Perspective
Prediction of RNA subcellular localization: Learning from heterogeneous data sources

Anca Flavia Savulescu,1,* Emmanuel Bouilhol,2,3 Nicolas Beaume,4 and Macha Nikolski2,3,*

SUMMARY
RNA subcellular localization has recently emerged as a widespread phenomenon, which may apply to the majority of RNAs. The two main sources of data for characterization of RNA localization are sequence features and microscopy images, such as obtained from single-molecule fluorescent in situ hybridization-based techniques. Although such imaging data are ideal for characterization of RNA distribution, these techniques remain costly, time-consuming, and technically challenging. Given these limitations, imaging data exist only for a limited number of RNAs. We argue that the field of RNA localization would greatly benefit from complementary techniques able to characterize location of RNA. Here we discuss the importance of RNA localization and the current methodology in the field, followed by an introduction on prediction of location of molecules. We then suggest a machine learning approach based on the integration between imaging localization data and sequence-based data to assist in characterization of RNA localization on a transcriptome level.

THE IMPORTANCE OF RNA SUBCELLULAR LOCALIZATION
Gene expression in eukaryotes is regulated at various stages of the life cycle of RNA molecules, including transcription initiation, RNA processing, stability, subcellular localization, and translation into protein in the case of mRNAs, stages which are often linked. Subcellular localization of RNA transcripts leads to restriction of translation in a spatial and temporal manner (Bashirullah et al., 1998; Besse and Ephrussi, 2008; Jansen, 2001; Kloc et al., 2002; Martin and Ephrussi, 2009; Mardakheh et al., 2015; Zappulo et al., 2017; Savulescu et al., 2021), as well as serving to avoid toxicity of protein products and to corroborate rapid cellular responses (Shahbaban and Chartrand, 2012). Early studies in the field of RNA subcellular localization investigated the subcellular location of one and up to a few mRNA transcripts in developmental models, such as the Drosophila embryo and Xenopus oocyte, determining morphogen gradients and cellular fates (Macdonald and Struhl, 1998; Tautz and Pfeifle, 1989), as well as polarized cells such as migrating fibroblasts, budding yeast, and neuronal cells (Mili et al., 2008; Martin and Ephrussi 2009; Battish et al., 2012; Buxbaum et al., 2014; Tzingounis and Nicoll, 2006; Yasuda et al., 2017). However, recent years have seen an increasing number of studies in various organisms/models of RNA subcellular localization, resulting in the consensus that RNA subcellular localization is indeed not limited to a handful of RNAs in a small number of systems but rather a widespread phenomenon, which may apply to the majority of cellular RNAs, including non-coding RNAs (Bouvette et al., 2017; La Manno et al., 2018; Lécuyer et al., 2007; Moor et al., 2017; Sharp et al., 2011; Weis et al., 2013; Cabili et al., 2015; Zappulo et al., 2017).

The subcellular location of the RNA can influence which proteins will bind the transcript, contributing to the RNA’s fate, such as directing it toward degradation, increasing/decreasing its rate of translation, and determining molecular interactions. Consequently, the spatial distribution of the RNA can potentially influence the cellular concentration and location of its protein product (Brangwynne et al., 2009; Katz et al., 2012, 2016; Moor et al., 2018; Savulescu et al., 2021), which can, in turn, affect the cell’s function and capacity to interact with adjacent cells or react to various environmental cues. Similarly, association of transcripts with specific cellular structures (Hughes and Simmonds, 2019; Khong et al., 2017; Padrón et al., 2019; Savulescu et al., 2021; Suter, 2018; Wilbertz et al., 2019 and others) may indicate a functional correlation between the transcript in the cell where this association occurs and broad cellular processes in the same cells, such as polarization, differentiation, etc. Taken together, this indicates that in addition to expression levels of RNA transcripts, subcellular spatial distribution of these transcripts may contribute to cell state and type (for example, Moor et al., 2018). Using solely single-cell genomics approaches, which are standard in cell...
state characterization studies, spatial information of RNA transcripts might be lost. For example, two neighboring cells in a tissue, which possess similar concentrations of the same RNA transcripts, would be classified by single-cell genomics approaches as the same cell type/state; however, the RNAs in these cells may exhibit marked patterns in subcellular dispersion (Savulescu et al., 2020) (Figure 1, Panel B). This again emphasizes the significance of characterizing subcellular distribution alongside expression levels of RNA and correlation of RNA with cellular markers, such as the MTOC or ER.

In addition to contributions in basic research, knowledge of RNA subcellular spatial localization can be beneficial in biomedical applications. For instance, Didiot et al. (2018) show that htt mRNA, which encodes the protein responsible for Huntington’s disease, is located both in the nucleus and cytoplasm (50/50) in neurons, while being purely cytoplasmic in non-neuronal cells. This has a therapeutic impact as nuclear htt mRNA is more stable and more resistant to oligonucleotide therapy than the cytoplasmic htt. Consequently, new therapeutic strategies can be envisioned, such as modifying the mRNA encoded by the abnormal allele of huntingtin to redirect it to the cytoplasm where oligonucleotide therapy is more efficient. Taken together, given the functional relevance of subcellular spatial distribution of RNA in both basic and clinical contexts, tools that are able to efficiently detect subcellular localizations at a transcriptome-wide level are urgently needed.

Visualization of RNA using single-molecule fluorescent in situ hybridization (smFISH) or MS2-system-based techniques have been the gold standard method to study RNA subcellular localization. In recent years, smFISH-based techniques and downstream in silico procedures for localization of mRNAs have been developed to cope with a large number of RNA transcripts. However, these methods can be technically challenging, be expensive, and often require sophisticated equipment. Machine learning, which has been successfully applied in various biological fields, including prediction of subcellular locations of proteins and of mRNAs, may be a suitable complementary approach to predict the precise subcellular localization of a large number of RNA species in a variety of systems on a transcriptome-wide level. In the perspective below, we provide a brief overview of techniques to study subcellular RNA localization and then introduce how machine learning can be applied to the study of RNA subcellular localization and discuss how machine learning can be harnessed to predict precise RNA subcellular localization. Finally, we discuss potential synergies between experimental and computational approaches of subcellular localization.

**CURRENT METHODS TO STUDY SUBCELLULAR LOCALIZATION OF RNA**

To date, the most popular approach to study subcellular localization of RNA is image based. A large number of studies make use of smFISH-based techniques followed by epifluorescent or confocal microscopy to visualize and quantify intracellular mRNAs (typical smFISH images in Figure 1, Panel A). Alternatively, an
MS2 tagging system followed by live cell imaging is applied (for example, Bertrand et al., 1998; Hocine et al., 2013 and others). smFISH makes use of multiple single-stranded DNA oligonucleotides, each labeled with a single fluorophore, which tile a specific RNA target (Raj et al., 2008). The signal obtained from multiple single fluorophores is sufficiently bright to be seen as a spot on an epifluorescent or confocal microscope and can easily be quantified (Raj et al., 2008). In recent years, a variety of smFISH-based techniques to increase the capacity of the method have emerged. These include the use of multiple rounds of labeling and imaging of the same sample to label a large number of RNAs. osmFISH is one such smFISH-based technique, in which the cellular organization of the mouse somatosensory cortex was mapped by labeling 33 RNAs over the course of 13 rounds. The number of labeled RNAs can be further increased by applying FISH probe barcoding and sequential labeling approaches, such as multiplexed error-robust FISH (MERFISH) (Moffitt and Zhuang, 2016) and sequential FISH (seqFISH and seqFISH+) (Lubeck et al., 2014; Eng et al., 2019). These methods possess the capacity to label hundreds to 1000 RNAs, nearly approaching full-transcriptome imaging. Additionally, Wang et al. (2018), used in situ amplification of RNA-specific probe barcode regions, decoding by 3D sequencing within samples converted to a hydrogel matrix, in a method termed spatially resolved transcript amplicon readout mapping (STARmap), to target hundreds of genes in various tissues (Wang et al., 2018). Branched DNA (bDNA)-based techniques rely on signal amplification using a series of non-isotopic DNA probes hybridized in a sequential manner to detect and quantify RNA (Player et al., 2001; Wang et al., 2012; Battich et al., 2013).

Following imaging of smFISH samples, data analysis is carried out. The analysis contains several steps, including segmentation, image processing to reduce background noise, readout of barcodes associated to the different transcripts in the case of barcoding-based smFISH methods, as well as precise localization and quantification of RNA FISH spots and in some cases, association of detected spots with cellular landmarks. A wide range of computational methods exist to extract quantified data for RNAs from smFISH images, and a number of widely used microscopy image analysis software programs are available for mRNA spot detection, such as the ICY spot detector (de Chaumont et al., 2012), ImageJ spot detection, FISH-Quant (Mueller et al., 2013), its more recent version FISH-quant v2 (Imbert et al., 2021), and CellProfiler (McQuin et al., 2018), among others. Many tools also provide cell and nucleus segmentation to apply on nuclear and cellular marker stainings that are often acquired with smFISH images. The extracted vectorized features can take different forms, such as RNA spot coordinates, or statistical features such as RNA counts and/or densities per cell or in different compartments as well as their positioning relative to cellular landmarks. See Imbert et al. (2021); Samacoits et al., 2018 for examples of possible features as well as their representation in relation to cellular landmarks (Savulescu et al., 2021).

In addition to image-based approaches, subcellular localization of RNA can be studied using methods such as biochemical cellular fractionation or physical separation of cellular compartments, followed by RNA detection using RNA sequencing or microarrays (for example, Mili et al., 2008; Cajigas et al., 2012; Bigler et al., 2017). Although these image-based and biochemical approaches have provided high-resolution data regarding the subcellular localization of RNA in cells and tissues, they each have limitations both in the spatial resolution of the method, its throughput, as well as in the reliance on effort-heavy computational analysis of the generated images to detect mRNA localizations.

To summarize, given the increasing number of studies that characterize spatial and temporal subcellular localization patterns of RNAs, as well as the functional relevance of variation in spatial distribution of RNA molecules, there is an increasing need to develop appropriate tools and technologies to capture fine-grained variations in spatial distribution of RNAs. These tools would be required to be quantitative, analyze data on a single-cell level, and be accustomed for high-throughput data.

RNA LOCALIZATION PREDICTION

The rise in studies emphasizing the importance of RNA spatial and temporal subcellular localization in biological processes calls for an increased access to such information. Although it is clear that wet-lab experiments are the most straightforward way to characterize subcellular localization of biomolecules, these experiments are typically time- and money consuming. Contrastingly, with sufficient data and powerful in silico methods, predictions of subcellular localization of biomolecules would be far less expensive and intrinsically high throughput. We argue that this makes place for computational prediction of RNA localizations.
Prediction of subcellular locations based on sequence

The task of predicting subcellular location of biomolecules is not new. In the previous decade, a significant number of methods have been developed to predict subcellular locations of proteins (Emanuelsson et al., 2007; Imai and Nakai, 2010; Wan and Mak, 2015). The methods are based on certain features extracted from protein sequences, such as specific motifs, and aim to predict a rough location in terms of cell regions/departments such as those defined by the UniProt controlled vocabulary for subcellular locations. Often these methods use the guilt by association approach, using the homology with proteins whose location has been experimentally confirmed, the main experimental methods to establish protein localization being mass spectrometry and immunofluorescence.

Some mRNA transcripts contain within their sequences distinct motifs, which on their own, or by forming distinct secondary structures, have been shown to be determinant of the RNA subcellular location (reviewed in Martin and Ephrussi, 2009; Shahbabian and Chartrand, 2012 and others). Hence, several methods for prediction of RNA subcellular location have started to be developed based on mRNA/cDNA sequence composition (Yan et al., 2019; Garg et al., 2020), as well as introducing the predicted secondary structure (Yan et al., 2016). Interestingly, both of these very recent approaches are leveraged by the machine learning technology: deep recurrent networks (convolutional neural network [CNN], long short-term memory network [LSTM], and attention layers) for the former and more classical support vector machine (SVM)-based methodology for the latter. As in the protein world, these methods require training data sets providing mRNA subcellular locations for each annotated human protein-coding gene, such as cytosol, nuclear, endoplasmic reticulum (ER), membrane, etc. However, in contrast to the protein world, where annotation of data that is essential for the development of supervised machine learning typically exists, the RNA data generally lacks annotation. This lack of annotation is reflected in the very recent introduction of such methods for the mRNA, as well as in the fact that there remain to be only a few methods. Consequently, their adoption by the wider scientific community remains to be seen.

Predicting the subcellular location of non-coding RNAs (ncRNA), such as long non-coding RNAs (lncRNAs), microRNAs (miRs), and others is generally a more difficult task than for mRNAs. This is due to several factors, including the complexity of intra-molecular organization that ncRNAs can exhibit (Yan et al., 2016), as well as their typical short lengths and fewer known localization-correlated motifs than mRNAs (Ross et al., 2021; Constanty and Shkumatava, 2021). However, some associations of sequence motifs in IncRNAs with their subcellular localization have been identified (Zhang et al., 2014) and databases containing information on cellular compartments to which the IncRNAs localize have been developed, including LncLocate (Mas-Ponte et al., 2017) and RNAlocate (Zhang et al., 2016). Given these recent developments, several methods have been proposed to predict subcellular location of lncRNAs. For example, Su et al. (2018) have combined a PseKNC and SVM to predict subcellular location of lncRNAs to the ribosome and exosome, Gudenas and Wang (2018) have developed a deep neural network to predict whether lncRNAs are nuclear or cytosolic based on their sequence, and Cao et al., (2018) used k-mer and high-level abstraction features generated by unsupervised deep models to construct four classifiers and predict five subcellular localizations of lncRNAs. Similarly, a number of tools exist to predict the location of miRNAs, such as an approach based on gene ontology (GO)-based functional similarity (Yang et al., 2018) and an SVM-based predictor (Meher et al., 2020), both relying on the mRNA locations from the RNAlocate database.

Prediction of subcellular location from microscopy data

More recently, a plethora of tools have been developed to predict the locations of molecules based on microscopy imaging data rather than sequence, in particular, in application to protein subcellular localization. For example, Newberg and Murphy (2008) have developed a framework for image-based protein subcellular location prediction and it has been successfully applied to the Human Protein Atlas database (Newberg and Murphy, 2008). This has paved the way for a whole set of work of microscopy-image-based subcellular localization prediction of proteins, which includes methods relying on k-NN classifiers, SVMs, artificial neural networks, decision trees, and deep CNNs (Xijie Lu and Moses, 2016; Jiang et al., 2019). Recently, a crowd citizen science effort has attracted participants to annotate the subcellular locations of proteins in images and resulted in a novel deep learning method based on transfer learning (Sullivan et al., 2018), capable of predicting distributions of proteins to major organelles. Not only the authors have achieved high accuracy of predictions but also they were able to construct a fully annotated data set (Sullivan et al., 2018). In the next section, we will discuss how sequence data could be combined with smFISH data to improve the predictive accuracy.
PERSPECTIVE ON PREDICTING RNA SUBCELLULAR LOCALIZATION FROM HETEROGENEOUS DATA

As discussed earlier, the two main sources of data for characterization of subcellular localization of RNA are RNA sequence features and smFISH images. In addition, as subcellular localization is considered as a means to restrict translation, among other functions, information regarding the subcellular localization of the encoded protein product, as well as cell-type-specific information may aid in subcellular localization characterization of RNA. However, in some cases, including embryonic development models, subcellular localization of proteins might not be correlated with the subcellular localization of the encoding mRNAs (Knaut et al., 2000; Little et al., 2015; Mardakheh et al., 2015) and, as such, might be misleading if considered as a sole parameter for training the model. Additional information regarding the cellular model should be considered to account for variations in subcellular distribution of an mRNA and its encoded protein.

Sequence features are easy to obtain; however, they do not provide straightforward information regarding subcellular localization. Moreover, current methods of prediction of molecules’ location remain limited in their predictive power. For example, precision can widely vary, as reported in the study by Garg et al. (2020), where the area under the curve lies between 0.7 and 0.98 for different compartments. Importantly, no method with good performance for most cellular compartments is available. On the other hand, fluorescent microscopy imaging data are the gold standard to determine the RNA subcellular localization; however, it is typically costly, time-consuming, and complex to perform for all known RNAs. In this part we will discuss how imaging data may be combined with sequence features to accurately predict RNA subcellular location.

Beyond the improvement of both observation techniques and predictive algorithms, we would like to suggest that a synergy between the existing two approaches — sequence and image-based — may assist in deciphering RNA localization (Figure 2). To illustrate our argument, consider the following manual steps to be followed when characterizing the subcellular localization of an RNA transcript that has not yet been visualized using smFISH or MS2-system based techniques.

1. The first step would be to extract specific features from the transcript sequence within the RNA, which would aid in characterization of its role and location. These include conserved motifs, regulatory sequences, domains that bind specific RNA binding proteins (RBPs), secondary structures, etc.

2. Second, one would have to collect the information on the role and subcellular localization of the protein product in the case of mRNAs and the role of the RNA, if known, in the case of long ncRNAs, as well as additional cell-type- and model-specific localization data.

3. Third, identify RNAs with similar sequences and collect the information from databases on RNA localization, protein localization, smFISH and IF images, etc, for these RNAs.

We argue that prediction of the subcellular localization of the RNA could be made by integrating these collected heterogeneous data of the RNA (sequence-based features of the RNA, role, and subcellular localization of the protein product/RNA, etc) as well as imaging data collected from databases of RNAs (Figure 2). A machine learning approach (e.g., using multimodal learning) of this objective would be built of the following steps:

1. Gather a database of RNAs for which both sequence features and imaging information are known (Figure 2, Panel A).

2. Extract features from sequence and images as detailed in the “RNA localization prediction” section, and vectorize them.

3. Train a model based on sequences, represented by their vector of feature. Figure 2 offers an example on how to train a model with heterogeneous data (Figure 2, Panel A).

4. Use this model to predict localization of RNA of interest (see the following part for details) (Figure 2, Panel B).

The critical part of this process is the use of a model created from RNAs for which both sequences and imaging data are available, to predict the localization of RNAs for which only sequences are available. Recent developments in machine learning and deep learning (siamese networks, adversarial training) may help to
address this challenge. For instance, a network from the output of two networks, each based on one data type, could be trained. Once trained, this “aggregate network” could be fed by one type of data and still retain its predicting power. These approaches would create models “boosted” by the availability of smFISH in the training phase but able to predict with sequence features only.

LIMITATIONS IN THE DATA THAT MIGHT CHALLENGE PREDICTION
Prediction greatly depends on the data used to train the model. First, a large amount of data might be required to train a model properly, especially for methods such as deep learning or a combination of classifiers, which might be data greedy. Thus, increasing the number of experimentally observed locations...
would be helpful to train better models. Second, it is clear that our knowledge of RNA subcellular localization is not uniform for different RNA species—some classes of RNA, e.g., IncRNA and miRNA, are under-represented in our knowledge map of RNA subcellular localization, potentially leading to bias in the models. Targeted experiments toward the “blind spots” of the RNA localization transcriptome map would help to correct these biases.

One of the avenues to identify the important missing data that has to be acquired is to look at the most distinguishing features, which are features in the model that contribute the most to the prediction. Examining the importance of different features in prediction can be helpful to decipher biological processes behind RNA subcellular localization; however, most importantly, by considering the distinguishing features, it is possible to design experiments to acquire missing data for the subcellular localization of some transcripts. These data in turn increase the impact of distinguishing features and improve prediction accuracy. Among features that are already known to have strong impact on prediction, we can cite k-mer composition, RNA-protein binding motifs, and other genomic sequence features as identified in the study by Gudenas and Wang (2018), who have built a model to classify IncRNAs into cytosolic or nuclear and performed the analysis of feature importance. In this work, the authors have found that the k-mer composition accounted for 90% of the decision. Because it is assumed that such distinguishing features are linked to the biological processes underlying the subcellular localization, guiding acquisition of new data by the principle of enriching the features that contribute to precision accuracy appears to be a promising avenue to increase the impact of machine learning approaches in the field.

An additional limitation concerns the difficulties in segmentation of specialized cells, such as neurons, and the related RNA quantification in relevant compartments. Still, a number of local morphological descriptors such as dendritic tree, radial extension, soma area, and branching complexity have been computed to date (Shefi et al., 2002). Additionally, invariant measures such as Hu’s moments can be included (Bhaskar et al., 2019). Such morphological features can be used by a downstream machine learning pipeline. Segmentation, although not always perfect, can be often well performed, and although the annotation of neurons is still not yet fully automated, significant progress has been made on this topic with very promising results (Li et al., 2019; Lin and Zheng, 2019; Schubert et al., 2019). Normalization of RNA quantities in different compartments can then be done by quantization methods, such as DypFISH (Savulescu et al., 2021) and others. Naturally, if the identification of certain compartments is imperfect, the corresponding normalized RNA quantities will consequently be skewed, which will inevitably impact predictions that would use these data. In conclusion, any machine learning localization prediction method is only as good as the data that it are built upon.

Finally, regarding features that may involve two-dimensional (2D) or three-dimensional (3D) features such as zip codes, machine learning approaches have been shown to be able to integrate complex features such as 2D and 3D structures from sequences (Singh et al., 2019; Jumper et al., 2021; Sweeney et al., 2021), and thus, machine learning methods may be able to take the zip code elements into account, at least indirectly.

**Addressing the challenge of variability in subcellular locations**

One of the major limitations in our understanding of RNA localization is the variability and dynamics of the subcellular localization. In many cases, a given RNA may be addressed to different subcellular compartments, depending on the circumstance (cell type, cell state, environmental conditions, various treatments, as well as temporally). This aspect remains poorly understood for the majority of cellular RNAs. Additionally, a given RNA’s subcellular localization patterns might not be clearly pronounced or their localized enrichment high enough for detection. This typically occurs in developmental systems, such as the early Drosophila embryo, where only up to 4% of a particular mRNA localizes to germ granules, while the remaining fraction disperses through the embryo (Bergsten and Gavis, 1999; Jambor et al., 2015; Trcek et al., 2015). The current practice is to consider only the localization with the highest probability to be the right one. Using highly efficient models, localizations with enough confidence could all be considered as correct, being the sign of a multi-localized RNA. The level of confidence could ultimately give an estimate of the tendency of this RNA to be addressed at different localizations.

Variability in subcellular location of a given RNA can also be addressed by performing smFISH on cells grown on microfabricated patterns (Savulescu et al., 2021). Micropatterning of cells reduces cell-to-cell variability and allows for a higher resolution, quantitative characterization of subcellular localization of RNAs (Savulescu et al., 2021). Comparison of the spatial localization of a given RNA in the same cell type on micropatterns under different conditions can be used to assess the variability in subcellular localization.
conditions or in different micropatterned cell types should allow for a thorough characterization of variability in subcellular localization of the given RNAs in these conditions/cell types. This should, in turn, assist in accounting for variability when predicting subcellular localization of RNAs.

To summarize, variability of RNA location dependent on cell types/conditions/cellular compartments etc is important to account for and should be integrated as parameters in the model. Although current models are limited in this respect, we would like to suggest that with growth of available annotated data, prediction models should gradually improve.

**DISCUSSION**

High-precision methods for determination and quantification of subcellular localization of RNAs, such as smFISH, are now well established; however, these techniques remain time-consuming and costly. To drive better understanding of cellular processes, there is a need for development of methods to cover the broad landscape of RNA subcellular localization, for a large range of RNAs and conditions, ideally at the whole-transcriptome level. We would like to advance the argument that observation and prediction of RNA subcellular localization are two complementary approaches that can be leveraged together. Although they remain largely disconnected, linking them has the potential to greatly increase knowledge in the field. As mentioned earlier, the quality of the prediction is directly linked both to the quantity and the quality of the available datasets. Thus, building robust models for the prediction of RNA location requires growth of the available and well-annotated data. This increase of relevant data can be driven both by the biological questions, as it is currently the case, the development of relevant data repositories, but also in a complementary fashion by the requirement to fill the gaps in the existing predictive features that are used to populate machine learning models.

As previously discussed, any bias in annotations and/or errors in the upstream analyses are inevitably propagated into predictions. A possible avenue to circumvent these biases would be a non-supervised machine learning approach that would make its own inferences about the structures it finds in the data instead of relying on its vectorized representation. However, unsupervised learning requires even more data than the supervised counterparts. An unsupervised approach would thus be an excellent way to circumvent annotation biases and errors and possibly provide a solution when the field is mature enough and more data are available.

We argue that cost, technical, and time considerations can be alleviated by designing robust predictive methods that take advantage of heterogeneous data, where RNA location prediction is based on both imaging and sequence data. Continued growth of available data sets containing both the data itself and its reliable annotations and covering the diversity of different RNA species in various contexts provides hope that the construction of robust models based on heterogeneous data—both imaging and sequences—for prediction of subcellular RNA localization is realistically feasible in the near future.

**ACKNOWLEDGMENTS**

We thank Jean-Baptiste Sibarita and Dana M. Savulescu for fruitful discussions and critical reading of the manuscript.

**AUTHOR CONTRIBUTIONS**

Conceptualization, A.F.S., M.N., N.B.; writing - original draft, A.F.S., E.B., M.N., N.B.; machine learning architecture for location prediction - A.F.S., E.B., M.N.; writing - reviewing and editing, A.F.S., E.B., M.N., N.B.

**DECLARATION OF INTERESTS**

The authors declare no competing interests.

**REFERENCES**

Bashirullah, A., Cooperstock, R.L., and Lipshitz, H.D. (1998). RNA localization in development. Annu. Rev. Biochem. 67, 335–394.

Bhaskar, D., Lee, D., Knútsdóttir, H., Tan, C., Zhang, M.H., Dean, P., Roskelley, C., and Edelstein-Keshet, L. (2019). A methodology for morphological feature extraction and unsupervised cell classification. BioRxiv. https://doi.org/10.1101/623793.

Battich, N., Stoeger, T., and Pelkmans, L. (2013). Image-based transcriptomics in thousands of single human cells at single-molecule resolution. Nat. Methods 10, 1127–1133.

Batish, M., van den Bogaarda, P., Kramer, F.R., and Tyagi, S. (2012). Neuronal mRNAs travel singly into dendrites. Proc. Natl. Acad. Sci. U S A 109, 4645–4650.
Bergsten, S. E., and Gavis, E. R. (1999). Role for mRNA localization in translational activation but not spatial restriction of nanos RNA. Development 126, 659–669.

Bertrand, E., Chartard, P., Schaefer, M., Shenoy, S. M., Singer, R. H., and Long, R. M. (1998). Localization of ASH1 mRNA particles in living yeast. Mol. Cell. Biol. 2, 437–445.

Besse, F., and Ephrussi, A. (2008). Translational control of localized mRNAs: restricting protein synthesis in space and time. Mol. Cell. Biol. 28, 971–981.

Boussou, P.-L., Bod, A., Baffi, T., and Ephrussi, A. (2007). Celfr-seq reveals broad asymmetric mRNA and non-coding RNA distribution profiles in Drosophila and human cells. Nucleic Acids Res. 35, W111–W115.

Brangwynne, C., Eckmann, C. R., Courson, D. S., Biedenbender, L., and Singer, R. H. (2014). Single-biactin mRNA detection in neurons reveals a mechanism for regulating its translatability. Science 345, 419.

Cabit, M. N., Dunagin, M. C., and McClanahan, P. D. (2019). Single-molecule analysis of human lncRNAs at single-cell and single-molecule resolution. Genom. Biol. 16, 20.

Caig, I. J., Tushev, G., Will, T. J., tom-Dick, F., and Tomancak, P. (2015). Systematic imaging reveals features and changing localization of mRNAs in Drosophila development. Elife 4, e05003.

Jambor, H., Suredranath, V., Kalinka, A. T., Mejstrik, P., Saalfeld, S., and Tomancak, P. (2015). Systematic imaging reveals features and changing localization of mRNAs in Drosophila development. Elife 4, e05003.

Jansen, R. P. (2001). mRNA localization: message on the move. Nat. Rev. Mol. Cell. Biol. 2, 247–256.

Jiang, Z., Wang, D., Wu, P., Chen, Y., Shang, H., Wang, L., and Xie, H. (2019). Predicting subcellular localization of multisite proteins using differently weighted multi-label k-nearest neighbors sets. Technol. Health Care 27(Suppl 1), 185–193.

Katz, Z. B., Wells, A. L., Park, H. Y., Wu, B., Shenoy, S. M., and Singer, R. H. (2012). β-Actin mRNA compartmentalization enhances focal adhesion stability and directs cell migration. Genes Dev. 26, 1885–1890.

Jumper, J., Evans, R., Pritzel, A., Green, T., Fidjeland, T., and Fergus, R. (2016). Mapping translation ‘hotspots’ in live cells by tracking single molecules of RNA and ribosomes. Elife 5, e10415.

Kong, A., Matheny, T., Jain, S., Mitchell, S. F., Wheeler, J. R., and Parker, R. (2017). The stress granule transcriptome reveals principles of mRNA accumulation in stress granules. Mol. Cell 68, 808–820.

Kloc, M., Zearfoss, N. R., and Etkin, L. D. (2002). Mechanisms of subcellular mRNA localization. Cell 108, 533–544.

Knaus, H., Pelegrí, F., Bohmann, K., Schwartz, H., and Nüsslein-Volhard, C. (2000). Zebrafish mRNA but not its protein is a component of the germ plasm and segregates asymmetrically before germlinespecification. J. Cell Biol. 149, 875–888.

La Manna, G., Soldatov, R., Zessel, A., Braun, E., Hochgerber, H., Petukhov, V., Lidschreiber, K., Kasstritt, M. E., Löchnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature 560, 494–498.

Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T. R., Tomancak, P., and Krause, H. M. (2007). Global analysis of mRNA localization reveals a prominent role in regulating cellular architecture and function. Cell 131, 174–187.

Lin, X., and Zheng, J. (2019). A neuronal morphology classification approach based on locally cumulative connected deep neural networks. Appl. Sci. 9, 3876.

Little, S. C., Snsimer, K. S., Lee, J. J., Wieschaus, E. F., and Gavis, E. R. (2015). Independent and coordinate trafficking of single Drosophila germ plasm mRNAs. Nat. Cell Biol. 17, 558–568.

Lubeck, E., Coskun, A. F., Zhyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell in situ RNA profiling by sequential hybridization. Nat. Methods 11, 360–361.

Macdonald, P. M., and Struhl, G. (1998). cis-acting sequences responsible for anterior localization of bicoid mRNA in Drosophila embryos. Nature 336, 595–598.

Mardalke, F. K., Paul, A., Kümper, S., Sadok, A., Paterson, H., Mccarthy, A., Yuan, Y., and Marshall, C. J. (2015). Global analysis of mRNA translation, and protein localization: local translation is a key regulator of cell protrusions. Dev.Cell 35, 344–357.

Martin, K. C., and Ephrussi, A. (2009). mRNA localization: gene expression in the spatial dimension. Cell 136, 719.

Mas-Ponte, D., Carlevaro-Fita, J., Palumbo, E., Pulido, T. H., Guigo, R., and Johnson, R. (2017). LncATLAS database for subcellular localization of long noncoding RNAs. RNA 23, 1080–1087.

McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B. A., Karhohs, K. W., Doan, M., Ding, L., Rafeels, S. M., Thirstrup, D., et al. (2018). CellProfiler 3.0: Next-generation image processing for biology. PLoS Biol. 16, e2005970.

Meher, P. K., Sapathy, S., and Rao, A. R. (2020). miRNAloc: predicting miRNA subcellular localizations based on principal component scores of physico-chemical properties and pseudo compositions of di-nucleotides. Sci. Rep. 10, 14557.
Mili, S., Moisissoglu, K., and Macara, I.G. (2008). Genome-wide screen reveals APC-associated RNAs enriched in cell protrusions. Nature 453, 115.

Moffitt, J.R., and Zhuang, X. (2016). RNA imaging with multiplexed error-resistant fluorescence in situ hybridization (MERFISH). Methods Enzymol. 572, 1–49.

Moor, A.E., Golan, M., Massasa, E.E., Lemze, D., Weisman, T., Shenhar, R., Baydatch, S., Mizrahi, O., Winkler, R., Golan, O., et al. (2017). Global mRNA polarization regulates translation efficiency in the intestinal epithelium. Science 357, 1999–1303.

Moor, A.E., Harnik, Y., Ben-Moshe, S., Massasa, E.E., Rozenberg, M., Elam, R., Bahar Halpern, K., and Yakoetz, S. (2018). Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis. Cell 175, 1156–1167.e15.

Mueller, F., Senechal, A., Tantale, K., Marie-Nelly, H., Ly, N., Collin, O., Basyuk, E., Bertrand, E., Darzacq, X., and Zimmer, C. (2013). FISH-quant: automatic counting of transcripts in 3D FISH images. Nat. Methods. 10, 277–278.

Newberg, J.Y., and Murphy, R.F. (2008). A framework for the automated analysis of subcellular patterns in human protein atlas images. J. Proteome Res. 7, 2300–2308.

Padrón, A., Iwasaki, S., and Ingolia, N.T. (2019). Proximity RNA labeling by APEX-seq reveals the organization of translation initiation complexes and repressive RNA granules. Mol.Cell 75, 875–887.e5.

Player, A.N., Shen, L.P., Kenny, D., Antao, V.P., and Kolberg, J.A. (2001). Single-copy gene detection using branched DNA (bDNA) in situ hybridization. J.Histochem.Cytochem. 49, 603–612.

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nat. Methods 5, 877–879.

Ross, C.J., Rom, A., Spinrad, G., Gelbard, Salodkin, D., Degani, N., and Ulitsky, I. (2021). Uncovering deeply conserved motif combinations in rapidly evolving noncoding sequences. Genome Biol. 22, 29.

Samacoits, A., Choob, R., Safieddine, A., Traboulsi, A.-M., Ouyang, W., Zimmer, C., Peter, M., Bertrand, E., Walter, T., and Mueller, F. (2018). A computational framework to study subcellular RNA localization. Nat.Commun. 9, 4584.

Savulescu, A.F., Brackin, R., Bouilhol, E., Dartigues, B., Warrel, J.H., Pimentel, M.R., Beumé, N., Cortonato, I.C., Dallongeville, S., Bouille, M., et al. (2021). Interrogating RNA and protein spatial subcellular distribution in smFISH data with DypFISH. Cell Rep. Methods. 100068.

Savulescu, A.F., Jacobs, C., Negishi, Y., Davignon, L., and Mahlangu, M.M. (2020). Pinpointing cell identity in time and space. Front. Mol. Biosci. 7, 209.

Shahbaban, K., and Chartrand, P. (2012). Control of cytoplasmic mRNA localization. Cell. Mol. Life Sci. 69, 535–552.

Sharp, J.A., Plant, J.J., Ohsumi, T.K., Borovjask, M., and Blower, M.D. (2011). Functional analysis of the microtubule-interacting transcriptome. Mol. Biol. Cell. 22, 4312–4323.

Schubert, P.J., Dorkenwald, S., Januszewski, M., Jain, V., and Konfeld. J. (2019). Learning cellular morphology with neural networks. Nat.Commun. 10, 2736.

Shefi, O., Golding, I., Segev, R., Ben-Jacob, E., and Ayali, A. (2002). Morphological characterization of in vitro neuronal networks. Phys. Rev. E 66, 021905.

Singh, J., Hanson, J., Paliwal, K., and Zhou, Y. (2019). RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat.Commun. 10, 5407.

Su, Z.-D., Huang, Y., Zhang, Z.-Y., Zhao, Y.-W., Wang, D., Chen, W., Chou, K.-C., and Lin, H. (2018). Loci-IncRNA: predict the subcellular location of IncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics. 34, 4196–4204.

Sullivan, D.P., Winsnes, C.F., Åkesson, L., Hjelmare, M., Wiking, M., Schutten, R., Campbell, L., Leifsson, H., Rhodes, S., and Nordgren, A. (2018). Deep learning is combined with massive-throughput microscopy images. PLoS One 11, e0158712.

Yan, K., Arfat, Y., Li, D., Zhao, F., Chen, Z., Yin, C., Sun, Y., Hu, L., Yang, T., and Qian, A. (2016). Structure prediction: new insights into decoding long noncoding RNAs. Int. J. Mol. Sci. 17, 132.

Yan, Z., Lécuyer, E., and Blanchette, M. (2019). Prediction of mRNA subcellular localization using deep recurrent neural networks. Bioinformatics 35, 333–342.

Yang, Y., Fu, X., Qi, W., Xiao, Y., and Shen, H.B. (2018). MIRGOF: a GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA–disease association. Bioinformatics 34, 3547–3556.

Yasuda, Y., Clatterbuck-Soper, S.F., Jackrel, M.E., Shorter, J., and Mili, S. (2017). FUS inclusions disrupt RNA localization by sequestering kinesin-1 and inhibiting microtubule detyrosination. J.Cell Biol. 216, 1015–1034.

Zhang, B., Gunawardane, L., Niazi, F., Jahanbani, F., Chen, X., and Valadkhani, S. (2014). A novel RNA motif mediates the strict nuclear localization of a long noncoding RNA. Mol. Cell. Biol. 34, 2318–2329.

Zhang, T., Tan, P., Wang, L., Jin, N., Li, Y., Zhang, L., Yang, H., Hu, Z., Zhang, L., Hu, C., et al. (2016). Rnaloicate: a resource for RNA subcellular localizations. Nucleic Acids Res. 45, D135–D138.

Zappulo, A., van den Bruck, D., Ciollimattioni, C., Franke, V., Imami, K., McShane, E., Moreno-Estelles, M., Calviello, L., Filipyck, A., Puggero-Sanchez, E., et al. (2017). RNA localization is a key determinant of neurite-enriched proteome. Nat.Commun. 8, 583.