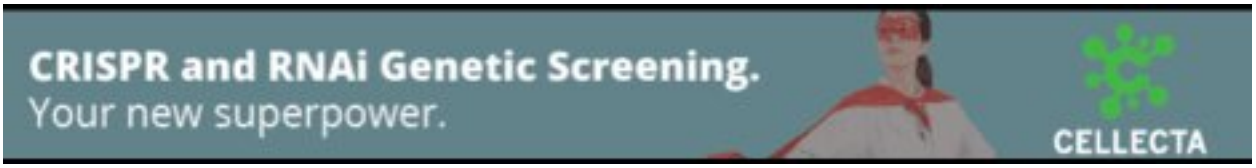
Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship

Marie A. Brunet, Sébastien A. Levesque, Darel J. Hunting, et al.

Genome Res. 2018 28: 609-624 originally published online April 6, 2018
Access the most recent version at doi:10.1101/gr.230938.117



Perspective

Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship

Marie A. Brunet,[1,2,3] Sébastien A. Levesque,[4] Darel J. Hunting,[5] Alan A. Cohen,[2] and Xavier Roucou[1,3]

[1]*Biochemistry Department, Université de Sherbrooke, Quebec J1E 4K8, Canada;* [2]*Groupe de recherche PRIMUS, Department of Family and Emergency Medicine, Quebec J1H 5N4, Canada;* [3]*PROTEO, Quebec Network for Research on Protein Function, Structure, and Engineering, Université Laval, Quebec G1V 0A6, Canada;* [4]*Pediatric Department, Centre Hospitalier de l'Université de Sherbrooke, Quebec J1H 5N4, Canada;* [5]*Department of Nuclear Medicine & Radiobiology, Université de Sherbrooke, Quebec J1H 5N4, Canada*

Technological advances promise unprecedented opportunities for whole exome sequencing and proteomic analyses of populations. Currently, data from genome and exome sequencing or proteomic studies are searched against reference genome annotations. This provides the foundation for research and clinical screening for genetic causes of pathologies. However, current genome annotations substantially underestimate the proteomic information encoded within a gene. Numerous studies have now demonstrated the expression and function of alternative (mainly small, sometimes overlapping) ORFs within mature gene transcripts. This has important consequences for the correlation of phenotypes and genotypes. Most alternative ORFs are not yet annotated because of a lack of evidence, and this absence from databases precludes their detection by standard proteomic methods, such as mass spectrometry. Here, we demonstrate how current approaches tend to overlook alternative ORFs, hindering the discovery of new genetic drivers and fundamental research. We discuss available tools and techniques to improve identification of proteins from alternative ORFs and finally suggest a novel annotation system to permit a more complete representation of the transcriptomic and proteomic information contained within a gene. Given the crucial challenge of distinguishing functional ORFs from random ones, the suggested pipeline emphasizes both experimental data and conservation signatures. The addition of alternative ORFs in databases will render identification less serendipitous and advance the pace of research and genomic knowledge. This review highlights the urgent medical and research need to incorporate alternative ORFs in current genome annotations and thus permit their inclusion in hypotheses and models, which relate phenotypes and genotypes.

The now irrefutable existence of "alternative" proteins

Recent work has revealed that genomes harbor many non-annotated open reading frames (ORFs) (Vanderperre et al. 2011; Bergeron et al. 2013; Anderson et al. 2015; Mouilleron et al. 2016; D'Lima et al. 2017; Plaza et al. 2017). Although two decades have passed since the first eukaryotic genome was sequenced, assigning translated ORFs to genetic loci remains a daunting task (Basrai et al. 1997; Claverie et al. 1997; Ladoukakis et al. 2011). Indeed, current genome annotations rely partly on ORF prediction algorithms that are only reliable for sequences beyond a certain length. Consequently, three main criteria are implemented to distinguish "true" ORFs from random events: the use of an ATG start codon, a minimum length of 100 codons, and a limit of a single ORF per transcript (Cheng et al. 2011; Andrews and Rothnagel 2014; Saghatelian and Couso 2015; Plaza et al. 2017). These criteria result in an important underestimation of translated ORFs in the genome (Andrews and Rothnagel 2014; Saghatelian and Couso 2015; Couso and Patraquim 2017; Plaza et al. 2017). With functional evidence for previously unannotated ORFs in bacteria (Wadler and Vanderpool 2007; Hemm et al. 2008, 2010; Storz et al. 2014; Lluch-Senar et al. 2015; Baek et al. 2017), *Drosophila* (Galindo et al. 2007; Kondo et al. 2007; Reinhardt et al. 2013; Aspden et al. 2014; Albuquerque et al. 2015; Li et al. 2016a; Pueyo et al. 2016a), plants (Hanada et al. 2013; Juntawong et al. 2014; Hsu et al. 2016; Hsu and Benfey 2017), and other eukaryotes (Oyama et al. 2007; Ingolia et al. 2011; Vanderperre et al. 2013; Ma et al. 2014), genome annotations will need to be revised.

These "hidden" ORFs are found in multiple places within RNA: within long noncoding RNAs (lncRNAs), within 5′ and 3′ "untranslated" regions (UTRs) of mRNAs, or overlapping canonical coding sequences (CDSs) in an alternative reading frame (Slavoff et al. 2013; Mouilleron et al. 2016). They are, in general, notably smaller than annotated CDSs, but they are not limited to small ORFs (smORFs—ORFs smaller than 100 codons) (Samandi et al. 2017). Here, we define alternative ORFs as any coding sequence with an ATG start codon encoded within any reading frame of either lncRNAs or known coding mRNAs (either in UTRs or overlapping the CDS). Such a definition of ORFs allows for a more exhaustive yet more complex view of the genomic landscape. As shown in Figure 1A, a gene inherently carries transcriptomic complexity (RNA splicing events leading to multiple

Corresponding authors: xavier.roucou@usherbrooke.ca, marie.brunet@usherbrooke.ca

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.230938.117. Freely available online through the *Genome Research* Open Access option.

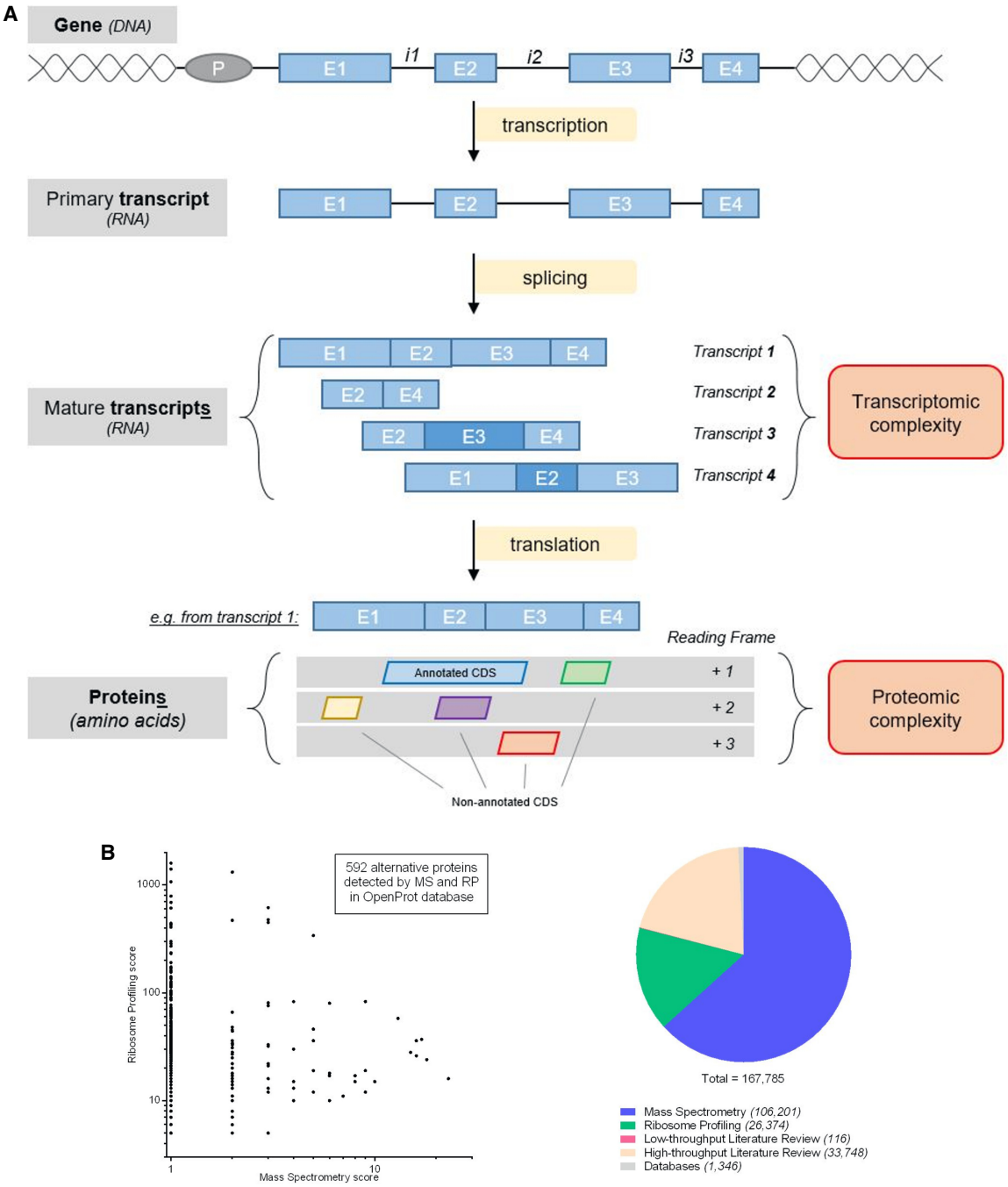


Figure 1. Schema of the transcriptomic and proteomic complexity inherent to a gene. (*A*) Genomic complexity representation. A gene is represented with a promoter (P) and introns (i) and exons (E). Splicing events lead to a suite of transcripts with frameshifted exons (darker blue shade), skipped exons, or retained introns. Then, proteomic complexity comes from each transcript with ORFs from any reading frames. However, now only one CDS is annotated per transcript, leaving an entire hidden proteome (unannotated CDSs). (*B*) Alternative ORFs databases. The OpenProt database predicts every ORF longer than 30 codons and reports experimental detection evidence for each of them. Five hundred ninety-two alternative ORFs were detected by both ribosome profiling (RP) and mass spectrometry (MS). The SmProt database reports smORFs (<100 codons) in different data sets (mass spectrometry, ribosome profiling, literature mining, and databases).

transcripts and thus to a suite of isoforms) and proteomic complexity (more than one protein per transcript). Even though the transcriptomic complexity is now widely accepted, consideration of proteomic complexity is usually restricted to one protein (and its splicing-derived isoforms) per gene. Protein complexity can arise from multiple sources, such as RNA splicing and editing, post-translational modifications, alternative initiation (internal ribosome entry site), stop codon read-through, or non-AUG initiation (Dunn et al. 2013; Venne et al. 2014; Ingolia 2016; Nishikura 2016; Blencowe 2017; Li et al. 2018). Notwithstanding, this review will focus on the proteomic complexity resulting from proteins encoded in alternative ORFs.

Recently, the development of new techniques or the optimization of existing ones has allowed for large-scale detection of alternative proteins and a more in-depth view of a cell proteomic landscape (Boekhorst et al. 2011; Aspden et al. 2014; Calviello et al. 2016; Hellens et al. 2016; Ma et al. 2016; Pueyo et al. 2016a; Delcourt et al. 2017; Hsu and Benfey 2017; Willems et al. 2017). Three detection methods—ribosome profiling, proteogenomics, and conservation signatures—have been used to identify likely translated and functional alternative ORFs, as further discussed later in this review. Such experimental data have been compiled in the sORFs repository, SmProt, and OpenProt databases (Olexiouk et al. 2016; Hao et al. 2017; openprot.org). The SmProt database reports 167,785 small proteins (mostly from lncRNAs) in the human genome, including 106,201 identified via mass spectrometry data and 26,374 via ribosome profiling (Fig. 1B; Hao et al. 2017). Comparatively, the OpenProt database currently reports 28,007 alternative proteins with experimental evidence, including 20,919 detected by mass spectrometry and 7680 by ribosome profiling; 592 alternative proteins were detected in both mass spectrometry and ribosome profiling experiments (Fig. 1B; openprot.org).

The number of reported alternative proteins varies between the databases, as they uphold different definitions of alternative ORFs (start codon other than ATG, length threshold, number of studies analyzed, and pipeline stringency). Indeed, the start codon use (restricted to ATG or not) and length threshold (<100 codons, or >30 codons) implemented will significantly alter the number of predicted alternative proteins. Subsequently, this will affect the sensitivity and specificity of detection methods, especially if the mass spectrometry identification pipeline is not adapted to an increase in the search space (Guthals et al. 2015). The various databases enforce different identification pipelines and stringencies on re-analysis of published data sets. This would inevitably lead to discrepancies in numbers of detected alternative ORFs. Here, we advocate for a cautious interpretation of the data, encouraging identifications made with different techniques and across several studies (identifications from mass spectrometry and ribosome profiling) (Fig. 1B).

These detection methods—ribosome profiling, proteogenomics, and conservation signatures—are revealing this hidden proteome, a novel repertoire for biomarkers and therapeutic strategies (Couso and Patraquim 2017; Karginov et al. 2017; Plaza et al. 2017). Here, we briefly review functional evidence for the biological roles of alternative proteins.

SmORFs: mRNA or lncRNA ORFs smaller than 100 codons

The field of smORFs is rapidly expanding. With the implementation of large-scale proteogenomics and ribosome profiling studies for smORF detection, their discovery is becoming less serendipitous (Ma et al. 2016; Delcourt et al. 2017; Willems et al. 2017). One of the first and most striking examples is that of the apelin (*APELA* smORF), 58 amino acids (aa), shown to bind apelin receptors (Pauli et al. 2014). Since then, even smaller apelin variants have been discovered (Huang et al. 2017). All isoforms originate from a 77-aa precursor, pre-proapelin (Fig. 2A; Lee et al. 2005; Castan-laurell et al. 2012). Apelin stimulates several metabolic pathways, such as glucose uptake, mitochondrial biogenesis, and fatty acid oxidation, while inhibiting lipolysis and insulin secretion (Boucher et al. 2005; Dray et al. 2008; O'Carroll et al. 2013; Alfarano et al. 2015; Bertrand et al. 2015). Rapidly, apelin went from an overlooked ORF in a lncRNA to a promising biomarker and therapeutic target in cardiovascular diseases, diabetes, and diabetic complications (Castan-Laurell et al. 2011, 2012; O'Carroll et al. 2013; Hu et al. 2016; Huang et al. 2018). Elevated apelinemia was found in obese patients across several studies, which was suggested to be a compensatory mechanism prior to insulin resistance (Boucher et al. 2005; Heinonen et al. 2005, 2009; Li et al. 2006; Castan-Laurell et al. 2008; Dray et al. 2008, 2010; Erdem et al. 2008; Soriguer et al. 2009; Telejko et al. 2010). Both short- and long-term apelin treatments in insulin-resistant obese mice were proven to improve insulin sensitivity (Dray et al. 2008; Castan-Laurell et al. 2011; Hu et al. 2016). *APELA* annotation has now changed from lncRNA (GRCh37) to mRNA (GRCh38), highlighting the dynamic nature of genome annotations (Delcourt et al. 2017). Other biological roles attributed to smORFs include sarcoendoplasmic reticulum calcium transport ATPase (SERCA) machinery regulation, regulation of ribosome-protein complexes, prevention of cell death, and regulation of transcription (Hashimoto et al. 2001; Galindo et al. 2007; Kondo et al. 2007; Hanyu-Nakamura et al. 2008; Escobar et al. 2010; Magny et al. 2013; Anderson et al. 2015; Pueyo et al. 2016b; D'Lima et al. 2017; Matsumoto et al. 2017a, b). Moreover, smORFs have been detected within the mitochondrial genome, encoding short circulating peptides acting in a hormone-like manner (Yen et al. 2013; Lee et al. 2016; Kim et al. 2017; Okada et al. 2017).

These multiple reports of smORFs, often encoded in lncRNAs, highlight the previously hidden coding potential of lncRNAs (Niazi and Valadkhan 2012; Slavoff et al. 2013; Ruiz-Orera et al. 2014; Ji et al. 2015). Admittedly, not all lncRNAs are misannotated, and evidence that these transcripts act as functional RNAs rather than protein coding RNAs is not to be dismissed (Guttman et al. 2013).

Upstream ORFs: ORFs encoded in the 5′ UTR of mRNAs

Advances in large-scale ribosome profiling led to the discovery of widespread translation events outside of annotated CDS (Ingolia 2014, 2016). A large portion of these events were observed upstream of annotated CDS, in the 5′ UTR (Ingolia et al. 2009). Translation of these upstream ORFs (uORFs) was first described as a regulatory mechanism for the translation machinery. Indeed, several examples show that mutations creating or suppressing an uORF led to a decrease or increase in the downstream canonical protein expression (Cabrera-Quio et al. 2016). One of the best studied examples of protein expression regulation by uORF translation is that of the GCN4 protein (Natarajan et al. 2001). The *GCN4* transcript contains four uORFs that ensure a tightly regulated expression of the CDS, a transcription factor. GCN4 protein targets most genes required for amino acid biosynthesis (Natarajan et al. 2001). Upon starvation, translation re-initiation at the multiple uORFs is down-regulated and the GCN4 protein expression level thus rises (Hinnebusch 2005; Gunišová et al. 2016). Multiple examples of uORF-mediated regulation of protein levels have been published; however, numerous studies also highlight the biological role of uORF-encoded peptides (Lee et al. 2014; Cabrera-Quio et al. 2016; Plaza et al. 2017). In 2004, a proteomics study detected 54 novel microproteins mapped back to uORFs (Oyama et al. 2004) and 40% of identified smORF-encoded peptides (SEPs) were from uORFs in Slavoff et al. (2013). At least two of these uORF peptides were shown to be functional proteins (on *SLC35A4* and *MIEF1* transcripts), and several others are conserved and likely to be of biological importance (Vanderperre et al. 2013; Andreev et al. 2015; Ebina et al. 2015; Young and Wek 2016). The *SLC35A4* transcript was shown to be resistant to stress (sodium arsenite), and uORF-

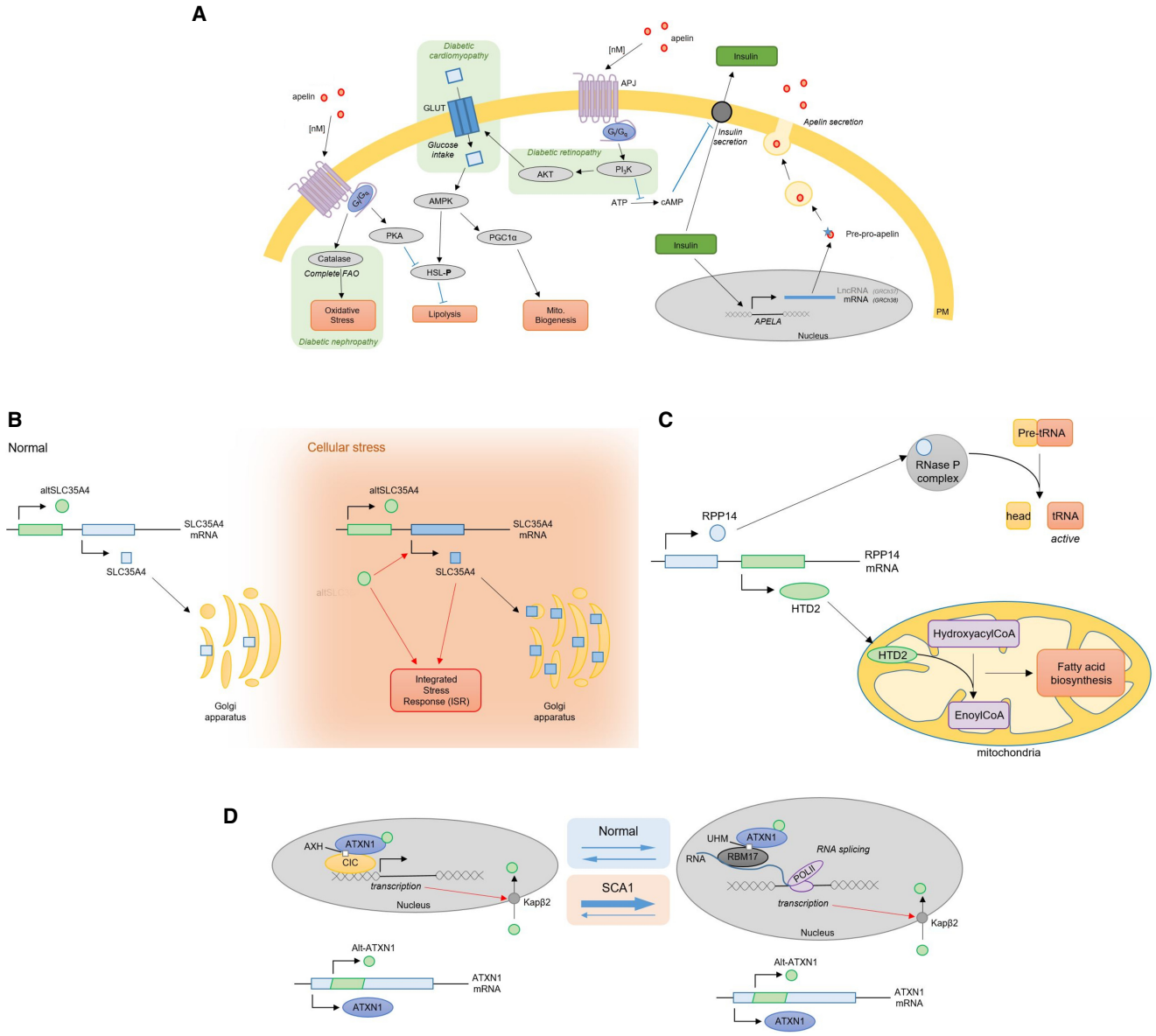


Figure 2. Examples of biologically important alternative ORFs. (*A*) Apelin, from overlooked to metabolic regulator. Apelin is encoded in an mRNA (GRCh38), previously annotated lncRNA (GRCh37), and subsequently secreted. Upon binding with APLNR (also known as APJ) receptor at a nanomolar range, it stimulates different metabolic pathways (glucose uptake, fatty acid oxidation, and mitochondrial biogenesis) and inhibits others (lipolysis and insulin secretion). These pathways are also involved in diabetic complications (cardiomyopathy, nephropathy, and retinopathy). Blue arrows represent inhibitory relationships, pathways involved in diabetic complications are highlighted in green. FAO: fatty acid oxidation; PM: plasma membrane; lncRNA: long non-coding RNA; Mito: mitochondrial. (*B*) SLC35A4 and its uORF-encoded protein, alt-SLC35A4. The SLC35A4 mRNA encodes two ORFs. Under physiological conditions, the canonical ORF, SLC35A4, is weakly expressed. The uORF-encoded protein alt-SLC35A4 is suspected to be the major protein product. Under cellular stress, both proteins are expressed. The alt-SLC35A4 expression level remains unchanged but positively regulates expression of SLC35A4. Both proteins are thought to be involved in the integrated stress response. ISR: integrated stress response. (*C*) RPP14 and its dORF-encoded protein, HTD2. The *RPP14* mRNA encodes two ORFs. The canonical ORF encodes a member of the ribonuclease P (RNase P) complex (RPP14) involved in tRNAs maturation. In the 3′ UTR, a second ORF encodes a mitochondrial dehydroxylase, HTD2. HTD2 is involved in mitochondria fatty acid synthesis. (*D*) *ATXN1* is a dual coding gene. *ATXN1* mRNA encodes two proteins, ataxin and alt-ataxin. Upon entry into the nucleus, ataxin binds the transcription factor capicua (CIC) and associates with DNA at transcription sites. Ataxin nuclear localization and transcription are necessary for alt-ataxin nuclear import and its interaction with ataxin in nuclear inclusions. Ataxin is thought to shuttle between CIC complexes and RNA-binding RBM17 complexes. Polyglutamine extensions in ataxin are responsible for spinocerebellar ataxia type 1 (SCA1) and alter the dynamics of ataxin localization, thereby altering gene expression.

encoded alt-SLC35A4 was shown to positively regulate SLC35A4 protein translation in the context of the integrative stress response (Fig. 2B). Alt-SLC35A4 expression levels remained unchanged following sodium arsenite treatment (Andreev et al. 2015; Ma et al. 2016).

Downstream ORFs: ORFs encoded in the 3′ UTR of mRNAs

Targeted proteomics for small peptides has also increased the detection of proteins mapped back to 3′ UTR ORFs, downstream from an annotated CDS (dORFs). Sixteen percent of the identified

SEPs in Slavoff's study were from dORFs (Slavoff et al. 2013). To the best of our knowledge, only one 3′ UTR encoded protein has been functionally characterized thus far (Autio et al. 2008). HTD2 (hydroxyacyl-thioester dehydratase type 2) is localized on *RPP14* mRNA, downstream from the canonical sequence (Fig. 2C). RPP14 is a component of the ribonuclease P (RNase P) complex necessary for tRNA maturation, while HTD2 is a mitochondrial protein involved in mitochondrial fatty acid biosynthesis (Autio et al. 2008). Evolutionary analysis of RPP14 and HTD2 sequences highlight a conserved bicistronic relationship over 400 million years and thus suggest a functional link between RNA processing and mitochondrial fatty acid synthesis (Autio et al. 2008). Numerous ribosome profiling and mass spectrometry studies highlight translation events in the 3′ UTR and dORF-encoded peptide detection (Slavoff et al. 2013; Ingolia 2016). For example, the OpenProt database reports 5180 predicted dORFs detected by mass spectrometry and 535 by ribosome profiling, including 41 detected by both techniques (openprot.org). The SmProt database reports no dORFs in their mass spectrometry data set but 2389 in their ribosome profiling data set (Hao et al. 2017). The small ORFs repository (sORFs.org) reports 44,163 dORFs detected by ribosome profiling (Olexiouk et al. 2016).

Polycistronic regions: overlapping ORFs on one mRNA

Finally, hundreds of unannotated ORFs overlapping a canonical CDS have been described as well (Vanderperre et al. 2011, 2013; Bergeron et al. 2013; Slavoff et al. 2013). These ORFs are at the same locus as an annotated CDS but encoded in an alternative reading frame and can either partially overlap the CDS or be completely nested within it. Several mammalian polycistronic mRNAs have been reported over the past decade (for review, see Karginov et al. 2017). These overlapping ORFs might be more common than previously thought, given that 30% of peptides from Slavoff's study were mapped back to overlapping ORFs (Slavoff et al. 2013). The OpenProt database reports 4916 alternative proteins from overlapping ORFs detected by mass spectrometry and 3756 by ribosome profiling, including 268 detected by both techniques (openprot.org). For example, the *ATXN1* gene, involved in spinocerebellar ataxia type 1, was identified as a dual coding gene (Fig. 2D). The canonical gene product, ataxin, is a chromatin-binding factor and is thought to have a role in RNA metabolism (Yue et al. 2001). The alternative protein product, alt-ataxin, directly interacts with ataxin and poly(A)$^+$RNAs (Bergeron et al. 2013). Polyglutamine extensions in ataxin are responsible for spinocerebellar ataxia type 1 (SCA1). Normally, upon entry into the nucleus, ataxin binds the transcription factor capicua (CIC). Ataxin-CIC complexes then associate with DNA at transcription sites (Lim et al. 2008). Alt-ataxin is diffusely localized in the nucleus in the absence of ataxin, but in the presence of ataxin it readily colocalizes in nuclear inclusions (Bergeron et al. 2013). Ataxin is thought to equilibrate between CIC complexes and RNA-binding RBM17 complexes, which regulates transcription and RNA processing, notably splicing. In the case of SCA1, the polyQ extensions favor ataxin-RBM17 complexes over those with CIC, thereby competing with CIC containing complexes and altering gene transcription (Lim et al. 2008; Paulson et al. 2017).

As illustrated by this last example, proteins encoded within the same mRNA often share a functional link. Most fall into three categories: (1) a direct protein interaction, either in a complex or as a chaperone (Quelle et al. 1995; Bergeron et al. 2013); (2) a positive functional interaction (involved in the same pathway but at distinct points and expression levels) (Abramowitz et al. 2004); and (3) a negative functional interaction (two proteins involved in the same pathway, with opposite roles) (Lee et al. 2014).

The clinical and research need for a better annotation system

This growing body of evidence for functional alternative ORFs calls attention to the need for a novel genome annotation (Yen et al. 2013; Pauli et al. 2014; Anderson et al. 2015; Lee et al. 2016; D'Lima et al. 2017; Huang et al. 2017). Indeed, the current overly restrictive definition of a gene inhibits research and clinical advances. The field is facing a vicious cycle phenomenon: Most alternative ORFs are not identified as new genetic drivers or pathological causes since they are not annotated. Yet, genome annotations, faced with the challenge of distinguishing functional ORFs from random events, do not include alternative ORFs. That is because of a lack of clinical importance and/or experimental evidence for alternative ORFs (Couso and Patraquim 2017). However, this paucity of evidence is largely due to their absence from current annotations (Cheng et al. 2011; Ladoukakis et al. 2011; Andrews and Rothnagel 2014; Saghatelian and Couso 2015; Couso and Patraquim 2017).

Genome annotations are the linchpin to today's research and clinical screening, and the practical impact of their incompleteness is thus substantial. With the development of time-efficient, reproducible, and cost effective Next Generation Sequencing (NGS), the amount of genome and exome sequencing data is no longer a major limit for today's clinical screening and research (Boycott et al. 2013; Goodwin et al. 2016). Indeed, an increasing number of genes have been related to pathological germline and somatic mutations since the use of NGS (Vogelstein et al. 2013; Amberger et al. 2015). Yet, only about 35% of exome sequencing tests result in the identification of a likely pathological mutation (Ku et al. 2016). This is partly due to the current recommendations from the American College of Medical Genetics and Genomics (ACMG), which considers likely pathological mutations from a uni-coding dogma point of view (Richards et al. 2015). The uni-coding dogma establishes that one gene encodes one protein and its splicing-derived isoforms. Thus, single nucleotide variants (SNVs) resulting in missense mutations are considered, but those resulting in synonymous mutations are often ignored unless they alter a splicing site or have a known functional consequence (Richards et al. 2015). Admittedly, a challenge coming with such a wealth of data is to distinguish single nucleotide polymorphisms (SNPs), or passenger mutations in cancer, from pathological SNVs (Makrythanasis and Antonarakis 2013; Vogelstein et al. 2013; Tokheim et al. 2016). So far, the response to this dilemma has been to use more stringent criteria for linking SNPs to pathologies, and synonymous mutations are often discarded and regarded as silent mutations under a uni-coding dogma (Nielsen et al. 2011, 2012; Olson et al. 2015). However, a synonymous mutation in one reading frame may be a missense in another and could thereby represent a pathological alteration for an alternative ORF (Fig. 3). In fact, synonymous mutations have been described in several pathologies, from cancer to neurological disorders (Sauna and Kimchi-Sarfaty 2011; Supek et al. 2014; Fahraeus et al. 2016; Li et al. 2016b; Waters et al. 2016; Austin et al. 2017; Batista et al. 2017; Soussi et al. 2017). The mechanisms put forward to explain a pathological outcome from a silent mutation mostly revolve around the stability of the mRNA, its splicing, or the

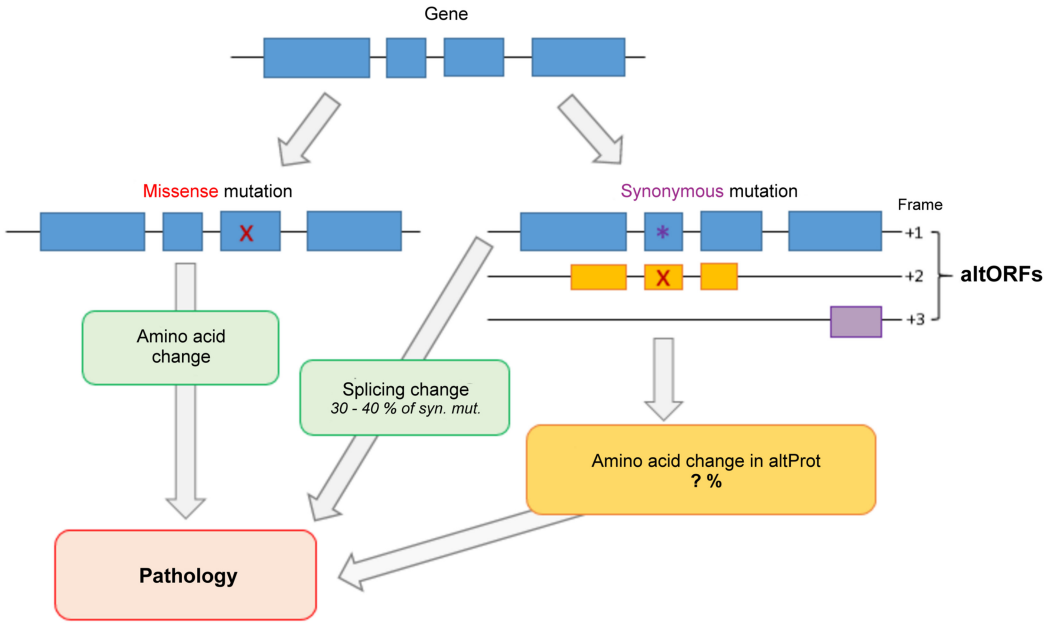


Figure 3. Graphical representation of ways a genetic mutation might cause pathology. Mutations from a single nucleotide variation (SNV) can result either in a missense mutation (red X) or in a synonymous mutation (purple star). Missense mutations are the most studied, as they lead to an amino acid change in the gene's annotated protein sequence. Synonymous mutations are studied mostly for their likelihood to alter splicing sites (about 30% of cases). However, a synonymous mutation in a gene's annotated protein sequence (in blue) might cause an amino acid change in a protein encoded in an alternative open reading frame (altORF/altProt—in yellow). These altered proteins might be a yet unexplored mechanism by which a SNV is pathological.

protein folding (Fahraeus et al. 2016). Yet, even these mechanisms only explain about a third of pathological synonymous SNVs (Supek et al. 2014).

Here, we suggest that alternative proteins might explain these additional pathological SNVs (Fig. 3). A silent mutation in an annotated protein might alter a second protein encoded in an alternative ORF in the same mRNA. The underlying pathological cause could thus be an amino acid change in that second protein, previously "hidden" because it is not annotated in genome databases. As an example, we explored Supek's study from 2014 (Table 1; Supek et al. 2014). In that study, synonymous mutations were identified as drivers in human cancers. For each gene found enriched in synonymous mutations in Supek's study, we gathered synonymous mutations coordinates from the The Cancer Genome Atlas (TCGA) database (The Cancer Genome Atlas Research Network 2013, 2016; Supek et al. 2014; Favazza et al. 2017). In the first data set, 25 oncogenes were found enriched in synonymous mutations in a tissue-specific manner. Synonymous SNVs coordinates from the TCGA database for each specific tissue were checked against genomic coordinates of predicted alternative ORFs (Table 1; openprot.org). Sixty-four percent of genes displayed at least one "synonymous" SNV altering the amino acid sequence of at least one predicted alternative protein. We consider here any type of alterations, be it missense, nonsense, frameshift, or point mutations. Of all listed synonymous SNVs within these 25 genes, in the specific tissues, 29.6% fell within a predicted alternative protein. Out of these, 7% have been detected in ribosome profiling and reanalysis of large-scale mass spectrometry studies (openprot.

Table 1. Overview of alternative ORFs altered by synonymous SNVs in TCGA database for genes of interest

Data set	Genes (#)	Predicted alternative ORFs (#)	Genes with at least one SNV affecting one alternative ORF	% of SNV affecting an alternative ORF *(any type of mutation)*	Median length of alternative ORFs (aa)	Expected by chance
Supek's data set	25	159	16 (*NTRK3*; *MSI2*; *TCF7L2*; *AKT2*; *SMO*; *KIT*; *BCL6*; *ERBB2*; *FGFR2*; *RUNX1*; *MLLT6*; *RET*; *JAK1*; *XPO1*; *PTPN11*; *PIK3CA*)	29.6%	48	13.4%
Census genes data set	20	329	20 (*KMT2C*; *FAT4*; *NCOR2*; *MYH11*; *KMT2D*; *PTPRB*; *SPEN*; *TRRAP*; *RNF213*; *POLE*; *FAT1*; *CAMTA*; *FLT4*; *ATP2B3*; *ZFHX3*; *ALK*; *ZNF521*; *KMT2A*; *GRIN2A*; *SETBP1*)	31.6%	44	24.25%
3′ UTR clustered data set	14	141	6 (*ETV6*; *PPM1D*; *CCND2*; *AR*; *BCL11B*; *BCL11A*)	36.4%	40	20.8%

org). The majority of these predicted alternative proteins are small proteins with a median length of 48 aa.

In the second data set, we gathered the top 20 Census genes harboring the most synonymous SNVs in the TCGA database and repeated the analysis (Futreal et al. 2004). All genes displayed at least one synonymous SNV altering at least one predicted alternative protein. About 30% of all listed synonymous SNVs within these 20 Census genes fell within a predicted alternative protein (Table 1). Out of these, 7.3% have been detected in re-analysis of large-scale mass spectrometry studies (openprot.org). The majority of these predicted alternative proteins are small proteins with a median length of 44 aa.

Finally, at the end of their publication, Supek et al. (2014) provided a list of clustered SNVs in the 3′ UTR of 14 different genes associated with human cancers. We checked this list of SNVs coordinates against that of alternative proteins within these genes. Of these mutations, 31.6% fell within a predicted alternative protein (Table 1). Again, the majority of these predicted alternative proteins are small proteins with a median length of 40 aa.

All of these percentages were higher than expected by chance (Table 1).

The absence of most of the alternative ORFs from genome annotations prevents us from identifying novel genetic drivers. As of today, only about 62% of Mendelian phenotypes have a known molecular basis, consistent with the hypothesis that at least some of these phenotypes result from defective alternative proteins (Amberger et al. 2015). A striking example is provided by one of the first discovered smORFs, apelin. Since its change in annotation from lncRNA to mRNA following its incidental discovery and functional characterization, genomic studies have identified polymorphisms linked to cardiovascular diseases and obesity risks (Zhao et al. 2010; Liao et al. 2011; Jin et al. 2012; Sentinelli et al. 2016). Finally, many published studies may need to be re-interpreted in light of the existence of more than one CDS per mRNA, and future overexpression and knockdown experiments will become technically more complex. For example, transfection of a CDS might actually result in the overexpression of two proteins, which are often functionally related (Klemke et al. 2001; Bergeron et al. 2013; Delcourt et al. 2017). This also means that the knockdown or knockout of genes could result in the absence of two or more proteins rather than one.

Proposition of a novel annotation framework

It is difficult to come up with a genome annotation pipeline that is both accurate and exhaustive, yet the need is evident. Different strategies have been adopted over the past years, which essentially regroup two goals: (1) to identify transcript structure (e.g., intron vs. exon); and (2) to identify the functional potential (e.g., contains a CDS) (Pruitt et al. 2009; Harrow et al. 2012; Aken et al. 2016; Mudge and Harrow 2016). These pipelines, however, invoke a uni-coding presumption. ORF-prediction algorithms apply the criteria of a single CDS per transcript, and a minimum length of 100 codons, unless the sequence bears high similarity to known proteins or domains (Furuno et al. 2003; Pruitt et al. 2012; Aken et al. 2016). As a result, the foreseen increase in smORF count in Swiss-Prot falls short, with an increase from 3.1% in 2009 to 3.3% in 2017 (Southan 2017). This means that despite the large number of smORF and alternative ORF discoveries, only a limited number make it through to genome annotation (Southan 2017). The current genome annotation system has been blamed for simplifying a transcript's definition, not taking into account their potential to hold multiple functional features (for review, see Mudge and Harrow 2016).

Here, we propose a framework for the incorporation of alternative ORFs into current genome annotations. With minimal modifications to the existing annotation pipelines (GENCODE, Ensembl, or NCBI for the human genome), alternative ORFs could be included (Harrow et al. 2012; Pruitt et al. 2012; Aken et al. 2016). As shown in Figure 4, most pipelines annotate ORFs and subsequent protein products from ab initio ORF prediction or sequence alignment with known proteins (from the UniProt or RefSeq databases) (Keller et al. 2011; The UniProt Consortium 2014). ORF prediction mostly relies on ORF size, codon usage, and the nonsynonymous to synonymous mutation ratio (Pruitt et al. 2009; Keller et al. 2011; Mudge and Harrow 2016). This means that current genome annotations are shaped by evolution-, prior knowledge-, and hypothesis-driven data. As proposed in Wright et al. (2016) for the emerging field of proteogenomics, protein sequences from alternative ORFs, reported in databases such as OpenProt (openprot.org), sORFs (Olexiouk et al. 2016), or SmProt (Hao et al. 2017), with detection evidence by ribosome profiling or mass spectrometry, could be downloaded for genome annotation (Fig. 4). Such an annotation pipeline would prevent some of today's pitfalls, abolishing the unique CDS presumption and empowering experimental data as well as conservation signatures (Mudge and Harrow 2016; Southan 2017). This would add a layer of experiment-driven data to genome annotation pipelines.

One of the biggest challenges for genome annotation will be to distinguish random ORFs from functional ones. Random ORFs are ORFs that could arise through evolutionary noise, e.g., a mutation causing a start codon to appear randomly within a transcript. Random ORFs could potentially be translated and thus be a source of translational noise but would not usually yield a functional detectable peptide (Brar and Weissman 2015). Purifying selection is expected to weed out detrimental random ORFs relatively quickly for dominant traits but more slowly for neutral random ORFs. It is not known what percentage of alternative ORFs predicted based on transcript sequences are random. Obviously, we would like to exclude random ORFs from annotations. However, the better we exclude random ORFs, the more functional ORFs will also be excluded, analogous to problems of true and false positives in medical diagnostics. The short length of alternative ORF sequences means that, for statistical reasons, either the false positive or false negative rate will be higher than for longer sequences.

While we believe that current annotation methods are too restrictive, there is also a real interest in avoiding false positives. The relative balance between inclusivity and exclusivity (sensitivity and specificity) will depend strongly on the experimental context and the questions being asked. To deal with these complexities, we propose a solution where annotations include filters that allow researchers to adjust the levels and types of evidence for annotated proteins. Evidence can be inferred from large-scale detection methods, either at the DNA (conservation signatures), the translation (ribosome profiling), or the protein level (mass spectrometry). And even though there is no perfect detection method for alternative proteins, one should be cognizant of each technique's strengths and pitfalls and strive to use and adapt them to better detect the entire proteomic landscape of a cell or tissue (Boekhorst et al. 2011; Aspden et al. 2014; Calviello et al. 2016; Hellens et al. 2016; Ma et al. 2016; Pueyo et al. 2016a; Delcourt et al. 2017; Hsu and Benfey 2017; Willems et al. 2017).

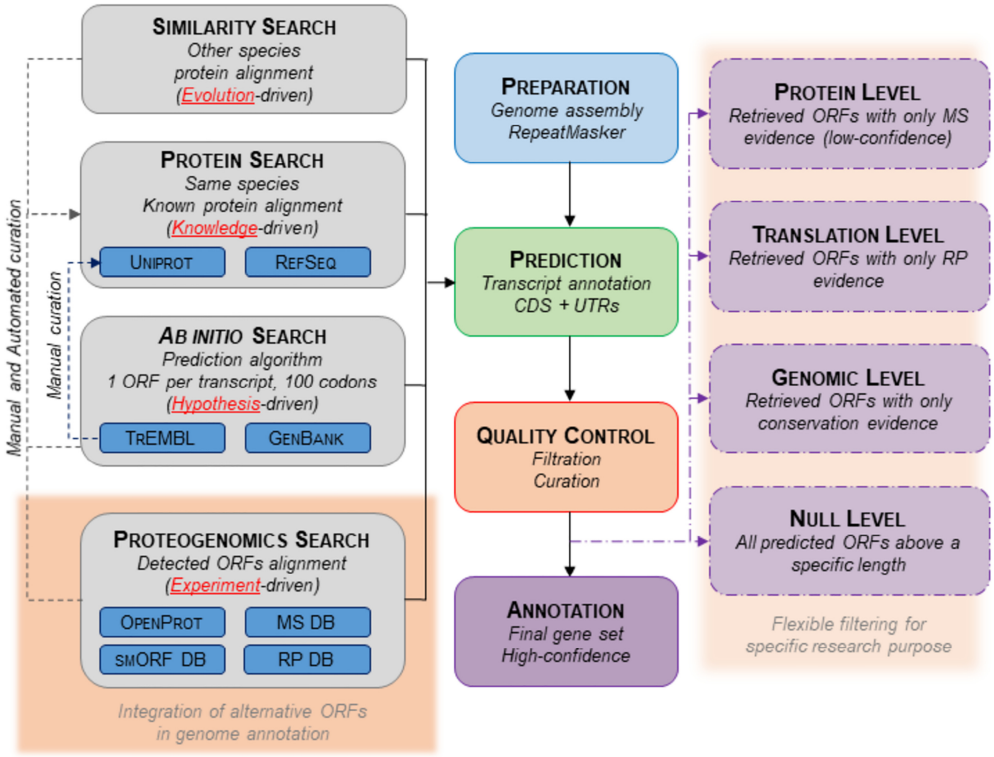


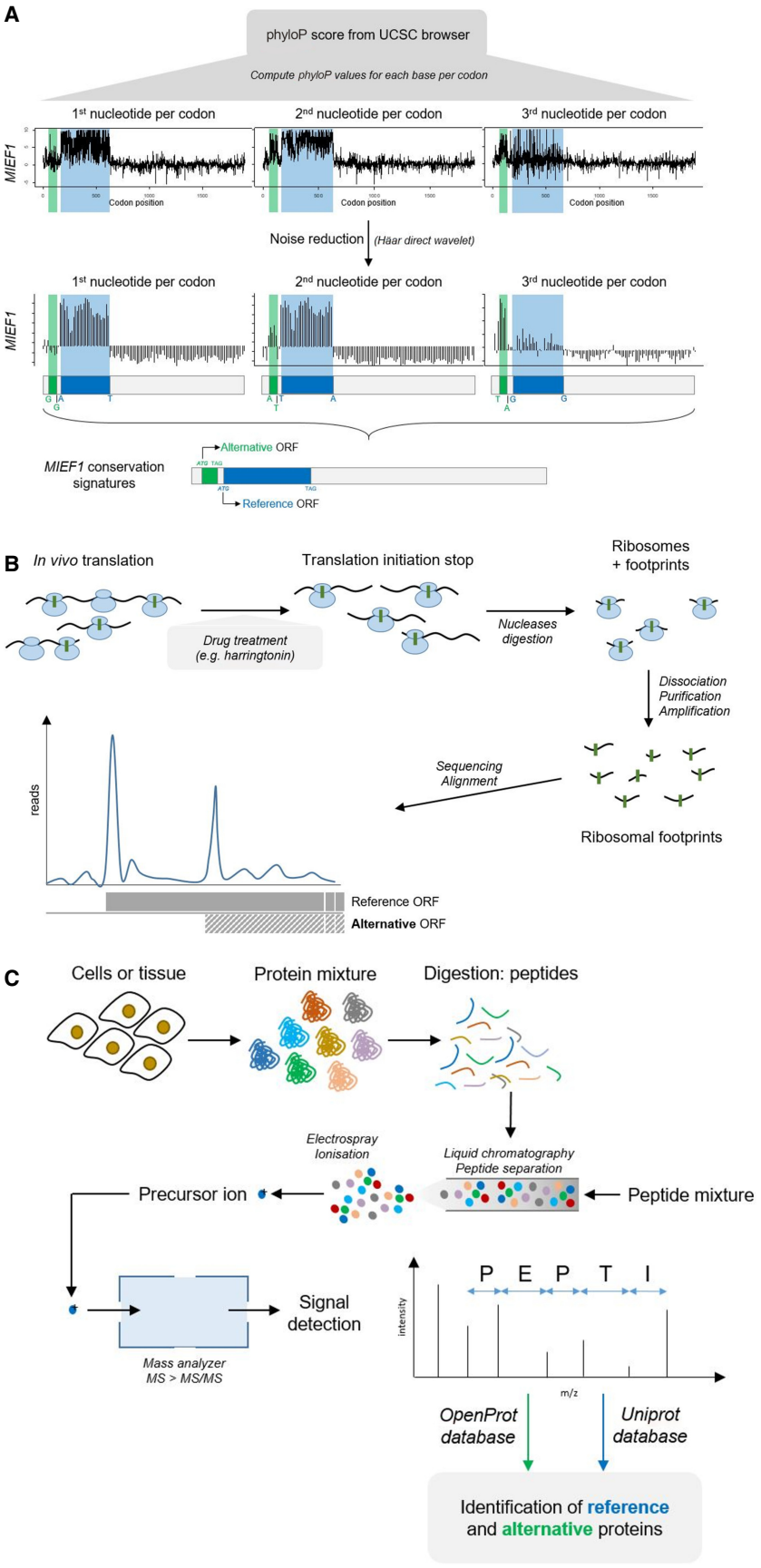
Figure 4. Proposed novel genome annotation framework. Current genome annotations' pipelines have four main steps: Preparation, Prediction, Quality Control, and Annotation. The Prediction step aims to annotate transcripts (exons, introns) and CDSs (with flanking UTRs). It mostly relies on three methods: a search by homology (different species known proteins are aligned to the genome assembly), a search by prior knowledge (same species known proteins are aligned to the genome assembly), and a search ab initio (prediction of ORFs by algorithms). Here, we suggest adding an experiment-driven search and including alternative ORFs with experimental detection. The output could also be flexible to fit different experimental purposes. The pipeline steps highlighted in red correspond to the suggested implementation.

Evidence at the genomic level

An indirect but potentially powerful piece of evidence of a protein's expression is its conservation signature. Conservation signatures are already used to distinguish functional ORFs in current ORF prediction algorithms (Mudge and Harrow 2016). Functional proteins are expected to be under purifying selection and the ratio of nonsynonymous to synonymous mutations highlights protein-coding sequences (Hughes 1999; Pál et al. 2006). The first and second nucleotide of a codon experience stronger selection than the third because of the genetic code's redundancy (Pollard et al. 2010; Samandi et al. 2017). This selection periodicity can allow for detection of conservation signatures in each of the reading frames (Fig. 5A). The phyloP score (a measure of probability to be under purifying selection) can be computed for every third base giving a triplet signal (three graphs corresponding, respectively, to the first, second, and third nucleotide for each codon) (Cooper et al. 2005; Samandi et al. 2017). After noise reduction, we can detect independent purifying selection signals in each of the reading frames, e.g., for the dual coding *MIEF1* gene (Fig. 5A). This method allows for annotation of genetic loci under purifying selection, but it relies on a good signal-to-noise ratio (and id facto on genome annotations for other species). However, this ratio may be biased by the phyloP score itself. Indeed, the phyloP score first evaluates the rate of neutral evolution for one locus based on empirical values of substitution rates, but these have been defined under a uni-coding gene presumption (Cooper et al. 2005). Moreover, some alternative ORFs could be the result of a more recent evolution and still be in a phase of adaptive selection (Ruiz-Pesini et al. 2004; Evans et al. 2014; McLysaght and Hurst 2016). Other measures of phylogenetic evolution can be used, such as PhyloCSF or CPC (Coding Potential Calculator), and Bazzini et al. combined evolutionary methods (PhyloCSF) to ribosome profiling (Kong et al. 2007; Lin et al. 2011; Bazzini et al. 2014). PhyloCSF uses the widely implemented phylogenetic analyses by maximum likelihood (Yang 1994, 2007), but it still relies on previously empirically determined matrices of codons' transition rates (ECMs; Empirical Codon Models). These ECMs were defined under a uni-coding gene presumption and could thereby bias the PhyloCSF score (Kosiol and Goldman 2005; Kosiol et al. 2007; Lin et al. 2011). The CPC score is designed to measure the coding potential of a transcript and uses machine-learning algorithms. However, the true nature (coding or noncoding) of the transcripts used in the training data set would be a critical element to the CPC's performance (Kong et al. 2007; Halevy et al. 2009). Conservation signatures may improve in the near future as new algorithms take into account the multicoding potential of mature mRNAs.

Evidence at the translational level

Ribosome profiling is a technique that measures ribosomal occupancy and initiation in vivo using deep sequencing of ribosome-protected mRNA fragments. First described by Steitz, ribosome profiling was recently adapted by Ingolia to make use of NGS



techniques and is now a widely used technique to describe the full coding potential of a genome (Steitz 1969; Ingolia et al. 2009, 2012; Ingolia 2014). In brief, the idea is to sequence ribosomal footprints, given that each ribosome encloses about 30 nucleotides when translating and thus protects them from nuclease digestion (Fig. 5B). These footprints can be amplified, sequenced, and mapped on the genome, thus identifying in vivo translation events (Ingolia et al. 2012). Ribosome profiling techniques have also been adapted to specifically isolate initiating ribosomes. Using drugs that stall the first step of elongation (harringtonine, lactimidomycin with puromycin), all initiation sites can be mapped on the genome (Ingolia et al. 2011; Ingolia 2016). However, the accuracy of ribosome profiling depends on fragment mapping on the genome, and since fragments are short, this creates a risk of multimapping (multiple match) and a bias against repetitive regions. There is also evidence that some genuine ribosome profiling identifications do not lead to the translation of functional proteins but rather are regulatory ribosome-RNA interactions (Ingolia 2016; Raj et al. 2016). Nonetheless, ribosome profiling offers a translation overview of the genome that is evolution-free, meaning that nonconserved or de novo translated ORFs would still be identified. There is also a dogma that function implies conservation, and

Figure 5. Large-scale detection methods for alternative proteins detection. (*A*) Conservation signatures of proteins encoded in different reading frame from the same mRNA. PhyloP scores can be computed from the UCSC Genome Browser, and noise filtration (by Haar direct wavelet) allows for the identification of distinct purifying selection signals in each reading frames. Here, the example of the dual coding *MIEF1* gene is represented and corroborates data from mass spectrometry and ribosome profiling with the detection of an alternative ORF upstream of the canonical CDS (reference ORF). (*B*) Schematic representation of the ribosome profiling technique. This technique allows for detection of ribosomal footprints, and subsequent mapping on the genome yields a map of translation events throughout. Translation initiation at alternative ORFs can then be detected. (C) Schematic representation of the mass spectrometry technique. The search space bears crucial consequences on peptide identification. Here, we represent the strategy used by the OpenProt database that re-analyzed published mass spectrometry studies adding their predicted alternative ORFs to the scope of possibilities.

accordingly, the possibility to identify nonconserved yet functional proteins arouses strong opinions (The ENCODE Project Consortium 2012; Graur et al. 2013; Han et al. 2014). Available online tools for visualization of ribosome profiling data are listed in Table 2.

Evidence at the protein level

Mass spectrometry (MS)-based proteomics has emerged as the gold standard technique to assess the protein landscape of a cell or tissue and thus can offer additional evidence beyond ribosome profiling (Aebersold and Mann 2003, 2016; Vogel and Marcotte 2012; Huttlin et al. 2015). Cells or tissue lysates are digested to peptides, subsequently identified by mass spectrometry (Fig. 5C). However, the scope of the search space has a substantial impact on the proportion of peptide identification (Aebersold and Mann 2003, 2016). Peptide identification relies on matching mass spectra to predicted peptides from CDS (e.g., from UniProt database). If the database does not contain the relevant peptide, the associated protein will never be identified since it is not included in the scope of possibilities (Samandi et al. 2017). As of today, <50% of all MS/MS spectra from a proteomics experiment are matched with high confidence (Heo et al. 2010; Chick et al. 2015). These unassigned peptides can correspond to peptide modifications or to proteins not in the database (Heo et al. 2010). In the recently developed proteogenomics approaches, addition of more inclusive databases to the search space allows for the discovery of novel proteins thus far undetected (Oyama et al. 2007; Saghatelian and Couso 2015; Samandi et al. 2017; openprot.org). Yet, not all proteins produce peptides detectable by mass spectrometry, owing to their subcellular localization, chemistry, and/or size. This is partly why false-discovery rates in proteomics experiments can be difficult to evaluate (Nesvizhskii 2014). Alternative ORFs are smaller than canonical CDS, with a median length of 45 aa (Samandi et al. 2017), and mass spectrometry detection of small and low-abundance proteins is challenging (Nesvizhskii 2014; Aebersold and Mann 2016). Identification of any protein by mass spectrometry relies heavily on good quality spectra, but this is particularly true for alternative proteins, as most smaller proteins will produce fewer peptides upon enzymatic digestion (Ma et al. 2016). There could also be cases where a protein might not produce any peptides from trypsin digestion (most used enzyme) or might produce highly hydrophilic peptides, rendering its identification by proteomics challenging (Young et al. 2017). Nonetheless, specific proteomics protocols to better detect small proteins are emerging and raise hopes for the future of proteogenomics in genome annotation (Ma et al. 2016). Table 2 contains a list of online tools available for alternative ORF mass spectrometry identification.

Available online resources

Several online tools either allow for raw data enquiry or provide a list of all alternative ORFs with corresponding evidence of expression for several species (see Table 2). Moreover, some tools also predict the translation of alternative ORFs, such as SPECtre,

Table 2. Online tools for alternative ORFs search within a gene of interest

	Tool	Description	Comments	Reference
Ribosome profiling	GWIPS-viz	Ribosome profiling data	Footprint and mRNA-seq genome browser	(Michel et al. 2014)
	RPFdb	Ribosome profiling data	Footprint genome browser and expression measurements	(Xie et al. 2016)
	TISdb	Ribosome profiling data	Translation initiation sites	(Wan and Qian 2014)
	sORFs.org	Ribosome profiling data	Short ORF annotations	(Olexiouk et al. 2016)
Mass spectrometry	MaxQB database	MS data repository	Referenced proteins only (UniProt db)	(Schaab et al. 2012)
	PRIDE	MS data repository	Downloadable raw data	(Martens et al. 2005)
	Global Proteome Machine Database	MS data repository	Downloadable raw data	(Beavis 2006)
	PeptideAtlas	MS data repository	Downloadable raw data	(Desiere et al. 2006)
	NIST libraries (peptide.nist.org)	MS data repository	Downloadable raw data	(Wallace et al. 2017)
Conservation	UCSC Genome Browser	Conservation signal browser	phyloP and PhastCons tracks	(Cheng et al. 2014; Miller et al. 2007)
	ECR Browser	Conservation signal browser	Identify evolutionary conserved regions (ECRs)	(Ovcharenko et al. 2004)
	GTB	SNV intolerance browser	Identify regions likely intolerant for mutations	(Shihab et al. 2017)
	GERP scores	Available on Ensembl genome browser	Genomic Evolutionary Rate Profiling identifies constraint element in multiple alignments	(Cooper et al. 2005; Davydov et al. 2010)
Databases	OpenProt	Database of alternative ORFs and reference ORFs	All ORFs (>30 codons), 13 species, with ribosome profiling or MS evidence	openprot.org
	sORFs.org	Repository of small ORFs detected by ribosome profiling	All ORFs (≤100 codons) detected by ribosome profiling	(Olexiouk et al. 2016)
	N/A	Identification of conserved small ORFs	All predicted conserved small ORFs (>27 codons) in Supplementary Tables for five species	(Mackowiak et al. 2015)
	tsORFdb	Theoretical short ORF database	Systematic six-frame translation, using ATG and non-ATG initiation codons	(Heo et al. 2010)
	smORFdb	Database of small ORFs	Small ORFs (<100 codons), several species, protein, transcript or prediction evidence level	immunet.cn/smorf
	smProt	Database of small proteins	Small proteins (<100 codons) with various evidence level (literature, MS, ribosome profiling)	(Hao et al. 2017)

RiboTaper, ORF-RATER and PROTEOFORMER (Crappé et al. 2015; Fields et al. 2015; Calviello et al. 2016; Chun et al. 2016), based on integration of ribosome profiling data. PROTEOFORMER and RiboTaper combine ribosome profiling data with an implemented construction of protein sequences from thus detected ORFs. Thereby, they build a database that can be used for proteomics without leading to a large increase in the proteomic search space (Jeong et al. 2012; Crappé et al. 2015; Guthals et al. 2015). Some databases of alternative ORFs also offer a freely downloadable FASTA file for proteomics experiments (openprot.org).

On the importance of filtration and curation

There is an undeniable close relationship between the quality of a genome annotation and experimental and clinical results. That is why all genome annotation pipelines include a step of database filtration and curation (Pruitt et al. 2012; The UniProt Consortium 2014; Aken et al. 2016; Tatusova et al. 2016). Often, the first step (Prediction on Fig. 4) emphasizes sensitivity over specificity. However, because false positives could burden variant-calling workflows, putative functional annotations are removed at the filtration and curation steps (Quality Control on Fig. 4; Koonin and Galperin 2003; Mudge and Harrow 2016). However, as discussed earlier, genome specificity needs might differ based on the experimental purpose (variant calling, novel protein identification, etc.). The RefSeq database offers some more putative functional annotation (XM_, XP_ annotations), and Ensembl reports to some extent less supported transcripts' annotations (Pruitt et al. 2012; Aken et al. 2016). Yet, these still rely on overly restrictive criteria (one CDS per transcript, longer than 100 codons) (Chung et al. 2007; Galindo et al. 2007; Saghatelian and Couso 2015; Pueyo et al. 2016a; Couso and Patraquim 2017). While adapting the framework of genome annotations to consider alternative ORFs will more likely yield significant advances, the need for a more flexible annotation for various purposes could be addressed (Fig. 4). The different levels of confidence suggested in Figure 4 are based on evidence levels discussed earlier: conservation, ribosome profiling, mass spectrometry, or none of the aforementioned.

The complexity behind the data sets

In the suggested annotation pipeline, we emphasize experiment-driven annotations. In that aspect, the pipeline would rely on the data quality of the databases used (OpenProt, sORF, SmProt, etc.) (Fig. 4). It is important to note that although implementation of identifications from these databases would be straightforward, the quality control of the data might not be. Indeed, as mentioned earlier, the various databases present discrepancies in numbers of identifications. They uphold different definitions of alternative ORFs, but they also enforce different identification pipelines. All methods suggested here—ribosome profiling, proteogenomics, and conservation signatures—are noisy and require adequate filtering and thresholding to minimize the risk of false positives (Guthals et al. 2015; Aebersold and Mann 2016; Ingolia 2016; Calviello and Ohler 2017; Wallace et al. 2017). We would recommend databases using raw data and an adequate pipeline of identification. For example, a two-stage FDR pipeline could be used for mass spectrometry, in order to minimize the impact of the increased search space (Woo et al. 2014, 2015; Pauli et al. 2015). The use of additional algorithms in order to control for misidentification of post-translational modifications would be encouraged (Kong et al. 2017). In ribosome profiling, multimapping should be filtered out, keeping only unique mappings, with an appropriate sequencing depth threshold (Calviello and Ohler 2017). Moreover, elongating reads and RNA-seq data will strengthen the observations (Calviello and Ohler 2017).

Is ORF length an appropriate filter?

The rationale behind the minimum ORF length of 100 codons is to avoid polluting annotations with random events (Pruitt et al. 2009). Yet, it is clear it also leads to numerous false negatives, i.e., functional ORFs shorter than 100 codons excluded from annotations (Andrews and Rothnagel 2014; Ma et al. 2014; Pauli et al. 2014; Couso and Patraquim 2017). Notwithstanding, we could also question the arbitrary cut-off taken by groups studying alternative ORFs. For instance, the smORF community only reports ORFs shorter than 100 codons, but they would then miss all longer alternative ORFs (Cabrera-Quio et al. 2016; Hellens et al. 2016). The OpenProt team does not limit itself to alternative ORFs shorter than 100 codons, but it still uses an arbitrary minimal cut-off of 30 codons (openprot.org). This 30-codon cut-off allows for prediction of multiple alternative ORFs (361,173 unique alternative ORFs predicted in the human genome) without overcrowding the search space for proteomics experiments (Jeong et al. 2012; Nesvizhskii 2014; Guthals et al. 2015). However, examples of smORFs shorter than 30 codons have been published, and it questions the adequacy of an ORF length threshold (Yosten et al. 2016). The aforementioned genome annotation framework would still rely on some arbitrary ORF length cut-offs. Users should be aware of it and, because accumulation of random events with a lower cut-off is a statistical reality, we would recommend using the "Null Level" data set only for bioinformatics studies (Fig. 4).

Accumulation of clinical reports as an evidence level?

The causal link from the quality of genome annotations to variant-calling misinterpretation is evident; hence, most putative annotations are removed to limit clinical false positives. Thinking about it backward, pathological family-specific variants clustered on genetic loci are a valuable yet overlooked resource. For example, in the case where no "likely pathological" variant is determined (about 65% of cases), a new variant-calling file could be generated using a less stringently filtered data set (for example, using the "Protein Level," "Translation level," or "Conservation level") (Fig. 4). Thereby, mutations altering alternative proteins could be retrieved. This could generate a positive feedback loop instead of the current vicious cycle phenomenon. Likely pathological mutations, especially in the case of severe or pediatric Mendelian phenotypes, could represent a source of functional evidence (same loci, several individuals, and same family). Alternative ORFs with clinical evidence could then be annotated in the next genome annotation release.

Foreseen consequences of implementing alternative ORFs in genome annotations

In Supek's study on cancer-driver silent mutations (Table 1), genes containing potential alternative proteins affected by so-called "silent" mutations were identified earlier (Supek et al. 2014). Considering three genes (the top mutated for each of the three data sets), all of them present at least one alternative protein affected by such "silent" mutation or clustered mutations in the 3′ UTR (Fig. 6). These alternative proteins from the *NTRK3* and *KMT2C*

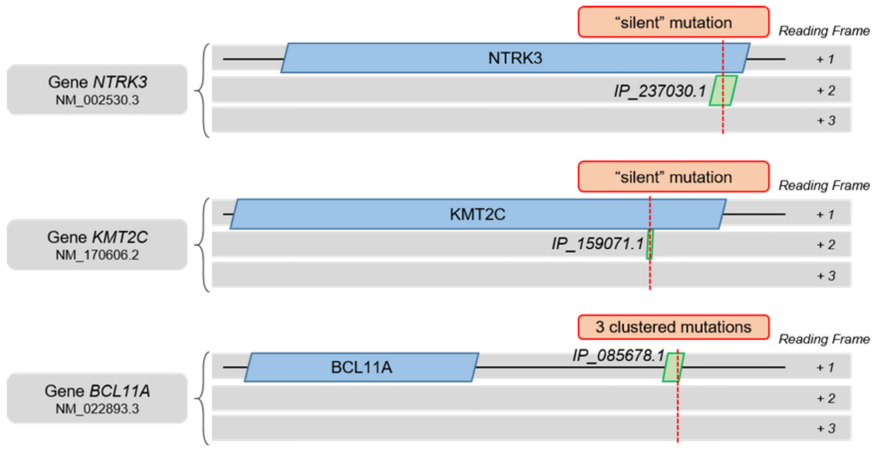


Figure 6. Graphical representation of alternative ORFs affected by "silent" and clustered 3′ UTR SNVs in *NTRK3*, *KMT2C*, and *BCL11A* genes. Length proportions between the full mRNA, the canonical CDS, and the alternative ORF are respected. The SNV position is represented by a red dotted line. The RefSeq transcript accession number (NM_) and the alternative ORF OpenProt accession number (IP_) are indicated.

genes were predicted in the OpenProt database and subsequently detected in at least one published mass spectrometry experiment re-analyzed with the OpenProt pipeline (openprot.org; Hein et al. 2015; Hurwitz et al. 2016). Thus, here are three new potential genetic drivers of human cancer.

These examples show how our current genome annotation approaches may have hidden functional proteins with pathological importance. The examples from Supek's study echo reports of human pathologies from disrupted or inserted uORFs, and studies of cellular consequences of smORF-encoded peptide disruption (Barbosa et al. 2013; Supek et al. 2014; Couso and Patraquim 2017; Plaza et al. 2017). The current body of evidence for functional alternative ORFs is but a small peek at the potential for future discoveries when their implementation in genome annotation will render identification less serendipitous. Identification of alternative ORFs will then increase, and with it, the pace of research in physiological and pathological pathways. As for the clinical side, the *APELA* gene annotation example highlights the foreseen gain. Alternative ORFs are an as-yet unexplored reservoir of genetic drivers, pathological causes, therapeutic targets, and/or biomarkers. The cooperation between fundamental and clinical research to implement and improve alternative ORFs annotation in the genome is pivotal, and it could well advance the pace of research and genomic knowledge.

Eventually, perhaps the best argument for incorporating alternative ORFs into genome annotation is to look at what might happen if we maintain the status quo. As of today, there is a dichotomy between genome annotations and experimental evidence. This gap will deepen, pulling apart genomics and proteomics. Currently, the emphasis is put on conservation signatures above all, and experimental evidence of a novel protein will not be considered if it is not followed up by a functional characterization. This means that current genome annotations provide a conceptual framework for research and medicine that is incomplete. One could question providing only partial information to the scientific community and ultimately to patients when a more exhaustive framework could be implemented. It would certainly be questionable to pollute it with random ORFs annotations. That is why we have proposed a strategy to annotate specific ORFs, with experimental evidence, rather than opening the floodgates to all alternative ORFs. Annotation censoring of alternative ORFs would likely hamper progress in alternative proteome investigations (detection, structure, and function) but also in understanding the relationship between genotype and phenotype.

Conclusions

Current genome annotations are the linchpin to and profoundly mold today's research and genetic medicine. However, by assuming one mature RNA encodes only one protein, these annotations are incomplete. The number of functional alternative ORFs within an mRNA or a lncRNA is rapidly increasing, yet a systemic incorporation of these novel proteins into genome annotations is awaited. Hence, we need a better annotation system that can regroup the whole of transcriptomic and proteomic information contained within a gene, as suggested in Figure 4. We foresee that the implementation of such a framework would help bring attention to alternative ORFs and their potential involvement in cellular functions, pathways, and/or pathological phenotypes. Although this review focused on human genome annotations, the observations are valid for all species. Claude Bernard wrote, "It is what we know already that often prevents us from learning." The evidence for alternative ORF translation and function is accumulating. We need to unlearn our misconception of the gene, accepting its polycistronic nature, to strive for a better understanding of the genomic complexity underlying physiological and pathological mechanisms.

Acknowledgments

This research was supported by Canadian Institutes of Health Research (CIHR) grants MOP-137056 and MOP-136962, and by a Canada Research Chair in Functional Proteomics and Discovery of Novel Proteins to X.R. A.A.C., D.J.H., and X.R. are members of the Fonds de Recherche du Québec Santé (FRQS)-supported Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke, and A.A.C. is also a member of the FRQS-supported Centre de recherche sur le vieillissement and is supported by a New Investigator fellowship from the CIHR.

References

Abramowitz J, Grenet D, Birnbaumer M, Torres HN, Birnbaumer L. 2004. XLαs, the extra-long form of the α-subunit of the Gs G protein, is significantly longer than suspected, and so is its companion Alex. *Proc Natl Acad Sci* **101:** 8366–8371.

Aebersold R, Mann M. 2003. Mass spectrometry-based proteomics. *Nature* **422:** 198–207.

Aebersold R, Mann M. 2016. Mass-spectrometric exploration of proteome structure and function. *Nature* **537:** 347–355.

Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, García Girón C, Hourlier T, et al. 2016. The Ensembl gene annotation system. *Database* **2016:** baw093.

Albuquerque JP, Tobias-Santos V, Rodrigues AC, Mury FB, da Fonseca RN. 2015. small ORFs: a new class of essential genes for development. *Genet Mol Biol* **38:** 278–283.

Alfarano C, Foussal C, Lairez O, Calise D, Attané C, Anesia R, Daviaud D, Wanecq E, Parini A, Valet P, et al. 2015. Transition from metabolic

adaptation to maladaptation of the heart in obesity: role of apelin. *Int J Obes* **39:** 312–320.
Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43:** D789–D798.
Anderson DM, Anderson KM, Chang C-L, Makarewich CA, Nelson BR, McAnally JR, Kasaragod P, Shelton JM, Liou J, Bassel-Duby R, et al. 2015. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* **160:** 595–606.
Andreev DE, O'Connor PBF, Fahey C, Kenny EM, Terenin IM, Dmitriev SE, Cormican P, Morris DW, Shatsky IN, Baranov PV. 2015. Translation of 5′ leaders is pervasive in genes resistant to eIF2 repression. *eLife* **4:** e03971.
Andrews SJ, Rothnagel JA. 2014. Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* **15:** 193–204.
Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, Couso J-P. 2014. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *eLife* **3:** e03528.
Austin F, Oyarbide U, Massey G, Grimes M, Corey SJ. 2017. Synonymous mutation in TP53 results in a cryptic splice site affecting its DNA-binding site in an adolescent with two primary sarcomas. *Pediatr Blood Cancer* **64:** e26584.
Autio KJ, Kastaniotis AJ, Pospiech H, Miinalainen IJ, Schonauer MS, Dieckmann CL, Hiltunen JK. 2008. An ancient genetic link between vertebrate mitochondrial fatty acid synthesis and RNA processing. *FASEB J* **22:** 569–578.
Baek J, Lee J, Yoon K, Lee H. 2017. Identification of unannotated small genes in *Salmonella*. *G3 (Bethesda)* **7:** 983–989.
Barbosa C, Peixeiro I, Romão L. 2013. Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet* **9:** e1003529.
Basrai MA, Hieter P, Boeke JD. 1997. Small open reading frames: beautiful needles in the haystack. *Genome Res* **7:** 768–771.
Batista RL, di Santi Rodrigues A, Nishi MY, Gomes NLRA, Faria JAD, de Moraes DR, Carvalho LR, Frade EMC, Domenice S, de Mendonca BB. 2017. A recurrent synonymous mutation in the human androgen receptor gene causing complete androgen insensitivity syndrome. *J Steroid Biochem Mol Biol* **174:** 14–16.
Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, et al. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33:** 981–993.
Beavis RC. 2006. Using the global proteome machine for protein identification. *Methods Mol Biol* **328:** 217–228.
Bergeron D, Lapointe C, Bissonnette C, Tremblay G, Motard J, Roucou X. 2013. An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. *J Biol Chem* **288:** 21824–21835.
Bertrand C, Valet P, Castan-Laurell I. 2015. Apelin and energy metabolism. *Front Physiol* **6:** 115.
Blencowe BJ. 2017. The relationship between alternative splicing and proteomic complexity. *Trends Biochem Sci* **42:** 407–408.
Boekhorst J, Wilson G, Siezen RJ. 2011. Searching in microbial genomes for encoded small proteins. *Microb Biotechnol* **4:** 308–313.
Boucher J, Masri B, Daviaud D, Gesta S, Guigné C, Mazzucotelli A, Castan-Laurell I, Tack I, Knibiehler B, Carpéné C, et al. 2005. Apelin, a newly identified adipokine up-regulated by insulin and obesity. *Endocrinology* **146:** 1764–1771.
Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* **14:** 681–691.
Brar GA, Weissman JS. 2015. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* **16:** rm4069.
Cabrera-Quio LE, Herberg S, Pauli A. 2016. Decoding sORF translation – from small proteins to gene regulation. *RNA Biol* **13:** 1051–1059.
Calviello L, Ohler U. 2017. Beyond read-counts: Ribo-seq data analysis to understand the functions of the transcriptome. *Trends Genet* **33:** 728–744.
Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, Landthaler M, Obermayer B, Ohler U. 2016. Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* **13:** 165–170.
The Cancer Genome Atlas Research Network. 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499:** 43–49.
The Cancer Genome Atlas Research Network. 2016. Comprehensive molecular characterization of papillary renal-cell carcinoma. *N Engl J Med* **374:** 135–145.
Castan-Laurell I, Vítkova M, Daviaud D, Dray C, Kováciková M, Kovacova Z, Hejnova J, Stich V, Valet P. 2008. Effect of hypocaloric diet-induced weight loss in obese women on plasma apelin and adipose tissue expression of apelin and APJ. *Eur J Endocrinol* **158:** 905–910.
Castan-Laurell I, Dray C, Attané C, Duparc T, Knauf C, Valet P. 2011. Apelin, diabetes, and obesity. *Endocrine* **40:** 1.
Castan-Laurell I, Dray C, Knauf C, Kunduzova O, Valet P. 2012. Apelin, a promising target for type 2 diabetes treatment? *Trends Endocrinol Metab* **23:** 234–241.
Cheng H, Chan WS, Li Z, Wang D, Liu S, Zhou Y. 2011. Small open reading frames: current prediction techniques and future prospect. *Curr Protein Pept Sci* **12:** 503–507.
Cheng Y, Ma Z, Kim B-H, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J, et al. 2014. Principles of regulatory information conservation between mouse and human. *Nature* **515:** 371–375.
Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, Gygi SP. 2015. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* **33:** 743–749.
Chun SY, Rodriguez CM, Todd PK, Mills RE. 2016. SPECtre: a spectral coherence–based classifier of actively translated transcripts from ribosome profiling sequence data. *BMC Bioinformatics* **17:** 482.
Chung W-Y, Wadhawan S, Szklarczyk R, Pond SK, Nekrutenko A. 2007. A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Comput Biol* **3:** e91.
Claverie JM, Poirot O, Lopez F. 1997. The difficulty of identifying genes in anonymous vertebrate sequences. *Comput Chem* **21:** 203–214.
Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15:** 901–913.
Couso J-P, Patraquim P. 2017. Classification and function of small open reading frames. *Nat Rev Mol Cell Biol* **18:** 575–589.
Crappé J, Ndah E, Koch A, Steyaert S, Gawron D, De Keulenaer S, De Meester E, De Meyer T, Van Criekinge W, Van Damme P, et al. 2015. PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res* **43:** e29.
Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6:** e1001025.
Delcourt V, Staskevicius A, Salzet M, Fournier I, Roucou X. 2017. Small proteins encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA. *Proteomics* doi: 10.1002/pmic.201700058.
Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. 2006. The PeptideAtlas project. *Nucleic Acids Res* **34:** D655–D658.
D'Lima NG, Ma J, Winkler L, Chu Q, Loh KH, Corpuz EO, Budnik BA, Lykke-Andersen J, Saghatelian A, Slavoff SA. 2017. A human microprotein that interacts with the mRNA decapping complex. *Nat Chem Biol* **13:** 174–180.
Dray C, Knauf C, Daviaud D, Waget A, Boucher J, Buléon M, Cani PD, Attané C, Guigné C, Carpéné C, et al. 2008. Apelin stimulates glucose utilization in normal and obese insulin-resistant mice. *Cell Metab* **8:** 437–445.
Dray C, Debard C, Jager J, Disse E, Daviaud D, Martin P, Attané C, Wanecq E, Guigné C, Bost F, et al. 2010. Apelin and APJ regulation in adipose tissue and skeletal muscle of type 2 diabetic mice and humans. *Am J Physiol Endocrinol Metab* **298:** E1161–E1169.
Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. 2013. Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *eLife* **2:** e01179.
Ebina I, Takemoto-Tsutsumi M, Watanabe S, Koyama H, Endo Y, Kimata K, Igarashi T, Murakami K, Kudo R, Ohsumi A, et al. 2015. Identification of novel *Arabidopsis thaliana* upstream open reading frames that control expression of the main coding sequences in a peptide sequence-dependent manner. *Nucleic Acids Res* **43:** 1562–1576.
The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.
Erdem G, Dogru T, Tasci I, Sonmez A, Tapan S. 2008. Low plasma apelin levels in newly diagnosed type 2 diabetes mellitus. *Exp Clin Endocrinol Diabetes* **116:** 289–292.
Escobar B, de Cárcer G, Fernández-Miranda G, Cascón A, Bravo-Cordero JJ, Montoya MC, Robledo M, Cañamero M, Malumbres M. 2010. Brick1 is an essential regulator of actin cytoskeleton required for embryonic development and cell transformation. *Cancer Res* **70:** 9349–9359.
Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G, et al. 2014. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nat Genet* **46:** 1089–1096.
Fahraeus R, Marin M, Olivares-Illana V. 2016. Whisper mutations: cryptic messages within the genetic code. *Oncogene* **35:** 3753–3760.

Favazza L, Chitale DA, Barod R, Rogers CG, Kalyana-Sundaram S, Palanisamy N, Gupta NS, Williamson SR. 2017. Renal cell tumors with clear cell histology and intact VHL and chromosome 3p: a histological review of tumors from the Cancer Genome Atlas database. *Mod Pathol* **30:** 1603–1612.

Fields AP, Rodriguez EH, Jovanovic M, Stern-Ginossar N, Haas BJ, Mertins P, Raychowdhury R, Hacohen N, Carr SA, Ingolia NT, et al. 2015. A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol Cell* **60:** 816–827.

Furuno M, Kasukawa T, Saito R, Adachi J, Suzuki H, Baldarelli R, Hayashizaki Y, Okazaki Y. 2003. CDS annotation in full-length cDNA sequence. *Genome Res* **13:** 1478–1487.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. *Nat Rev Cancer* **4:** 177–183.

Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. 2007. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol* **5:** e106.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17:** 333–351.

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* **5:** 578–590.

Gunišová S, Beznosková P, Mohammad MP, Vlčková V, Valášek LS. 2016. In-depth analysis of cis-determinants that either promote or inhibit reinitiation on GCN4 mRNA after translation of its four short uORFs. *RNA* **22:** 542–558.

Guthals A, Boucher C, Bandeira N. 2015. The generating function approach for peptide identification in spectral networks. *J Comput Biol* **22:** 353–366.

Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. 2013. Ribosome profiling provides evidence that large non-coding RNAs do not encode proteins. *Cell* **154:** 240–251.

Halevy A, Norvig P, Pereira F. 2009. The unreasonable effectiveness of data. *IEEE Intell Syst* **24:** 8–12.

Han P, Jin FJ, Maruyama J, Kitamoto K. 2014. A large nonconserved region of the tethering protein leashin is involved in regulating the position, movement, and function of Woronin bodies in *Aspergillus oryzae*. *Eukaryot Cell* **13:** 866–877.

Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K, Nishi R, Ohashi C, Iida K, Tanaka M, et al. 2013. Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc Natl Acad Sci* **110:** 2395–2400.

Hanyu-Nakamura K, Sonobe-Nojima H, Tanigawa A, Lasko P, Nakamura A. 2008. *Drosophila* Pgc protein inhibits P-TEFb recruitment to chromatin in primordial germ cells. *Nature* **451:** 730–733.

Hao Y, Zhang L, Niu Y, Cai T, Luo J, He S, Zhang B, Zhang D, Qin Y, Yang F, et al. 2017. SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief Bioinform* doi: 10.1093/bib/bbx005.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774.

Hashimoto Y, Niikura T, Tajima H, Yasukawa T, Sudo H, Ito Y, Kita Y, Kawasumi M, Kouyama K, Doyu M, et al. 2001. A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Aβ. *Proc Natl Acad Sci* **98:** 6336–6341.

Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y, Gak IA, Weisswange I, Mansfeld J, Buchholz F, et al. 2015. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163:** 712–723.

Heinonen MV, Purhonen AK, Miettinen P, Pääkkönen M, Pirinen E, Alhava E, Akerman K, Herzig KH. 2005. Apelin, orexin-A and leptin plasma levels in morbid obesity and effect of gastric banding. *Regul Pept* **130:** 7–13.

Heinonen MV, Laaksonen DE, Karhu T, Karhunen L, Laitinen T, Kainulainen S, Rissanen A, Niskanen L, Herzig KH. 2009. Effect of diet-induced weight loss on plasma apelin and cytokine levels in individuals with the metabolic syndrome. *Nutr Metab Cardiovasc Dis* **19:** 626–633.

Hellens RP, Brown CM, Chisnall MAW, Waterhouse PM, Macknight RC. 2016. The emerging world of small ORFs. *Trends Plant Sci* **21:** 317–328.

Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. 2008. Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol* **70:** 1487–1501.

Hemm MR, Paul BJ, Miranda-Ríos J, Zhang A, Soltanzad N, Storz G. 2010. Small stress response proteins in *Escherichia coli*: proteins missed by classical proteomic studies. *J Bacteriol* **192:** 46–58.

Heo H-S, Lee S, Kim JM, Choi YJ, Chung HY, June Oh S. 2010. tsORFdb: theoretical small Open Reading Frames (ORFs) database and massProphet: Peptide Mass Fingerprinting (PMF) tool for unknown small functional ORFs. *Biochem Biophys Res Commun* **397:** 120–126.

Hinnebusch AG. 2005. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol* **59:** 407–450.

Hsu PY, Benfey PN. 2017. Small but mighty: functional peptides encoded by small ORFs in plants. *Proteomics* doi: 10.1002/pmic.201700038.

Hsu PY, Calviello L, Wu H-YL, Li F-W, Rothfels CJ, Ohler U, Benfey PN. 2016. Super-resolution ribosome profiling reveals unannotated translation events in *Arabidopsis*. *Proc Natl Acad Sci* **113:** E7126–E7135.

Hu H, He L, Li L, Chen L. 2016. Apelin/APJ system as a therapeutic target in diabetes and its complications. *Mol Genet Metab* **119:** 20–27.

Huang SK, Shin K, Sarker M, Rainey JK. 2017. Apela exhibits isoform- and headgroup-dependent modulation of micelle binding, peptide conformation and dynamics. *Biochim Biophys Acta* **1859:** 767–778.

Huang Z, Wu L, Chen L. 2018. Apelin/APJ system: a novel potential therapy target for kidney disease. *J Cell Physiol* **233:** 3892–3900.

Hughes AL. 1999. *Adaptive evolution of genes and genomes*. Oxford University Press, Oxford, UK.

Hurwitz SN, Rider MA, Bundy JL, Liu X, Singh RK, Meckes DG. 2016. Proteomic profiling of NCI-60 extracellular vesicles uncovers common protein cargo and cancer type-specific biomarkers. *Oncotarget* **7:** 86999–87015.

Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, Tam S, Zarraga G, Colby G, Baltier K, et al. 2015. The BioPlex network: a systematic exploration of the human interactome. *Cell* **162:** 425–440.

Ingolia NT. 2014. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet* **15:** 205–213.

Ingolia NT. 2016. Ribosome footprint profiling of translation throughout the genome. *Cell* **165:** 22.

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324:** 218–223.

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.

Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. 2012. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7:** 1534–1550.

Jeong K, Kim S, Bandeira N. 2012. False discovery rates in spectral identification. *BMC Bioinformatics* **13:** S2.

Ji Z, Song R, Regev A, Struhl K. 2015. Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4:** e08890.

Jin W, Su X, Xu M, Liu Y, Shi J, Lu L, Niu W. 2012. Interactive association of five candidate polymorphisms in Apelin/APJ pathway with coronary artery disease among Chinese hypertensive patients. *PLoS One* **7:** e51123.

Juntawong P, Girke T, Bazin J, Bailey-Serres J. 2014. Translational dynamics revealed by genome-wide profiling of ribosome footprints in *Arabidopsis*. *Proc Natl Acad Sci* **111:** E203–E212.

Karginov TA, Pastor DPH, Semler BL, Gomez CM. 2017. Mammalian polycistronic mRNAs and disease. *Trends Genet* **33:** 129–142.

Keller O, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27:** 757–763.

Kim S-J, Xiao J, Wan J, Cohen P, Yen K. 2017. Mitochondrially derived peptides as novel regulators of metabolism. *J Physiol* **595:** 6613–6621.

Klemke M, Kehlenbach RH, Huttner WB. 2001. Two overlapping reading frames in a single exon encode interacting proteins—a novel way of gene usage. *EMBO J* **20:** 3849–3860.

Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, Kageyama Y. 2007. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat Cell Biol* **9:** 660–665.

Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, Wei L, Gao G. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35:** W345–W349.

Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. 2017. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* **14:** 513–520.

Koonin EV, Galperin MY. 2003. Genome annotation and analysis. In *Sequence - evolution - function: computational approaches in comparative genomics*, Chapter 5. Kluwer Academic, Boston.

Kosiol C, Goldman N. 2005. Different versions of the Dayhoff rate matrix. *Mol Biol Evol* **22:** 193–199.

Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. *Mol Biol Evol* **24:** 1464–1479.

Ku C-S, Cooper DN, Patrinos GP. 2016. The rise and rise of exome sequencing. *Public Health Genomics* **19:** 315–324.

Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. 2011. Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol* **12:** R118.

Lee DK, Saldivia VR, Nguyen T, Cheng R, George SR, O'Dowd BF. 2005. Modification of the terminal residue of apelin-13 antagonizes its hypotensive action. *Endocrinology* **146:** 231–236.

Lee C, Lai H-L, Lee Y-C, Chien C-L, Chern Y. 2014. The A2A adenosine receptor is a dual coding gene: a novel mechanism of gene usage and signal transduction. *J Biol Chem* **289:** 1257–1270.

Lee C, Kim KH, Cohen P. 2016. MOTS-c: a novel mitochondrial-derived peptide regulating muscle and fat metabolism. *Free Radic Biol Med* **100:** 182–187.

Li L, Yang G, Li Q, Tang Y, Yang M, Yang H, Li K. 2006. Changes and relations of circulating visfatin, apelin, and resistin levels in normal, impaired glucose tolerance, and type 2 diabetic subjects. *Exp Clin Endocrinol Diabetes* **114:** 544–548.

Li H, Hu C, Bai L, Li H, Li M, Zhao X, Czajkowsky DM, Shao Z. 2016a. Ultradeep sequencing of ribosome-associated poly-adenylated RNA in early *Drosophila* embryos reveals hundreds of conserved translated sORFs. *DNA Res* **23:** 571–580.

Li X, Chen Y, Qi H, Liu L, Shuai J. 2016b. Synonymous mutations in oncogenesis and apoptosis versus survival unveiled by network modeling. *Oncotarget* **7:** 34599–34616.

Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, Pritchard JK. 2018. Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet* **50:** 151.

Liao Y-C, Chou W-W, Li Y-N, Chuang S-C, Lin W-Y, Lakkakula BV, Yu M-L, Juo S-HH. 2011. Apelin gene polymorphism influences apelin expression and obesity phenotypes in Chinese women. *Am J Clin Nutr* **94:** 921–928.

Lim J, Crespo-Barreto J, Jafar-Nejad P, Bowman AB, Richman R, Hill DE, Orr HT, Zoghbi HY. 2008. Opposing effects of polyglutamine expansion on native protein complexes contribute to SCA1. *Nature* **452:** 713–718.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27:** i275–i282.

Lluch-Senar M, Delgado J, Chen W-H, Lloréns-Rico V, O'Reilly FJ, Wodke JA, Unal EB, Yus E, Martínez S, Nichols RJ, et al. 2015. Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Mol Syst Biol* **11:** 780.

Ma J, Ward CC, Jungreis I, Slavoff SA, Schwaid AG, Neveu J, Budnik BA, Kellis M, Saghatelian A. 2014. Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J Proteome Res* **13:** 1757–1765.

Ma J, Diedrich JK, Jungreis I, Donaldson C, Vaughan J, Kellis M, Yates JR, Saghatelian A. 2016. Improved identification and analysis of small open reading frame encoded polypeptides. *Anal Chem* **88:** 3967–3975.

Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, Mastrobuoni G, Rajewsky N, Kempa S, Selbach M, et al. 2015. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol*. **16:** 179.

Magny EG, Pueyo JI, Pearl FMG, Cespedes MA, Niven JE, Bishop SA, Couso JP. 2013. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* **341:** 1116–1120.

Makrythanasis P, Antonarakis SE. 2013. Pathogenic variants in non-protein-coding sequences. *Clin Genet* **84:** 422–428.

Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R. 2005. PRIDE: the proteomics identifications database. *Proteomics* **5:** 3537–3545.

Matsumoto A, Pasut A, Matsumoto M, Yamashita R, Fung J, Monteleone E, Saghatelian A, Nakayama KI, Clohessy JG, Pandolfi PP. 2017a. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* **541:** 228–232.

Matsumoto A, Clohessy JG, Pandolfi PP. 2017b. SPAR, a lncRNA encoded mTORC1 inhibitor. *Cell Cycle* **16:** 815–816.

McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: what, how and why. *Nat Rev Genet* **17:** 567–578.

Michel AM, Fox G, M Kiran A, De Bo C, O'Connor PBF, Heaphy SM, Mullan JPA, Donohue CA, Higgins DG, Baranov PV. 2014. GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res* **42:** D859–D864.

Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, et al. 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* **17:** 1797–1808.

Mouilleron H, Delcourt V, Roucou X. 2016. Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Res* **44:** 14–23.

Mudge JM, Harrow J. 2016. The state of play in higher eukaryote gene annotation. *Nat Rev Genet* **17:** 758–772.

Natarajan K, Meyer MR, Jackson BM, Slade D, Roberts C, Hinnebusch AG, Marton MJ. 2001. Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Mol Cell Biol* **21:** 4347–4368.

Nesvizhskii AI. 2014. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **11:** 1114–1125.

Niazi F, Valadkhan S. 2012. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3′ UTRs. *RNA* **18:** 825–843.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12:** 443–451.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. 2012. SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. *PLoS One* **7:** e37558.

Nishikura K. 2016. A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev Mol Cell Biol* **17:** 83.

O'Carroll A-M, Lolait SJ, Harris LE, Pope GR. 2013. The apelin receptor APJ: journey from an orphan to a multifaceted regulator of homeostasis. *J Endocrinol* **219:** R13–R35.

Okada AK, Teranishi K, Lobo F, Isas JM, Xiao J, Yen K, Cohen P, Langen R. 2017. The mitochondrial-derived peptides, HumaninS14G and small humanin-like peptide 2, exhibit Chaperone-like activity. *Sci Rep* **7:** 7802.

Olexiouk V, Crappé J, Verbruggen S, Verhegen K, Martens L, Menschaert G. 2016. sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res* **44:** D324–D329.

Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM. 2015. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front Genet* **6:** 235.

Ovcharenko I, Nobrega MA, Loots GG, Stubbs L. 2004. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res* **32:** W280–W286.

Oyama M, Itagaki C, Hata H, Suzuki Y, Izumi T, Natsume T, Isobe T, Sugano S. 2004. Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome Res* **14:** 2048–2052.

Oyama M, Kozuka-Hata H, Suzuki Y, Semba K, Yamamoto T, Sugano S. 2007. Diversity of translation start sites may define increased complexity of the human short ORFeome. *Mol Cell Proteomics* **6:** 1000–1006.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet* **7:** 337–348.

Pauli A, Norris ML, Valen E, Chew G-L, Gagnon JA, Zimmerman S, Mitchell A, Ma J, Dubrulle J, Reyon D, et al. 2014. Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* **343:** 1248636.

Pauli A, Valen E, Schier AF. 2015. Identifying (non-)coding RNAs and small peptides: challenges and opportunities. *BioEssays* **37:** 103–112.

Paulson HL, Shakkottai VG, Clark HB, Orr HT. 2017. Polyglutamine spinocerebellar ataxias—from genes to potential treatments. *Nat Rev Neurosci* **18:** 613–626.

Plaza S, Menschaert G, Payre F. 2017. In search of lost small peptides. *Annu Rev Cell Dev Biol* **33:** 391–416.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20:** 110–121.

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, et al. 2009. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19:** 1316–1323.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40:** D130–D135.

Pueyo JI, Magny EG, Couso JP. 2016a. New peptides under the s(ORF)ace of the genome. *Trends Biochem Sci* **41:** 665–678.

Pueyo JI, Magny EG, Sampson CJ, Amin U, Evans IR, Bishop SA, Couso JP. 2016b. Hemotin, a regulator of phagocytosis encoded by a small ORF and conserved across metazoans. *PLoS Biol* **14:** e1002395.

Quelle DE, Zindy F, Ashmun RA, Sherr CJ. 1995. Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell* **83:** 993–1000.

Raj A, Wang SH, Shim H, Harpak A, Li YI, Engelmann B, Stephens M, Gilad Y, Pritchard JK. 2016. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife* **5:** e13328.

Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. 2013. De novo ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet* **9:** e1003860.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17:** 405–424.

Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. 2014. Long non-coding RNAs as a source of new peptides. *eLife* **3:** e03523.

Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V, Wallace DC. 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* **303:** 223–226.

Saghatelian A, Couso JP. 2015. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat Chem Biol* **11:** 909–916.
Samandi S, Roy AV, Delcourt V, Lucier J-F, Gagnon J, Beaudoin MC, Vanderperre B, Breton M-A, Jacques J-F, Brunelle M, et al. 2017. Deep transcriptome annotation suggests that small and large proteins encoded in the same genes often cooperate. bioRxiv doi: 10.1101/142992.
Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* **12:** 683–691.
Schaab C, Geiger T, Stoehr G, Cox J, Mann M. 2012. Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol Cell Proteomics* **11:** M111.014068.
Sentinelli F, Capoccia D, Bertoccini L, Barchetta I, Incani M, Coccia F, Manconi E, Lenzi A, Cossu E, Leonetti F, et al. 2016. Search for genetic variant in the apelin gene by resequencing and association study in European subjects. *Genet Test Mol Biomarkers* **20:** 98–102.
Shihab HA, Rogers MF, Ferlaino M, Campbell C, Gaunt TR. 2017. GTB – an online genome tolerance browser. *BMC Bioinformatics* **18:** 20.
Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A. 2013. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* **9:** 59–64.
Soriguer F, Garrido-Sanchez L, Garcia-Serrano S, Garcia-Almeida JM, Garcia-Arnes J, Tinahones FJ, Garcia-Fuentes E. 2009. Apelin levels are increased in morbidly obese subjects with type 2 diabetes mellitus. *Obes Surg* **19:** 1574–1580.
Soussi T, Taschner PEM, Samuels Y. 2017. Synonymous somatic variants in human cancer are not infamous: a plea for full disclosure in databases and publications. *Hum Mutat* **38:** 339–342.
Southan C. 2017. Last rolls of the yoyo: assessing the human canonical protein count. *F1000Research* **6:** 448.
Steitz JA. 1969. Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. *Nature* **224:** 957–964.
Storz G, Wolf YI, Ramamurthi KS. 2014. Small proteins can no longer be ignored. *Annu Rev Biochem* **83:** 753–777.
Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. 2014. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156:** 1324–1335.
Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* **44:** 6614–6624.
Telejko B, Kuzmicki M, Wawrusiewicz-Kurylonek N, Szamatowicz J, Nikolajuk A, Zonenberg A, Zwierz-Gugala D, Jelski W, Laudański P, Wilczynski J, et al. 2010. Plasma apelin levels and apelin/APJ mRNA expression in patients with gestational diabetes mellitus. *Diabetes Res Clin Pract* **87:** 176–183.
Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. 2016. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci* **113:** 14330–14335.
The UnitProt Consortium. 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **42:** D191–D198.
Vanderperre B, Staskevicius AB, Tremblay G, McCoy M, O'Neill MA, Cashman NR, Roucou X. 2011. An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. *FASEB J* **25:** 2373–2386.
Vanderperre B, Lucier J-F, Bissonnette C, Motard J, Tremblay G, Vanderperre S, Wisztorski M, Salzet M, Boisvert F-M, Roucou X. 2013. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS One* **8:** e70698.
Venne AS, Kollipara L, Zahedi RP. 2014. The next level of complexity: crosstalk of posttranslational modifications. *Proteomics* **14:** 513–524.
Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13:** 227–232.
Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. 2013. Cancer genome landscapes. *Science* **339:** 1546–1558.
Wadler CS, Vanderpool CK. 2007. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc Natl Acad Sci* **104:** 20454–20459.
Wallace WE, Ji W, Tchekhovskoi DV, Phinney KW, Stein SE. 2017. Mass spectral library quality assurance by inter-library comparison. *J Am Soc Mass Spectrom* **28:** 733–738.
Wan J, Qian S-B. 2014. TISdb: a database for alternative translation initiation in mammalian cells. *Nucleic Acids Res* **42:** D845–D850.
Waters AM, Bagni R, Portugal F, Hartley JL. 2016. Single synonymous mutations in KRAS cause transformed phenotypes in NIH3T3 cells. *PLoS ONE* **11:** e0163272.
Willems P, Ndah E, Jonckheere V, Stael S, Sticker A, Martens L, Van Breusegem F, Gevaert K, Van Damme P. 2017. N-terminal proteomics assisted profiling of the unexplored translation initiation landscape in *Arabidopsis thaliana*. *Mol Cell Proteomics* **16:** 1064–1080.
Woo S, Cha SW, Guest C, Na S, Bafna V, Liu T, Smith RD, Rodland KD, Payne S. 2014. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale Next Generation Sequencing data. *Proteomics* **14:** 2719–2730.
Woo S, Cha SW, Bonissone S, Na S, Tabb DL, Pevzner PA, Bafna V. 2015. Advanced proteogenomic analysis reveals multiple peptide mutations and complex immunoglobulin peptides in colon cancer. *J Proteome Res* **14:** 3555–3567.
Wright JC, Mudge J, Weisser H, Barzine MP, Gonzalez JM, Brazma A, Choudhary JS, Harrow J. 2016. Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat Commun* **7:** 11778.
Xie S-Q, Nie P, Wang Y, Wang H, Li H, Yang Z, Liu Y, Ren J, Xie Z. 2016. RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res* **44:** D254–D258.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* **39:** 306–314.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586–1591.
Yen K, Lee C, Mehta H, Cohen P. 2013. The emerging role of the mitochondrial-derived peptide humanin in stress resistance. *J Mol Endocrinol* **50:** R11–R19.
Yosten GLC, Liu J, Ji H, Sandberg K, Speth R, Samson WK. 2016. A 5′-upstream short open reading frame encoded peptide regulates angiotensin type 1a receptor production and signalling via the β-arrestin pathway. *J Physiol* **594:** 1601–1605.
Young SK, Wek RC. 2016. Upstream open reading frames differentially regulate gene-specific translation in the Integrated Stress Response. *J Biol Chem* **291:** 16927–16935.
Young C, Podtelejnikov AV, Nielsen ML. 2017. Improved reversed phase chromatography of hydrophilic peptides from spatial and temporal changes in column temperature. *J Proteome Res* **16:** 2307–2317.
Yue S, Serra HG, Zoghbi HY, Orr HT. 2001. The spinocerebellar ataxia type 1 protein, ataxin-1, has RNA-binding activity that is inversely affected by the length of its polyglutamine tract. *Hum Mol Genet* **10:** 25–30.
Zhao Q, Gu D, Kelly TN, Hixson JE, Rao DC, Jaquish CE, Chen J, Huang J, Chen C-S, Gu CC, et al. 2010. Association of genetic variants in the apelin-APJ system and ACE2 with blood pressure responses to potassium supplementation: the GenSalt study. *Am J Hypertens* **23:** 606–613.

Received November 6, 2017; accepted in revised form March 27, 2018.