Computational Identification of Novel Genes: Current and Future Perspectives

Steffen Klasberg, Tristan Bitard-Feildel and Ludovic Mallet

Institute for Evolution and Biodiversity, Westfalian Wilhelms University Muenster, Huefferstrasse 1, Muenster, Germany.

ABSTRACT: While it has long been thought that all genomic novelties are derived from the existing material, many genes lacking homology to known genes were found in recent genome projects. Some of these *novel genes* were proposed to have evolved de novo, ie, out of noncoding sequences, whereas some have been shown to follow a duplication and divergence process. Their discovery called for an extension of the historical hypotheses about gene origination. Besides the theoretical breakthrough, increasing evidence accumulated that novel genes play important roles in evolutionary processes, including adaptation and speciation events. Different techniques are available to identify genes and classify them as novel. Their classification as novel is usually based on their similarity to known genes, or lack thereof, detected by comparative genomics or against databases. Computational approaches are further prime methods that can be based on existing models or leveraging biological evidences from experiments. Identification of novel genes remains however a challenging task. With the constant software and technologies updates, no gold standard, and no available benchmark, evaluation and characterization of genomic novelty is a vibrant field. In this review, the classical and state-of-the-art tools for gene prediction are introduced. The current methods for novel gene detection are presented; the methodological strategies and their limits are discussed along with perspective approaches for further studies.

KEYWORDS: novel genes, de novo genes, novel domains, gene detection, evolutionary genomics

CITATION: Klasberg et al. Computational Identification of Novel Genes: Current and Future Perspectives. *Bioinformatics and Biology Insights* 2016:10 121–131 doi: 10.4137/BBI.S39950.

TYPE: Review

RECEIVED: April 15, 2016. **RESUBMITTED:** May, 31, 2016. **ACCEPTED FOR PUBLICATION:** June 05, 2016.

ACADEMIC EDITOR: Thomas Dandekar, Associate Editor

PEER REVIEW: Six peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,340 words, excluding any confidential comments to the academic editor.

FUNDING: LM is funded by the grant B02544/7-2 from the program Schwerpunktprogramme SPP1399 of the Deutsche Forschungsgemeinschaft. SK is supported by Leibnitz Graduate School on Genomic Biodiversity Research. We acknowledge support by Open Access Publication Fund of University of Muenster. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: s.klasberg@uni-muenster.de

Introduction

To build and run living cells, organisms store a great part of their necessary biological program into genes, segments of their DNA sequences. In ever-changing environments, organisms competing for resources must adapt and evolve. The emergence of novel functions at the gene level provides such a possibility. Changes in the biological program can mediate unprecedented alternatives that can lead to adaptation. Changes will be submitted to natural selection, may contribute to the emergence of innovative traits, and lead to the adaptation and evolution of organisms.[1]

The understanding of relationships between genes is one of the fundamental bases of molecular evolution analyses, phylogeny, and many other fields, from pure theoretical biology to applied biotechnology.[1]

Based on the sequence similarity, or homology, between genes, it is possible to establish associative links between genes from different genomes. The analysis of the yeast genome[2] uncovered for the first time a set of genes without homologs in other species. These genes were named *novel genes*, and new hypotheses were proposed to explain their presence regarding the molecular evolution theory. After their initial discovery, novel protein-coding genes were found in most, if not all, newly sequenced genomes in amounts of 10%–20%.[3–6] The importance of the functions encoded by novel genes has been underestimated for a long time. They play important roles in crucial biological functions, including developmental processes, sexual reproduction, behavior, or morphological phenotypic traits.[7,8] Furthermore, it has been shown that novel genes can become essential in a short time span in *Drosophila melanogaster*.[9,10]

In this study, a review of the current theories of novel gene emergence is provided with a special emphasis on the methodological and computational challenges brought by a special type of novel genes, the de novo genes.

The first part of this review also provides an introduction on the biological background to define what genes are consisting of, and by extension what a novel gene is. Several types of genes exist, the first described and the most well known being the protein-coding gene. As a result, the first novel genes studied were also protein-coding genes, but more and more attention is currently given to nonprotein-coding genes. With several evidences from different research groups of the presence of novel genes in various species appeared several definitions of what a novel gene is emerged. The historical way novel genes were thought to originate was driven by duplication of ancient, also called parental, genes followed by independent divergence.[11] However, this theory is nowadays

regarded as incomplete, as it excludes the concept of de novo genes. De novo genes are previously nonfunctional genomic sequence that acquired enough modifications to become transcribed. The number of de novo genes is not as low as one would expect, for instance, Wu et al.[3] found 60 genes classified as de novo in the human genome.

The second part of this review is dedicated to the methodology for gene detection/prediction. These methods can be listed into two groups, either relying primarily on biological experiments, such as high throughput sequencing, combined with software to analyze them, or fully computational methods relying on existing biological knowledges and datasets. Prior to the classification and study of novel genes, the classical annotation of the whole set of genes in one or several genomes in a considered clade is the first primordial step. Gene prediction can either be performed ab initio, by a comparative approach using annotations of known genes in other genomes, or relies on biological experiments. Ab initio gene prediction methods make use of different intrinsic properties of the genomic sequence of a species, such as the nucleotide or k-mer composition of a sequence or the length of an *open reading frame (ORF)*. Additional evidences for gene prediction can be extrinsic, such as gene predictions from sister species, genic features, syntax analysis, or known regulatory elements. Biological experiments such as RNA sequencing (RNA-seq) provide a common material to improve gene prediction of transcribed genes.[12] Nowadays, the majority of the gene models are computationally predicted and while they may be supported by high-throughput sequencing, they are rarely validated experimentally.

What are Novel Genes?

Proving that a gene is novel is a difficult task. In this section, we present a more detailed description of which characteristics a gene need to exhibit to be qualified as novel.

What is a gene?. First, a novel gene need to be categorized as a gene, ie, a genomic nucleotide sequence, which possesses the features of a gene. Several types of genes exist, the first type described being the protein-coding gene. A protein-coding gene is characterized based on the presence of some specific elements at the genomic level.[13] As a protein-coding gene, the gene must have a coding sequence (CDS). The CDS is a crucial part of a gene as it is the sequence transcribed into RNA and later on translated into an amino acid sequence.

However, a protein-coding gene displays other features that are almost as important. It is debated whether those elements are actual part of the gene or, more precisely, only adjuncts enabling a gene to fulfill its function.[14] For instance, *cis*-regulatory elements, in the untranslated regions (UTR) of the gene, before the CDS (5′-UTR) or after (3′-UTR), may contribute to many different roles[15]: (a) a polyadenylation signal (3′-UTR) that protects the messenger RNA (mRNA) from degradation and in eukaryotes mediates its transportation outside of the nucleus, (b) a promoter that may contain the transcription start site (TSS) and will mediate the initiation of transcription, and (c) transcription factor binding sites (TFBS) that are necessary for a guided transcription to start. For most eukaryotic genes, their structure is not continuous but made of successive exons and introns.[16,17] Upon gene expression, the transcription machinery produces an RNA copy of the gene from which introns are spliced out. Alternative splicing may also subtract or combine targeted exons, enabling optional combination of different exons potentially altering the function of the product. The remaining concatenated exons form the mRNA that is later translated into a peptide by the ribosomal machinery.

Rules and structures that define classical genes are already flexible and need to be even more flexible to define novel genes.

What is novel?. Primarily, novelty is defined as the absence of similarity with other sequences. This definition is mostly assessed by comparison of the primary sequences to the knowledge to date, archived in public databases.

Different terms associated with slightly different definitions are currently used regarding the study of novel genes in biology. Thereafter, a definition is proposed for the terms "orphan gene", "taxonomically restricted gene", "novel gene", and "de novo gene"[6,18] (see definition boxes). The "orphan" adjective was originally associated with genes that are specific to a single organism. However, with the emerging *next-generation sequencing* techniques and a rapidly increased coverage of sequenced species, many formerly assumed species-specific genes were subsequently found in other organisms. The term "taxonomically restricted" was then introduced to

Orphan genes
Orphan (or taxonomically restricted) genes are classified based on a given phylogeny. A gene that is only found inside a single species or a branch, but not outside, is orphan in that specific branch.

Novel genes
Novel genes are classified by their age. Genes that have emerged inside a defined time frame are novel genes. The time frame is not fixed and need to be defined for each study. All novel genes are orphan in a specific clade, but, depending on the time frame, not all orphan genes are classified as novel.

De novo genes
De novo genes are defined based on their mechanism of emergence, ie, out of previously noncoding DNA. This might, eg, occur via acquisition of transcriptional regulation, consecutive point mutations, or genomic rearrangements.

refer to genes not only limited to a single organism but also found in phylogenetically related species.[19]

In this paper, the term "orphan gene" is used substitutionary for both cases, original "orphans" and "taxonomically restricted genes". The emergence of orphan genes can be explained by at least three events.[20] (a) Fast-evolving genes can diverge beyond the level of recognition of homology searches. The homology of such fast-evolving genes cannot be traced back to their ancestral genes, but they are not entirely novel genes. (b) Genes can be lost in other species they are compared with, leading to a wrong backdating and their false classification as novel. A gene that is lost in other species might be found as a pseudogene, ie, a gene-like fragment with homology to the novel gene that is not transcribed could survive in the DNA sequence.[21] (c) True orphan genes appear on a specific phylogenetic branch and evolved in a specific lineage. This evolution of orphan genes might be mediated by recombination methods based on previously available DNA or de novo.

Therefore, all orphan genes are not necessarily novel genes, whereas all novel genes can be qualified as orphan in a specific phylogenetic tree. The distinction between the more general group of orphan genes, ie, novel and ancient *taxonomically restricted genes*, and the narrower group of novel genes is based on the emergence time of the studied gene.[22] It has to be kept in mind that no universal time limit is clearly defined to make the distinction between orphan and novel genes. This time limit is strongly dependent on the set of species studied, in particular the sampling of the species and how representative of the true evolutionary path the set is. An important subset of novel genes are grouped under the term de novo.[23] De novo genes are characterized by their mechanism of origin, which consists of their creation out of previously noncoding sequences.[8] In this paper, a separation will be made between the terms de novo and orphan, based on the mechanism of origin of the gene, out of noncoding sequence or not. The general term novel will be used to refer to the emergence of a gene in a phylogenetic tree, keeping in mind that the emergence should be recent. Figure 1 shows the different terms explained on a phylogenetic tree.

Origins of novel genes. Many other remarkable features can be noted regarding the origin and outcome of novel genes. Several events can mediate novelty at the gene level, including, but not limited to, duplication, truncation, elongation, juxtaposition, fusion, and translocation of genes, mediated by recombination methods or nucleotide mutations.[24,25] Prevalent recombination methods are (nonallelic) homologous or illegitimate recombination. Transposable element activity is able to include RNA sequences in the genome by retroposition[26] and can shuffle the structure of the genome and sequences containing genes or parts of genes. Further mechanisms, ie, all mechanisms that are able to insert any sequence in the genome, can lead to the formation of novel genes indirectly with various degrees of implication. Sequences can be inserted at proximity of present regulatory elements, which

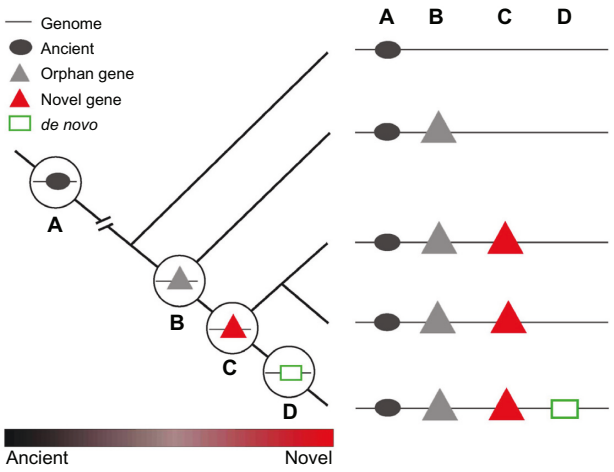

Figure 1. Definition of novel genes based on phylogeny. Circles in the tree show gain of orphan (gray), novel (red), and de novo (green) genes. (**A**) Ancestral genome, ancestral genes are widely spread on the phylogenetic tree. (**B**) Ancestral acquisition of a gene, the orphan character is defined by its uniqueness among all currently known sequences. (**C**) Acquisition of a gene, the novel character is defined by hitherto absence on a considered phylogeny and is attributed by experts of the local taxonomy. (**D**) A genomic fragment gains transcriptional activity, constituting a de novo gene.

enables their transcription. Reciprocally, regulatory motifs can be inserted in the vicinity of untranscribed regions or rewire the transcriptional response.[27] One example of such an indirect supporter of gene emergence is horizontal gene transfer, which results in the transfer of a gene or parts of a gene from one organism (eg, a virus) to another organism. Horizontal gene transfer has been shown to play important roles in prokaryotes,[28] but is far less characterized in eukaryotes.[29,30]

Different outcome scenarios are possible for genes impacted by a duplication or translocation event. (a) Duplicated genes can undergo a subfunctionalization process where the original function of a gene is separated into subfunctions. Both genes are needed to provide the original function.[31] (b) Neofunctionalization describes a process where one duplicated gene evolves and acquires a new function that might lead to its classification as a novel gene.[31] The original gene is still able to fulfill its original function while the duplicate can accumulate mutations under relaxed selective pressure. (c) Another common case is gene fusion, denoted by the emergence of a new gene by joining two neighboring genes followed by intergenic splicing.[24] (d) An existing gene can be extended with nongenic DNA through the loss of a stop codon or modification of splicing sites. (e) Novelty can also encompass the expression profile of the genes when regulatory elements are modified. It is possible that novel features make use of the existing regulation from neighboring genes or benefit from steady transcription happening in peculiar genomic regions.[32]

Novel or mutated sequences can have profound implications, such as lowered selection pressure. Signaling reprogramming can occur when a novel gene product disturbs

the network of protein interactions responsible for signal transmission or transduction. Newly transcribed sequences or changes in regulation can alter protein dosage, enzyme activity, or specificity. Selection can also act toward noncoding RNA genes, ie, not translated into proteins, *trans*-regulatory elements, or de novo gene creation. Noncoding RNA genes, historically more complicated to study, gained interest upon availability of computational methods and power to achieve predictions. The contribution of functional noncoding RNA genes, often understated in genome projects, might nevertheless encompass a wide panel of functions: structural contribution, protein complex formation or stabilization, catalysis, regulation, immune defense, protein synthesis, or self-propagating elements like retrotransposons. The number of characterized noncoding RNA genes is constantly increasing, most of them were linked to regulatory functions.[33]

The distinction whether a gene is newly derived from an ancient gene after a duplication process or evolved de novo can be answered by the evolutionary process involved. Duplicated, or shuffled, genes that use existing regulatory elements of the parental gene and share the classical features of gene structure might be detected with conventional methods of gene prediction. Though novel genes that emerged after duplication have not evolved out of a noncoding sequence, they might have diverged beyond the level of recognition as paralogs, resulting in their false classification as de novo. Buljan et al.[24] showed fusion of exons from neighboring genes as the most common mechanism of novel genes that emerged after a duplication process.

New genetic material can also be created fragment-wise as protein domains and such create a novel gene by modification of the old domain layout. Domains are evolutionary or structurally conserved units that can be used as *building blocks* for proteins[34] and can subsequently be integrated into proteins. There are, at least, three advantages regarding the annotation of novel genes by using a domain-centric view on proteins as follows. (a) Many novel genes are not derived from previously noncoding sequences, but from rearrangements or reuse of existing protein domains. This modular use of domains can explain many novel proteins when an ancient domain is found in the gene. The domain might also be found within a context of low sequence similarity to other genes that include that domain. (b) Protein domains are able to gain multiple copies in a genome and a single novel domain can enable the emergence of multiple novel genes in a short time span.[34] (c) Protein domains are usually described by probabilistic models, namely, hidden Markov models (HMMs). The search with HMMs is far more sensitive than it would be directly with amino acid sequences. Furthermore, breaking down genes into domains is also a more robust way to explain novel genes and compare them with the existing ones based on domain content. A domain-centric view on gene or exon shuffling leads to a similar finding of domain fusion as the main factor of novel domain arrangements, followed by fission of a domain arrangement into two distinct arrangements.[35,36]

De novo gene creation, in contrast, can be defined as the emergence of regulated transcription of a hitherto untranscribed DNA fragment. The acquiring of transcription might be achieved when a promoter sequence is newly created by mutations or inserted from DNA rearrangement events. Two competing models propose mechanisms for the de novo emergence of protein-coding genes. In the *transcription-first* model, a DNA sequence is transcribed to RNA. The now genetic DNA can acquire an ORF afterward under evolutionary pressure.[10] An ORF of a certain length is not needed before transcription occurs in the transcription-first model. The second model that is considered here is the evolution of new genes through *proto-ORFs*.[37] The first step in the proto-ORF model is an ORF that is occasionally transcribed afterward by acquiring regulatory elements. Noncoding RNA genes might arise in an analogous way to the two models for protein-coding genes. In the case of noncoding RNA genes, conservation is not given for an amino acid sequence or protein structure, but for the secondary structure of the RNA sequence. De novo genes, in comparison to ancient genes, tend to share common properties such as a fast evolution, shortness, and fewer exons,[38] as well as a lower transcription level,[39] a tissue-specific transcription,[40] and a higher abundance of transposable elements. The fast evolution and simplicity of de novo genes supports the concept of initial transcription and translation before a gene needs to gain more complex elements of regulation or splicing.

The functional characterization of novel genes is still at the beginning. Functional annotation of sequences is mostly based on homology to genes of known function and the subsequent annotation with, eg, *Gene Ontology* terms,[41] which is not possible in most cases by the definition of novel genes. The annotation of novel genes with protein domains might be possible, but many domains also lack functional annotation. The remaining possibilities of functional characterization involve experiments such as knock out or knock down of those genes, or the analysis of their expression on different conditions. Li and Wurtele,[42] for example, showed a successful characterization of an *Arabidopsis thaliana* orphan gene by expression in soybean.

Methods to Detect Novel Genes

Biological/analytical methods for gene detection. High-throughput generation of transcriptomic data has been an essential contribution to the annotation of known genes and the detection of novel genes. While assessing the gene expression in a biological sample, transcriptomic methods provide evidences of which portions of the genome are effectively transcribed. The comparison of transcriptomic data with known gene models of an organism can result in the prediction of transcribed sequences hitherto unknown as genes. Here, two major technologies based directly on experimental biological evidence are reviewed: RNA-seq and ribosome profiling.

RNA-Seq consists of the determination of the RNA sequences in a given biological sample coming from cells under normal activity or cells under a controlled stimulus. In the perspective of novel protein-coding gene detection, efforts are put on sequencing mRNA.[43] Several RNA-seq solutions exist. The most popular RNA-seq methods are the *Illumina* platforms, relying on massively parallel sequencing of short DNA fragments. While affordable and offering the possibility of deep sequencing to accurately assess the expression level or mutations, the technology requires computational processing to reconstruct best-effort, full-length RNA sequences and extract the biological sense of the sequences. In the perspective of novel gene detection, the long-read technology from Pacific Biosciences displays valuable merits regarding identification of isoforms and improved sensitivity.[44]

Different methods exist to handle and interpret short reads data, of which mapping of the reads on a reference sequence and de novo assembly of the reads in transcripts are the most advanced. The choice of a method mostly depends on the availability of a reference genome, its quality, completeness, accuracy, and its distance to the studied subject. When available, the genome sequence can be used to place the short fragments in correct order by a mapper.[45] De novo assembly of the RNA sequences, on the contrary, will only use the read overlaps to merge and extend them into continuous consensus sequences. Both strategies have strengths and drawbacks. Mapping methods will likely produce better quality gene models if the genomic sequence of the studied species is available, but will likewise fail to detect signals in individual-specific structural variations. Pure de novo methods will sample isoforms and individual-specific transcripts, but might fail on accurate boundary predictions. Few methods allow a combination and fusion of the two strategies, resulting in overall improved gene models.[46,47]

However, not all the transcripts do necessarily encode a protein,[33] and therefore, novel gene can be missed when only targeting mRNA. The global cell RNA pool, along with protein-coding mRNAs, is populated with several other RNA classes such as ribosomal, transfer, long noncoding, small nuclear, micro, and small interfering RNAs (rRNA, tRNA, lncRNA, snRNA, miRNA, and siRNA, respectively). Though the roles of noncoding RNA now receive a growing interest, the experimental and computational challenges that underlie the prediction of their functions had led researchers to mainly focus on protein-coding genes.

The ribosome profiling technique,[48–51] by directly focusing on RNA fragments protected by ribosome, is well adapted to the detection of ORFs and novel protein-coding genes. The protected fragments are sequenced and a direct reading of which genes are translated is possible. The two techniques, namely, RNA-seq and ribosome profiling, need the use of a read mapper utility such as TopHat[52] to align the sequenced reads against the reference genome. Transcriptional properties inferred with RNA-seq methods can help to identify further novel genes. The overall lower expression of novel genes in general can give hints to hitherto unknown genes, as well as a tissue-specific expression of novel genes in, eg, testis for animals.[39,40]

Another promising technique for completing genome annotation is the use of mass spectrometry (MS). The advances on high-throughput MS have opened a new field termed proteogenomics.[53–55] Proteogenomics typical application consists of mapping short peptides produced by MS techniques to protein sequence databases. The technique and its implications are reviewed in the study by Nesvizhskii.[56] The method of proteogenomics has, for example, been successfully applied in *A. thaliana*,[53] mouse, or human[57] to create a precise annotation of known genes and to discover novel protein-coding genes. Many microbial genomes that lack a high quality of annotation can benefit from proteogenomics to improve gene prediction.[58] Moreover, the impact of proteogenomics can be of significant importance in nonmodel organisms. However, accessing the correctness of a peptide identification is a challenging problem and is highly sensitive to false discovery.[56]

Mixing technologies seems also to be a promising perspective. For example, in their study, Sun et al.[59] used a combination of tandem MS and RNA-seq data to detect ORFs in wrongly annotated noncoding RNA sequences.

All biological data need to be processed computationally, where the mapping of short sequence reads to the genome is the most crucial part.

Computational methods for gene detection. When no biological data are available, algorithms are used to predict genes directly from the genomic sequence. Most genome databases, like *Ensembl*, have their own pipeline for gene annotation,[60] but they are not designed for the detection of novel genes.[61] In this section, a review of the classical gene annotation methodology is provided along with the methods aiming at detecting novel genes.

Gene prediction can be accomplished with different types of information. External genetic data can be used to annotate genomic DNA sequences using candidate genes found by similarity of known features of other species. Homologs may be provided by databases of DNA, cDNA, or amino acid sequences.[62] *Projector*[63] and *GeneWise*[64] are programs that compare and align two related DNA sequences and predict the gene structure of one sequence based on the gene structure of the second sequence, assuming that similar sequences share a similar gene structure. Gene prediction based on homology usually makes use of fast heuristic alignment programs such as *Blast*[65] or *Exonerate*.[66]

Ab initio gene prediction software uses intrinsic properties of the sequence to find genes. Thus, for the prediction of a protein-coding gene, a strong emphasis is put on the detection of an intact ORF. ORF detection can be done with tools such as *getorf* included in *EMBOSS*[67] or *OrfPredictor*.[68] The nucleotide hexamers, or in general k-mers, frequency is a good predictor for coding and noncoding sequences. Accordingly,

k-mer frequency is used by several gene prediction programs, eg, *SORFIND*[69] or *Genview*.[70] HMM approaches with the k-mer composition are used by more recent programs such as *GeneMark*[71] and *Eugene*.[72] Other approaches look for recognizable parts in the sequence. As most protein-coding genes consist of more than one exon, the splice site prediction can be used to determine the exon boundaries of a gene and so the gene structure. An example of a splice site predictor based on multiple sequence alignments is *SPLICEVIEW*,[73] whereas *NetGene2*[74] uses a neural network approach. Splice site prediction can be achieved using RNA-seq data with the *KisSplice*[75] or *FineSplice* tools.[76] The *RSVP* tool predicts splice variants of genes based on a genomic sequence and incorporates information from RNA-seq reads.[77] A comparison of computational effort and dependency to biological data of these tools are given in Table 1.

The *Augustus*[78] program combines both intrinsic and extrinsic methods of gene prediction. Modern gene prediction pipelines like *Maker*[79] or *Ensembl*[80] use a combination of all methods to get the best possible evidences for each predicted gene.

Few properties can give clues about the coding potential of a sequence. The most widely used properties to predict coding potential are the sequence homology to known databases, the presence of a stop codon, and the amino acid composition of a sequence. These properties need to be inferred and processed with statistical analysis or machine learning approaches used by programs such as *CPC*[81] or *CPAT*.[82] Other techniques, based on statistical scores or evolutionary simulation, are also giving promising results to detect the coding potential of a nucleotide sequence such as *PhyloCSF*,[83] *ReEVOLVER*,[84] or the $t_{1/2}$ statistic.[85] However, methods based on evolutionary simulation

Table 1. List of tools and methods used for novel gene prediction.

METHOD	COMP. EFFORT	INPUT	COMMENTS
Primary methods RNA-Seq Mapping	**oo**	r G	RNA-Seq experiment; Transcribed sequences Mapped to a reference genome
RNA-Seq *de novo* assembly	**o o o**	r	RNA-Seq experiment; Transcribed sequences
Ribosome profiling	**o o o**	r	RNA-Seq on ribosomes, likely translated
Proteogenomics	**o o o**	RP	Mass spectrometry of peptides
In silico methods	**o o**	D	Gene structure prediction based on homology
Projector/Genewise			eg, Emboss" getorf, OrfPredictor
Simple ORF finding	**o**	D	Based on Codon usage
SORFIND/Genview	**o**	D	Exon prediction partly based on Hexamer/K-mer frequency
Gene mark	**o**	D	Hidden Markov Model based coding region prediction
Eugene	**o**	D R	Gene prediction pipeline using Hidden Markov Models
SPLICEVIEW	**o**	D	Splice site prediction based on homology
NetGene2	**o o**	D	Splice site prediction using neural network
KissSplice/FineSplice	**o o**	r D	Splice site prediction using primary RNA-seq data
RSVP	**o o**	r D	Splice variants predictor using RNA-seq data
GlimmerHMM/Genescan	**o**	D	Gene structure prediction using *GHMM*
FragGeneScan	**o**	r D G	Gene structure prediction using *HMM* on reads
TWINSCAN/N-SCAN	**o**	D R	Gene structure prediction using *GHMM* and external sequences
CONTRAST	**o o**	D	Gene prediction using machine learning and homology
CPC	**o**	D R A	Coding potential prediction using *SVM*
CPAT/PhyloCSF	**o**	D R A	Coding potential prediction using statistic scores
ReEvolver/	**o**	D	Coding potential prediction based on evolutionary simulation
Maker/Augustus	**o o**	R D A	Gene prediction pipelines
Seg-HCA	**o**	A	Domain prediction based on hydrophobic clusters
Classification methods	**o o**	R D A	Sequence homology searches
Blastp/Exonerate			Classification based on
Domain trees	**o**	D R A	Phylogeny and Parsimony approaches
Comparative genomics	**o o o**	D R A	Gene classification based on clustering by homology
Phylostratigraphy	**o**	D R A	Gene classification based on phylogeny and homology

Note: The computational effort (Comp. effort) and required input types are shown.
Abbreviations: R, RNA sequence; r, RNA-reads; D, DNA sequence; A, amino acid sequence; P, peptides; G, reference genome; Comp. effort, computational effort.

require the targeted potential genes to have accumulated enough mutations to be distinguished from other genes. These methods need improvement to be used in the context of novel genes, because the recent appearance of a gene is not necessary followed by accumulation of mutations.

The first accurate and widely used gene prediction tools were *GENSCAN*[86] for eukaryotes and *GLIMMER*[87] for prokaryotes. *GENSCAN* and the successor of GLIMMER, GLIMMERHMM, detect the presence of a gene in a DNA sequence using only the input genome sequence and a generalized hidden Markov model (GHMM) to define a general gene structure. New methods that use other sources of information, such as sister genomes, multiple sequence alignments (MSA), expressed sequence tags (EST, short sequenced cDNA fragments), or combinations of them, are now outperforming *GENSCAN,* examples for those tools are *TWINSCAN* or *N-SCAN.*[88] *FragGeneScan* is a tool that can predict protein-coding regions from short-read data, using sequencing error models and codon usage with a HMM model.[89] GHMM is a popular machine learning methodology among these tools, but has the disadvantage of needing the development of a statistical model a priori. The *CONTRAST*[90] software, using supervised machine learning algorithms such as support vector machine and conditional random field, has been shown to outperform the previously cited methodologies.

Gene classification. The previously listed methods of gene detection are canonical and not specific to the detection of novel genes. The classification of a gene as novel can only be made based on comparison with other species. After the detection of a gene, one needs to find if this gene is novel by searching for potential homologs. Sequence databases represent current biological knowledge and are used to classify a gene as a novel gene. Most sequence databases are synchronized among each other, but might differ in their treatment of features such as splicing variants, including sources, or just in their composition or meta content. *Refseq* and *Uniprot* are two major curated protein sequence databases, publicly available and maintained, which are suitable for purposes of gene prediction and classification. A common basis for the classification of annotated genes as novel or orphan is the known gene repertoire of a species. Orphans can be detected by homology searches of known genes to gene databases. A gene that lacks homology to any other gene in another clade can be classified as orphan. *NCBIs blastp* program has been shown to be sufficient at the identification of homology[91] based on sequence similarity. Different expectation value (*E*-value) cutoffs for *blastp* are used in literature for the classification of homology, ranging from 10^{-10} to 10^{-3} *E*-value cutoffs[5,91] scale with the size of the used target database and should be adapted to the performed study. A drawback of orphan finding by homology is fast evolving genes, which can be falsely classified as novel by lack of detectable homology to their ancestors. Methods of homology detection with increased sensitivity that are not only based on sequence similarity would help circumvent the problem of hidden homology and support novel gene detection.

Novel gene finding can be computed during species comparison analyses by a clustering of known genes based on orthology.[92] Programs like *OrthoMCL*[93] or *ProteinOrtho*[94] are able to use the annotated gene sets of different species as an input and cluster them into families. Genes that are not clustered with any other gene, ie, that have no orthologous genes in these other species, can be used as potential novel genes. The treatment of paralogs is important when designing a clustering approach and has to be interpreted under the second important consideration: how many and how are manually selected species included in the analysis. The number of genes lacking homology to other genes that are found by a clustering is directly linked to the number and relatedness of the species used for the clustering. Species that are closely related to the species of interest can yield reliable results with more orthologs than more distantly related species; however, more distantly related species with better gene annotation should also be considered.

Phylogenetic approaches can be used after a gene clustering to find orphan genes that are restricted to a clade. A defined set of species, that represents a clade of interest, and a phylogenetic tree of these species, can serve as an input for a phylogenetic analysis. The definition of orthologous genes at branches of the tree and so the time of emergence of orphan genes can be achieved after a clustering by gene homology. *Phylostratigraphy* is a similar approach to estimate the age of genes and finding orphan genes and their point of emergence.[95,96] *Phylostratigraphy* uses a set of species defining outgroups at important ancestral nodes in a phylogenetic tree. The gene content of each node is inferred by comparing the gene content of descendant nodes with that of the corresponding outgroup. Subsequently, the branch where a gene emerged can be assigned and defines its clade and time of emergence.

The decision whether a putative novel gene is actually novel or has been lost in other species might be clarified by the existence of a pseudogene. Pseudogenes can be detected by homology of putative novel genes on the DNA level and might be compared to a database of collected pseudogenes like *Pseudogene.org.*[97] The *PseudoPipe* pipeline predicts pseudogenes in the genome using *Blast* and a clustering algorithm, but is limited to mammalian genomes.[98] Alternatively, the classification might be baffled when the observed loss of the gene ensue from errors or lack of evidence at the genome assembly or gene prediction steps. Such issues can be resolved at a rather small scale by simple experiments such as targeted PCR or targeted resequencing.

Novel genes can also be defined based on their domain content. The coverage of proteins with domains is nowadays usually quite good and ever improving. The amount of proteins annotated with at least one protein domain ranges from ~50% in plants to ~75% in insects.[35,99–101] The abovementioned domain coverage is derived from *PFAM,* one of the most

widely used protein domain databases.[102] Novel genes, in terms of domains, can be engendered by the emergence of a novel domain, and/or the rearrangement of ancient domains. The challenging criteria in detecting genetic novelty reside in the quality of the domain models as it deeply impacts the domain annotation of a protein. To detect and link the appearance of a novel gene to domain rearrangements or the emergence of a novel domain, the domain content of several sister species need to be compared. Comparison with sister species facilitates a better phylum coverage of sequenced genomes and proteomes, which is crucial for the analysis. Protein domains present in sister species are assigned to a phylogenetic tree and the content of the ancestral nodes are predicted by using a method of parsimony reconstruction, such as the *Dollo* parsimony. Dollo parsimony limits the number of times that a character, in this case a domain, can be gained in a tree and therefore is well adapted to the study of domain gains that are supposed to be rare events. The domains present in the ancestral nodes can be used to detect the time frame of emergence of novel domains and genes with a program like *ProteinHistorian*.[103] It has been shown that different mechanisms can enable the origination of new domain arrangements, such as domain gain or loss, fusion, or fission. However, the comparative analysis to reliably determine the presence of new domain arrangements requires genomes of high quality, good domain annotations, and reliable phylogenetic trees. Domain annotations are using HMM models that are based on MSAs of parts of protein sequences. Due to their intrinsic definition, a novel sequence or a de novo sequence is less likely to have known sequences to align to in an MSA and subsequently less likely to be annotated by a HMM model than a well-studied sequence. Therefore, the detection of novel domains only based on MSA is limited to clades with either a high number of species or to genes with many orthologs or paralogs.

Other methodologies can be used to circumvent the intrinsic MSA problem and detect true novel domains.[104] The Seg-HCA method has been proposed to annotate domains on whole proteomes.[105] The method is based on the hydrophobic cluster analysis (HCA) of protein sequences, which uses the hydrophobic pattern of a sequence.[106,107] Seg-HCA discriminates hydrophobic clusters to delineate sequences with domains. These annotated portions of sequences are closely related to the structural definition of a protein domain with the presence of a hydrophobic core.

Discussion

Recently developed sequencing techniques are able to give new biological insights into the biology of species. Sequencing of single strains, either DNA or RNA, can help to detect very recent population-specific genes and transcripts. Zhao et al.[40] found 142 putative de novo candidate genes in six *D. melanogaster* strains with RNA-seq methods. The analysis of population data might enable to find very recent changes in sequences that lead to the creation of genes, as well as support for previously detected putative genes. Further perspectives in sequencing techniques are, for example, sequencing of a single cell, new sequencing devices for long reads or strand-specific sequencing. Single-cell sequencing is a promising technique for sequencing bacterial species that cannot be cultivated or tissues that currently cannot be sequenced with reliable results. Different cells of an individual can be studied to find genes and their possible evolution, for example, in cells involved in the immune system or tumor cells, in which the transcription of genetic elements that correspond to novel genes could be activated in a deleterious way.[108] The enhancement of current gene prediction with antisense transcript, assisted by strand-specific RNA-seq can lead to the detection of further ancient and novel genes.[109] A review of current RNA-seq technologies and future perspectives is provided in Ref. 43. The *MinION*, a small novel sequencing device by Oxford Nanopore, is theoretically able to produce sequence reads as long as the underlying DNA or RNA chain.[110] Longer sequence reads, as produced by novel techniques such as the MinION or Pacific Biosciences technologies,[111] can help to improve mapping quality and circumvent problems with repeated sequences or structural variations in the genome, leading to better evidence for gene predictions. The sequencing of more closely related species will also help in the classification of novel genes.

The study of novel genes should not be limited to novel protein-coding gene. The minimal size of genetic features, either protein-coding genes or RNA genes, has been reevaluated in the past few years. Genes, previously thought to need a minimal size of 200–300 base pairs, have been shown to be functional as smaller units such as small secreted proteins acting as *trans*-effectors, antimicrobial peptides, and small RNAs. Several small types of peptides have been assigned to different functional profiles, for example, enriched in signaling or regulation.[112] The loss of minimal size requirements and the finding of other than protein-coding genes highly affects the prediction of novel genes, as, for example, a minimum gene length or presence of an ORF are not crucial properties of genes. De novo genes might lack recognizable regulatory elements, depending on their current state of evolution.

Additionally, gene structure is not always canonical and can hold or miss features that are currently used for gene prediction. Several RNA transcripts might be reliably mapped to the same genomic location that can be the result of different mechanisms. While alternative splicing leads to more than one transcript per gene, usually of different length, it is possible that two genes overlap on the DNA level, either on the same or the opposite strand. The phenomenon of overprinting describes the case when a gene is "embedded" in another one, accomplished by different start and stop sites in a genetic sequence.[113] Other cases of diffuse gene structure might be caused by differently fused genes or genes in different reading frames. The mapping of transcript evidence, as well as the identification of gene boundaries, is getting even more complex through genes where one transcript contains

more than one gene (polycistronic genes). Polycistronic genes are common in prokaryotes, but have also been reported in eukaryotic genomes.[114] The occurrence of different transcripts of the same DNA sequence makes the definition of a gene fuzzy and interferes with its prediction. Interpretation of transcriptomic data has to take all the cases mentioned above into account while constructing a reliable gene structure.

The detection of novel genes is as diverse as the processes leading to their emergence. The classification of a gene as novel, defined by missing sequence homology to other known genes outside the lineage where it emerged, highly depends on the sequence alignment tools that are used to detect homology and parameters of these tools, and even when a gene is classified as novel, its mechanism of emergence is not easily accessible.

Novelty at the domain level needs to be considered in gene prediction as it appears that many genes evolved by recombination of ancient domains.[115] The impact of novel domains on functional innovation and changes in phenotypes are a promising research field for novel gene prediction and understanding evolution. Computational methods that use evolutionary reconstruction, statistical models, or machine learning algorithms are showing very promising results for the annotation of genomes. However, de novo detection is still a very challenging problem as sequences from sister species are crucial to perform a confident prediction and time frame events.

In summary, computational methods, predicting novel genes are based on sequence properties, the detection of known elements, and comparison with sister species. Novel computational techniques to predict correct structures for genes with the unusual properties of novel genes are on the onset of development and will open new perspectives.

Conclusions

The analysis of emerging de novo genes only recently became a topic in genomic research. Whereas most novel genes emerged from ancient genetic material, an amount of up to 20% evolved de novo from noncoding intergenic or intronic sequences.[3] Current methodology to detect de novo genes is using conventional gene prediction workflows. The first basis for predicting genes is the intrinsic properties of the DNA sequence itself, ie, the prediction of features that a gene can consist of such as ORFs, known promoters, TFBS, splicing sites, or features held by the sequence like nucleotide or k-mer composition. A second basis for gene prediction is external data. External data include, but are not limited to, comparison with known genes of related species, and can also be new biological data from experiments. A major drawback in the classification of novel genes is the uncertainty of false positive as the prediction will be only as good as the completeness of the set of genes that the novel genes are compared to.

Finding generic gene descriptors, at least for certain species, and adaptations of known methods for gene prediction to those generic descriptors is a great goal for gene annotations, either for ancient or for novel genes. However, known properties of de novo genes often differ from that of ancient ones and are also not necessarily consistent within or across species.

The detection and classification of (novel) genes can be improved in several ways. (1) The ever better taxonomically coverage of important clades enable better comparisons of genomes for the classification of predicted genes as novel. (2) Better sequence comparison tools can help finding hidden homologies to circumvent wrong gene predictions and classifications by combining DNA/RNA and amino acid sequences together with profile-based or machine learning methods. (3) The finding of generic gene descriptors to avoid overfitting of gene prediction methods. (4) The development of further biological techniques to improve identification of specifically transcribed or translated sequences.

Novel genes are thought to impact organisms and their aptitude to adapt by providing, at the population level, a varied set of tinkering and novelties. The prediction of novel genes, the classification of known genes as novel, and possible explanations of mechanisms of emergence are crucial for understanding recent evolutionary traits. Development of new computational and experimental methods is necessary to build atop of the existing knowledge of genomes the tools to unravel the genesis and impact of the novel and de novo genes on species and evolution.

Author Contributions

Wrote the first draft of the manuscript: SK. Contributed to the writing of the manuscript: SK, LM, TB-F. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Lynch M. *The Origins of Genome Architecture*. Sunderland, MA: Sinauer Associates, Inc. Publishers; 2007.
2. Dujon B. The yeast genome project: what did we learn? *Trends Genet*. 1996;12:263–70.
3. Wu DD, Irwin DM, Zhang YP. De novo origin of human protein-coding genes. *PLoS Genet*. 2011;7:e1002379.
4. Ashburner M, Misra S, Roote J, et al. An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the Adh region. *Genetics*. 1999;153:179–219.
5. Ekman D, Elofsson A. Identifying and quantifying orphan protein sequences in fungi. *J Mol Biol*. 2010;396:396–405.
6. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet*. 2009;25:404–13.
7. Chen S, Krinsky BH, Long M. New genes as drivers of phenotypic evolution. *Nat Rev Genet*. 2013;14:645–60.
8. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. *Genome Res*. 2010;20:1313–26.
9. Chen S, Zhang YE, Long M. New genes in *Drosophila* quickly become essential. *Science*. 2010;330:1682–5.
10. Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. De novo ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet*. 2013;9:e1003860.
11. Ohno S. *Evolution by Gene Duplication*. New York, NY: Springer; 1970.
12. Zhang Michael Q. Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet*. 2002;3:698–709.
13. Fogle T. The dissolution of protein coding genes in molecular biology. In: Beurton PJ, Falk R, Rheinberger HJ, eds. *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives*. (Chap. 1). Cambridge: Cambridge University Press; 2000:3–25.

14. Gerstein MB, Bruce C, Rozowsky JS, et al. What is a gene, post-ENCODE? History and updated definition. *Genome Res*. 2007;17:669–81.
15. Stepanova M, Tiazhelova T, Skoblov M, Baranova A. A comparative analysis of relative occurrence of transcription factor binding sites in vertebrate genomes and gene promoter areas. *Bioinformatics*. 2005;21:1789–96.
16. Berget SM, Moore C, Sharp PA. Spliced segments at the 5" terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA*. 1977;74:3171–5.
17. Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5" ends of adenovirus 2 messenger RNA. *Cell*. 1977;12:1–8.
18. Tautz D, Domazet-Loso T. The evolutionary origin of orphan genes. *Nat Rev Genet*. 2011;12:692–702.
19. Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D. Orphans as taxonomically restricted and ecologically important genes. *Microbiology*. 2005;151:2499–501.
20. Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res*. 2003;13:2213–9.
21. Mighell AJ, Smith NR, Robinson PA, Markham AF. Vertebrate pseudogenes. *FEBS Lett*. 2000;468:109–14.
22. Palmieri N, Kosiol C, Schlötterer C. The life cycle of Drosophila orphan genes. *Elife*. 2014;3:e01311.
23. Keese PK, Gibbs A. Origins of genes: "big bang" or continuous creation? *Proc Natl Acad Sci USA*. 1992;89:9489–93.
24. Buljan M, Frankish A, Bateman A. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol*. 2010;11:R74.
25. Bornberg-Bauer E, Alba MM. Dynamics and adaptive benefits of modular protein evolution. *Curr Opin Struct Biol*. 2013;23:459–66.
26. Cordaux R, Udit S, Batzer MA, Feschotte C. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci U S A*. 2006;103:8101–6.
27. Ellison CE, Bachtrog D. Dosage compensation via transposable element mediated rewiring of a regulatory network. *Science*. 2013;342:846–50.
28. Koonin Eugene V, Makarova Kira S, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*. 2001; 55:709–42.
29. Hall C, Brachat S, Dietrich FS. Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. *Eukaryot Cell*. 2005;4:1102–15.
30. Moran NA, Jarvik T. Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science*. 2010;328:624–7.
31. Assis R, Bachtrog D. Neofunctionalization of young duplicate genes in *Drosophila*. *Proc Natl Acad Sci USA*. 2013;110:17409–14.
32. Thomas BJ, Rothstein R. Elevated recombination rates in transcriptionally active DNA. *Cell*. 1989;56:619–30.
33. Mattick John S, Makunin Igor V. Non-coding RNA. *Human Mol Genet*. 2006;15 Spec No:R17–29.
34. Moore AD, Björklund AK, Ekman D, Bornberg-Bauer E, Elofsson A. Arrangements in the modular evolution of proteins. *Trends Biochem Sci*. 2008;33:444–51.
35. Kersting AR, Bornberg-Bauer E, Moore AD, Grath S. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. *Genome Biol Evol*. 2012;4:316–29.
36. Moore AD, Grath S, Schüler A, Huylmans AK, Bornberg-Bauer E. Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. *Biochim Biophys Acta*. 2013;1834:898–907.
37. Carvunis A-R, Rolland T, Wapinski I, et al. Proto-genes and de novo gene birth. *Nature*. 2012;487:370–4.
38. Alba MM, Castresana J. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol Biol Evol*. 2005;22:598–606.
39. Yu D, Shi W, Zhang YE. Underrepresentation of active histone modification marks in evolutionarily young genes. *Insect Science*. 2016;00:1–13.
40. Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science*. 2014;343:769–72.
41. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet*. 2000;25:25–9.
42. Li L, Wurtele ES. The QQS orphan gene of *Arabidopsis* modulates carbon and nitrogen allocation in soybean. *Plant Biotechnol J*. 2015;13:177–87.
43. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genetics*. 2011;12:87–98.
44. Au KF, Sebastiano V, Afshar PT, et al. Characterization of the human ESC transcriptome by hybrid sequencing. *Proc Natl Acad Sci USA*. 2013;110:E4821–30.
45. Engström PG, Steijger T, Sipos B, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods*. 2013;10:1185–91.
46. Haas BJ, Zeng Q, Pearson MD, Cuomo CA, Wortman JR. Approaches to fungal genome annotation. *Mycology*. 2011;2:118–141.
47. Rhind N, Chen Z, Yassour M, et al. Comparative functional genomics of the fission yeasts. *Science*. 2011;332:930–6.
48. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science*. 2009;324:218–23.
49. Ingolia NT, Lareau L, Weissman J. Ribosome profiling of mouse embryonic stem cells reveals complexity of mammalian proteomes. *Cell*. 2012;147:789–802.
50. Michel AM, Baranov PV. Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley Interdiscip Rev RNA*. 2013;4:473–90.
51. Ingolia Nicholas T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet*. 2014;15:205–13.
52. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14:R36.
53. Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP. Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc Natl Acad Sci U S A*. 2008;105:21034–8.
54. Castellana N, Bafna V. Proteogenomics to discover the full coding content of genomes: a computational perspective. *J Proteomics*. 2010;73:2124–35.
55. Krug K, Nahnsen S, Macek B. Mass spectrometry at the interface of proteomics and genomics. *Mol Biosyst*. 2011;7:284–91.
56. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat Methods*. 2014;11:1114–25.
57. Branca Rui MM, Orre Lukas M, Johansson Henrik J, et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat Methods*. 2014;11:59–62.
58. Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM. Non-model organisms, a species endangered by proteogenomics. *J Proteomics*. 2014;105:5–18.
59. Sun H, Chen C, Shi M, et al. Integration of mass spectrometry and RNA-Seq data to confirm human ab initio predicted genes and lncRNAs. *Proteomics*. 2014;14:2760–8.
60. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet*. 2012;13:329–42.
61. Zhang YE, Landback P, Vibranovski M, Long M. New genes expressed in human brains: implications for annotating evolving genomes. *Bioessays*. 2012;34:982–91.
62. Mathe C, Sagot M-F, Schiex T, Rouze P. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*. 2002;30:4103–17.
63. Meyer IM, Durbin R. Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Res*. 2004;32:776–83.
64. Birney E, Durbin R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res*. 2000;10:547–8.
65. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10.
66. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005;6:31.
67. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet*. 2000;16:276–7.
68. Min XJ, Butler G, Storms R, Tsang A. OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res*. 2005;33:677–80.
69. Hutchinson GB, Hayden MR. The prediction of exons through an analysis of spliceable open reading frames. *Nucleic Acids Res*. 1992;20:3453–62.
70. Milanesi L, Kolchanov N. GenViewer: a computing tool for protein-coding regions prediction in nucleotide sequences. In: Proceedings of the 2nd International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis. Singapore: World Scientific Publishing; 1993:573–88.
71. Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res*. 2005;33:W451–4.
72. Schiex T, Moisan A, Rouze P. EugEne: an eukaryotic gene finder that combines several sources of evidence. *Lect Notes Comput Sci*. 2001;2066:11–125.
73. Rogozin IB, Milanesi L. Analysis of donor splice sites in different eukaryotic organisms. *J Mol Evol*. 1997;45:50–9.
74. Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouzé P, Brunak S. Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res*. 1996;24:3439–52.
75. Sacomoto GAT, Kielbassa J, Chikhi R, et al. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics*. 2012;13(Suppl 6):55.
76. Gatto A, Torroja-Fungairiño C, Mazzarotto F, et al. FineSplice, enhanced splice junction detection and quantification: a novel pipeline based on the assessment of diverse RNA-Seq alignment solutions. *Nucleic Acids Res*. 2014;42:e71.
77. Majoros William H, Lebeck N, Ohler U, Li S. Improved transcript isoform discovery using ORF graphs. *Bioinformatics*. 2014;30:1958–64.
78. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*. 2008;24:637–44.
79. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. 2011;12:491.
80. Flicek P, Amode MR, Barrell D, et al. Ensembl 2014. *Nucleic Acids Res*. 2014;42:D749–55.

81. Kong L, Zhang Y, Ye Z-Q, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*. 2007;35:W345–9.
82. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res*. 2013;41:e74.
83. Lin Michael F, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*. 2011;27:i275–82.
84. Dupanloup I, Kaessmann H. Evolutionary simulations to detect functional lineage-specific genes. *Bioinformatics*. 2006;22:1815–22.
85. Zhang J, Webb DM. Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates. *Proc Natl Acad Sci U S A*. 2003;100:8337–41.
86. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. 1997;268:78–94.
87. Delcher Arthur L, Bratke Kirsten A, Powers Edwin C, Salzberg Steven L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*. 2007;23:673–9.
88. Baren Marijke J, Koebbe Brian C, Brent Michael R. Using N-SCAN or TWINSCAN to predict gene structures in genomic DNA sequences. *Curr Protoc Bioinformatics*. 2007;Chapter 4:Unit4.8.
89. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res*. 2010;38:1–12.
90. Gross Samuel S, Do Chuong B, Sirota M, Batzoglou S. CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol*. 2007;8:R269.
91. Alba MM, Castresana J. On homology searches by protein Blast and the characterization of the age of genes. *BMC Evol Biol*. 2007;7:53.
92. Capra JA, Pollard KS, Singh M. Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biol*. 2010;11:R127.
93. Fischer S, Brunk BP, Chen F, et al. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinformatics*. 2011;Chapter 6:Unit 6.12.1–12.19.
94. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*. 2011;12:124.
95. Domazet-Lošo T, Brajković J, Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet*. 2007;23:531–3.
96. Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics*. 2013;14:117.
97. Karro John E, Yan Y, Zheng D, et al. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res*. 2007;35:55–60.
98. Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M. PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics*. 2006;22:1437–9.
99. Barrera A, Alastruey-Izquierdo A, Martín MJ, Cuesta I, Vizcaíno JA. Analysis of the protein domain and domain architecture content in fungi and its application in the search of new antifungal targets. *PLoS Comput Biol*. 2014;10:e1003733.
100. Moore Andrew D, Bornberg-Bauer E. The dynamics and evolutionary potential of domain loss and emergence. *Mol Biol Evol*. 2012;29:787–96.
101. Ekman D, Björklund AK, Elofsson A. Quantification of the elevated rate of domain rearrangements in metazoa. *J Mol Biol*. 2007;372:1337–48.
102. Punta M, Coggill PC, Eberhardt RY, et al. The Pfam protein families database. *Nucleic Acids Res*. 2012;40:D290–301.
103. Capra JA, Williams AG, Pollard KS. ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. *PLoS Comput Biol*. 2012;8:el002567.
104. Bitard-Feildel T, Heberlein M, Bornberg-Bauer E, Callebaut I. Detection of orphan domains in *Drosophila* using "hydrophobic cluster analysis". *Biochimie*. 2015;119:244–53.
105. Faure G, Callebaut I. Comprehensive repertoire of foldable regions within whole genomes. *PLoS Comput Biol*. 2013;9:el003280.
106. Gaboriaud C, Bissery V, Benchetrit T, Mornon JP. Hydrophobic cluster analysis: an efficient new way to compare and analysis amino acid sequences. *FEBS Lett*. 1987;224:149–55.
107. Callebaut I, Labessea G, Duranda P, et al. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci*. 1997;53:621–45.
108. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet*. 2013;14:618–30.
109. Levin JZ, Yassour M, Adiconis X, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods*. 2010;7:709–15.
110. Quick J, Quinlan AR, Loman NJ. Erratum: a reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *Gigascience*. 2015;4:6.
111. Zhang X, Davenport KW, Gu W, et al. Improving genome assemblies by sequencing PCR products with PacBio. *Biotechniques*. 2012;53:61–2.
112. Bazzini AA, Johnstone TG, Christiano R, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J*. 2014;33:981–93.
113. Veeramachaneni V, Makałowski W, Galdzicki M, Sood R, Makałowska I. Mammalian overlapping genes: the comparative perspective. *Genome Res*. 2004;14:280–6.
114. Tautz D. Polycistronic peptide coding genes in eukaryotes – how widespread are they? *Brief Funct Genomic Proteomic*. 2009;8:68–74.
115. Toll-Riera M, Alba MM. Emergence of novel domains in proteins. *BMC Evol Biol*. 2013;13:47.