De Novo Genes Arise at a Slow but Steady Rate along the Primate Lineage and Have Been Subject to Incomplete Lineage Sorting

Daniele Guerzoni and Aoife McLysaght*

Smurfit Institute of Genetics, Department of Genetics, Trinity College Dublin, University of Dublin, Ireland

*Corresponding author: E-mail: aoife.mclysaght@tcd.ie.

Accepted: March 26, 2016

Abstract

De novo protein-coding gene origination is increasingly recognized as an important evolutionary mechanism. However, there remains a large amount of uncertainty regarding the frequency of these events and the mechanisms and speed of gene establishment. Here, we describe a rigorous search for cases of de novo gene origination in the great apes. We analyzed annotated proteomes as well as full genomic DNA and transcriptional and translational evidence. It is notable that results vary between database updates due to the fluctuating annotation of these genes. Nonetheless we identified 35 de novo genes: 16 human-specific; 5 human and chimpanzee specific; and 14 that originated prior to the divergence of human, chimpanzee, and gorilla and are found in all three genomes. The taxonomically restricted distribution of these genes cannot be explained by loss in other lineages. Each gene is supported by an open reading frame-creating mutation that occurred within the primate lineage, and which is not polymorphic in any species. Similarly to previous studies we find that the de novo genes identified are short and frequently located near pre-existing genes. Also, they may be associated with Alu elements and prior transcription and RNA-splicing at the locus. Additionally, we report the first case of apparent independent lineage sorting of a de novo gene. The gene is present in human and gorilla, whereas chimpanzee has the ancestral noncoding sequence. This indicates a long period of polymorphism prior to fixation and thus supports a model where de novo genes may, at least initially, have a neutral effect on fitness.

Key words: de novo genes, incomplete lineage sorting, primates, human, new genes.

Introduction

Taxonomically restricted genes are important for the evolution of lineage-specific traits (Khalturin et al. 2009). Considering protein-coding genes, the greatest genetic novelty occurs when genes originate de novo from previously noncoding DNA because the resultant proteins bear no similarity to pre-existing proteins. There is a large amount of potential for de novo genes within large eukaryotic genomes both from the large number of nonexpressed open reading frames (ORFs) (i.e., random ORFs in noncoding regions) (Carvunis et al. 2012; McLysaght and Guerzoni 2015) and pervasive transcription of noncoding regions of the genome (ENCODE Project Consortium 2012).

Surveys of de novo genes have been carried out in diverse lineages including insects, yeasts, ciliates, mammals, and plants (e.g., Begun et al. 2006, 2007; Levine et al. 2006; Donoghue et al. 2011; Yang and Huang 2011; Carvunis et al. 2012; Murphy and McLysaght 2012; Zhao et al. 2014; McLysaght and Guerzoni 2015). De novo genes are consistently discovered but usually in small numbers. De novo genes have been shown to be involved in important processes such as promoting vegetative growth in yeast (Li et al. 2010), and pathogen defense and starch biosynthesis in plants (Li et al. 2009; Xiao et al. 2009). In the human genome, some de novo genes are associated with disease (Knowles and McLysaght 2009; Toll-Riera et al. 2009; Samusik et al. 2013; Suenaga et al. 2014) and a novel transcript with protein-coding potential was recently shown to be required for maintenance of pluripotency (Wang et al. 2014).

Previous reports of de novo genes in primates have found differing results, probably due to different stringencies in the search criteria as well as volatility in genome annotations

(Knowles and McLysaght 2009; Guerzoni and McLysaght 2011; Wu et al. 2011). Thus, it is likely that two independent groups working on identifying de novo genes in the same lineage could end up with different or noncompletely overlapping findings. Considering the small numbers of robustly supported de novo originated genes, there is a risk with genome-wide studies that false positives outnumber true positives. In particular, the phylostratigraphic approach, which relies on sequence similarity searches (usually BLAST [Basic Local Alignment Search Tool]) to detect homologs and infers a young age for genes without detectable sequence similarity in more distant lineages, suffers from systematic underestimation of the age of genes, particularly for short and quickly evolving genes (Moyers and Zhang 2015; Moyers and Zhang 2016). For this reason, in this study we use a strict set of parameters designed to avoid annotation and data errors while at the same time account for alternative evolutionary explanations for the apparently taxonomically restricted distribution of a gene, such as gene loss and accelerated evolution.

We identify 36 de novo originated genes in the Homininae (human, chimpanzee, and gorilla) since their divergence from orangutan (~16.5 Ma; Perelman et al. 2011) including one gene that has experienced incomplete lineage sorting (ILS) and is present only in human and gorilla. Eighteen of these genes are supported by peptide evidence. This is the most up-to-date survey of de novo genes in our lineage and takes advantage of the large amount of available data and employs rigorous search criteria to produce reliable de novo gene identifications.

Features of the genome which may contribute to the origin of de novo protein-coding genes have been previously suggested, including the presence of transposable elements (Chen et al. 2007; Toll-Riera et al. 2009), the close proximity of pre-existing genes (Knowles and McLysaght 2009; Siepel 2009), and the prior transcription of the region, perhaps as RNA genes (Xie et al. 2012; Reinhardt et al. 2013; Ruiz-Orera et al. 2014). We find each of these associated with de novo genes reported here.

Our results suggest a relatively constant rate of origin of new genes de novo. In terms of the speed of establishment of such genes, this is likely to vary according to the biological activity, if any, of the new gene. Nonetheless, the case of ILS is at least one example where a novel gene that originated de novo in the human–chimp–gorilla ancestor remained polymorphic for an extended period, past two speciation events, which is suggestive of neutral evolution (Dutheil et al. 2015).

Materials and Methods

Sequence Data

Genome annotations and genomic sequences were obtained from EnsEMBL (Flicek et al. 2012). Data for human, chimpanzee, gorilla, orangutan, and macaque were initially downloaded from EnsEMBL v60. The full human proteome contained 81,860 proteins corresponding to 52,580 protein-coding genes. The remaining primates combined accounted for 123,532 proteins.

Subsequent updates to EnsEMBL (v61–69) were incorporated into our analyses by identifying 3,036 newly added human genes (being careful to distinguish cases where it is merely a new database ID, or a minor annotation modification). Of these, 440 protein-coding genes satisfied gene structure plausibility criteria (described below).

Data Set Refinements

Annotated genes where the coding sequence was not a multiple of 3 were excluded as implausible (>13% of the proteome in EnsEMBL v60), as were those with nonstandard start or stop codon.

The smallest known introns are 18 bp long (Gilson and McFadden 1996; Deutsch and Long 1999) so we excluded cases with introns smaller than this. Unlikely small introns are abundant in gene structures of automatically annotated genomes, such as chimpanzee, gorilla, orangutan, and macaque.

Sequence Similarity Searches

We performed a BLASTp search of the human proteins against the merged primate protein data set using an *e*-value threshold of 1×10^{-4}. These results formed the basis for the list of initial candidate genes.

We used tBLASTn to search the protein sequences of interest against the genomes of up to five outgroup genomes (chimpanzee, gorilla, orangutan, gibbon, and macaque). Candidate human-specific de novo genes were searched against all five, and human + chimpanzee genes were searched against the other four, etc. We only considered cases with tBLASTn hits with sequence identity (SI) > 60% and coverage > 0.4 (length of the hit/length of the human protein).

We excluded cases where we could not detect the orthologous sequence in the outgroup genomes, with the exception that we retained cases where the orthologous DNA was unidentifiable in only one of gibbon or macaque. We discarded cases where more than one possible homologous sequence was found in one or more outgroups or where the human protein had highly similar copies (SI > 90%) in the human genome itself.

Examination of Outgroup Sequence Coding Potential and Inference of Ancestral State

We examined the conceptual translation of the orthologous DNA sequence from outgroup genomes with particular attention paid to frameshifts and premature stop codons.

Multiple nucleotide sequence alignments were constructed of the candidate de novo genes and the orthologous outgroup DNA using MUSCLE (Edgar 2004). These were examined for the presence of a stop codon or frameshift located in the first 60% of the alignment and shared the outgroups. In the case where either gibbon or macaque did not share the disabler, the gene was still considered de novo if the ORF in that genome was interrupted by other disablements.

Quality Controls

All candidate genes were compared with both the GenBank nonredundant data set and RefSeq using BLASTp and none had any additional hits. We also confirmed that none of the candidates had any plausible EnsEMBL annotated orthologs (i.e., without unlikely small introns and whose protein product shared both SI and coverage over 40% with the human protein).

We examined the synteny conservation around the candidate de novo genes. We could not carry out this step for Gibbon due to the poor organization of its currently available genome assembly. The number of neighboring genes selected varied depending on the gene density thus we chose a minimum of 4 to maximum of 12 neighboring genes both upstream and downstream. We searched for their orthologs in each genome (human, chimpanzee, gorilla, and macaque) and retained candidates if we could confidently identify orthologs of at least two upstream and two downstream neighbors. We observed strong synteny conservation for all but two cases, which were excluded from further analysis.

Supporting Evidence

Transcription evidence was obtained from Unigene (Wheeler et al. 2003) through crosslink provided by EnsEMBL (Flicek et al. 2012). Short sequenced peptides were obtained from PRIDE (Vizcaíno et al. 2013), PeptideAtlas (Deutsch et al. 2008), and gpmDB (Craig et al. 2004).

RNAseq Data and Analysis

We obtained RNAseq data from European Nucleotide Archive (Leinonen et al. 2011). We downloaded unaligned reads of human, chimpanzee, and gorilla from a single study (SRP007412). The data were mapped against the respective genomes using Tophat2 (Kim et al. 2013). The alignment files produced were visualized and analyzed using IGV (Thorvaldsdóttir et al. 2013) to reveal the presence of intron-spanning in outgroups.

Results and Discussion

Detection of De Novo Genes in Primate Genomes

There is quite wide variability in the estimates of de novo genes due to different approaches to their detection (McLysaght and Guerzoni 2015). We take the view that it is important to adopt a conservative methodology. Permissive methods are susceptible to classifying any genes with difficult-to-detect-homologs as de novo genes. In particular, it is incorrect to infer that failure to detect a BLAST hit in a given lineage is evidence of the absence of the gene, because such a situation frequently arises with short and quickly evolving genes which can easily be mistaken for young genes (Moyers and Zhang 2015, 2016). The method we use here builds on the approach of Knowles and McLysaght (2009) where initially plausible de novo genes are examined for evidence of the absence of the gene in the ancestor, as well as for supporting evidence. This approach requires the detection of the orthologous DNA sequence in the outgroup lineage, otherwise the gene is excluded as ambiguous. In order for a gene to be considered novel, the orthologous DNA must be identifiable and must be shown to lack coding capacity (i.e., to lack an intact ORF). This avoids the problem of misattributing failure to detect a BLAST hit as evidence of gene novelty, as in all cases we require a BLAST hit. However, it is worth noting that all such studies are subject to fluctuations in the databases, with poorly characterized de novo genes perhaps more susceptible than others to being excluded from database updates, often without much explanation.

We compared the complete human proteome with that of chimpanzee, gorilla, orangutan, and macaque using BLAST. Candidate de novo genes were those where none of the potential proteins of the gene had hits in orangutan or macaque. These were classified as human-specific (H), human + chimpanzee specific (HC), or human + chimpanzee + gorilla specific (HCG) depending on the presence of BLAST hits in those genomes. This resulted in 734 candidate de novo genes from EnsEMBL v60 and an additional 67 genes from subsequent EnsEMBL versions (v61–69).

For tractability reasons, genes with more than one coding exon were excluded. This is because, in multiple-coding-exon genes, during the assessment of outgroup genomes it is difficult to distinguish the absence of coding potential due to frameshifts and stop codons (which supports the inference of de novo origins) from the alternative explanation of evolutionary change of intron–exon boundaries (which does not). Seeing as intron–exon boundary changes can be invoked to accommodate any stop codon or frameshift, more direct evidence of gene structure (such as RNAseq data) is required from all lineages under investigation (ingroups and outgroups). In the absence of sufficient depth of such data, we restricted our search to include only uninterrupted ORFs. It is unlikely that this exclusion will have a large impact on the results as most de novo genes are initially structurally simple (Knowles and McLysaght 2009; Siepel 2009; Abrusán 2013; Zhao et al. 2014).

In order to unambiguously show that a given gene has arisen de novo it is necessary to demonstrate that the ancestral sequence was noncoding. We used tBLASTn to search for the

orthologous DNA in outgroup primate genomes. The outgroup orthologous DNA was identifiable for 233 genes.

This orthologous DNA was then examined to determine whether it was potentially coding or not. Any cases that are potentially coding in an outgroup are no longer considered plausible recent de novo genes (i.e., having originated in the ape lineage after the divergence of orangutan). Only cases where the primate outgroup genomes had no potential ORF longer than 60% of the length of the human ORF were considered. Furthermore, in order to exclude the alternative hypothesis of independent gene loss/inactivation in the outgroups, we also required that there was shared disabler (premature stop codon, or frameshift causing a premature stop codon) in the primate outgroup genomes. This analysis reduced the number of candidates to 37 genes. We further confirmed that none of these has a BLASTp hit in any other genomes. Two of these candidate genes were in regions of poor synteny conservation, and we could not exclude the possibility that they were created as the result of genome rearrangements, which is not the phenomenon of interest here, so they were excluded from further analysis.

There is only a small amount of polymorphism data for nonhuman primates, with data for a small number of unrelated individuals available for chimpanzee, gorilla, orangutan, and macaque (Gokcumen et al. 2013; Scally et al. 2013). Nonetheless, we examined these data and found no polymorphism at the disabler site.

Ideally, candidate de novo genes should be supported by transcription and translation evidence. However, these data are volatile and are themselves usually dependent on the genome annotation being present first. Nonetheless, we searched Unigene for evidence of transcription, and three peptide databases (PRIDE, PeptideAtlas, and gpmDB) for evidence of translation. All but 5 of the 35 de novo genes had transcription or translation evidence (table 1). Fifteen genes were supported by both transcripts and short peptides. In all cases the available data are mainly from human, so even for candidate genes shared with other apes we could only search for supporting evidence of activity in human. We compared our results with those of Ruiz-Orera et al. (2015) who searched for novel genes based on transcriptome sequencing of human, chimpanzee, macaque, and mouse and we found no overlap in the lists. However, Ruiz-Orera et al. filtered out all intronless genes, which automatically excludes almost half of our cases; and only 8 of the 2,714 human- and/or chimpanzee-specific genes initially identified by them were annotated as protein-coding and only 20 had some evidence of translation, further limiting the opportunity for overlap in the two approaches.

The fluctuations in the genome annotations and supporting data are easily apparent. These are changes in the database status of the gene that reflect annotation uncertainty, but of course the true biological status does not change. One case (ENSG00000187461) was initially associated with a Unigene cluster that has since been retired from the database leaving this gene somewhat paradoxically with translation but not transcription evidence. ENSG00000196273 was identified in version 60 and had no Unigene cluster. This gene remains annotated in version 70 and is currently associated with two Unigene clusters. Moreover, it still retains associated translational evidence even though EnsEMBL v70 classifies it as a lincRNA gene.

Similarly, we can consider the stability of the gene annotation in the EnsEMBL database. We found that 22 genes of 35 are still annotated in EnsEMBL v70. However, six of these are no longer classified as "protein-coding" and one has new gene identifier (ENSG00000255766 became ENSG00000259498). Thus, it is clear that even with a conservative approach the results obtained will depend on external factors, particularly database changes.

The 35 de novo genes include 16 human-specific, 5 human + chimpanzee specific, and 14 human + chimpanzee + gorilla specific genes (table 1). Consistent with previous studies the de novo genes identified here code for short proteins (155 ± 53 amino acids). Most of the genes are uncharacterized. *GR6* (ENSG00000198685) is the only gene that has been previously studied; it is normally expressed during fetal development but ectopic expression has been observed in some cancers (Pekarsky et al. 1997).

The approximate rate of de novo gene origin can be calculated as the number of events per million years. We observe an average rate of 2.12 de novo gene origins per million years. For the different branches of the tree we obtain approximate rates of 2.42 genes per million years for the H set, a rate of 2.94 genes per million years for the HC set and a rate of 1.71 genes per million years for the HCG set, which are not significantly different from each other (chi-square test). This differs from origin of new genes by duplication where more recent branches have a proportionately larger number of gain events (Lynch and Conery 2000). However, even if such a pattern were true for de novo genes it would be difficult to observe considering the small numbers of events.

Evidence for ILS of De Novo Genes

Our de novo gene detection protocol requires that the enabling sequence difference that establishes the ORF of interest is monophyletic (i.e., found only in humans, only in humans and chimpanzees, or only in humans, chimpanzees and gorillas). This is a pragmatic criterion intended to maximize the reliability of the reported de novo genes, but it may exclude some biologically interesting cases. In particular, ILS describes a scenario where a polymorphism present in an ancestral species survives past two speciation events after which it may become differentially fixed. One outcome of ILS is that for some loci the genetic relatedness is different from the species relatedness. ILS is a well-documented phenomenon in ape genomes where it is responsible for about 15% of the

Table 1

De Novo Genes that Originated Recently in the Primate Lineage

Gene Name	EnsEMBL ID (Human)	Lineages	Exons	Length (aa)	Alu Elements Found within Exons	Overlap with Other Genes	Transcriptional Evidence[a]	Peptide Evidence[b]
AC012366.1	ENSG00000226452	H	1	65	Yes—overlapping the CDS	Opposite strand overlap	Hs.617350	No
AL079342.1	ENSG00000203863	H	1	144	Yes—overlapping the CDS	No	Hs.640013	gpmDB, PRIDE
AP002380.2	ENSG00000214780	H	2	195	Yes—In UTR regions	Same strand overlap	Hs.676126	No
AC125494.1	ENSG00000219410[c]	H	4	139	No	Opposite strand overlap	Hs.714839	gpmDB
DNAH10OS (RP11-380L11.1)	ENSG00000250091	H	2	163	Yes—In UTR regions	Opposite strand overlap	Hs.548335*, Hs.679261, Hs.728379	gpmDB, PRIDE
AC016251.1	ENSG00000205148	H	1	126	No	No	Hs.58690	PRIDE
C14orf70	ENSG00000196273[c]	H	2	105	No	No	Hs.379802, Hs.662255	gpmDB, PRIDE
AC011497.1	ENSG00000213904[c]	H	5	138	No	Opposite strand overlap	Hs.600453*, Hs.624933	No
RP11-429E11.3	ENSG00000179253	H	2	140	Yes—In UTR regions	Opposite strand overlap	Hs.683806	gpmDB, PRIDE
C18orf56	ENSG00000176912	H	2	123	No	Opposite strand overlap	No	gpmDB, PRIDE
AL353698.1	ENSG00000233889	H	1	75	No	No	Hs.573631	No
AC005262.1	ENSG00000255869	H	1	140	No	No	Hs.654784	No
AP001468.2	ENSG00000256842	H	2	158	No	Same strand overlap	Hs.721335	No
AL022578.1	ENSG00000256707	H	1	243	No	No	Hs.496083	No
RP11-326A13.2	ENSG00000258961	H	1	181	No	No	Hs.531264	PRIDE
AC132186.1	ENSG00000247270	H	1	201	No	Opposite strand overlap	Hs.730232*, Hs.97805	PRIDE
GR6 (C3orf27)	ENSG00000198685	HC	3	149	Yes—In UTR regions	No	Hs.194283	PRIDE
TMEM133	ENSG00000170647	HC	1	129	No	No	Hs.44004	PRIDE
AC007608.1	ENSG00000205414[c]	HC	2	140	Yes—In UTR regions	Opposite strand overlap	Hs.689579	gpmDB
AL358252.1	ENSG00000256831	HC	2	170	No	Same strand overlap	No	No
AC079328.1	ENSG00000255766[c,d]	HC	1	266	No	Opposite strand overlap	Hs.602995, Hs.712217	No
AC011239.1	ENSG00000216839	HCG	1	153	No	Same strand overlap	No	No
C6orf114	ENSG00000187461	HCG	2	136	No	Same strand overlap	Hs.674313*	No
AL132661.1	ENSG00000176424	HCG	1	234	No	Opposite strand overlap	Hs.708964	gpmDB, PRIDE
AC124781.1	ENSG00000227273	HCG	1	117	No	Same strand overlap	No	No
AC005071.3	ENSG00000229429	HCG	2	158	No	Both strands overlap	No	No
C10orf111	ENSG00000176236	HCG	2	155	Yes—In UTR regions	Opposite strand overlap	No	gpmDB, PRIDE
AL589787.1	ENSG00000203779	HCG	7	152	No	Opposite strand overlap	Hs.646701	No
AP000679.1	ENSG00000176984	HCG	2	323	Yes—In UTR regions	No	Hs.638417	gpmDB
KRTAP20-4	ENSG00000206105	HCG	1	44	No	No	Hs.580879	gpmDB, PRIDE
AL360294.1	ENSG00000255646	HCG	1	182	Yes—In UTR regions	No	No	No
AC073439.1	ENSG00000256345	HCG	1	181	No	Opposite strand overlap	Hs.610961	No
C11orf39	ENSG00000255953[c]	HCG	2	140	Yes—In UTR regions	Opposite strand overlap	Hs.730330, Hs.730455*	No
AL844908.1	ENSG00000257100	HCG	2	163	No	Same strand overlap	Hs.534504	No
RP11-1127D7.1	ENSG00000259119	HCG	2	114	No	No	Hs.631462	gpmDB
AP001052.1	ENSG00000256247	ILS (H+G)	2	162	Yes—on junction	Same strand overlap	No	gpmDB

[a]Transcriptional evidence by displaying the identifier of associated Unigene clusters. Currently retired clusters are marked with *.

[b]Peptide evidence is shown by displaying the name of the repository in which it can be currently found.

[c]The gene is still annotated and has the same exonic structure, but it is not considered as protein coding in e70.

[d]e70 gene ID is ENSG00000259498.

human genome being more similar to that of gorilla than of chimpanzee (Rogers and Gibbs 2014). ILS is of particular interest in the case of de novo genes because it provides indirect evidence of the population genetics dynamics of these genes. Observing ILS at genome-typical rates supports the inference of neutral evolution of these loci (Dutheil et al. 2015).

There were 322 (289 from EnsEMBL v60 and 33 from v61–v69) human proteins that had BLASTp hits in gorilla but not in chimpanzee and other primates. Similarly to the above analysis, for these to be plausible cases of ILS the enabling sequence difference should be shared by human and gorilla and the disabler should be shared by the other primates (including chimpanzee) so that we can reliably infer that the gene is de novo and that it has not been subject to differential gene loss.

In order to check for these cases, we used the same genome and proteome data but selected human proteins that had BLAST hits only in gorilla (either to annotated proteins or to an unannotated ORF of similar length). We also carried out the complementary search (gorilla proteins that only have plausible hits in the human genome).

We mapped these against chimpanzee, orangutan, gibbon, and macaque to identify the orthologous DNA. We searched all of these cases for evidence of an intact ORF in other primates (start and stop codon present and the predicted ORF at least 60% of the length of the human or gorilla ORF). For those with no intact ORF in the other primates, we also searched for a disabler shared by chimpanzee, orangutan, gibbon, and macaque as before. For three cases (ENSG00000256247, ENSG00000256109, and ENSGGOG000028018), we could identify a disabler shared across chimpanzee and the outgroups.

We mapped these three genes against the Bonobo (*Pan paniscus*) genome (Prüfer et al. 2012) which diverged from the common chimpanzee about 2 Ma. In two of the three cases the Bonobo sequence shared the disabler with chimpanzee, as expected given their close relationship. In one case (ENSGGOG000028018), an ORF-enabling mutation was found in Bonobo exactly like the one observed in human and gorilla.

However, in this case and in one other (ENSG00000256109) the enabling difference is a 1-bp insertion found within a small repetitive region (6–7 identical base pairs). In such cases, independent mutation in two lineages or, alternatively, sequencing errors cannot be confidently excluded as the explanation for the pattern of sequence similarity.

For the third case (ENSG00000256247) the situation is different. First, the DNA sequence does not have such low complexity: The enabling mutation is a single base-pair insertion that does not occur within a string of identical base pairs and we thus infer no increased probability of sequencing errors or independent mutation at this locus. Additionally, although the DNA sequence around the enabler/disabler site is generally well conserved across primates, human and gorilla uniquely share the sequence "GTG" where all other species have "CCC." This three base-pair shared difference is located very shortly downstream of the enabler (fig. 1). The concordance of the presence of the enabler mutation and other sequence differences supports the inference that this mutation arose in an individual in the common ancestor species of human, chimpanzee and gorilla, possibly on the "GTG" allele. This locus remained polymorphic for the presence/absence of the new ORF through two speciation events, before it was eventually fixed for the presence in human and gorilla and for the absence in chimpanzee. This prolonged period of polymorphism suggests that this de novo gene had a neutral effect on fitness at least until after the human–chimpanzee speciation.

Evidence for De Novo Origin of Protein-Coding Genes in Noncoding RNA Loci

Even though we selected candidate de novo genes with monoexonic ORFs, only 16 of the 35 genes are actually monoexonic. The remaining 19 have at least one intron in an untranslated region (UTR). DNA sequence conservation of the genes (including the splice sites) across all outgroups is high (supplementary table S1, Supplementary Material online). The conserved splice sites may be cryptic or may be part of an ancestral transcribed multiexonic noncoding locus.

The gene ENSG00000176912 is of particular interest because it has an 8-kb intron. The protein-coding gene appears to be human-specific and the annotation is stable across the EnsEMBL versions investigated here. The sequence of both exons and of the splice junctions is well conserved in chimpanzee and gorilla where the orthologous region is noncoding. The validity of the human intron is supported by uniquely mapping RNAseq reads overlapping the exons including over a dozen intron-spanning reads.

We tested whether the conserved splice sites sequence in chimpanzee and gorilla are actual splice locations by searching for RNAseq data from those genomes that span the intron location. Even though the number of reads from those genomes is much smaller, we found uniquely mapped reads on the exon homologs and some intron-spanning reads. The read coverage is low, which may reflect the threadbare nature of the database, or could be spurious transcription. Nonetheless, these data indicate that at least the ancestral sequence already carried splice signals, be they active or cryptic.

The fact that most genes have conserved splice sites in chimpanzee and gorilla and that for at least one of them we have evidence of transcription and splicing taking place in lineages that do not have the ORF provides examples of the "RNA-first" model of de novo gene origination. In the "RNA-first" model the ORF arises at a transcribed locus, as opposed to "ORF-first" where a locus containing an ORF becomes transcribed (McLysaght and Guerzoni 2015). The RNA-first model provides a simple explanation for splicing signals that

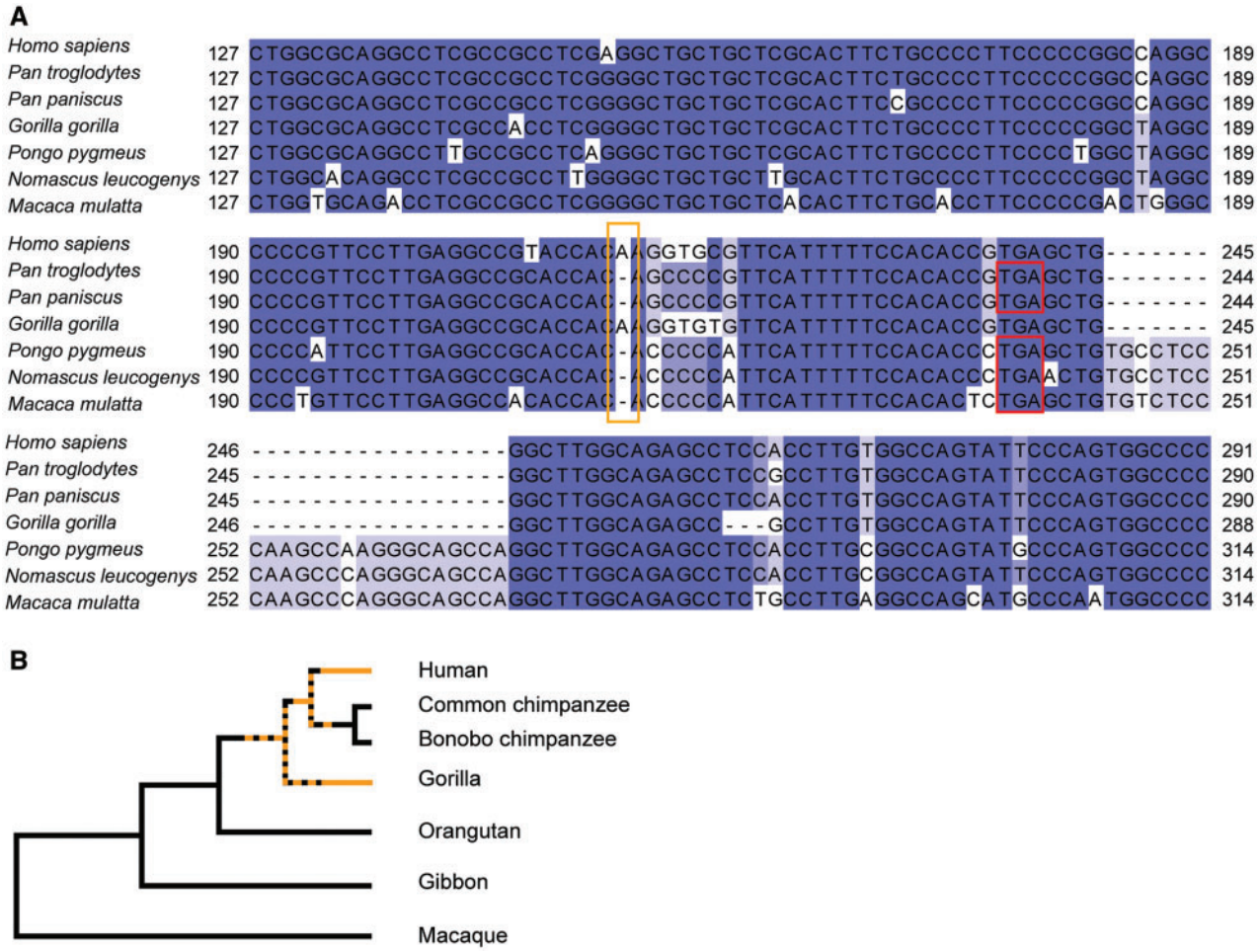

Fig. 1.—ILS of a de novo gene. (*A*) Segment of alignment of the de novo gene *ENSG00000256247* with the orthologous region from other primates. The ORF is present only in human and gorilla. The ORF was created by a single base-pair insertion uniquely found in human and gorilla (indicated by an orange box). This frameshift means that the TGA stop codon (boxed in red) is no longer in frame in human and gorilla. These two species also uniquely share a three base-pair difference (GTG vs. CCC) very close to the insertion site. The start and stop codons in human and gorilla are not pictured in this segment. Numbers at the side of the alignment indicate base-pair positions starting from the human start codon. (*B*) Inferred evolutionary history of this de novo gene: The one base-pair insertion occurred in the ancestor of the great apes. The substitutions resulting in the downstream "GTG" were either already present in that individual, or occurred later in an individual carrying the insertion. The ORF thus created remained polymorphic (indicated by the dashed orange lines) until after the human–chimpanzee divergence. Subsequent independent lineage sorting saw the fixation of the original locus lacking the gene (black) in the chimpanzee lineage and the de novo gene (orange) was independently fixed in human and gorilla. Alignment visualized using JalView (Waterhouse et al. 2009). Species Latin names are shown in the alignment and the corresponding common names are shown in the phylogenetic tree.

predate the acquisition of the ORFs (Li et al. 2010; Yang and Huang 2011) possibly due to functional RNAs at the locus (Xie et al. 2012).

Features Neighboring De Novo Genes

Being complex entities, genes require more than the presence of an ORF to be transcribed and translated. Lineage-specific genes have a tendency be relatively close or to overlap existing genes (Makałowska et al. 2007), and this remains true of de novo genes. This is particularly interesting because it provides a route for a de novo gene to acquire regulated transcriptional activity relatively easily (Siepel 2009; Gotea et al. 2013). Twenty-two of 35 candidates overlap with at least one other annotated gene of which 14 overlap with genes on the opposite strand (including one de novo gene that has overlapping genes on both strands; table 1). Same strand overlaps are always either in alternative reading frames or in noncoding regions of the gene.

Transposable elements, in particular Alus and other Short Interspersed Elements, may contribute to de novo gene origin (Toll-Riera et al. 2009), including by providing start codons or by catalyzing RNA-editing (Schmitz and Brosius 2011). Almost half of the de novo genes described here (18 of 35) have Alu elements embedded in their gene structure; however in most cases, these were located within introns or in UTRs (table 1).

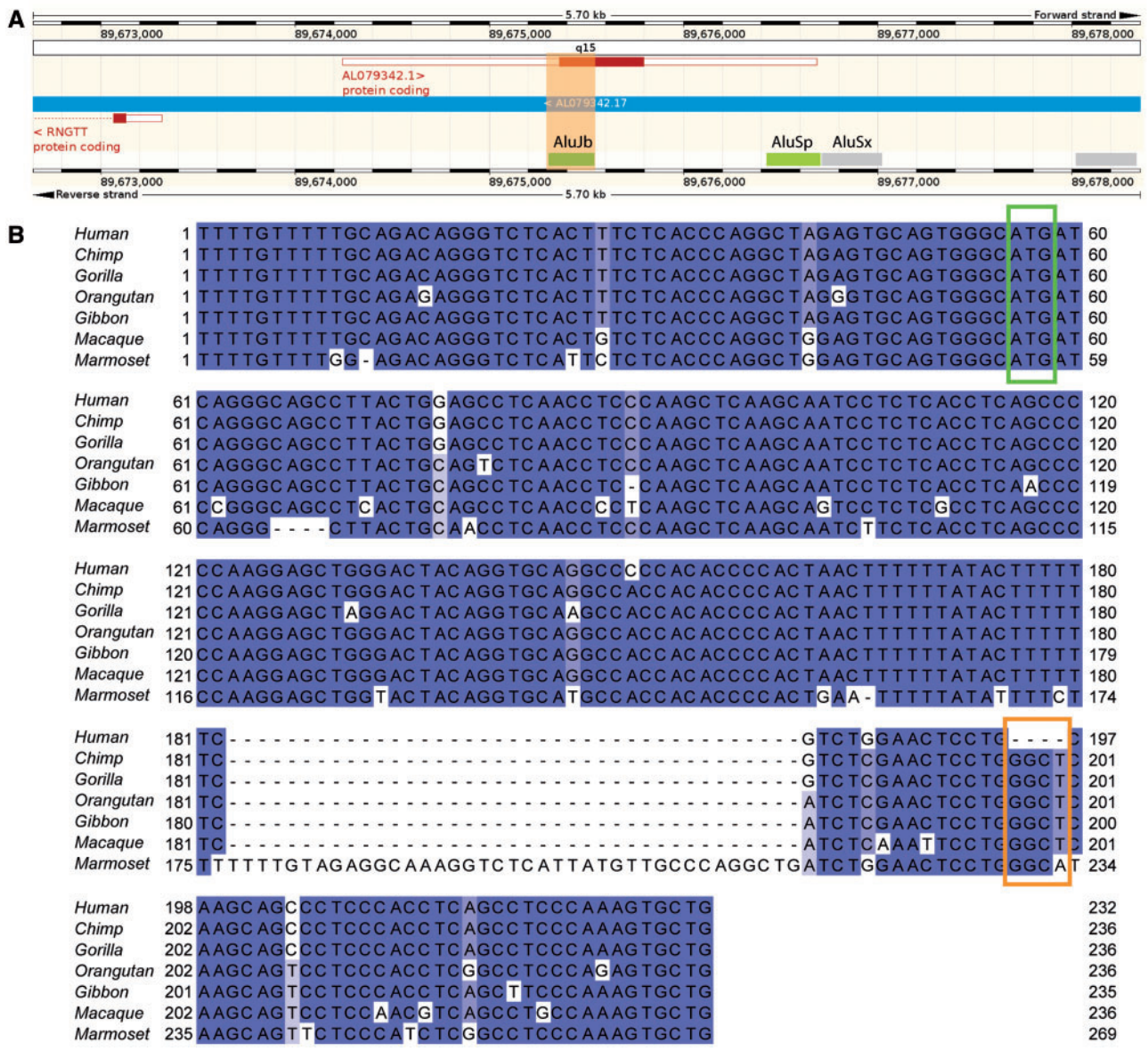

Fig. 2.—Alu elements and the de novo gene origins. The de novo gene AL079342 (ensembl ID ENSG00000203863) is overlapping with two Alu elements. (*A*) Schematic of the region on chromosome 6 that includes ENSG00000203863 (coding sequence shown in red). Two Alu elements (shaded green) overlap the gene sequence. The area shaded orange is shown in detail in part (*B*) of the figure. (*B*) Multiple sequence alignment of the orthologous region in several primates. AluJb provides the start codon for the ORF in human and is present cryptically in all other species examined (boxed in green). A human-specific frameshift is caused by the deletion of four bases (boxed in orange). The human ORF continues beyond the alignment segment shown.

Two genes have Alu elements overlapping their coding sequence. Gene ENSG00000226452 has an Alu element (AluSx) overlapping both the 5′-UTR and the beginning of its coding sequence, thus including the start codon. Similarly, gene ENSG00000203863 has an AluJb element overlapping the start of the coding sequence (fig. 2). The human ORF is 144 codons long whereas the longest possible ORF is much shorter in other primates ranging from 75 to 76 codons long (in order to reach the 60% threshold it should be 85 codons long). However, the timing of the Alu insertion and the ORF origin do not coincide because the AluJb elements were active around 87–90 Ma (Schmitz and Brosius 2011) whereas the 4-bp deletion that creates the long ORF is human-specific. Thus, the presence of the transposable element at this locus predates the gene. In fact, we can find this element in all of the considered outgroups and their sequences all cryptically possess the "ATG" base pairs that would become the start codon of the human-specific de novo gene.

Human Polymorphism

We searched the 1000 genomes data (1000 Genomes Project Consortium et al. 2012) for evidence of polymorphism within the ORF of these de novo genes (supplementary table S2, Supplementary Material online). In all 36 cases (including the ILS gene), the enabler mutation is not polymorphic, suggesting that the genes are fixed in human populations. We found a total of 256 variants and only 49 of these have observed total frequencies of 5% or greater. The vast majority of these variants are either silent or nonsynonymous (respectively, 73 and 167 single nucleotide polymorphisms [SNPs]). There is a small number of polymorphisms that disrupt the ORF, either nonsense SNPs (seven cases) or indels (nine cases) found in 12 genes. Only three of these cases (all of which are indels) have frequencies $\geq$0.05 and are in genes ENSG00000226452, ENSG00000256707, and ENSG00000255766.

Of the 12 genes with disruptive variants within the ORF, the majority (eight) are human-specific genes while the older ORFs of the HC and HCG sets include three and one disruptive variant, respectively.

For six of the de novo genes, we found (presumed healthy) individuals who were homozygous for ORF-disrupting alleles. In four cases, there were only one or two homozygous individuals out of the 1,089 examined. On the other hand for both ENSG00000255766 and ENSG00000226452, we observed a relatively higher number of homozygous individuals for the ORF-disrupting allele (respectively, 52 and 67 of 1,089) indicating that these genes are neither fixed nor essential in human.

Denisovan hominins diverged from anatomically modern humans about 800,000 years ago and the genome has been sequenced and assembled to high quality (Meyer et al. 2012). We examined the Denisova assembly in the UCSC genome browser (Kent et al. 2002) considering only those differences identified by multiple reads. We examined the regions orthologous to the 36 de novo gene ORFs. We identified 20 differences compared with the human reference assembly corresponding to regions orthologous to 15 of the human de novo genes. Eighteen of 20 differences are present as alleles within the human population and none of these is ORF-disabling. Only one nonsense substitution was observed and that is present in the region orthologous to the ORF of ENSG00000256831.

Concluding Remarks

We report a set of conservatively defined de novo genes that originated recently in the great ape lineage. Among these we identified 16 human-specific de novo genes, which is very close to a previous estimation of 18 such cases based on a similar methodology (Knowles and McLysaght 2009). However, of the three genes identified in that older study, only one (*DNAH10OS*) appears in this new list because the others have been excluded from the databases. Nonetheless, it is possible to say that the overall trend in terms of frequency of events is stable under similarly conservative search criteria. Not surprisingly, studies that employed more lenient search criteria also reported larger numbers of genes (Wu et al. 2011).

Aside from the low numbers of events, other features that are consistent across multiple studies of de novo genes are the initial simplicity of the genes and the recycling of pre-existing components or features of the genome (Carvunis et al. 2012; Abrusán 2013; Palmieri et al. 2014).

One interesting question concerns the dynamics of fixation of de novo genes. Here, we report the first case of independent lineage sorting of a de novo originated gene. This de novo gene originated prior to the gorilla divergence and remained polymorphic until after the chimpanzee–human divergence: A period of 3–4 Myr (Perelman et al. 2011). This extended period of polymorphism indicates a very slow pace of fixation where drift rather than selection is responsible (Dutheil et al. 2015).

How de novo genes become functional, and sometimes even essential, remains mysterious. It will be very interesting to explore the evolutionary dynamics that allow a new gene to integrate into a pre-existing and central processes. De novo genes are a potentially important contributor to evolutionary innovation. In some rare cases their functionality, and even essentiality, has been documented, but in general these genes and the mechanisms surrounding their establishment are poorly understood.

Acknowledgments

This work was supported by a Science Foundation Ireland research grant. The research leading to these results has received funding from the ERC under the European Union's Seventh Framework Programme (FP7/2007–2013)/ERC Grant Agreement 309834. The authors are grateful to all members of the McLysaght research group for valuable discussion.

Supplementary Material

Supplementary tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Literature Cited

1000 Genomes Project Consortium, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65.

Abrusán G. 2013. Integration of new genes into cellular networks, and their structural maturation. Genetics 195(4):1407–1417.

Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007. Evidence for de novo evolution of testis-expressed genes in the drosophila yakuba/drosophila erecta clade. Genetics 176(2):1131–1137.

Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. Genetics 172(3):1675–1681.

Carvunis A-R, et al. 2012. Proto-Genes and de novo gene birth. Nature 487(7407):370–374.
Chen S-T, Cheng H-C, Barbash DA, Yang H-P. 2007. Evolution of hydra, a recently evolved testis-expressed gene with nine alternative first exons in drosophila melanogaster. PLoS Genet. 3(7):e107.
Craig R, Cortens JP, Beavis RC. 2004. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res. 3(6):1234–1242.
Deutsch EW, Lam H, Aebersold R. 2008. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep. 9(5):429–434.
Deutsch M, Long M. 1999. Intron-Exon structures of eukaryotic model organisms. Edited by Chris P Ponting. Nucleic Acids Res. 27(15):3219–3228.
Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. 2011. Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. BMC Evol Biol. 11(1):47.
Dutheil JY, Munch K, Nam K, Mailund T, Schierup MH. 2015. Strong selective sweeps on the X chromosome in the human-chimpanzee ancestor explain its low divergence. Edited by Nick H Barton. PLoS Genet. 11(8):e1005451.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput.—PubMed—NCBI. Nucleic Acids Res. 32(5):1792–1797.
ENCODE Project Consortium 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74.
Flicek P, et al. 2012. Ensembl 2012. Nucleic Acids Res. 40(Database issue):D84–D90.
Gilson PR, McFadden GI. 1996. The miniaturized nuclear genome of eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns. Proc Natl Acad Sci U S A. 93(15):7737–7742.
Gokcumen O, et al. 2013. Primate genome architecture influences structural variation mechanisms and functional consequences. Proc Natl Acad Sci U S A. 110 (39):15764–15769.
Gotea V, Petrykowska HM, Elnitski L. 2013. Bidirectional promoters as important drivers for the emergence of species-specific transcripts. PLoS One 8(2):e57323.
Guerzoni D, McLysaght A. 2011. De novo origins of human genes. Edited by David J Begun. PLoS Genet. 7(11):e1002381.
Kent WJ, et al. 2002. The human genome browser at UCSC. Genome Res. 12(6):996–1006.
Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 25(9):404–413.
Kim D, et al. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14(4):R36.
Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. Genome Res. 19(10):1752–1759.
Leinonen R, et al. 2011. The European nucleotide archive. Nucleic Acids Res. 39(Database issue):D28–D31.
Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci U S A. 103(26):9935–9939.
Li C-Y, et al. 2010. A human-specific de novo protein-coding gene associated with human brain functions. PLoS Comput Biol. 6(3):e1000734.
Li D, et al. 2010. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Res. 20(4):408–420.
Li L, et al. 2009. Identification of the novel protein QQS as a component of the starch metabolic network in *Arabidopsis* leaves. Plant J. 58(3):485–498.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290(5494):1151–1155.
Makałowska I, Lin C-F, Hernandez K. 2007. Birth and death of gene overlaps in vertebrates. BMC Evol Biol. 7:193.
McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philos Trans R Soc B Biol Sci. 370(1678):20140332.
Meyer M, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338(6104):222–226.
Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. Mol Biol Evol. 32(1):258–267.
Moyers BA, Zhang J. 2016. Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. Mol Biol Evol. Advance Access published January 11, 2016; doi:10.1093/molbev/msw008.
Murphy DN, McLysaght A. 2012. De novo origin of protein-coding genes in murine rodents. PLoS One 7(11):e48650.
Palmieri N, Kosiol C, Schlötterer C, Tautz D. 2014. The life cycle of drosophila orphan genes. eLife 3(0):e01311–e01311.
Pekarsky Y, Rynditch A, Wieser R, Fonatsch C, Gardiner K. 1997. Activation of a novel gene in 3q21 and identification of intergenic fusion transcripts with ecotropic viral insertion site I in leukemia. Cancer Res. 57(18):3914–3919.
Perelman P, et al. 2011. A molecular phylogeny of living primates. Edited by Jürgen Brosius. PLoS Genet. 7(3):e1001342.
Prüfer K, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature 486(7404):527–531.
Reinhardt JA, et al. 2013. De novo ORFs in drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet. 9(10):e1003860.
Rogers J, Gibbs RA. 2014. Comparative primate genomics: emerging patterns of genome content and dynamics. Nat Rev Genet. 15(5):347–359.
Ruiz-Orera J, et al. 2015. Origins of de novo genes in human and chimpanzee. Edited by James Noonan. PLoS Genet. 11(12):e1005721.
Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. 2014. Long non-coding RNAs as a source of new peptides. eLife 3:e03523.
Samusik N, Krukovskaya L, Meln I, Shilov E, Kozlov AP. 2013. PBOV1 is a human de novo gene with tumor-specific expression that is associated with a positive clinical outcome of cancer. Edited by Ludmila Prokunina-Olsson. PLoS One 8(2):e56162.
Scally A, et al. 2013. A genome-wide survey of genetic variation in gorillas using reduced representation sequencing. Edited by John Hawks. PLoS One 8(6):e65066.
Schmitz J, Brosius J. 2011. Exonization of transposed elements: a challenge and opportunity for evolution. Biochimie. 93(11):1928–1934.
Siepel A. 2009. Darwinian alchemy: human genes from noncoding DNA. Genome Res. 19(10):1693–1695.
Suenaga Y, et al. 2014. 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(1):e1003996.
Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 14(2):178–192.
Toll-Riera M, et al. 2009. Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol. 26(3):603–612.
Vizcaíno JA, et al. 2013. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. 41(Database issue):D1063–D1069.
Wang J, et al. 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature 516(7531):405–409.
Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189–1191.

Wheeler DL, et al. 2003. Database resources of the national center for biotechnology. Edited by Chris P Ponting. Nucleic Acids Res. 31(1):28–33.
Wu D-D, Irwin DM, Zhang Y-P. 2011. De novo origin of human protein-coding genes. Edited by David J Begun. PLoS Genet. 7(11):e1002379.
Xiao W, et al. 2009. A rice gene of de novo origin negatively regulates pathogen-induced defense response. Edited by Hany A El-Shemy. PLoS One 4(2):e4603.
Xie C, et al. 2012. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. PLoS Genet. 8(9):e1002942.
Yang Z, Huang J. 2011. De novo origin of new genes with introns in *Plasmodium vivax*. FEBS Lett. 585(4):641–644.
Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in drosophila melanogaster populations. Science 1248286.

Associate editor: George Zhang