New genes expressed in human brains: Implications for annotating evolving genomes

*Yong E. Zhang[1)], Patrick Landback[2)], Maria Vibranovski[3)] and Manyuan Long[2)3)]**

New genes have frequently formed and spread to fixation in a wide variety of organisms, constituting abundant sets of lineage-specific genes. It was recently reported that an excess of primate-specific and human-specific genes were upregulated in the brains of fetuses and infants, and especially in the prefrontal cortex, which is involved in cognition. These findings reveal the prevalent addition of new genetic components to the transcriptome of the human brain. More generally, these findings suggest that genomes are continually evolving in both sequence and content, eroding the conservation endowed by common ancestry. Despite increasing recognition of the importance of new genes, we highlight here that these genes are still seriously under-characterized in functional studies and that new gene annotation is inconsistent in current practice. We propose an integrative approach to annotate new genes, taking advantage of functional and evolutionary genomic methods. We finally discuss how the refinement of new gene annotation will be important for the detection of evolutionary forces governing new gene origination.

Keywords:
brain evolution; gene annotation; new gene origination; pseudogene identification; ribosome profiling

Introduction

The term ''new gene'' [1, 2] refers to a novel genetic locus, a physically distinct and derived segment of DNA that encodes a new functional transcript(s). The young genes are new genes, which have arisen recently and thus exist in one or a few lineages. They are valuable in investigating the early stages of new gene evolution, as the sequence and structural signatures of their origins have not yet been obscured by subsequent mutations. The field has focused mainly on new protein-coding genes, although a few new non-coding genes have also been described (e.g. [3, 4]). Since the first discovery of a young gene, *jingwei*, two decades ago, the knowledge of the origination mechanisms and functional roles of new genes has been quickly accumulating.

Numerous works have shown that the predominant mechanism of new gene origination is duplication, including DNA-level duplication and retroposition (as reviewed in [1, 5]). In addition, genes without any parental gene are de novo genes, which originate from non-coding DNAs and may contribute up to 10% of all new genes [6, 7]. Compared to earlier work that often focused on the origination mechanisms of new genes (e.g. [8, 9]), the field is moving toward surveying the functional roles, phenotypic effects, and evolution of new genes. As expected, new genes are often involved in biological processes under tremendous selective pressure such as male reproduction or stress response (as reviewed in [1]). However, a few recent reports show counter-intuitive evidence that new genes quickly assume critical roles in the brain and developmental

DOI 10.1002/bies.201200008

[1)] Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, P.R. China
[2)] Committee on Genetics, Genomics, and Systems Biology, The University of Chicago, Chicago, IL, USA
[3)] Department of Ecology and Evolution, The University of Chicago, Chicago, IL, USA

***Corresponding author:**
Manyuan Long
E-mail: mlong@uchicago.edu

Abbreviations:
EST, expressed sequence tag; **GO**, gene ontology; **MSA**, multiple sequence alignments; **mya**, million years ago; **NGS**, next generation sequencing; **ORF**, open reading frame; **PFC**, prefrontal neocortex; ***SRGAP2***, Slit-Robo Rho GTPase-activating protein 2.

pathways of primates and *Drosophila*, indicating that the genetics of these essential structures and processes can evolve rapidly [10–13]. For example, an excess of primate- and human-specific genes were expressed in the early developing neocortex of humans, notably in the prefrontal neocortex (PFC) [11]. The striking correspondence between a spike in the rate of new genes being incorporated into the transcriptomes of the neocortex and PFC, and the estimated time of their emergence as distinct structures of the brain suggests that anatomical novelties in the brain have recruited new genes into their genetic control system [11]. The rapid changes in protein sequences of new genes and their overrepresentation among transcription factors upregulated in fetal and infant brains may suggest the action of natural selection in shaping the human brain [11].

These works provide a starting point for investigations into gene functions involved in lineage-specific biology (e.g. [14]). More generally, the new gene studies demonstrate that genomes evolve in a flux, continuously picking up genetic novelties ranging from sequence substitutions to new genetic components. The corresponding accumulation of evolutionary novelties often breaks or eclipses the conservation of genes and genomes, eroding the widely held assumption of stasis often employed in genome annotations and functional studies.

In this review, we begin with a brief summary of the evidence for frequent protein-coding gene origination in *Drosophila* and mammals, which suggests the profound consequences of this innovative process in genome evolution. We demonstrate that these new genes have been often neglected in experimental studies, contrary to their importance demonstrated in case studies. We argue that the practice of taking conservation as the sole indicator of functional importance has contributed to the widespread inconsistencies of new gene annotation, which in turn blunt the motivation of researchers to perform functional studies of these genes. We then describe three ways to better annotate new genes by taking advantage of new advances in mass spectrometry, ribosome profiling, and gene prediction based on evolutionary genetics. We end the review with a discussion of how refining new gene annotation will advance the study of one essential and unaddressed question: whether or not adaptive selection drives the emergence of new genes in the genomes of various organisms.

Abundant new genes in the genomes of flies and mammals

Although new gene origination is easy to understand in an abstract sense, it is not trivial to detect. As its definition implies, one must determine whether a locus is a recent addition to the genome in order to identify it as a new gene. Historically, the new genes generated by retroposition, i.e. retrogenes, have been a convenient system for investigation of new genes, since intron loss serves as an unambiguous marker between parental and daughter genes [8, 15]. With the completion of numerous whole genome sequencing projects, the origination times of genes could be dated based on the parsimony rule by examining the phylogenetic distribution of orthologs for a gene of interest (reviewed in [16, 17]). Such an approach has revealed a large number of new genes, which originated in recently diverged lineages of *Drosophila*, primates, and mice (Table 1). Specifically, in *D. melanogaster*, more than 100 genes emerged after its split with *D. yakuba* 10 million years ago (mya) [7, 18, 19]. In human, for its 20,000 protein-coding genes, 1,500–3,000 genes (7.5–15%) postdated the split of human and mouse 90 mya [20–23], and among them ~400–700 genes are human-specific or absent in chimpanzee [20, 21]. Interestingly, the human-specific genes originated more rapidly than primate specific genes (64–114 per my vs. 17–33 per my, Table 1). This rate is even higher for the mouse, with more than 3,000 rodent-specific genes [20–22]. Notably, non-coding genes such as microRNAs also have a high birth rate. For example, out of 147 annotated microRNAs in *D. melanogaster*, 29 (20%) originated after its divergence with *D. pseudoobscura* 55 mya [7].

However, these estimates of gene gain are difficult to appraise. Firstly, the underlying genome assembly and gene annotation seriously affect inference of gene origination. As demonstrated in [22], an additional 1,000 genes were annotated in the finished mouse genome assembly MGSCv3 compared to the old assembly MGSCv2. Many of these new genes are encoded by recently duplicated regions, which were not assembled in MGSCv2 [22]. A second supporting example is from [21], which is one of the pioneering efforts to estimate the rate of gene gain and gene loss in mammals. However, the result may be overestimated due to incomplete annotation: some genes may be annotated in humans, but not in chimpanzee, or vice versa. Thus, a gene would be counted falsely as a gene gain or loss event (as of Aug., 2012: http://www.plosone.org/annotation/listThread.action?root=8729). Secondly, different bioinformatic methods can lead to significantly different estimates of the rate of gene gain. A maximum number of primate- or rodent-specific genes were identified in [22], which may be simply caused by the small number of outgroups, i.e. they compared only human and mouse. In contrast, multiple outgroups have been used in other studies (e.g. [20]) which help to avoid the misidentification of new genes caused by random gene loss or sequencing gap in a single outgroup.

Table 1. Four groups of species-specific or lineage-specific genes.

Categories	Number	Citation
D. melanogaster-specific or	131	[18]
D. melanogaster/*D. simulans* specific	128	[7]
	124	[19]
Human-specific	389	[20]
	689	[21]
Primate-specific	1,828	[20]
	1,557	[21]
	2,941	[22]
	1,817	[23]
Rodent-specific	3,111	[20]
	3,178	[21]
	3,767	[22]

In any case, the large numbers of new genes presented in Table 1 imply that gene origination is a widespread aspect of genome evolution. Furthermore, as mentioned above, recent case studies have revealed that new genes could be essential for fly development [10] and contribute to the evolution of human brain development [12, 24]. Despite these striking individual studies, new genes remain under-characterized by functional studies, and we highlight below our analysis of current gene annotations.

New genes have been under-represented in experimental characterization

The paucity of functional studies for new genes can be demonstrated by examining Gene Ontology (GO) annotation [25] terms across genes with different ages. An unexpected and striking correlation exists between the age of genes in the human genome and whether or not they have been experimentally studied (Fig. 1). In this linear relationship, more than 50% of the oldest genes which emerged before the vertebrate split have at least one experimental GO annotation entry assigned, while this proportion drops to 8% for human-specific genes (Fig. 1A and B). Genes in the mouse also show a similar linear ascertainment bias (Fig. 1A and C).

Both technical and logistic constraints can contribute to this observed bias. The majority (90%) of new genes were generated by gene duplication [7, 18, 20]. Differentiating new genes from their parental loci can be quite challenging, given their significant sequence similarity. In this case, the term ''multicopy gene'' refers to clusters of highly similar copies (e.g. [26, 27]). Although some multicopy gene families such as tRNA or rRNA genes could increase the copy number of identical loci to optimize dosage [28], reckless use of this term may confound the possible uniqueness of different copies caused by different sequence structure or sequence identity. This potential for distinct roles of highly similar duplicates is vividly demonstrated by recent studies [12, 24]: the Slit-Robo Rho GTPase-activating protein 2 (*SRGAP2*) was recently duplicated in the human lineage with gene family members sharing more than 98% sequence identity. Despite this similarity, the parental locus, *SRGAP2A*, is involved in neuronal migration and morphogenesis, while the partial derived copy, *SRGAP2C*, induced neoteny by antagonistically inhibiting *SRGAP2A* [12, 24]. On the other hand, change of even a very few amino acids could lead to remarkably different functions. For example, *forkhead box P2* (*FOXP2*) has only two amino acid differences in human compared to chimp and such a small change is capable of conferring differential regulation for dozens of downstream genes possibly contributing to human speech [29]. These two lines of evidence make it necessary to differentiate copies. It is feasible to make such a call considering a uniquely identifying nucleotide often exists even for polymorphic duplicates within the human population [27]. With improvement of high throughput techniques, the number of new genes covered unambiguously is increasing. For example, both the old design of the Affymetrix microarray, Affymetrix HGU 133 Plus 2 (133p2) and the new design, Affymetrix Exon Array 1.0 ST (1.0 ST), cover more than 90% of vertebrate-shared old genes with uniquely mapping probes (Fig. 2). In contrast, out of 389 human-specific genes identified in [20], 133p2 covered only 14 of these with unique probes, while this number increases to 280 (72%) in 1.0 ST [30]. Interestingly, similarly to Fig. 1, a linear regression can be modeled to fit the data significantly in both panels ($R^2 = 0.75$ and 0.85 for the left and right panels, respectively, both are significant at $p = 0.001$). In addition, a well fitted more complex model, the exponential model, of the age variable suggests that the coverage of genes in arrays dropped all the periods, even disproportionally in the recent lineages, e.g. placental mammals.

Besides these technical challenges, an in-depth functional study is costly in terms of time and resources. Investigators often prioritize genes of significance by following two practical heuristics. First, it has been held for decades that conservation is the primary indicator of functional importance (e.g. [31]). Second, because functional characterization of conserved genes could shed light on the functions of orthologs in other species, for example, humans, they are favored in experimental studies [32]. The immediate advantage is to allow inference into the functionality of orthologous genes in humans for medical research. By definition, new genes are not deeply conserved, not shared among many species, and thus deprioritized for studies based on such strategic considerations. The first rule is based on an assumption that since sequences are maintained by millions of years' purifying selection, the genes are therefore more likely functional than those genes that were recently generated [33]. This leads to numerous efforts hunting for conserved sequences in the genome (e.g. [34, 35]). However, as commented on in [31], the related implication that the less-conserved or fast-evolving loci are unimportant or even impotent has been shown to be incorrect. Evidence for the functional importance of new genes is quickly emerging, such as the aforementioned reports on new genes implicated in human brain evolution [11, 12, 24] and *Drosophila* development [10], or those related with origin of cnidocytes in *Hydra* [36]. Compared to the first heuristic of assuming that critical genes must be deeply conserved, the second used for prioritizing conserved genes appears more problematic given the following two reasons: (i) the well known ''divergence without duplication'' [37] has indicated that species accumulate biological differences between each other, e.g. human and mouse orthologs often show functional differences [38] with more than 20% of human essential genes found to be not essential in mouse [39]; (ii) widespread gene duplication can distort functional projection between orthologs, because the projection depends on how the ancestral function is maintained across paralogous copies [40].

Current new gene annotation is error-prone

Besides the technical challenges, the error-prone nature of current new gene annotation can also lead to their under-characterization. In order to understand how and why new gene annotation is more problematic than that for the old genes, we dissected the software pipelines used in gene annotation. Many efforts have been dedicated to gene annotation, such as Ensembl [40], NCBI RefSeq [41], UCSC KnownGene [42], etc. They act as the cornerstones supporting almost all

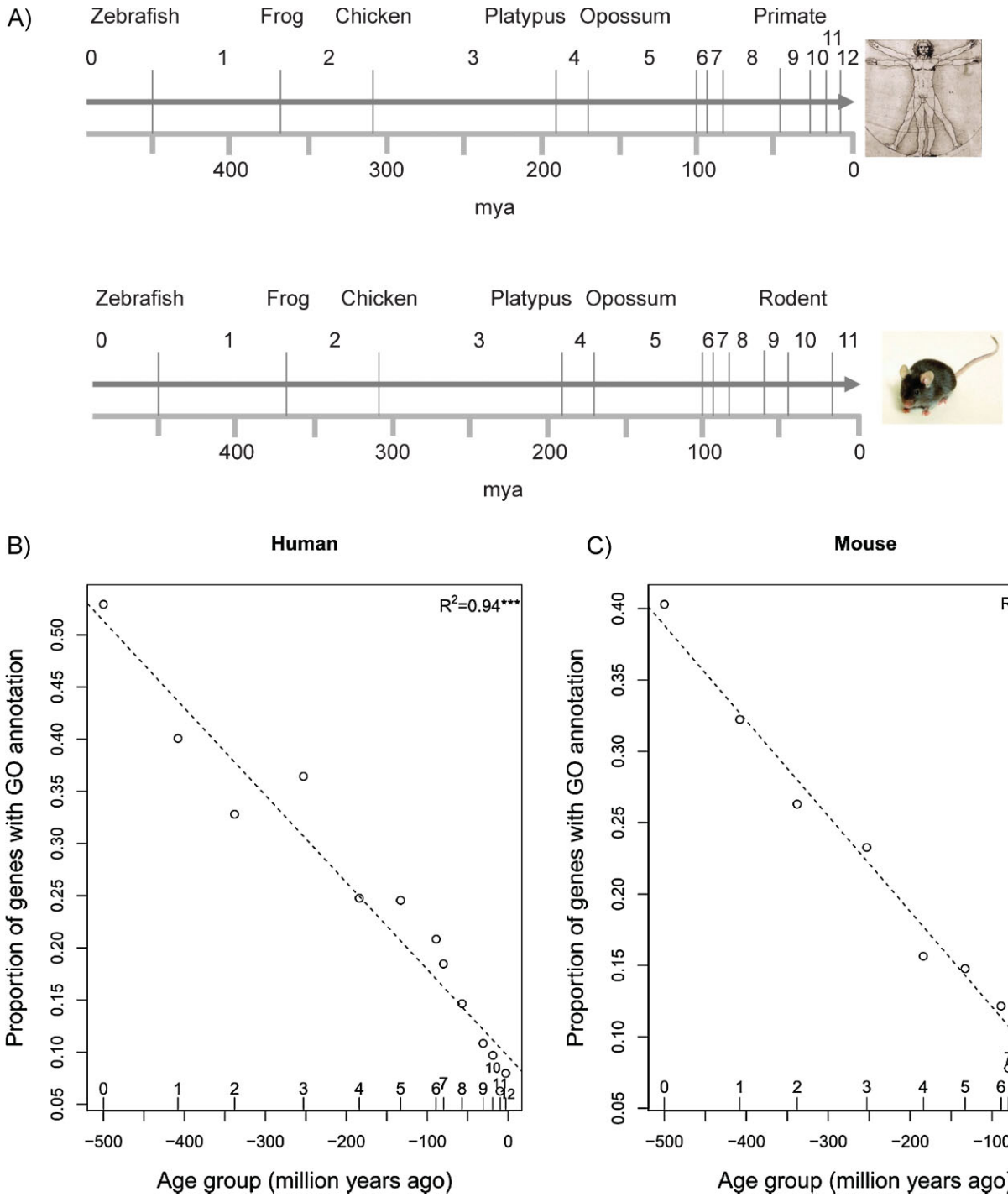

Figure 1. Functional characterization is highly biased towards old genes. **A:** The age groups for human and mouse were depicted based on the phylogenic relationship of extant species. Groups 0–7 are shared between the human and mouse, with group 0 indicating the oldest genes shared by vertebrates. Groups 8–12 in human and 8–11 in mouse refer to primate-specific and rodent-specific genes, respectively. The details of the analysis can be seen in [20]. **B:** The GO annotations of human genes originating in ancestral stages. The *X*-axis shows different age groups together with corresponding time in terms of mya, while the *Y*-axis refers to the proportion of genes with experimental characterization data demonstrated by GO evidence tags including EXP, IDA, IPI, IMP, IGI, IEP, TAS and IC out of all genes in the same age group. For group 0, the origination time is arbitrarily set as -500. A linear regression curve is shown as a dashed line with the R^2 marked in the top-right corner. ''***'' indicates a *p* smaller than 0.001 for both the intercept and slope (or age). The significant linear regression reveals a systematic and quantitatively continuous bias against the genes of reduced ages in the GO annotation. Both gene information and GO annotation were downloaded from Ensembl v51. **C:** An analogous plot has been generated for mouse.

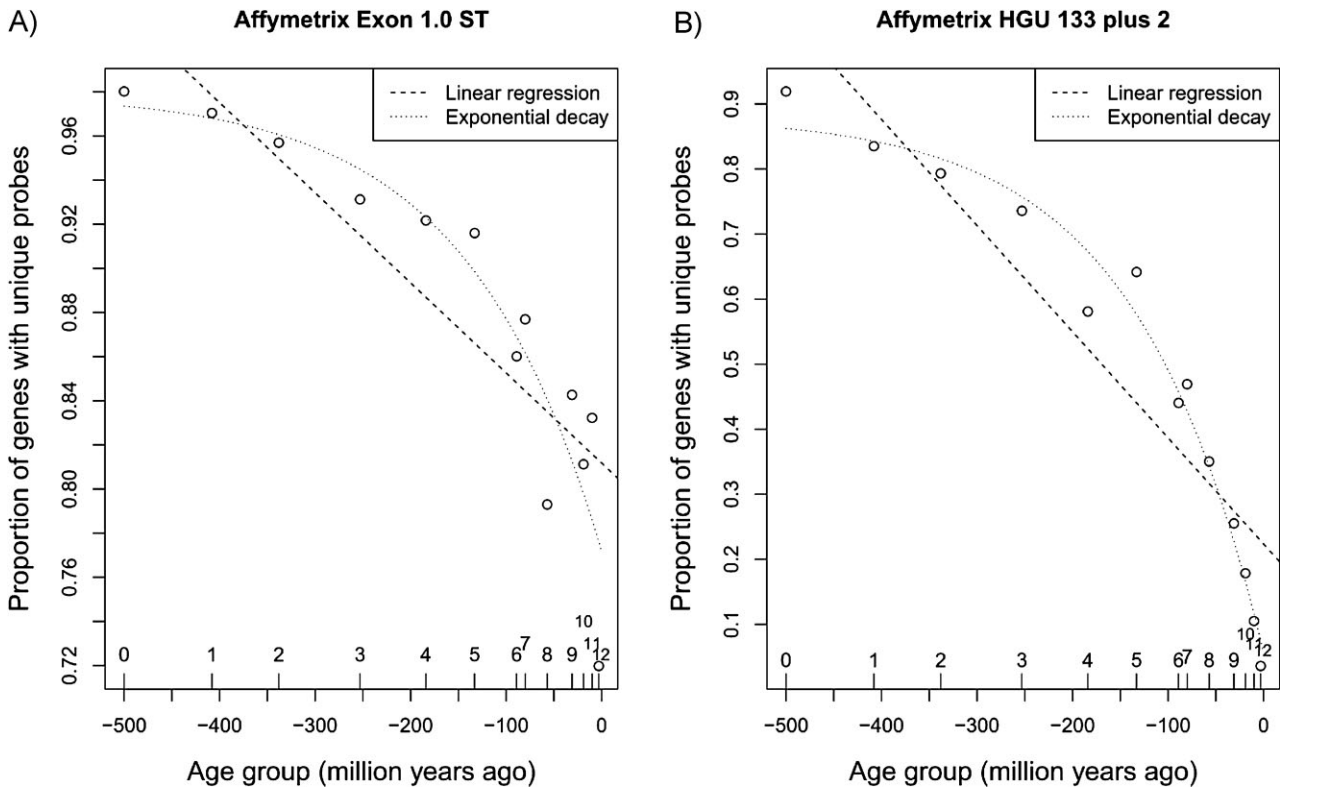

Figure 2. Proportion of human genes covered by unique Affymetrix probes across different age groups. Age information is from [20]. The mappings between probes and genes were downloaded from Aroma Affymetrix website [30]. Panels A and B show Affymetrix Human Exon Array 1.0 ST and Affymetrix HGU 133 Plus 2, respectively. Herein, two models, a simple linear model and a more complex exponential model, have been fitted well, revealing a systematic and quantitatively continuous bias against the genes of the reduced ages, especially the young genes in the lineages toward primates and other placental mammals (the exponential model reveals more rapid decay of the rate being used in and after nodes 5 and 6).

fields of molecular biology. Ensembl has been widely used and its annotation strategy has been adopted by many other systems (e.g. gene annotation for *Arabidopsis* in TAIR, http://www.arabidopsis.org/help/faq.jsp#annotation as of Aug., 2012). In the subsequent discussion, therefore, we will take the widely used Ensembl gene annotations as an example and exemplar to discuss the challenges of annotating evolutionary novelties.

As documented in [43], Ensembl implemented an evidence-based gene annotation pipeline [44], which can be divided into five stages: (1) targeted alignment: for the species of interest, proteins collected by the background databases (e.g. SwissProt/TrEMBL [45]) are aligned to the genome, which is termed *cis* alignment [46]; (2) similarity alignment: for genomic regions not covered by the first stage, protein sequences from closely related species are aligned, termed *trans* alignment [46]; (3) cDNA alignment: for the species of interest, the full-length cDNAs from the EMBL [47] and RefSeq [41] databases are aligned back to the genome; (4) GeneBuilder: all the alignments are merged and clustered in order to infer the possible gene structure; (5) post processing: in which the gene types, for example pseudogene or functional genes, are determined.

Such a pipeline cannot avoid two challenges when used to annotate a genome for new genes, given their specific biological features. First, besides the aforementioned under-characterization of new genes, the narrow transcription profile (often testis-specific) for many duplicated or de novo new genes [1] will cause them to be missed in the cDNA sequencing projects, which are usually unable to cover most tissues, and thus they often lack cDNA sequence used for annotation evidence. Figure 3 demonstrates how gene transcription profiles change across different age groups: primate-specific genes usually transcribe only about a median of four expressed sequence tags (EST) in two tissues while the vertebrate-shared oldest genes transcribe more than 100 ESTs in 20 tissues. In this scenario, Ensembl will have to rely on the very few template sequences collected by the underlying databases to annotate new genes. If these sequences are subject to some changes in the background databases, which is especially likely for the automatic annotations in the TrEMBL database [45], it will subsequently affect the gene status in Ensembl. For example, ENSG00000204626 is one of the only three human-specific de novo genes identified in [48] based on Ensembl version 47 (v47). It has only one protein template in the TrEMBL database, Q6ZSR2, which is predicted based on one mRNA sequence, AK127211. Since this protein entry was removed in TrEMBL update 39.1 (as of Aug., 2012: http://www.uniprot.org/uniprot/Q6ZSR2?version=*), ENSG00000204626 was discontinued (or retired) in Ensembl v55.

Second, 90% of new genes are generated by duplication in humans and flies [7, 18, 20] and it is often difficult to differentiate

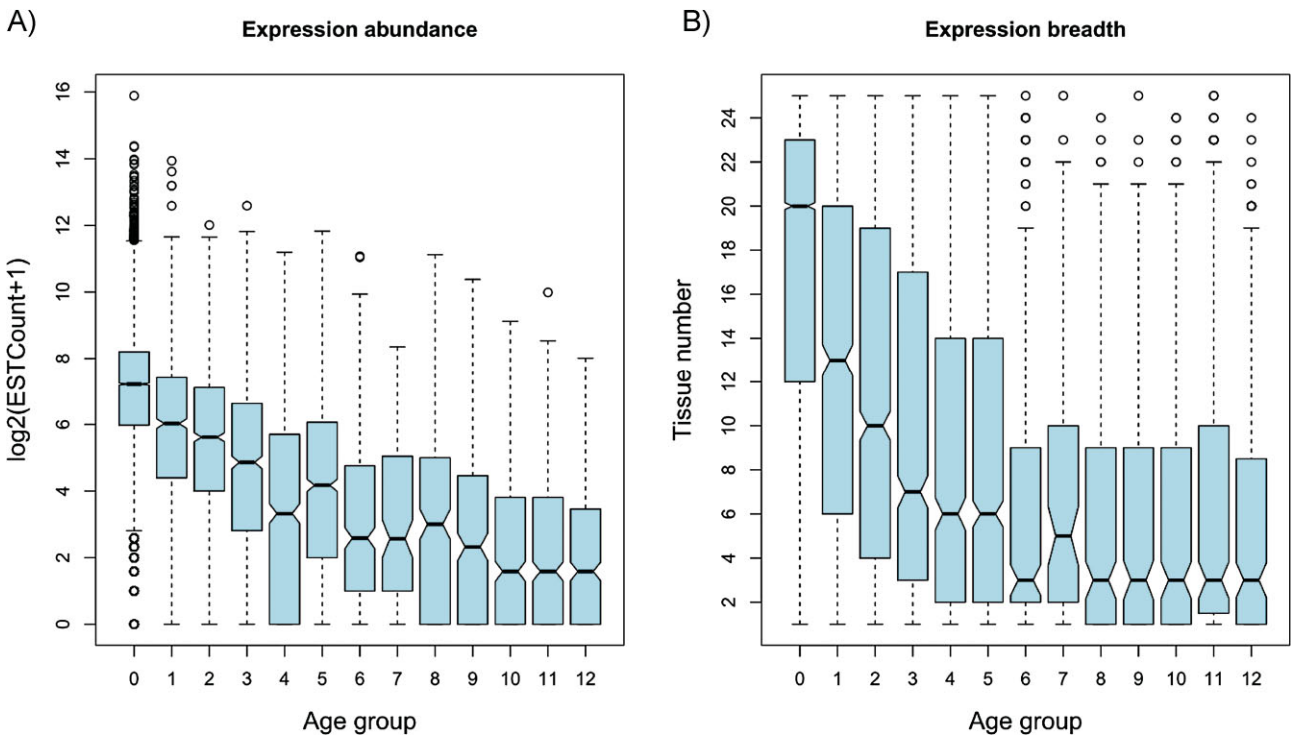

Figure 3. Boxplot-based expression profile of human genes across different age groups. Panels A and B describe the log2-based UniGene [82] EST count and the number of tissues where these ESTs are expressed. The blue boxes indicate hinges of 25% to 75% percentile and median, while the upper/lower whiskers indicate the largest/smallest observations unless these observations are beyond the 1.5 times the hinge size treated as outliers (circles). UniGene libraries lacking specific tissue information such as ''mixed'', ''uncharacterized'' or ''–'' have been removed. The EST-to-gene mapping procedure has been previously described in [11, 83].

a functional duplicate from a functionless pseudogene. This issue is demonstrated by the observation that various programs designed to search for pseudogenes often have very small overlap in their predictions (as of Aug., 2012: http://genome.cshlp.org/content/17/6/839/F1.expansion.html). Furthermore, transposons or repetitive elements could constitute significant portions of exons for new genes [1, 49]. These two facts may lead to the ever-changing criteria for calling pseudogenes by Ensembl. In 2004, compared to 24,261 protein coding genes in the human genome, Ensembl only annotated a small dataset of 962 processed pseudogenes which are single-exonic and encode at least one in-frame stop codon, inferred based on the functional paralog [43]. However, in Ensembl v65 released in 2011, the number of predicted protein coding genes decreased to 21,160; while the number of pseudogenes increased by more than 10-fold to 12,950. This striking difference is because genes matching any of the following four standards were defined as pseudogenes now (as of Aug., 2012: http://www.ensembl.org/info/docs/genebuild/genome_annotation.html): (1) single-exon genes with a multi-exon paralog; (2) genes fully masked by repeats defined by RepeatMasker; (3) single-exon genes with multiple frameshifts; and (4) genes encoding frameshifts and introns, where all of the introns are >80% covered by repeats. The inconsistency between these two sets of pseudogenes highlights the challenges faced by Ensembl in drafting a gold standard for reliably and consistently annotating pseudogenes.

Lack of annotation evidence together with ambiguity in calling pseudogenes will place new gene annotation in a flux from version to version, with many gene models expired or modified with different types. This situation can be demonstrated by examining how many new genes identified in the old version of Ensembl survived in the new version. As demonstrated by Table 2, only 43% (781) of 1,828 primate-specific genes annotated in Ensembl v51 (Nov. 2008) were still annotated as protein-coding genes in Ensembl v65 (Dec. 2011). In contrast, for genes emerging before the primate split, the majority (90%) stayed as protein-coding genes. Eight years ago, around 100 retrogenes in Ensembl Version 10 of the human and mouse genomes were identified with multiple rigorous criteria [50]. Although the majority (75%) of parental (relatively old) genes remains in the Ensembl v65, the annotation of 74% of the human retrogenes and of 90% of mouse retrogenes were subsequently removed (Table 2).

Thus, Ensembl works remarkably well for old genes, while problematic for new genes. Such an inconsistence of new gene annotation will certainly make researchers uncomfortable in conducting costly in-depth studies. Moreover, because practical considerations lead to the use of model organisms to study human biology, new genes are less likely to be characterized than the old, phylogenetically shared genes, especially the genes that have orthologs in the human genome. These two inadequacies have compromised functional and evolutionary studies of new genes. To remedy this undesirable situation, the first step is to better annotate the genome with direct and more reliable measurements of protein-coding potential, a step that may be attainable given the progress in the functional genomics and evolutionary studies.

Table 2. Statistics of new genes identified in previous studies in the current Ensembl annotation v65 (Dec. 2011).

Studies	Gene categories	Coding genes	Noncoding genes	Retired
Retrogene scan in human and mouse based on Ensembl v10 [50]	Human retrogene	21	5	73
	Human parental gene	74	0	20
	Mouse retrogene	11	0	97
	Mouse parental gene	78	2	25
Human specific de novo genes based on Ensembl v47 [48]		1	1	1
Human specific de novo genes based on Ensembl v40-v57 [51]		6	10	44
Gene age dating in human and mouse based on Ensembl v51 [20]	Primate-specific genes	781	660	387
	Non-primate-specific genes	16,762	550	795
	Rodent-specific genes	2,323	278	510
	Non-rodent-specific genes	16,836	382	793

Notes: ''non-coding genes'' refer to all other gene categories in Ensembl such as pseudogene, antisense transcripts, etc; ''Retired'' means that genes annotated previously have been removed from v65 where the corresponding location is replaced with new gene models or no gene model exists any more.

Inferring the coding potential with an integrative approach

Possibly the most relevant advances in the field of functional genomics are proteogenomics and ribosome profiling [52–55]. In contrast to conventional mass spectrum data analyses, which directly map peptides against annotated proteins, next generation proteogenomics interrogates peptides against all possible six-frame translations across a genome [52, 53]. This method could help to uncover previously undetected new genes or misclassified pseudogenes. A striking example has been provided by a proteogenomics study in the best annotated plant genome, *Arabidopsis* [56], which generated 144,079 distinct peptides and identified 778 novel genes not overlapping with the previously annotated coding genes in TAIR. These genes appear to be evolutionarily young genes ignored by conventional annotation since the majority (653 or 84%) are not conserved and a significant proportion (498 or 64%) were formerly annotated as pseudogenes or transposons [56]. Given the capability of proteogenomics to assess the coding potential of genomic loci in a high-throughput way, it has been proposed to be integrated into the initial annotation stage rather than being a purely downstream tool [52].

Different from proteogenomics, but with similar promise, the recently developed technique of ribosome profiling indirectly infers which genomic loci are being translated by sequencing the mRNAs within a ribosome [54, 55]. One surprising result is that a majority of previously annotated long intergenic noncoding RNAs are bound by ribosomes in mouse, suggesting these loci may actually be coding genes [54]. Although ribosome binding is not equivalent to protein-coding capability [57], it nevertheless could be used as supporting evidence, especially considering that such a new technique could precisely locate the initiation codon and measure the translation speed [54, 58]. Moreover, proteogenomics is hobbled in that it could only detect proteins in amounts across 3–4 orders of magnitudes, and faces the huge search space imposed by the six-frame translation of the genome [52]. The short peptide (by average 14 amino-acids) generated by mass spectrometry may not be enough to differentiate derived duplicates and their parental genes. In contrast, ribosome profiling takes advantage of next generation sequencing (NGS), which could identify transcripts varying in amount by at least five orders of magnitude [59]. Moreover, mRNAs could be directly mapped back to the genome without the need of translation. Finally, the read length of the mainstream NGS platform, i.e. Illumina, is continuously increasing, with a current read length of 150 bp (as of Aug., 2012: http://www.illumina.com/systems/hiseq_systems.ilmn), which can be mapped back to different paralogs with less ambiguity. Thus, these two techniques are complementary, which could help to increase the coverage of protein-coding gene annotation.

However, high-quality, high-coverage functional genomic data has been unavailable for most species. This situation makes it necessary to use evidence-based or de novo prediction of genes using DNA-sequence alone. Two programs are particularly noticeable. CONTRAST predicted a perfect open reading frame (ORF) structure for up to 58% of human protein-coding genes by recognizing coding region boundaries in multiple sequence alignments (MSA) across different species [46, 60]. PhyloCSF could accurately determine a peptide as short as five amino acids by modeling codon substitution frequency in MSA [57, 61]. However, as these packages are based on alignments across diverged species (mammals or 12 *Drosophila* [60, 61]) and target conserved proteins, new genes may be neglected. Thus, a couple of tools have been developed to infer the coding potential of lineage-specific genes. One of the most popular tools detects whether replacement and synonymous sites have different substitution rates using the *Ka/Ks* test (the ratio between non-synonymous substitution rate and synonymous substitution rate) implemented in PAML [62] and the McDonald–Kreitman test implemented in DnaSP or PGEToolbox [63–65]. Another series of tools including ReEVOLVER [66] and $t_{1/2}$ test [67] estimate whether the candidate ORF is selectively constrained by simulating neutral evolution. All these tools have been successfully applied in the study of new genes (e.g. [8, 68–70]). However, all these tools require that new genes of interest have accumulated enough substitutions for the statistical tests. For example, a new gene in humans would need to have emerged at least eight million

years ago for ReEVOLVER to have the power [69] to perform the statistical analysis. In this case, human-specific genes (absent in chimpanzee, younger than six million years) could not be tested.

It is still a great challenge to infer the complete gene structure (even the ORF only) using short peptides or short reads compared to protein or cDNA based annotation implemented in Ensembl. However, the current protein evidence (especially proteins annotated in TrEMBL) used in Ensembl is unreliable and the criteria used to differentiate functional genes from pseudogenes are fairly arbitrary. Thus, the Ensembl annotation can be significantly improved if proteogenomics, ribosome profiling and evolutionary inference are incorporated into the final step of distinguishing functional genes from pseudogenes. Thus, all these new techniques complement and enrich conventional annotation practice, such as that used by Ensembl.

Annotation refinement will help understanding of the evolutionary forces acting on new genes

For around a decade, it has remained unsettled whether new gene emergence is driven by neutral evolution or positive selection [16, 17, 71–73]. Refining annotations of new genes will certainly help address this issue considering the key is to determine whether functions played by new genes are evolutionarily novel. Actually, although new genes are novel genetic components, they do not necessarily play new functional roles. For example, in the subfunctionalization model of how new genes escape degeneration, a new gene merely subdivides the functions of the ancestral gene with its parents, with the whole process driven by neutral evolution [73]. In this respect, the following three types of new genes will be particularly illuminating.

First, as reviewed in [16], chimeric new genes have a structure different from their parental genes and thus they are most likely to play new functions. Second, a partial duplicate may result in vastly distinct functions. As in the case of the previously mentioned *SRGAP2*, experimental study revealed that a partial copy, *SRGAP2C*, antagonizes its parental gene *SRGAP2A*, which is a dimer in its functional form [24]. Thus, immediately after its birth, *SRGAP2B* gained a new function [12]. Finally, the de novo genes may represent the clearest example of neofunctionalization, since they have no related ancestors with functions to partition. For example, it has been reported repeatedly that numerous primate-specific or even human-specific de novo genes are transcribed in the human brain [11, 51, 74]. Further investigation of the functions of these de novo genes may shed light on the process of brain evolution by the addition of novel gene functions.

For all these three types of new genes, accurate gene annotation is essential. Without precise exon-intron structure, it is impossible to detect whether or not a new gene is subject to chimerism in evolution. Both *SRGAP2B* and *SRGAP2C* are annotated as pseudogenes by Ensembl v66 (*SRGAP2P1*, *SRGAP2P2*). If the research community had simply passively accepted such annotation, the exciting discovery of their function in neuronal migration and spines [12, 24] would have been unlikely. The remarkable discovery of de novo genes is another similar case [51] in which authors have examined all Ensembl releases in the last four years and identified 60 human-specific de novo genes. If they had focused only on the latest version of Ensembl, they could have found only 27 cases [51].

Conclusions and prospects

The new genes are attracting more attention from both molecular and evolutionary studies for their origination mechanisms, evolutionary patterns, and functional importance [1, 16, 17]. The large number of new genes that have been identified suggests the generality of this evolutionary process. However, two interlinked inadequacies in study have hampered progress in the field: new genes are generally under-characterized in functional studies and they are often misannotated. We argue here that the new genes should be paid more attention while studying rapidly evolving phenotypes (e.g. the brains of human). While *cis*-regulatory changes have been proposed as major evolutionary shortcuts in morphological evolution [75–77], the origination of new genes may provide a complementary and important route toward evolution of phenotypes, shown by the research implying that new genes contributed to the genetic novelty of human brains [11, 12].

Regarding gene annotation, various new techniques and evolutionary approaches can better infer the coding potential of genes and thus provide a more consistent dataset of new genes. This will help the study of new genes tremendously, especially regarding the evolutionary forces that govern their origination process. More excitingly, if such annotation could scale up to the population level, i.e. to annotate polymorphic gene gains across populations or strains such as 1,000 genomes of humans [78] or the *Arabidopsis* 1001 genomes [79], we could trace how new genes have evolved in recent evolutionary time and then infer selective forces acting on the newborn genes. For example, a retrogene polymorphism survey demonstrated using a generalized McDonald–Kreitman framework that natural selection may drive retrogene fixation onto autosomes, which had parental loci on the X chromosome [65, 80]. These lines of research are not only important for addressing theoretical questions of genome evolution, but will also help to identify more new gene duplicates underlying specific instances of phenotypic divergence, like the increase of *Amylase* copy number in humans [81]. All these productive approaches require a new vision of genomes and genes that differs from the traditional conservation-based dogma.

Acknowledgments

We are grateful for the critical reading and editing of Robin Bush. M.L. was supported by the NSF MCB1051826 and NIH R01 GM100768-01A1 in USA; Y.E.Z. was supported by the ''Thousand Young Talents'' program of China's government.

References

1. **Kaessmann H.** 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**: 1313–26.

2. **Long M**, **Betran E**, **Thornton K**, **Wang W.** 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**: 865–75.
3. **Dai H**, **Chen Y**, **Chen S**, **Mao Q**, et al. 2008. The evolution of courtship behaviors through the origination of a new gene in Drosophila. *Proc Natl Acad Sci USA* **105**: 7478–83.
4. **Heinen T**, **Staubach F**, **Häming D**, **Tautz D.** 2009. Emergence of a new gene from an intergenic region. *Curr Biol* **19**: 1527–31.
5. **Siepel A.** 2009. Darwinian alchemy: human genes from noncoding DNA. *Genome Res* **19**: 1693–5.
6. **Zhou Q**, **Wang W.** 2008. On the origin and evolution of new genes – a genomic and experimental perspective. *J Genet Genomics* **35**: 639–48.
7. **Zhang YE**, **Vibranovski MD**, **Krinsky BH**, **Long M.** 2010. Age-dependent chromosomal distribution of male-biased genes in Drosophila. *Genome Res* **20**: 1526–33.
8. **Long M**, **Langley CH.** 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science* **260**: 91–5.
9. **Nurminsky DI**, **Nurminskaya MV**, **De Aguiar D**, **Hartl DL.** 1998. Selective sweep of a newly evolved sperm-specific gene in Drosophila. *Nature* **396**: 572–5.
10. **Chen S**, **Zhang YE**, **Long M.** 2010. New genes in Drosophila quickly become essential. *Science* **330**: 1682–5.
11. **Zhang YE**, **Landback P**, **Vibranovski MD**, **Long M.** 2011. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol* **9**: e1001179.
12. **Dennis Megan Y**, **Nuttle X**, **Sudmant Peter H**, **Antonacci F**, et al. 2012. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**: 912–22.
13. **Chen S**, **Spletter M**, **Ni X**, **White KP**, et al. 2012. Frequent recent origination of brain genes shaped the evolution of foraging behavior in drosophila. *Cell Rep* **1**: 118–32.
14. **Konopka G.** 2012. The fetal brain provides a Raison d'etre for the evolution of new human genes. *Brain Behav Evol* **79**: 213–4.
15. **Brosius J.** 2003. The contribution of RNAs and retroposition to evolutionary novelties. *Genetica* **118**: 99–115.
16. **Ranz JM**, **Parsch J.** 2012. Newly evolved genes: moving from comparative genomics to functional studies in model systems. *BioEssays* **34**: 477–83.
17. **Cardoso-Moreira M**, **Long M.** 2012. The origin and evolution of new genes. *Methods Mol Biol* **856**: 161.
18. **Zhou Q**, **Zhang G**, **Zhang Y**, **Xu S**, et al. 2008. On the origin of new genes in Drosophila. *Genome Res* **18**: 1446–55.
19. **Hahn MW**, **Han MV**, **Han SG.** 2007. Gene family evolution across 12 Drosophila genomes. *PLoS Genet* **3**: e197.
20. **Zhang YE**, **Vibranovski MD**, **Landback P**, **Marais GAB**, et al. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol* **8**: e1000494.
21. **Demuth JP**, **De Bie T**, **Stajich JE**, **Cristianini N**, et al. 2006. The evolution of mammalian gene families. *PLoS ONE* **1**: e85.
22. **Church DM**, **Goodstadt L**, **Hillier LW**, **Zody MC**, et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* **7**: e1000112.
23. **Cai JJ**, **Petrov DA.** 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol Evol* **2**: 393–409.
24. **Charrier C**, **Joshi K**, **Coutinho-Budd J**, **Kim JE**, et al. 2012. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**: 923–35.
25. **Ashburner M**, **Ball CA**, **Blake JA**, **Botstein D**, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25–9.
26. **Mueller J**, **Mahadevaiah S**, **Park P**, **Warburton P**, et al. 2008. The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. *Nat Genet* **40**: 794–9.
27. **Sudmant PH**, **Kitzman JO**, **Antonacci F**, **Alkan C**, et al. 2010. Diversity of human copy number variation and multicopy genes. *Science* **330**: 641–6.
28. **Ohno S.** 1970. *Evolution by Gene Duplication*. Berlin, Heidelberg and New York: Springer-Verlag.
29. **Konopka G**, **Bomar JM**, **Winden K**, **Coppola G**, et al. 2009. Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* **462**: 213–7.
30. **Robinson MD**, **Speed TP.** 2007. A comparison of Affymetrix gene expression arrays. *BMC Bioinform* **8**: 449.
31. **Monroe D.** 2009. Genetics. Genomic clues to DNA treasure sometimes lead nowhere. *Science* **325**: 142–3.
32. **Koonin EV.** 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**: 309–38.
33. **Kimura M.** 1985. *Neutral Theory of Molecular Evolution*. New York: Cambridge University Press.
34. **Lindblad-Toh K**, **Garber M**, **Zuk O**, **Lin MF**, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**: 476–82.
35. **Bejerano G**, **Pheasant M**, **Makunin I**, **Stephen S**, et al. 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321–5.
36. **Khalturin K**, **Hemmrich G**, **Fraune S**, **Augustin R**, et al. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* **25**: 404–13.
37. **Soskine M**, **Tawfik DS.** 2010. Mutational effects and the evolution of new protein functions. *Nat Rev Genet* **11**: 572–82.
38. **Gharib WH**, **Robinson-Rechavi M.** 2011. When orthologs diverge between human and mouse. *Brief Bioinform* **12**: 436–41.
39. **Liao BY**, **Zhang J.** 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci USA* **105**: 6987–92.
40. **Flicek P**, **Amode MR**, **Barrell D**, **Beal K**, et al. 2012. Ensembl 2012. *Nucleic Acids Res* **40**: D84–90.
41. **Pruitt KD**, **Tatusova T**, **Maglott DR.** 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**: D61–5.
42. **Hsu F**, **Kent WJ**, **Clawson H**, **Kuhn RM**, et al. 2006. The UCSC known genes. *Bioinformatics* **22**: 1036–46.
43. **Curwen V**, **Eyras E**, **Andrews TD**, **Clarke L**, et al. 2004. The Ensembl automatic gene annotation system. *Genome Res* **14**: 942–50.
44. **Yandell M**, **Ence D.** 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**: 329–42.
45. **Boeckmann B**, **Bairoch A**, **Apweiler R**, **Blatter MC**, et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**: 365–70.
46. **Brent MR.** 2008. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat Rev Genet* **9**: 62–73.
47. **Kanz C**, **Aldebert P**, **Althorpe N**, **Baker W**, et al. 2005. The EMBL nucleotide sequence database. *Nucleic Acids Res* **33**: D29–33.
48. **Knowles DG**, **McLysaght A.** 2009. Recent de novo origin of human protein-coding genes. *Genome Res* **19**: 1752–9.
49. **Li CY**, **Zhang Y**, **Wang Z**, **Cao C**, et al. 2010. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput Biol* **6**: e1000734.
50. **Emerson JJ**, **Kaessmann H**, **Betran E**, **Long M.** 2004. Extensive gene traffic on the mammalian X chromosome. *Science* **303**: 537–40.
51. **Wu DD**, **Irwin DM**, **Zhang YP.** 2011. De novo origin of human protein-coding genes. *PLoS Genet* **7**: e1002379.
52. **Castellana N**, **Bafna V.** 2010. Proteogenomics to discover the full coding content of genomes: a computational perspective. *J Proteomic* **73**: 2124–35.
53. **Krug K**, **Nahnsen S**, **Macek B.** 2011. Mass spectrometry at the interface of proteomics and genomics. *Mol Biosyst* **7**: 284–91.
54. **Ingolia NT**, **Lareau LF**, **Weissman JS.** 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**: 789–802.
55. **Ingolia NT**, **Ghaemmaghami S**, **Newman JR**, **Weissman JS.** 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218–23.
56. **Castellana NE**, **Payne SH**, **Shen Z**, **Stanke M**, et al. 2008. Discovery and revision of Arabidopsis genes by proteogenomics. *Proc Natl Acad Sci USA* **105**: 21034–8.
57. **Guttman M**, **Rinn JL.** 2012. Modular regulatory principles of large non-coding RNAs. *Nature* **482**: 339–46.
58. **Weiss RB**, **Atkins JF.** 2011. Translation goes global. *Science* **334**: 1509–10.
59. **Mortazavi A**, **Williams BA**, **McCue K**, **Schaeffer L**, et al. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621–8.
60. **Gross SS**, **Do CB**, **Sirota M**, **Batzoglou S.** 2007. CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol* **8**: R269.
61. **Lin MF**, **Jungreis I**, **Kellis M.** 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**: i275–82.
62. **Yang Z.** 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–91.
63. **Rozas J**, **Sanchez-DelBarrio JC**, **Messeguer X**, **Rozas R.** 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–7.

64. **Cai JJ.** 2008. PGEToolbox: a Matlab toolbox for population genetics and evolution. *J Hered* **99**: 438–40.
65. **McDonald JH**, **Kreitman M.** 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**: 652–4.
66. **Dupanloup I**, **Kaessmann H.** 2006. Evolutionary simulations to detect functional lineage-specific genes. *Bioinformatics* **22**: 1815–22.
67. **Zhang J**, **Webb DM.** 2003. Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates. *Proc Natl Acad Sci* **100**: 8337–41.
68. **Betran E**, **Thornton K**, **Long M.** 2002. Retroposed new genes out of the X in Drosophila. *Genome Res* **12**: 1854–9.
69. **Marques AC**, **Dupanloup I**, **Vinckenbosch N**, **Reymond A**, et al. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol* **3**: e357.
70. **Zhang Y**, **Lu S**, **Zhao S**, **Zheng X**, et al. 2009. Positive selection for the male functionality of a co-retroposed gene in the hominoids. *BMC Evol Biol* **9**: 252.
71. **Lynch M**, **Conery JS.** 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–5.
72. **Long M**, **Thornton K.** 2001. Gene duplication and evolution. *Science* **293**: 1551.
73. **Innan H**, **Kondrashov F.** 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11**: 97–108.
74. **Airavaara M**, **Pletnikova O**, **Doyle ME**, **Zhang YE**, et al. 2011. Identification of novel GDNF isoforms and *cis*-antisense GDNFOS gene and their regulation in human middle temporal gyrus of alzheimer disease. *J Biol Chem* **286**: 45093–102.
75. **Carroll SB.** 2005. Evolution at two levels: on genes and form. *PLoS Biol* **3**: e245.
76. **Carroll SB.** 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**: 25–36.
77. **Wittkopp PJ**, **Kalay G.** 2012. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**: 59–69.
78. **Durbin RM**, **Abecasis GR**, **Altshuler DL**, **Auton A**, et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–73.
79. **Weigel D**, **Mott R.** 2009. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol* **10**: 107.
80. **Schrider DR**, **Stevens K**, **Cardeño CM**, **Langley CH**, et al. 2011. Genome-wide analysis of retrogene polymorphisms in *Drosophila melanogaster*. *Genome Res* **21**: 2087–95.
81. **Perry GH**, **Dominy NJ**, **Claw KG**, **Lee AS**, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**: 1256–60.
82. **Sayers EW**, **Barrett T**, **Benson DA**, **Bolton E**, et al. 2010. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **38**: D5–16.
83. **Zhang Y**, **Li J**, **Kong L**, **Gao G**, et al. 2007. NATsDB: Natural Antisense Transcripts DataBase. *Nucleic Acids Res* **35**: D156–61.