is still sufficient synteny conservation and is harder to scale.

Are you sure it is a functioning protein-coding gene? In studying the origin of *de novo* genes, we are interested in the evolution of functional protein-coding genes. This leaves two questions: how do you know it is actually translated, and how do you know it is functional? The former can be ascertained directly through proteomics and ribosome protection assays. But how can 'functional' be defined? This issue is not unique to this debate; for example, it is also central to understanding the importance of the numerous non-coding RNAs and alternative splice forms.

Glossary

Apert syndrome
A congenital disorder that is caused by the failure of appropriate apoptosis to occur during fetal development, resulting in malformed skull, face, hands and feet.

Biased gene conversion
Gene conversion (that is, the replacement of a DNA sequence by homologous sequence from the other allele at the same locus, or from elsewhere in the genome) involving a process that repairs mismatches in a non-random fashion. Gene conversion in mammals is thought to be weakly biased towards GC residues over AT residues.

***dN*/*dS* analysis**
(Also known as *Ka*/*Ks* analysis). Analysis to determine the ratio of nonsynonymous substitutions per nonsynonymous site (*dN*) to synonymous substitutions per synonymous site (*dS*), which is indicative of the mode of evolution acting on a protein-coding gene. This is interpreted as purifying selection if less than 1; positive selection if greater than 1; and neutral evolution if effectively 1. The numbers of substitutions are estimated by counting the observed differences in orthologous genes identified in at least two different species.

Domesticated genes
Exogenous genetic material that has become incorporated into a genome and subsequently adapted for a host function.

Fixed
An allele is said to be fixed in a population once it rises to 100% frequency.

Independent lineage sorting
This phenomenon is observed when population polymorphism segregating in an ancestral species is maintained past two (or more) speciation events, such that the descendent species each contains alleles that date back to before the speciation events. The descendent species may each independently fix one or other of the ancestral alleles (independently sorting the alleles). When sister species fix different alleles to each other the phylogenetic relationship of the genes is different from the phylogenetic relationship of the species.

Maternal effect lethals
Loci in which the maternal genotype determines the viability of the zygote.

Meiotic drive
Any process that causes a given allele to be overrepresented in the gametes following meiosis. Most commonly, the term is restricted to cases in which the distorted segregation ratios affect whole chromosomes rather than just a particular chromosomal location.

Neofunctionalization
Evolution of a novel function, which may exist alongside an ancestral function or replace an ancestral function.

Open chromatin
Decondensed chromosomal structure associated with gene expression.

Orthologues
Homologous genes that diverged following a speciation event.

Paralogues
Homologous genes that diverged following a gene duplication event; duplicated genes.

Phylostratigraphic approach
An approach for estimating gene age based on its phylogenetic distribution. Commonly, genes are inferred to have been present in the common ancestor of any organisms in which they are detectable by sequence similarity search (such as BLAST), and their origin is assigned to the branch on the tree just prior to the node corresponding to that common ancestor. The term is a portmanteau of phylogenetics and stratigraphy, the latter being the study and dating of rock layers.

Purifying selection
(Also known as negative selection). Removal of deleterious mutations from a population by selection.

Red Queen co-evolution
Named after the Red Queen in Lewis Carroll's *Alice Through the Looking Glass* who is continually running to stay in the same place. This describes an evolutionary scenario in which two interacting loci (often one from a parasite and one from a host) are both rapidly evolving but the relationship (the interaction) has no qualitative change.

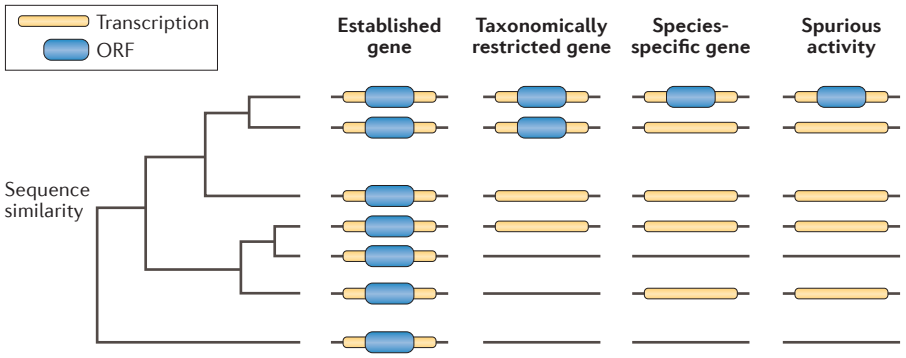
Selective sweeps
Positive selection on a DNA mutation that incidentally carries closely linked variation to high frequency, thus reducing the genetic diversity in the surrounding region of the genome.

Site frequency spectrum
Distribution of allele frequencies at a set of loci. The shape of the distribution can be used to infer demography and natural selection (for example, through hitchhiking).

Spurious gene expression
Gene expression with no selective advantage. A necessary concept but one that in practice is hard to demonstrate, not least because the strength of selective effects relevant to the evolutionary process is typically more subtle than the effects measured in the laboratory. Selective advantage may be defined with respect to the bearer genome or to a selfish element.

Subfunctionalization
Partitioning of functions of an ancestral, multifunctional gene between daughter paralogues.

Synteny
Meaning 'same chromosome', this describes the physical genetic linkage of two or more loci on a chromosome. A region of shared synteny between genomes (where the orthologous genes have an equivalent relative location) is indicative of genome arrangement conservation since their most recent common ancestor. In the context of *de novo* genes, the syntenic location in the ancestral genome is the expected location of origin of the gene.



	Established gene	Taxonomically restricted gene	Species-specific gene	Spurious activity
Evolutionary conservation	✓	✓or✗	✗	✗
Purifying selection ($dN/dS < 1$)	✓	✓	NA	NA
Positive selection ($dN/dS > 1$)	✓or✗	✓	NA	NA
Transcription	✓	✓	✓	✓
Translation	✓	✓	✓	✓
Knockdown or knockout phenotype	✓or✗	✓or✗	✓or✗	✗

Figure 2 | **Validation of novel genes.** There are significant challenges in verifying lineage-specific *de novo* genes owing to the unavailability of some tools and the uninformative nature of others. Well-established genes receive support from sequence similarity across many genomes and evolutionary patterns of selection, and a certain fraction of them will show a phenotype in a knockout or knockdown experiment. For most of these tests, species-specific genes are indistinguishable from spurious expression in the genome. Functional characterization through observation of knockdown or knockout phenotypes is one way to distinguish genuine novel genes from spurious expression. Note that the distinction between 'taxonomically restricted' and 'species-specific' need not be absolute in that the latter can become the former when more closely related genomes are added. In the lower panel, ticks indicate observed and crosses indicate not observed. NA, not available; ORF, open reading frame.

Is it functional? The Encyclopedia of DNA Elements (ENCODE) project advocated a definition of functionality based on evidence of physical interaction (loosely defined as any biochemical activity or interaction between chromosomal regions or between proteins and DNA) or proximity to physical interaction[25]. Anything transcribed is functional according to this definition, rendering most of the human genome functional. But by the same logic, the bonnet (or hood) of a car would be ascribed the function of propelling pedestrians several metres[26], this interaction being observed in traffic accidents. Put differently, the ENCODE definition sidesteps the null of accidental interaction (for example, spurious protein–DNA binding). Many argue against such a definition on principle[27,28]. Moreover, as *de novo* gene origination requires the previously unselected expression of a section of DNA as a step in the process, this definition is particularly unhelpful in the context of *de novo* genes.

Evidence that 'accidents' commonly occur comes from genomic features that make sense as evolved responses to such accidents. For example, the observations of in-frame stop codons in small introns[29], avoidance of codons that are one transcription error from a stop codon[30] and a high density of exonic splice enhancers near larger introns and weak splice sites[31] are suggestive of the prevalence of spurious gene products. All of these make sense if accidents happen that result in sufficient selective pressure to generate quality-control measures: mis-splicing is mitigated by an in-frame stop codon in an intron; mistranscription must be sufficiently common to select for residues less damaging if mistranscribed; and splice enhancers are common at weak splice sites and long introns because they are prone to being accidentally mis-spliced.

How then might we differentiate accidents (for example, spurious gene expression) from biologically relevant cellular events? A powerful approach is to examine the mode of sequence evolution, looking for evidence of selection[27,28]. We can therefore reformulate functional relevance as 'visibility to selection' either to enhance or to maintain fitness and function. Although visibility to selection need not imply that the product, per se, is functional — the process of transcription and/or splicing could be the selectively relevant function[32] — in the case of *de novo* protein-coding genes, we are explicitly interested in the evolution of a functional product, the protein.

To determine whether a protein is visible to selection, typically the orthologous genes are first identified and compared with each other to determine whether faster rates of evolution occur at synonymous sites and intronic sites than at non-synonymous sites. We can also test whether promoters and splice sites are conserved. Unfortunately, with such a methodology, there is an inverse relationship between the age of a putative gene (and hence, our inability to define it as *de novo*) and our ability to discern function through sequence analysis alone. Long-established protein-coding genes, for example, are easy to validate as functional, but their origins are commonly ambiguous.

For the validation of *de novo* genes that are recent but not species-specific, these same criteria can be applied (for examples, see REFS 33,34) (FIG. 2). For such genes, the signals may be of positive selection (characteristic of genes that are acquiring novel or improved functionality) rather than purifying selection[35]. Although recent but not species-specific genes are perhaps the most tractable for *en masse* sequence-based discovery and validation of *de novo* genes, misinterpretation of the signals of accelerated evolution must be avoided; for example, due to biased gene conversion rather than positive selection[36].

Validation of the functionality of species-specific genes presents the most difficult challenge, owing to an absence of an orthologous ORF. However, they are the easiest to validate as being truly *de novo* using synteny approaches. For species-specific ORFs, false positives are an important issue owing to the abundance of small ORFs, which are present even in randomized sequences. In the human genome, there are more than 13.5 million ORFs of at least 33 codons in length[11]. Reasons such as strength of selection and genetic load alone render it unlikely that all of these ORFs are subject to selection to be preserved. More directly, ribosome protection assays have detected evidence of translation for only approximately 1% of the 108,000 unannotated ORFs in yeast that are longer than 30 bp[19]. Given this prevalence of presumed false positives, conservative tests for functionality are preferable.

How then can functional recent *de novo* genes be distinguished from non-functional artefacts? Methods that rely on sequence

analysis alone are unlikely to be definitive. Although evidence of selective sweeps[12] is suggestive of functionality, it is hard to pinpoint the specific gene or mutation that is being selected for. Moreover, absence of evidence of a selective sweep is not definitive because the spread to fixation of the *de novo* gene (or mutations in the gene) could pre-date the last common ancestor of the individuals studied. A possible alternative is to rely on the analysis of single-nucleotide polymorphisms (SNPs) and look for deviations from neutral expectations in, for example, the site frequency spectrum. However, small new genes are unlikely to have a sufficiently high SNP density for such methods to be conclusive.

Given the possibility of spurious gene expression, data indicative of expression provide support for functionality but cannot be considered definitive evidence. If sequence analysis is not definitive (for example, in the case of species-specific genes), then we suggest that the only way to be confident that an apparent *de novo* gene is actually functional is to conduct experimental functional assays (FIG. 2). As knockdowns and CRISPR–Cas9 knockouts have made functional testing easier, the prospects for expansion in this direction are evident[37,38]. So far, functional characterization has only been carried out in a handful of cases[6,39–42].

Unfortunately, definition based on an observable fitness effect in the laboratory usually identifies many fewer classical protein-coding genes as functional than suggested by *dN*/*dS* analysis[43,44]. This apparent abundance of well-described protein-coding genes with little or no phenotype upon knockout may reflect irrelevance of the tested environment[43], genotype effects[45], functional compensation (for example, by paralogues)[37] or our inability to detect small but selectively relevant effects. For the case of *de novo* genes, it may also reflect their selective irrelevance. If such a definition is adopted as a gold standard, many genes will be left in a classification limbo. Given the high false positive expectations, we suggest that a significant false negative rate is the better of two evils.

Translation is not evidence of functionality. Tests for functionality might be considered unnecessary if an ORF can be demonstrated to be translated. Unfortunately, such evidence alone need not be a reliable indicator. The analysis of pseudogenes (genes with disrupted ORFs or presumed loss of promoters) is especially enlightening in this context. A recent proteomics study identified protein products from 107 so-called pseudogenes in humans[46]. Subsequent analysis using macaque orthologues found evidence for significant purifying selection in 5 out of 34 translated pseudogenes, which may call into question their designation as pseudogenes. For the remainder, the null hypothesis of neutrality could not be rejected, suggesting — but not proving — lack of functionality. Knockdown of these 34 genes would be a valuable next step, comparable to what has been attempted in some studies[47]. If the molecular evolutionary analysis is correct, we expect that knockdown of the five conserved genes will show a phenotype, whereas knockdown of the others should be of little or no consequence.

The above results also lend credence to the idea that spurious transcription and translation cannot be dismissed. The extent of spurious transcription, despite having been discussed for more than two decades[48], remains unknown. Bioinformatics analyses suggest that in a random sequence there is a high density of possible transcription factor-binding sites[49,50], which leads to a possible null test: if 5,000 random base pairs of sequence were inserted into the human genome, how often would some of it be transcribed and, if transcribed, would it be translated? Would it be spliced? If expressed, would this be repeatedly time- and tissue-specific?

The converse possibility, that the signatures of functionality trump the absence of direct evidence for protein production, can be more reasonable. A gene with a $dN/dS << 1$ but with no evidence of translation is highly likely to be a functional protein, the expression of which has yet to be resolved. In this case, knockouts are less definitive as they cannot differentiate functional effects at the RNA and protein levels.

How do *de novo* genes arise?

A continuing debate in the field of evolutionary genomics concerns whether phenotypic change is caused by changes in expression or changes in the protein sequence[51]. Another issue is the extent to which selection acts on established genes versus new genes (*de novo* or otherwise). As always, the answer will depend on two factors: the rate at which new variants arise and the likelihood that these variants will be selectively favoured. Here we consider the former.

It is important to understand these mutational processes to better understand the patterns that cannot be explained by such processes, and thus implicate selection. If, for example, a burst in *de novo* gene creation is observed in a given lineage, we might suspect selection for innovation in that lineage. However, this may instead reflect a change (for example, a new transcription factor) that increases the net rate of *de novo* gene creation. Conversely, if we witness a propensity for *de novo* genes to arise in transcriptionally active genomic domains, we might suspect a mutation bias; however, this might reflect selective filtering. For example, in *Drosophila* spp., there seems to be a bias towards the origin of *de novo* genes on the X chromosome[1,52]. Is this because the X chromosome is more likely to give rise to *de novo* genes, or because the chances of fixation are higher for X-linked *de novo* genes? Many features of the X chromosome are likely to affect fixation probability, including different effective population size, recombination rate and mean penetrance. That this trend is not observed for genes expressed in the testes[35] suggests a selective filter of some kind.

The determination of factors affecting the rate of *de novo* gene creation is in its infancy. In this section we highlight several open questions (FIG. 3).

How important is intergene distance? Are *de novo* genes more likely to arise in genomes with large intergenic spaces (as these have more 'evolutionary playground' in which genes can emerge)? As the expression of one gene affects that of its neighbours[53], the expression of a new ORF sequence may be mainly limited to the borders of intergene spaces. Consequently, relatively compact genomes may have more *de novo* genes per base pair (though not in absolute terms). Notably, close proximity or even an overlapping arrangement of *de novo* genes with pre-existing genes seems to be commonplace[4,54,55]. The prior existence of a gene may thus facilitate the origination of a *de novo* gene nearby[54]. Plausible mechanisms include transcription read-through, bidirectional promoter activity[56,57] or spurious transcription in regions of open chromatin. The contributions of these effects have yet to be examined fully, although recent data suggest that read-through is surprisingly common; the previous estimate[58,59] of 4–6% of genes being subject to transcriptional read-through is now considered an underestimate[60]. This

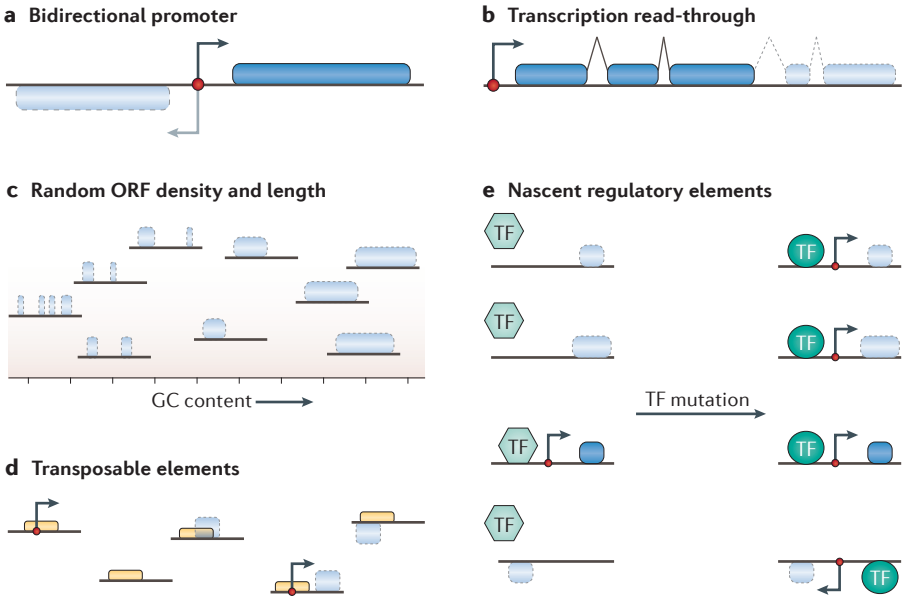


Figure 3 | **Features of genome anatomy that alter the likelihood of novel gene origination.** **a** | A pre-existing gene (dark blue box) may facilitate the origin of a *de novo* gene (light blue box) through the re-use of the promoter, perhaps in a bidirectional conformation. **b** | A pre-existing gene may also contribute to novel gene origination through transcriptional read-through, which can occur in as many as 11% of cases. **c** | Start (ATG) and stop (TAA, TAG and TGA) codons are AT-rich and are thus less frequent in GC-rich regions of the genome. This means that the distance between any start and stop codon is longer in GC-rich regions of genome, resulting in longer random open reading frames (ORFs; light blue boxes). Conversely, although stop codons are common in GC-poor regions, so are start codons; thus, random ORFs will be short but so will the inter-ORF distances, resulting in high ORF density. **d** | Transposable elements (yellow boxes) can facilitate the origin of new genes by providing regulatory sequences (arrows) or by contributing to the ORF of the novel gene (light blue boxes). **e** | Mutation of a transcription factor (TF) may alter its DNA-binding specificity, thus activating expression at many previously unexpressed loci (light blue boxes) with or without affecting the expression of pre-existing genes (dark blue boxes).

suggests that particularly large intergene distances need not make a genome especially prone to new gene evolution. Conversely, recent evidence suggests that over a very limited evolutionary span, all genomic sequence is transcribed[61,62], possibly opening large intergenic spaces for gene origination.

Is GC content important? The evidence discussed above suggests that *de novo* genes are more likely to be found in proximity to active genes. In the case of a chromatin accessibility model, the origination of a *de novo* gene still requires *ad hoc* expression. How likely is this expression, and what determines whether a reasonably sized ORF might be found? As many *cis* factors influence the transcription of a locus[63], different regions of the genome will have different propensities towards *ad hoc* expression, and perhaps even contain arbitrary sequence that includes cryptic regulatory elements. As stop codons are AT-rich, GC-rich regions of the genome are also more likely to include long arbitrary ORFs (the same logic explains why frame-shifting runs are longer in GC-rich genomes[64]). Similarly, GC-rich segments of the human genome are more transcriptionally active[65], and transcription factor-binding sites tend to be GC-rich (in mammals, but not other organisms)[66]. Conversely, AT-rich regions are likely to have a higher density of possible start (ATG) codons.

How important are transposable elements? An emerging theme is the extent to which *de novo* genes co-opt segments of transposable elements and generate *de novo* genes in *cis*. Such elements come preloaded with functional elements, such as transcription factor-binding sites[67], regulatory sequences[68,69] and splice sites that are amenable to the generation of chimeric genes[70,71] and short ORFs[71]. For this reason, it is highly likely that transposable elements are generators of many new transcripts. As a case in point, recent evidence indicates that sequences derived from human endogenous retrovirus subfamily H (HERVH) produce new transcripts in humans. HERVH-derived sequences generate many non-coding RNAs (including some that are key to pluripotency; for example, *LINC-ROR* (long intergenic non-protein-coding RNA, regulator of reprogramming), which acts as a microRNA sponge[72]), chimeric transcripts and, occasionally, new protein-coding genes (for example, embryonic stem cell related (*ESRG*)[39]). Alternatively, if a genome, such as that of *Neurospora crassa*[73], can rid itself of transposable elements, does this limit its ability to evolve new genes? Similarly, there are suggestive links with retroviral DNA[66]. The contribution of coding sequence to novel genes from transposable elements[8] may often be better regarded as domestication or horizontal transfer.

Are non-coding RNAs launch pads for* de novo *genes? Given that evolution is more likely to occur in an incremental, stepwise fashion than in a sudden transition from functionless sequence to functional protein, can expressed and functional or neutral non-coding RNAs provide a launch pad for *de novo* gene origination by extending the opportunity for an ORF to originate? Consistent with this idea, several studies have reported the origination of functional protein coding genes from RNA genes[34–36,74–76]. *De novo* gene origination might also be more likely in some genomic regions, for example GC-rich domains where spontaneous ORFs are likely to be longer[35].

Do changes in gene regulation cause* de novo *gene evolution? Which new transcripts would be created following a mutation in a transcription factor? Mutations in these genes might indirectly create a proto-gene, owing to transcription factor-binding and activation of transcription at a previously untranscribed locus. At first sight, the genome-wide disruption of such *trans* effects would probably result in strong negative selection. It is curious, then, that some transcription factors have fast evolving DNA-binding domains (for example, retinal homeobox 1 (RX1)[77]), and some of the fastest evolving gene families are transcription factors. Indeed, 13% of the human transcription factor repertoire seems to be primate-specific, compared with just 2% for metabolic genes[78]. The relationship between transcription factor expansion and/or mutation, and *de novo* gene origination is worth further analysis.

From rare to common

The considerations discussed above concern the 'how' question in relation to mutations: how does a *de novo* gene arise from a non-transcribed non-coding sequence to become a transcribed, translated ORF-containing gene? Any key mutation enabling the final step in this process must have started as rare and become common. Why does this occur? The issues in the following subsections relate not only to *de novo* genes but to new genes in general. However, as highlighted, many of the issues are at their most acute when considering *de novo* genes.

Essentiality and established functionality; two coupled problems. At first sight, a shift from a random to a functional sequence looks hard to achieve, but in practice this does not seem to be the case. Indeed, it has been shown that a random sequence that is fused to an expressed functional sequence can readily evolve beneficial effects[79]. Assuming that a transition to functionality is not such a major hurdle, we highlight two enigmatic and coupled issues.

First is the problem of how new genes integrate into an established functionality. In the case of a *de novo* gene performing a novel function, a rare mutation that confers some fitness advantage may be relatively commonplace, as it only has to be better than nothing, especially in the context of a change in the environment. Conversely, if a *de novo* gene serendipitously has functionality that overlaps with a pre-existing gene (or genes), it is harder to see how it would confer any fitness advantage, as it would have to do better than an already adapted network.

Second, and perhaps most mysterious, is how some *de novo* genes undergo a transition to not only become functional but essential[6,11,37,39,42,47,80]. This transition is particularly intriguing, not least because both *de novo* genes[80] and horizontally transferred genes[81] tend to be found at the periphery of networks and have low connectivity. By contrast, there tends to be an association between the highly connected central hubs of gene networks and essentiality[37,82]. The problem of evolving essentiality is more acute for *de novo* genes than for other types of new genes. Indeed, unlike duplicated genes, which can become essential through the passive division of labour of essential functions between the two duplicates (subfunctionalization[83]), *de novo* genes can only become essential through adaptive evolution of new functionality (neofunctionalization[47]), as they have no functions to lose.

The first of these issues may be of little concern if incorporation into an established functionality is just a rare oddity. Indeed, phylostratigraphic plots of the dates of origins of new genes show a possible coincidence between the timing of peaks of gene origin activity and major evolutionary innovations[10]. However, simulations of a null model can also provide artefactual signals of bursts in *de novo* gene creation[17]. Robust testing will require quantitative definition of innovation rate (for example, expansion in the number of cell types), rather than *post hoc* interpretation. Nonetheless, prima facie evidence supports the notion of *de novo* genes for new functions.

Although the balance of 'new genes for new functions' to 'new genes for old functions' remains unknown, the problem of *de novo* genes inserting into prior functionality cannot be ignored, not least because it is repeatable. For example, the human type II *de novo* gene *ESRG* (FIG. 1), derived in part from an endogenous retrovirus, is involved in the maintenance of pluripotency in human naive stem cells[39]. This human-specific gene performs a function that is not new, in so much as we presume primates other than humans also have pluripotent stem cells. Most curiously, in mice, transcripts derived from a different endogenous retrovirus (murine endogenous retrovirus with leucine tRNA (*MERVL*; also known as *ERV4*)) control pluripotency[84,85].

Systematic analysis also suggests that both evolution of essentiality and incorporation into extant networks is detected easily. A recent CRISPR study of essential genes in human cell lines found that putatively new essential genes (loosely defined by the authors as genes with sequence matches only in metazoans or younger taxonomic groups) tend to operate with older essential genes (those found in opisthokonts; that is, those found in yeast and metazoans), suggesting that some essential genes are younger than the molecular pathways in which they operate[37]. This leads to the possibility that the functions of an old essential gene are, for reasons unknown, handed over to the new gene — much like the passing of a baton in a relay race[86] — potentially leading to the new gene adopting essential functions and possibly to the future demise of the older gene. As with a relay, if the baton is dropped, the race is over; thus, non-functional intermediates must be avoided[87].

Perhaps the best-studied example of this 'baton pass' phenomenon is the syncytin genes of mammals, which are best classed as domesticated genes rather than *de novo* genes. Syncytin genes establish the syncytium — a multi-nucleated cell that forms the fetal boundary of the placenta — by promoting cell–cell fusion and may also be involved in suppression of the maternal immune reaction. Mammalian syncytin genes are domesticated envelope glycoprotein (*env*) genes of retroviral origin that function to mediate fusion of the virus particle and the host cell, and also to suppress the host immune reaction. The similarity between the native function of *env* in retroviruses and the domesticated function in mammals makes *env* genes an unsurprising source for syncytins. More surprising is that the mammalian syncytins are not all orthologues of each other. In primates, the syncytium is established by syncytin 1 and syncytin 2 (encoded by *ERVW1* and *ERVV2*, respectively), but in mice the responsible genes are the non-orthologous genes syncytin A (*Syna*) and *Synb*, although they are also domesticated from a retroviral *env* gene (and are essential for placentation). Thus, the placenta is established by different, lineage-specific genes[40,88–91]. In the opossum (a marsupial), the syncytin gene turnover process has apparently been caught in the act, and it is possible to detect not only the functional syncytin gene but also a degraded syncytin that has evidence of past functionality, as well as a newer *env* that has acquired placenta- and uterus-specific expression and may yet emerge as a replacement syncytin[90]. Similarly, there is evidence for a relic of a predecessor of the human syncytin[92]. Other aspects of placentation in mammals have also experienced convergent evolution[93,94]. This repeatability, like that of pluripotency control, suggests that an explanation beyond happenstance is needed.

Can antagonistic co-evolution explain the two enigmas? These violations of the 'if it isn't broken, don't fix it' rule raise the question of why any organism might swap an existing functional gene for a new one. A recent study suggests that the new syncytin gene confers some increased benefit to the host compared with the older 'primitive' one[90]. However, why this might be is not obvious; in addition, this hypothesis does not explain how the emergent *env* can become adapted to function as a better syncytin than its predecessor. We propose that antagonistic co-evolution (BOX 1) may well explain the apparent paradoxes, just as it explains why immune system genes

Box 1 | **Antagonistic co-evolution**

In many circumstances, selection on a given allele is a result of the prevalence of an antagonistic biological agent. Such selection can in turn result in counter selection on the antagonist. For example, the dispersal of a damaging parasite creates the context in the host population for the spread of alleles offering some resistance (or otherwise reducing negative effects of infection). In turn, the parasite may counter-evolve, which can initiate further host evolution. The classic example of this phenomenon is the rapid evolution of immune-related genes in the host to 'catch up' with the rapid evolution of antigens in a parasite. Similarly, the rapid evolution of parasite immunogenic genes is both expected and observed. Although there is a large amount of evolutionary sequence change, the effect is to maintain a relationship (that is, immune recognition of a parasite) not to add a new one. This phenomenon has been dubbed 'Red Queen co-evolution' after the queen in Lewis Caroll's *Alice Through the Looking Glass* who insisted that, "…it takes all the running you can do, to keep in the same place."

The pair of biological agents may be different species (as in host–parasite, plant–herbivore and predator–prey interactions), different members of the same species (as in sexual conflict, sibling conflict and parent–offspring conflict), or different genes in the same genome (intragenomic or genetic conflict). The spread of a meiotic drive gene and the subsequent spread of resistance alleles is a case in point.

In the genetic conflict leading to the spread of resistance alleles, the initiating allele is commonly termed a selfish genetic element. The interaction strongly resembles host–parasite evolution in which the parasite receives only vertical transmission; for example, by being part of the same genome. The finding that cancer-associated genes are regularly associated with positive selection may reflect selection on alleles that enable germline cells to resist apoptosis. If the reduced organismal fitness that is associated with cancer is not too great, then the increased reproductive success of such alleles, which behave as selfish elements, is sufficiently strong to enable their increase in frequency in the population. Whether there is antagonistic co-evolution in this instance is unknown but it is theoretically plausible.

Different manifestations of antagonism are neither theoretically identical nor equally potent selective forces. For example, host–parasite co-evolution is considered more acute than plant–herbivore interactions, as parasite infection can be more damaging than herbivory and the number of generations of the parasite per host generation is usually considerable compared with the number of generations of the herbivore per plant generation, thus potentiating faster evolution. Note too that many antagonisms (for example, host–parasite) have a density-dependent component, whereas selfish genetic elements typically do not. Such density dependence can promote an ongoing antagonism; for example, when host populations grow in size the host density goes up, making parasite transmission more likely.

tend to be fast evolving. Genetic conflict was also proposed in the evolution of a new essential function in the *Drosophila melanogaster* duplicate gene *Umbrea* (also known as *HP6*)[95]. It remains to be seen what proportion of the evolution of essentiality is due to such processes.

It may prove to be the case that new genes, and *de novo* genes in particular, are mostly associated with external environmental change. In this case, our hypothesis may be of little relevance. However, antagonistic interactions have at least two characteristics that create fertile territory for the evolution of novelty, of which *de novo* genes are extreme examples. First, invasion of one allele creates the conditions for invasion of a response allele. More generally, antagonism can lead to perpetual Red Queen co-evolution with ongoing selection to incrementally increase the functionality of the invader and selection on new response functionality (suppression). Second, even small effect mutations can come under relatively strong selection. A meiotic drive gene that receives even weakly biased transmission (for example, a 51/49 segregation ratio) can still very easily invade. In population genetics, mutations of relatively strong effect are deemed to be any with $s > (1/2)N_e$, where s is the selection coefficient and N_e is the effective population size. A weak meiotic drive gene will have $s \approx 0.01$, orders of magnitude higher than this threshold for most taxa. Similarly, the strength of selection on even a weak suppressive modifier of meiotic drive is also likely to be orders of magnitude higher, just because the selective coefficients associated with conflicts can be remarkably large.

These are conditions that we predict would favour both the continuous rapid evolution of established proteins and the selection for novel proteins, *de novo* proteins included. In this framework, paradoxical features such as baton-pass evolution can be provided with parsimonious rationale. For example, in maternal–fetal interactions, the mother may have evolved mechanisms to suppress the activity of an established gene, providing the context for the counter-response: a new gene. The suppressed gene is then expected to be lost and replaced by the new unsuppressed gene.

This suggestion raises a simple prediction: there should be an association between gene classes in which we see the classical hallmarks of positive selection ($dN/dS > 1$) — a good indicator of antagonism — and the biological processes associated with new genes. Indeed, this seems to be the case. The usual culprits for antagonistic evolution (sexual antagonism, parent–offspring conflict and immune systems) are indeed foci for new genes (BOX 2). More specifically, an involvement of new genes in the maternal–fetal relationship is to be expected as it is a classically antagonistic relationship with ample opportunity for 'tug-of-war' evolution. Consistent with this, the placenta is one of the most variable tissues across mammalian species[96,97] despite its essential role.

Other instances of new genes adopting the function of existing genes may be a consequence of more subtle antagonisms by which the new gene is positively selected as it enables an advantage in a conflict situation to be passed to one of the parties. It may be relevant that pluripotency genes originated from ERVs in mammals. Is there a conflict between the advantage for the ERV to retrotranspose in a 'selfish gene' manner and the need of the organism to silence regions of the genome as part of the cellular differentiation process? Might this explain why, in both mice and humans, ERV-derived transcripts (from different retroviruses) are needed for the maintenance of pluripotency, thus extending the phase of transposability[98]?

Antagonisms can generate new essential genes. Although the discussion above suggests that *de novo* gene integration into extant functioning systems is not necessarily so enigmatic, why might essentiality also be part of the process? In some cases, we see that, for instigators of conflict, the inviability (or infertility) that defines essentiality is central to their spread in a population. Consider, for example, cytoplasmic incompatibility, a common phenotype in insects that is a consequence of the action of a cytoplasmically inherited bacterium, typically *Wolbachia* spp. Conceptually, in male insects the bacteria leave a toxin in the sperm, whereas the bacteria in zygotes that are inherited from the mothers provide

the antidote and neutralize the toxin. The precise mechanism is unclear, but it involves altered histone deposition in sperm and improper condensation of the paternal chromosomes in the first zygotic division[99]. A zygote without the bacterium, and thus the antidote, dies. Owing to this death, the bacterium increases in frequency in the population and can reach fixation[100]. Note how embryonic death is central to the spread. At fixation, loss of the antidote gene results in embryonic inviability (that is, it is an essential gene). So a new, possibly *de novo,* gene (the antidote) has been introduced into the population and, when the toxin is sufficiently common, has become a gene essential for the viability of the host.

We propose that this process, both narrowly and broadly viewed, is an exemplar for the evolution of essentiality of novel genes. Here, we refer to novel genes more generally (*de novo* being a subclass) but the logic applies to all. In a narrow view, the two-locus (toxin–antidote) system is seen in many contexts, including meiotic drive[101] and plasmid maintenance systems[102], and appears to explain others: such as maternal effect lethals (for example, *Medea* (maternal effect dominant embryonic arrest) in *Tribolium castaneum* and *Scat* (severe combined anaemia and thrombocytopenia) in mice[103]), both of which also seem to be species-specific. This type of system can drive the fixation of a new gene, even from a low starting frequency[104]. Indeed, this mechanism has been exploited for the introduction of new genes for pest control[105]. Take the case of *Medea* and *Scat*: conceptually, the mother puts a toxin into her embryos and those with the linked antidote gene survive, whereas those without it die. The gene pair can rapidly spread to fixation, at which point (as above) loss of the antidote will be lethal as all mothers will have the toxin. Similarly, two-locus lineage-specific meiotic drive genes are also well described[106,107] and can have large phenotypic effects in hybrids, indicating how rapidly such conflict-related processes can give rise to processes that, if disrupted or in a novel context (such as a hybrid), can have major phenotypic effects.

These two-locus toxin–antitoxin systems are, when viewed broadly, exemplars of a set of genes involved in genetic conflicts and commonly antagonistic co-evolution, which need not all be mediated by toxin–antidote systems. Embryonically expressed genes that 'manipulate' maternal resources will be similar in that their spread reduces fitness of some individuals, thereby instigating antagonisms. Similarly, a new gene that blocks a prevalent parasite from infecting the host is part of an ongoing antagonism. We also expect subtle antagonism between cell fate decisions, such as between vegetative growth and sexual reproduction. Indeed, the yeast *de novo* gene *MDF1* (mating depressing factor 1; also known as *FYV5*) suppresses mating and favours vegetative growth[41,42]. Similarly, there is an antagonism inherent between cell growth and cellular senescence, which can sometimes occur between tumour suppressor genes and oncogenes. This antagonism is evident in patterns of positive selection on protein-coding sequences in which genes involved in tumour suppression, apoptosis and spermatogenesis have a high incidence of positive selection, and many are shared between the processes[108,109]. This has been suggested to be due to conflicts between increased spermatogenesis (that is, reduced apoptosis) and the integrity of adult tissues (that is, avoidance of cancer)[108]. The persistence of Apert syndrome due to a mutation that confers selective sperm-level advantage is a related example[110]. Consistent with this antagonism and our hypothesis, a

Box 2 | **Processes experiencing positive selection and novel gene gain**

Positive selection (usually measured as a *dN/dS* ratio >1) of protein sequences is characteristic of genes that are undergoing adaptive evolution. Such evolution is comparatively rare, with the prevalent signature in genes being conservative purifying selection (*dN/dS* < 1). Genes that experience positive selection are thus interesting and are often associated with new or enhanced functions in the organism or adaptation to a new niche. Positive selection is also associated with groups of genes that experience antagonistic co-evolution (BOX 1).

Surveys of positive selection in mammalian genomes have identified genes that are experiencing adaptive evolution, and some of these are functionally related and are grouped into the same or related pathways. Notably, genes under positive selection in humans and chimpanzees are often involved in tumour suppression, apoptosis and spermatogenesis, leading to the suggestion of an antagonistic relationship between spermatogenesis and cancer[108]. The relationship between cell death and gametogenesis was also found in a survey of positively selected genes in mammalian genomes[109].

Interestingly, these functional categories are also observed among novel genes. Although most novel genes remain uncharacterized, some have been identified with functions in sperm motility (polymorphic derived intron containing (*Pldi*) in mouse[118]), apoptosis (BH3-like motif-containing cell death inducer (*BLID*; also known as BRCC2) in primates[119,120]) and cancer (see TABLE 1). Similar to the overlapping functional categories found in positively selected genes, *BLID* is a pro-apoptotic gene that also contributes to cancer outcomes.

In addition, of the 27 gene ontology (GO) categories overrepresented in positively selected genes in mammals[109], 9 were also found to be overrepresented in 'orphan' genes (genes without identifiable progenitors) that originated during vertebrate evolution (phylostrata 11–19 in REF. 23) (see the table). This suggests that the same antagonistic process can result in either adaptive evolution of protein sequence or selection for new genes. As quickly evolving genes are likely to be susceptible to the BLAST-search bias, interpretation of simple associations between positive selection and new genes can be problematic. This bias could be avoided by requiring *de novo* designation through synteny analysis.

GO term*	Phylostrata	Description
GO:0007186	19	G protein-coupled receptor protein signalling pathway
GO:0007338	19	Single fertilization
GO:0045087	15 and 16	Innate immune response
GO:0006968	11, 12 and 16	Cellular defence response
GO:0019882	11 and 14	Antigen processing and presentation
GO:0006959	14	Humoral immune response
GO:0050778	14	Positive regulation of immune response
GO:0050909	13	Sensory perception of taste
GO:0009615	12	Response to virus

*Table GO terms are enriched for both significant *dN/dS* (that is, >1)[107] and new gene origin along the vertebrate tree[23].

Table 1 | **Novel human and primate genes associated with cancer**

Gene	Product	Description
De novo *genes*		
CLLU1	CLLU1	This gene was first annotated from a screen for upregulated genes in CLL[123] and was one of the first identified human-specific *de novo* genes[4]
PART1	PART1	This primate-specific *de novo* gene has been implicated as a tumour suppressor gene[124]
MYCNOS (also known as *NCYM*)	Ncym	*De novo* human gene that stabilizes its antisense gene, the oncogene *MYCN*, in neuroblastomas[111]
PBOV1	PBOV1	*De novo* human gene associated with positive clinical outcomes in cancer[112]
GR6 (also known as *LINC01565*)	GR6	*De novo* human- and chimpanzee-specific gene that is normally expressed early in fetal development. Ectopic expression is associated with leukaemia[13,125]
Other novel genes		
KLKP1	KRIP1	Primate-specific gene duplication; expression of this gene was found to be upregulated in prostate cancer[126]
BLID	BLID	Primate-specific gene, probably originated by gene duplication. This gene is an inducer of apoptosis, and when downregulated is associated with poor breast cancer outcomes[119,120]

BLID, BH3-like motif containing, cell death inducer; *CLLU*, chronic lymphocyte leukaemia upregulated 1; *KLKP1*, kallikrein pseudogene 1; *MYCNOS*, MYCN opposite strand; *PART1*, prostate androgen-related transcript 1; *PBOV1*, prostate and breast cancer overexpressed 1.

substantial proportion of the few human or primate *de novo* genes with any functional clues has been implicated in cancer and cancer outcomes[4,8,111,112], as have other novel primate genes (TABLE 1). Analysis of the genes expressed exclusively or selectively in testes and malignant tumours[113–115] will be informative.

Future prospects

As is common in genomics, there is an inevitable trade-off between scalable pangenomic analysis, with its associated false positive issues, and closely studied case histories, in which ancestry and functionality are both resolved. We have argued that for *de novo* gene analysis such broad-brush approaches need to provide at least strong evidence that candidate genes and claimed trends are not method artefacts. With regard to trends, simulations of realistic null models help. For all *de novo* genes, especially the youngest, functional characterization in addition to sequence-based analysis is desirable.

These issues should not be consigned to the molecular evolution community. For example, determining why one sequence becomes expressed but a similar sequence elsewhere in the genome is silent, is central to understanding between-genome trends in rates of *de novo* gene origination, but should also flag potential side effects of gene therapies and transgenesis. Indeed, the cancellation of the first gene therapy trials owing to *cis* activation of an oncogene[116], provides as much a warning about genomic manipulation as it tells us about the possible consequences of introducing novel transcripts in parts of the genome where they were not previously located. Similarly, we can ask how often, if we disrupt genomes, we might cause the expression of previously unexpressed sequence. The problem of how likely it is that this will occur, how often the product would be translatable and whether a new peptide could fold[117] or be toxic is central to *de novo* gene analysis and is potentially relevant to understanding both the pathology of chromosomal mutations and the unintended consequences of genome manipulations. In these contexts, studies of the expression of inserted random sequences and the gain of function of expressed random sequence will be valuable.

Aoife McLysaght is at The Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland.

Laurence D. Hurst is at The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, Somerset BA2 7AY, UK.

Correspondence to A.McL.
aoife.mclysaght@tcd.ie

doi:10.1038/nrg.2016.78
Published online 25 Jul 2016;
corrected online 27 Jul 2016

1. 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).
2. Begun, D. J., Lindfors, H. A., Thompson, M. E. & Holloway, A. K. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* **172**, 1675–1681 (2006).
3. Xiao, W. *et al.* A rice gene of *de novo* origin negatively regulates pathogen-induced defense response. *PLoS ONE* **4**, e4603 (2009).
4. Knowles, D. G. & McLysaght, A. Recent *de novo* origin of human protein-coding genes. *Genome Res.* **19**, 1752–1759 (2009).
5. Li, L. *et al.* Identification of the novel protein QQS as a component of the starch metabolic network in *Arabidopsis* leaves. *Plant J.* **58**, 485–498 (2009).
6. Cai, J., Zhao, R., Jiang, H. & Wang, W. *De novo* origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179**, 487–496 (2008).
7. Zhou, Q. & Wang, W. On the origin and evolution of new genes — a genomic and experimental perspective. *J. Genet. Genom.* **35**, 639–648 (2008).
8. Toll-Riera, M. *et al.* Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* **26**, 603–612 (2009).
9. Wu, D.-D., Irwin, D. M. & Zhang, Y.-P. *De novo* origin of human protein-coding genes. *PLoS Genet.* **7**, e1002379 (2011).
10. Tautz, D. & Domazet-Loso, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
11. McLysaght, A. & Guerzoni, D. New genes from non-coding sequence: the role of *de novo* protein-coding genes in eukaryotic evolutionary innovation. *Phil. Trans. R. Soc. B* **370**, 20140332 (2015).
12. Schlötterer, C. Genes from scratch — the evolutionary fate of *de novo* genes. *Trends Genet.* **31**, 215–219 (2015).
13. 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).
14. Domazet-Loso, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
15. Wolfe, K. Evolutionary genomics: yeasts accelerate beyond BLAST. *Curr. Biol.* **14**, R392–R394 (2004).
16. Elhaik, E., Sabath, N. & Graur, D. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. *Mol. Biol. Evol.* **23**, 1–3 (2006).
17. Moyers, B. A. & Zhang, J. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol. Biol. Evol.* **32**, 258–267 (2015).
18. Moyers, B. A. & Zhang, J. Evaluating phylostratigraphic evidence for widespread *de novo* gene birth in genome evolution. *Mol. Biol. Evol.* **33**, 1245–1256 (2016).
19. Carvunis, A.-R. *et al.* Proto-genes and *de novo* gene birth. *Nature* **487**, 370–374 (2012).
20. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution. *BMC Genomics* **14**, 117 (2013).
21. Alba, M. M. & Castresana, J. On homology searches by protein Blast and the characterization of the age of genes. *BMC Evol. Biol.* **7**, 53 (2007).
22. Alba, M. M. & Castresana, J. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol. Biol. Evol.* **22**, 598–606 (2005).
23. Domazet-Loso, T. & Tautz, D. An ancient evolutionary origin of genes associated with human genetic diseases. *Mol. Biol. Evol.* **25**, 2699–2707 (2008).
24. Smith, N. G. C. & Eyre-Walker, A. Human disease genes: patterns and predictions. *Gene* **318**, 169–175 (2003).
25. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
26. Hurst, L. D. Open questions: a logic (or lack thereof) of genome organization. *BMC Biol.* **11**, 58 (2013).
27. 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).

28. Doolittle, W. F. Is junk DNA bunk? A critique of ENCODE. *Proc. Natl Acad. Sci. USA* **110**, 5294–5300 (2013).
29. Jaillon, O. *et al.* Translational control of intron splicing in eukaryotes. *Nature* **451**, 359–362 (2008).
30. Cusack, B. P., Arndt, P. F., Duret, L. & Roest Crollius, H. Preventing dangerous nonsense: selection for robustness to transcriptional error in human genes. *PLoS Genet.* **7**, e1002276 (2011).
31. Dewey, C. N., Rogozin, I. B. & Koonin, E. V. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* **7**, 311 (2006).
32. Schüler, A., Ghanbarian, A. T. & Hurst, L. D. Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol. Biol. Evol.* **31**, 3164–3183 (2014).
33. 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).
34. 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).
35. 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).
36. Galtier, N., Duret, L., Glémin, S. & Ranwez, V. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.* **25**, 1–5 (2009).
37. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092–1096 (2015).
38. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
39. Wang, J. *et al.* Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**, 405–409 (2014).
40. Lavialle, C. *et al.* Paleovirology of '*syncytins*', retroviral *env* genes exapted for a role in placentation. *Phil. Trans. R. Soc. B* **368**, 20120507 (2013).
41. Li, D., Yan, Z., Lu, L., Jiang, H. & Wang, W. Pleiotropy of the *de novo*-originated gene *MDF1*. *Sci. Rep.* **4**, 7280 (2014).
42. Li, D. *et al.* A *de novo* originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Res.* **20**, 408–420 (2010).
43. Ghysen, A. Debatable issues. Interview with L Wolpert and A García-Bellido. *Int. J. Dev. Biol.* **42**, 511–518 (1998).
44. Tautz, D. A genetic uncertainty problem. *Trends Genet.* **16**, 475–477 (2000).
45. Chalfin, L. *et al.* Mapping ecologically relevant social behaviours by gene knockout in wild mice. *Nat. Commun.* **5**, 4569 (2014).
46. Xu, J. & Zhang, J. Are human translated pseudogenes functional? *Mol. Biol. Evol.* **33**, 755–760 (2016).
47. Chen, S., Zhang, Y. E. & Long, M. New genes in *Drosophila* quickly become essential. *Science* **330**, 1682–1685 (2010).
48. Bird, A. P. Gene number, noise reduction and biological complexity. *Trends Genet.* **11**, 94–100 (1995).
49. Hurst, L. D. Evolutionary genomics and the reach of selection. *J. Biol.* **8**, 12 (2009).
50. Prestridge, D. S. & Burks, C. The density of transcriptional elements in promoter and non-promoter sequences. *Hum. Mol. Genet.* **2**, 1449–1453 (1993).
51. Hoekstra, H. E. & Coyne, J. A. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61**, 995–1016 (2007).
52. Begun, D. J., Lindfors, H. A., Kern, A. D. & Jones, C. D. Evidence for *de novo* evolution of testis-expressed genes in the *Drosophila yakuba*/*Drosophila erecta* clade. *Genetics* **176**, 1131–1137 (2007).
53. Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat. Cell Biol.* **10**, 1106–1113 (2008).
54. Siepel, A. Darwinian alchemy: human genes from noncoding DNA. *Genome Res.* **19**, 1693–1695 (2009).
55. Murphy, D. N. & McLysaght, A. *De novo* origin of protein-coding genes in murine rodents. *PLoS ONE* **7**, e48650 (2012).
56. Gotea, V., Petrykowska, H. M. & Elnitski, L. Bidirectional promoters as important drivers for the emergence of species-specific transcripts. *PLoS ONE* **8**, e57323 (2013).
57. Wu, X. & Sharp, P. A. Divergent transcription: a driving force for new gene origination? *Cell* **155**, 990–996 (2013).
58. Akiva, P. *et al.* Transcription-mediated gene fusion in the human genome. *Genome Res.* **16**, 30–36 (2006).
59. Parra, G. *et al.* Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.* **16**, 37–44 (2006).
60. Nacu, S. *et al.* Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med. Genom.* **4**, 11 (2011).
61. Ruiz-Orera, J. *et al.* Origins of *de novo* genes in human and chimpanzee. *PLoS Genet.* **11**, e1005721 (2015).
62. Neme, R. & Tautz, D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence. *eLife* **5**, e09977 (2016).
63. Necsulea, A. & Kaessmann, H. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat. Rev. Genet.* **15**, 734–748 (2014).
64. Warnecke, T., Huang, Y., Przytycka, T. M. & Hurst, L. D. Unique cost dynamics elucidate the role of frameshifting errors in promoting translational robustness. *Genome Biol. Evol.* **2**, 636–645 (2010).
65. Lercher, M. J., Urrutia, A. O., Pavlícek, A. & Hurst, L. D. A unification of mosaic structures in the human genome. *Hum. Mol. Genet.* **12**, 2411–2415 (2003).
66. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).
67. Wang, T. *et al.* Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. **104**, 18613–18618 (2007).
68. Gotea, V. & Makałowski, W. Do transposable elements really contribute to proteomes? *Trends Genet.* **22**, 260–267 (2006).
69. Thornburg, B. G., Gotea, V. & Makałowski, W. Transposable elements as a significant source of transcription regulating signals. *Gene* **365**, 104–110 (2006).
70. Göke, J. *et al.* Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16**, 135–141 (2015).
71. Denli, A. M. *et al.* Primate-specific ORF0 contributes to retrotransposon-mediated diversity. *Cell* **163**, 583–593 (2015).
72. Wang, Y. *et al.* Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev. Cell* **25**, 69–80 (2013).
73. Galagan, J. E., & Selker, E. U. RIP: the evolutionary cost of genome defense.*Trends Genet.* **20**, 417–413 (2004).
74. Xie, C. *et al.* Hominoid-specific *de novo* protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* **8**, e1002942 (2012).
75. Palmieri, N., Kosiol, C. & Schlötterer, C. The life cycle of *Drosophila* orphan genes. *eLife* **3**, e01311 (2014).
76. Neme, R. & Tautz, D. Evolution: dynamics of *de novo* gene emergence. *Curr. Biol.* **24**, R238–R240 (2014).
77. Kamijyo, A., Yura, K. & Ogura, A. Distinct evolutionary rate in the eye field transcription factors found by estimation of ancestral protein structure. *Gene* **555**, 73–79 (2015).
78. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).
79. Hayashi, Y., Sakata, H., Makino, Y., Urabe, I. & Yomo, T. Can an arbitrary sequence evolve towards acquiring a biological function? *J. Mol. Evol.* **56**, 162–168 (2003).
80. Zhang, W., Landback, P., Gschwend, A. R., Shen, B. & Long, M. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol.* **16**, 202 (2015).
81. Lercher, M. J. & Pál, C. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol. Biol. Evol.* **25**, 559–567 (2008).
82. Batada, N. N., Hurst, L. D. & Tyers, M. Evolutionary and physiological importance of hub proteins. *PLoS Comp. Biol.* **2**, e88 (2006).
83. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
84. Schoorlemmer, J., Pérez-Palacios, R., Climent, M., Guallar, D. & Muniesa, P. Regulation of mouse retroelement MuERV-L/MERVL expression by REX1 and epigenetic control of stem cell potency. *Front. Oncol.* **4**, 14 (2014).
85. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
86. Imakawa, K., Nakagawa, S. & Miyazawa, T. Baton pass hypothesis: successive incorporation of unconserved endogenous retroviral genes for placentation during mammalian evolution. *Genes Cells* **20**, 771–788 (2015).
87. Aakre, C. D. *et al.* Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594–606 (2015).
88. Esnault, C., Cornelis, G., Heidmann, O. & Heidmann, T. Differential evolutionary fate of an ancestral primate endogenous retrovirus envelope gene, the EnvV *syncytin*, captured for a function in placentation. *PLoS Genet.* **9**, e1003400 (2013).
89. Cornelis, G. *et al.* Retroviral envelope *syncytin* capture in an ancestrally diverged mammalian clade for placentation in the primitive Afrotherian tenrecs. *Proc. Natl Acad. Sci. USA* **111**, E4332–E4341 (2014).
90. Cornelis, G. *et al.* Retroviral envelope gene captures and *syncytin* exaptation for placentation in marsupials. *Proc. Natl Acad. Sci. USA* **112**, E487–E496 (2015).
91. Cornelis, G. *et al.* Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. *Proc. Natl Acad. Sci. USA* **110**, E828–E837 (2013).
92. Dupressoir, A., Lavialle, C. & Heidmann, T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured *syncytins* in placentation. *Placenta* **33**, 663–671 (2012).
93. Emera, D. *et al.* Convergent evolution of endometrial prolactin expression in primates, mice, and elephants through the independent recruitment of transposable elements. *Mol. Biol. Evol.* **29**, 239–247 (2012).
94. Maston, G. A. & Ruvolo, M. Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. *Mol. Biol. Evol.* **19**, 320–335 (2002).
95. Ross, B. D. *et al.* Stepwise evolution of essential centromere function in a *Drosophila* neogene. *Science* **340**, 1211–1214 (2013).
96. Elliot, M. G. & Crespi, B. J. Phylogenetic evidence for early hemochorial placentation in eutheria. *Placenta* **30**, 949–967 (2009).
97. Elliot, M. G. & Crespi, B. J. Genetic recapitulation of human pre-eclampsia risk during convergent evolution of reduced placental invasiveness in eutherian mammals. *Phil. Trans. R. Soc. B* **370**, 20140069 (2015).
98. Izsvák, Z., Wang, J., Singh, M., Mager, D. L. & Hurst, L. D. Pluripotency and the endogenous retrovirus HERVH: conflict or serendipity? *Bioessays* **38**, 109–117 (2016).
99. Landmann, F., Orsi, G. A., Loppin, B. & Sullivan, W. *Wolbachia*-mediated cytoplasmic incompatibility is associated with impaired histone deposition in the male pronucleus. *PLoS Pathog.* **5**, e1000343 (2009).
100. Fine, P. E. On the dynamics of symbiote-dependent cytoplasmic incompatibility in culicine mosquitoes. *J. Invertebr. Pathol.* **31**, 10–18 (1978).
101. Merrill, C., Bayraktaroglu, L., Kusano, A. & Ganetzky, B. Truncated RanGAP encoded by the *Segregation Distorter* locus of *Drosophila*. *Science* **283**, 1742–1745 (1999).
102. Gerdes, K. *et al.* The hok killer gene family in gram-negative bacteria. *New Biol.* **2**, 946–956 (1990).
103. Hurst, L. D. *scat*$^{+}$ is a selfish gene analogous to *Medea* of *Tribolium castaneum*. *Cell* **75**, 407–408 (1993).
104. Marshall, J. M. The toxin and antidote puzzle: new ways to control insect pest populations through manipulating inheritance. *Bioeng. Bugs* **2**, 235–240 (2011).
105. Chen, C.-H. *et al.* A synthetic maternal-effect selfish genetic element drives population replacement in *Drosophila*. *Science* **316**, 597–600 (2007).
106. Phadnis, N. & Orr, H. A. A single gene causes both male sterility and segregation distortion in *Drosophila* hybrids. *Science* **323**, 376–379 (2009).
107. Hurst, L. D. & Pomiankowski, A. Causes of sex ratio bias may account for unisexual sterility in hybrids: a new explanation of Haldane's rule and related phenomena. *Genetics* **128**, 841–858 (1991).

108. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170 (2005).
109. Kosiol, C. *et al.* Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4**, e1000144 (2008).
110. Goriely, A. *et al.* Gain-of-function amino acid substitutions drive positive selection of *FGFR2* mutations in human spermatogonia. *Proc. Natl Acad. Sci. USA* **102**, 6051–6056 (2005).
111. 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).
112. 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).
113. Zendman, A. J. W., Ruiter, D. J. & Van Muijen, G. N. P. Cancer/testis-associated genes: identification, expression profile, and putative function. *J. Cell. Physiol.* **194**, 272–288 (2003).
114. Simpson, A. J. G., Caballero, O. L., Jungbluth, A., Chen, Y.-T. & Old, L. J. Cancer/testis antigens, gametogenesis and cancer. *Nat. Rev. Cancer* **5**, 615–625 (2005).
115. Hofmann, O. *et al.* Genome-wide analysis of cancer/testis gene expression. **105**, 20422–20427 (2008).
116. Kohn, D. B., Sadelain, M. & Glorioso, J. C. Occurrence of leukaemia following gene therapy of X-linked SCID. *Nat. Rev. Cancer* **3**, 477–488 (2003).
117. Bornberg-Bauer, E. & Alba, M. M. Dynamics and adaptive benefits of modular protein evolution. *Curr. Opin. Struct. Biol.* **23**, 459–466 (2013).
118. Heinen, T. J. A. J., Staubach, F., Häming, D. & Tautz, D. Emergence of a new gene from an intergenic region. *Curr. Biol.* **19**, 1527–1531 (2009).
119. Broustas, C. G. *et al.* BRCC2, a novel BH3-like domain-containing protein, induces apoptosis in a caspase-dependent manner. *J. Biol. Chem.* **279**, 26780–26788 (2004).
120. Broustas, C. G. *et al.* The proapoptotic molecule BLID interacts with Bcl-X_L and its downregulation in breast cancer correlates with poor disease-free and overall survival. *Clin. Cancer Res.* **16**, 2939–2948 (2010).
121. Andrews, S. J. & Rothnagel, J. A. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* **15**, 193–204 (2014).
122. 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).
123. Buhl, A. M. *et al.* Identification of a gene on chromosome 12q22 uniquely overexpressed in chronic lymphocytic leukemia. *Blood* **107**, 2904–2911 (2006).
124. Lin, B. *et al. PART-1*: a novel human prostate-specific, androgen-regulated gene that maps to chromosome 5q12. *Cancer Res.* **60**, 858–863 (2000).
125. Pekarsky, Y., Rynditch, A., Wieser, R., Fonatsch, C. & Gardiner, K. Activation of a novel gene in 3q21 and identification of intergenic fusion transcripts with ecotropic viral insertion site I in leukemia. *Cancer Res.* **57**, 3914–3919 (1997).
126. Kaushal, A. *et al.* A novel transcript from the *KLKP1* gene is androgen regulated, down-regulated during prostate cancer progression and encodes the first non-serine protease identified from the human kallikrein gene locus. *Prostate* **68**, 381–399 (2008).

Acknowledgements
A.M. and L.D.H. are supported by funding from the European Research Council grant agreements 309834 and 669207, respectively.

Competing interests statement
The authors declare no competing interests.

ERRATUM

Open questions in the study of de novo genes: what, how and why

Aoife McLysaght & Laurence D. Hurst

Nature Reviews Genetics http://dx.doi.org/10.1038/nrg.2016.78 (2016)

In Table 1 of the original version of this article the gene name *NCYM* was incorrectly written as *NYCM*. This has now been corrected. The editors apologize for this error.