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, Alan A. Cohen, and Xavier Roucou

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; Saghateliat and Couso 2015; Plaza et al. 2017). These criteria result in an important underestimation of translated ORFs in the genome (Andrews and Rothnagel 2014; Saghateliat 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
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.

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).
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—ribose 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 IncRNAs) 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—ribose profiling, proteogenomics, and conservation signatures—are revealing this hidden proteome, a novel repertoire for biomarkers and therapeutic strat- nomics, 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 IncRNAs) 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—ribose profiling, proteogenomics, and conservation signatures—are revealing this hidden proteome, a novel repertoire for biomarkers and therapeutic strategies (Cousso 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 IncRNA 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 an IncRNA 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; Heinenon 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 IncRNA (GRCh37) to mRNA (GRCh38), highlighting the dynamic nature of genome annotations (Delcourt et al. 2017). Other biological roles attributed to smORFs include sarcoplasmic 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 IncRNAs, highlight the previously hidden coding potential of IncRNAs (Niazi and Valadkhan 2012; Slavoff et al. 2013; Ruiz-Orera et al. 2014; Ji et al. 2015). Admittedly, not all IncRNAs 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-
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 5′ UTR of mRNAs
Targeted proteomics for small peptides has also increased the detection of proteins mapped back to 5′ UTR ORFs (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 highlighted 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 ATAXT1 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 CIC, 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; Saghateilian 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 unicoding 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
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; MS12; TCF712; 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; NCO12; MYH11; KMT2D; PTR1B; SPP1; TRAP4; RNF213; POLE; FAT1; CAMTA1; FL14; ATP2B3; ZFHX3; ALK; ZNF521; KMT2A; GRIN2A; SETBP1) | 31.6%                                                       | 44                                   | 24.25%            |
| 3′ UTR clustered data set | 14        | 141                           | 6 (ETV6; PPM11D; CCND2; AR; BCL11B; BCL11A)               | 36.4%                                                       | 40                                   | 20.8%             |
The current genome annotation system has been blamed for number make it through to genome annotation (Southan 2017). number of smORF and alternative ORF discoveries, only a limited 3.3% in 2017 (Southan 2017). This means that despite the large Swiss-Prot falls short, with an increase from 3.1% in 2009 to et al. 2016). As a result, the foreseen increase in smORF count in proteins or domains (Furuno et al. 2003; Pruitt et al. 2012; Aken et al. 2016). These pipelines, however, invoke strategies have been adopted over the past years, which essentially 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 (Harlow 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).
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; Hutlin 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,
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_ 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...
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 pathophysiological 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. 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.

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. XLes, 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.

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

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

Allen BL, Avling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez-Banet J, Billis K, García Girón C, Houliher T, et al. 2016. The Ensembl gene annotation system. Database 2016: baw093.

Alfarano C, Foussal C, Lairol O, Calise D, Attané C, Anesia R, Daviaud D, Waneçq E, Partini 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, Andrews SJ, Rothnagel JA. 2014. Emerging evidence for functional peptides 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, Cornican P, Morris DW, Shatsky IN, Baranov PV. 2015. Translation of S’ leaders is pervasive in genes resistant to eIF2 repression. eLife 4: e09791.

Andrees S, 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 JP. 2014. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. eLife 3: e03532.

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: 1764–1771.

Autio KJ, Kastaniotis AJ, Pospiech H, Milinailen IJ, Schonauer MS, Dieckmann CL, Hiltunen JK. 2008. An ancient genetic link between vertebrate mitochondrial fatty acid synthase and RNA processing. FASEB J 22: S69–S78.

Baek J, Lee J, Yoon K, Lee H. 2017. Identification of unannotated small genes in Salmo sciatus. 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 RJ, di Santi Rodrigues A, Nishi MY, Gomes NLRA, Faría JAD, de Moraes DR, Carvalho LR, Frade EMC, Domenico 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 168: 16–16.

Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Veijnar CE, Lee MT, Rajewsky N, Walthier 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, Pelletier J, Roucou X. 2015. 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, Hu W, Wu W, Cating 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: 571–576.

Chick JM, Kohli A, Emergider D, Nutinon DP, Zhai R, Rad R, Hurtlin EL, Gygi SP. 2015. A mass-tolerable database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol 33: 743–749.

Chen 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, Pord SK, Nekuteno A. 2007. A first look at ARFome: dual-coding genes in mammalian genomes. PLoS Comput Biol 3: e91.

Claverie JM, Poirat O, Lopez F. 1997. The difficulty of identifying genes in anonymous vertebrate sequences. Comput Chem 21: 203–214.

Cooper GM, Stone EA, Anderson DM, Andrews SJ, Rothnagel JA. 2005–2007. Sequence profiling Program, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901–913.

Couso JP, Patraquin P. 2017. Classification and function of small open reading frames. Nat Rev Mol Cell Biol 18: 575–589.

Crépée J, Jadhav E, Koch A, Steyaert S, Gwiron D, De Keulenaer S, De Meester É, De Mey E, Van Criekinge W, Van Damme P, et al. 2015. PROTEOMFORMER: deep proteome coverage through ribosome profiling and MS integration. Nucleic Acids Res 43: e29.

Davydov EV, Goede 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++ with an updated human ancestral sequence. PLoS Comput Biol 6: e1001025.

Delascour 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 view of an mRNA. Proteomics doi: 10.1002/pmic.201700858.

Desiere F, Deutsch EW, King NL, Nevizhiskal AI, Mällick P, Eng J, Chen S, Eddes J, Lovenenish SN, Aebersol R. 2006. The PeptideAtlas project. Nucleic Acids Res 34: D655–D658.

D’Lima NG, MaJ, Winkler L, Chou Q, Loh KH, Corpuz EB, Budnik BA, Lykke-Andersen J, Saphattele A, Slavova SA. 2017. A human microprotein that interacts with the mRNA decapping complex. Nat Biol 13: 174–180.

Dray C, Knauf C, Daviaud D, Waget A, Boucher J, Baléon M, Cani PD, Attané C, Guigné C, Carpené 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, Disses E, Daviaud D, Martin P, Attané C, Waneucq C, Gost 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, Gaviss ER, Weissman JS. 2013. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. eLife 2: e01179.

Ebihn I, Takemoto-Tsutsui M, Watanabe S, Koyama H, Endo Y, Kimata K, Igashira T, Murakami K, O random tube R, Ohtsumi 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 type 2 diabetes mellitus. Exp Clin Endocrinol Diabetes 116: 289–292.

Escobar B, de Cancer G, Fernández-Miranda G, Casco C, Bravo-Cordero JJ, Montoya AG, 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, Muchovzo W, Brunner AM, Schackwitz W, Gunter L, Chen JG, et al. 2014. Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. Nat Genet 46: 1089–1096.

Fahraeus R, Marín M, Olivares-Illana V. 2016. Whisker mutations: cryptic messages within the genetic code. Oncogene 35: 3753–3760.
Saghatelian A, Couso JP. 2015. Discovery and characterization of smORFs-encoded bioactive polypeptides. Nat Chem Biol 11: 909–916.
Samandi S, Roy AV, Delcour V, Lucier J-F, Gagnon J, Beaudon 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.
Senticelli F, Capoccia D, Bertocci L, Barchetta I, Incani M, Coccia F, Manconi E, Lenzi A, Cosu E, Leonetti F, et al. 2016. Search for genetic variant in the apelin gene by sequencing and association study in European subjects. Genet Test Mol Biomarkers 20: 99–102.
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, Rudnik BA, Rinn JL, Saghatelian A. 2013. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 9: 59–64.
Soriger G, 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 KRAS cause transformed phenotypes in NIH3T3 cells. PLoS ONE 11: e0163272.
Willems P, Ndhah J, 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. Proteogeneomic 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 proteogeneomic analysis reveals multiple peptide mutations and complex immunoglobulin peptides in colon cancer. J Proteome Res 14: 3555–3567.
Wright JC, Mudge J, Weisser H, Batzine 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 genome-wide database 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 SW, 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, Podtelezhnikov 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, ataxon-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, Hisse J, Rao DC, Jaquish CE, Chen J, Huang J, Chen C-S, Gu CC, et al. 2010. Association of genetic variants in the aelin-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.