Prediction of RNA subcellular localization: learning from heterogeneous data sources

Anca Flavia Savulescu, Emmanuel Bouilhol, Nicolas Beaume, Macha Nikolski

Version of Record: https://www.sciencedirect.com/science/article/pii/S2589004221012670
Manuscript_9f6a28d0b9a74f2ff03018f4fa1784ab

Prediction of RNA subcellular localization: learning from heterogeneous data sources

Anca Flavia Savulescu[1,±], Emmanuel Bouilhol[2,3], Nicolas Beaume[4] and Macha Nikolski[2,3,±]

[1]Division of Chemical, Systems & Synthetic Biology, Institute for Infectious Disease & Molecular Medicine, Faculty of Health Sciences, 7925, University of Cape Town, Cape Town, South Africa.

[2]Université de Bordeaux, Bordeaux Bioinformatics Center, Bordeaux, France

[3]Université de Bordeaux, CNRS, IBGC, UMR 5095, Bordeaux, France

[4]Division of Medical Virology, Faculty of Health Sciences, 7925, University of Cape Town, Cape Town, South Africa

Lead contact: Anca Flavia Savulescu ankasi100@gmail.com

[±] To whom correspondence should be addressed email: ankasi100@gmail.com, macha.nikolski@u-bordeaux.fr

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 FISH-based techniques. Although such imaging data is ideal for characterization of RNA distribution, these techniques remain costly, time consuming and technically challenging. Given these limitations, imaging data exists 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 (Shahbabian 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, 1988, 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, Batish 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 noncoding RNAs (Bouvrette et al., 2017, La Manno et al., 2018, Lecuyer 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 towards 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, Katz *et al*., 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 transcripts for a more thorough classification of cell subtype or/and state.

In addition to contributions in basic research, knowledge of RNA subcellular spatial localization can be beneficial in biomedical applications. For instance, Didiot *et al* 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 in order 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, 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, we 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) (Moffit et al., 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 localisation 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 is 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 DAPI 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 e.g. (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). While 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 emphasising the importance of RNA spatial and temporal subcellular localization in biological processes calls for an increased access to such information. While 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 St Johnston, 1995, 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 (CNN, LSTM and attention layers) for the former and more classical SVM-based methodology for the latter. As in the protein world, these methods require training datasets 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 noncoding RNAs (ncRNA), such as long noncoding 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 compared to mRNAs (Ross et al., 2021, Constanty and Shkumatava, 2021). However, some associations of sequence motifs in lncRNAs with their subcellular localization

have been identified (Zhang et al., 2014) and databases containing information on cellular compartments to which the lncRNAs 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 la., 2018 have combined a PseKNC and SVM to predict subcellular location of lncRNAs to the ribosome and exosome, Gudenas et al., 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 GO-based functional similarity (Yang et al., 2018) and an SVM-based predictor (Meher et al., 2020), both relying on the miRNA 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, support vector machines, artificial neural networks, decision trees and deep convolutional neural networks (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 they were also able to construct a fully annotated dataset (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 above, the two main sources of data for characterization of subcellular localization of RNA are RNA sequence features and single molecule FISH images. In addition, as subcellular localization is considered as 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 Garg et al (2020) where AUC lies between 0.7 to 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 is 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 RBPs, secondary structures, etc'.

(2) Secondly, 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 noncoding RNAs, as well as additional cell type- and model-specific localization data.
(3) Thirdly, 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 below 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. Secondly, it is clear that our knowledge of RNA subcellular localization is not uniform for different RNA species - some classes of RNA, e.g. lncRNA and miRNA, are underrepresented 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 increasing the impact of distinguishing features and improving 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 Gudenas et al, 2018, who have built a model to classify lncRNAs 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. Since 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 can be computed to date (Shefi et al., 2005). 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, while 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). Normalisation 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 normalised

RNA quantities will consequently be skewed, which will inevitably impact predictions that would use this data. In conclusion, any machine learning localization prediction method is only as good as the data that it is built upon.

Finally, regarding features that may involve 2D or 3D features such as zipcodes, 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 zipcode 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, 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 above, 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 is 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 is 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 datasets 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.

Figure Legends

Figure 1: **Subcellular RNA localization and visualization. A.** Typical smFISH images. On the left, a raw image showing *Arhgdia* mRNA smFISH spots; on the right, a superimposed and denoised image containing the following stains: DAPI for DNA in blue, anti-tubulin antibody in green, *Arhgdia* mRNA smFISH spots in red. Organelles of interest, such as the Microtubule Organizing Center (MTOC) and cell contour are extrapolated from z-slices of the anti-tubulin stain. Scale bar 10 µm. **B**: A variety of subcellular spatial distribution features of the RNA can be observed and subsequently analyzed (RNA spots are depicted in red): specific enrichment of RNA in various subcellular locations, random versus clustered distribution of RNA and correlation of RNA with cellular markers, such as the MTOC or ER.

Figure 2: **An illustration on how deep learning could use smFISH data and RNA sequence features to build a model, followed by prediction of the subcellular localization of an mRNA using the model. A.** Training: RNAs for which both smFISH images and sequence features are available are collected to form a training set. For sequences, the relevant features are extracted (A, top), followed by training network 1 with images and network 2 with sequence features (A, bottom). The output of these networks is used as input to train network 3, which makes the

prediction (A, bottom). During the training process, errors are back propagated to improve all networks (A, bottom). **B**. Prediction: the localization of a new RNA, for which only sequence features are available, needs to be predicted. The sequence is processed by network 2, followed by network 3, which outputs the prediction.

References

Bashirullah, A., Cooperstock, R.L. & Lipshitz, H.D. RNA localization in development. *Annu. Rev. Biochem.* 1998 *67*:335-94.

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

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

Batish, M., van den Bogaarda, P., Kramer, F.R. & Tyagi, S. Neuronal mRNAs travel singly into dendrites. *Proc. Natl. Acad. Sci.* 2012 109, 4645–4650.

Bergsten, S.E., Gavis, E.R. Role for mRNA localization in translational activation but not spatial restriction of nanos RNA. *Development.* 1999 Feb;126(4):659-69.

Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S.M., Singer, R.H., Long, R.M. Localization of *ASH1* mRNA Particles in Living Yeast. *Molecular Cell* 1998 Oct;2(4):437-45.

Besse, F. & Ephrussi, A.Translational control of localized mRNAs: restricting protein synthesis in space and time. *Mol. Cell Biol.* 2008 *9*:971.

Bigler, R.L., Kamande, J.W., Dumitru, R., Niedringhaus, M., Taylor, A.M. Messenger RNAs localized to distal projections of human stem cell derived neurons. *Scientific Reports* 2017 7:611

Bouvrette, L.P., Cody, N.A.L., Bergalet, J., Lefebvre, F.A., Diot C., Wang X., Blanchette, M., Lécuyer, E. CeFra-seq reveals broad asymmetric mRNA and non-coding RNA distribution profiles in Drosophila and human cells. *RNA.* 2017 Oct 27 rna.063172.117.

Brangwynne, C.P., Eckmann, C.R., Courson, D.S., Rybarska, A., Hoege, C., Gharakhani, J., Jülicher, F., Hyman, A.A. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science.* 2009 Jun 26;324(5935):1729-32.

Buxbaum, A., Wu, B. & Singer, R.H. Single β-Actin mRNA Detection in Neurons Reveals a Mechanism for Regulating Its Translatability. *Science.* 2014 343, 419.

Cabili, M. N., Dunagin, M. C., McClanahan, P. D., Biaesch, A., Padovan-Merhar, O., Regev, A. & Raj, A. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome biology. *2015* 1*6*(1), 20.

Cajigas, I.J., Tushev, G., Will, T.J., tom Dieck, Fuerst, N., Schuman, E.M. SThe local transcriptome in the synaptic neuropil revealed by deep sequencing and high-resolution imaging. *Neuron.* 2012 May 10;74(3):453-66.

Cao, Z., Pan, X., Yang, Y., Huang, Y., Shen, H.B. The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. *Bioinformatics.* 2018 34(13):2185-2194.

Constanty, F., Shkumatava, A. lncRNAs in development and differentiation: from sequence motifs to functional characterization. *Development.* 2021 Jan 13;148(1):dev182741.

de Chaumont, F., Dallongeville, S., Chenouard, N., Hervé, N., Pop, S., Provoost, T., Vannary, M-Y, Pankajakshan, P, Lecomte, T., Le Montagner, Y., Lagache, T., Dufour, A., et al.Olivo-Marin, J-C., Icy: an open bioimage informatics platform for extended reproducible research. *Nat. Methods*. 2012 9, 690–696.

Didiot MC, Ferguson CM, Ly S, et al. Nuclear Localization of Huntingtin mRNA Is Specific to Cells of Neuronal Origin. *Cell Rep*. 2018;24(10):2553-2560.e5.

Emanuelsson, O., Brunak., S., von Heijne, G., Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protoc.* 2007 2, 953–971.

Eng, C.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.C., Cai, L. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature.* 2019 Apr;568(7751):235-239

Foster, L.J., de Hoog, C.L., Zhang, Y., Zhang, Y., Xie, X., Mootha, V.K., Mann, M. A mammalian organelle map by protein correlation profiling. *Cell.* 2006 Apr 7;125(1):187-99.

Garg, A., Singhal, N., Kumar, R., & Kumar, M. mRNALoc: a novel machine-learning based in-silico tool to predict mRNA subcellular localization. *Nucleic Acids Research.* 2020 Volume 48, Issue W1, pp W239–W243

Gudenas, B.L., and Wang, L. Prediction of LncRNA subcellular localization with deep learning from sequence features. *Scientific reports.* 2018 8.1 1-10.

Hocine, S., Raymond, P., Zenklusen, D., Chao, J.A., Singer, R.H. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nat Methods*. 2013 Feb;10(2):119-21.

Hughes, A.C., Simmonds, A.J. *Drosophila* mRNA Localization During Later Development: Past, Present, and Future. *Front Genet.* 2019 10: 135.

Imai, K., Nakai, K. Prediction of subcellular locations of proteins: where to proceed? *Proteomics*. 2010 10, 3970–3983.

Jambor, H., Surendranath, V., Kalinka, A.T., Mejstrik, P., Saalfeld, S., Tomancak, P. Systematic imaging reveals features and changing localization of mRNAs in *Drosophila* development. *eLife.* 2015 4: e05003.

Jansen, R.P. mRNA localization: message on the move. *Nat. Rev. Mol. Cell Biol.* 2001 *2*:247-56.

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

Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021 596, 583–589.

Katz, Z.B., Wells, A.L., Park, H.Y., Wu, B., Shenoy, S.M., Singer, R.H. β-Actin mRNA compartmentalization enhances focal adhesion stability and directs cell migration. *Genes Dev.* 2012 Sep 1;26(17):1885-90.

Katz, Z.B., English, B.P., Lionnet, T., Yoon, Y.J., Monnier, N., Ovryn, B., Bathe, M., Singer, R.H. Mapping translation 'hot-spots' in live cells by tracking single molecules of mRNA and ribosomes. *Elife*. 2016 Jan 13;5. pii: e10415.

Khong, A., Matheny, T., Jain, S., Mitchell, S.F., Wheeler, J.R., Parker, R. The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress Granules. *Mol Cell.* 2017 Nov 16;68(4):808-820.e5

Kislauskis, E.H., Zhu, X., and Singer, R.H. Sequences responsible for intracellular localization of beta-actin messenger RNA also affect cell phenotype. *J. Cell Biol.* 1994 27, 441–451.

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

Knaut, H., Pelegri, F., Bohmann, K., Schwartz, H., Nusslein-Volhards, C. Zebrafish *vasa* RNA but Not Its Protein Is a Component of the Germ Plasm and Segregates Asymmetrically before Germline Specification. *J Cell Biol.* 2000 May 15; 149(4): 875–888.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L.E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., Kharchenko, P.V. RNA velocity of single cells. *Nature.* 2018 Aug;560(7719):494-498

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

Li, R., Zhu, M., Li, J., Bienkowski, M.S., Foster, N.N., Xu, H., Ard, T., Bowman, I., Zhou, C., Veldman, M.B., Yang, X.W., Hintiryan, H., Zhang, J., Dong, H-W. Precise segmentation of densely interweaving neuron clusters using G-Cut. *Nature Communications*. 2019 10, Article number: 1549

Lin, X., Zheng, J. A Neuronal Morphology Classification Approach Based on Locally Cumulative Connected Deep Neural Networks. *Appl. Sci.* 2019, 9(18), 3876

Little, S.C., Sinsimer, K.S., Lee, J.J., Wieschaus, E.F., Gavis, E.R. Independent and coordinate trafficking of single *Drosophila* germ plasm mRNAs. *Nat Cell Biol.* 2015 May; 17(5): 558–568.

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

Macdonald, P.M. & Struhl, G. cis-acting sequences responsible for anterior localization of bicoid mRNA in Drosophila embryos. *Nature.* 1998 *336*:595-598.

Macdonald, P.M., Kerr, K., Smith, J.L., and Leask, A. RNA regulatory element BLE1 directs the early steps of bicoid mRNA localization. *Development.* 1993 118, 1233–1243.

Mardakheh, F.K., Paul, A., Kümper, S., Sadok, A., Paterson, H., Mccarthy, A., Yuan, Y., Marshall, C.J. Global Analysis of mRNA, Translation, and Protein Localization: Local Translation Is a Key Regulator of Cell Protrusions. *Dev Cell.* 2015 35(3): 344–357.

Martin, K.C. & Ephrussi, A. mRNA Localization: Gene Expression in the Spatial Dimension. *Cell.* 2009 *136*:719.

Mas-Ponte, D., Carlevaro-Fita, J., Palumbo, E., Pulido, T.H., Guigo, R., and Johnson, R. LncATLAS database for subcellular localization of long noncoding RNAs. RNA. 2017 vol. 23, no. 7, pp. 1080–1087.

Meher, P.K., Satpathy, S. & Rao, A.R. miRNALoc: predicting miRNA subcellular localizations based on principal component scores of physico-chemical properties and pseudo compositions of di-nucleotides. *Sci Rep.* 2020 10, 14557

Mili, S., Moissoglu, K. & Macara, I.G. Genome-wide screen reveals APC-associated RNAs enriched in cell protrusions. *Nature*. 2008 *453*:115.

Moffitt, J.R., and Zhuang, X. RNA Imaging with Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH). *Methods Enzymol.* 2016 572:1-49.

Moor, A.E., Golan M., Massasa E.E., Lemze D., Weizman T., Shenhav R., Baydatch S., Mizrahi O., Winkler R., Golani O., Stern-Ginossar N., Itzkovitz S. Global mRNA polarization regulates translation efficiency in the intestinal epithelium. *Science.* 2017 Sep 22;357(6357):1299-1303.

Moor, A.E., Harnik, Y., Ben-Moshe, S., Massasa, E.E., Rozenberg, M., Eilam, R., Bahar Halpern, K., Itzkovitz, S. Spatial Reconstruction of Single Enterocytes Uncovers Broad Zonation Along the

Intestinal Villus Axis. *Cell.* 2018 Nov 1;175(4):1156-1167.e15.

Mowry, K.L., and Melton, D.A. Vegetal messenger RNA localization directed by a 340-nt RNA sequence element in Xenopus oocytes. *Science.* 1992 255, 991–994.

Munro, T.P., Magee, R.J., Kidd, G.J., Carson, J.H., Barbarese, E., Smith, L.M., and Smith, R. Mutational analysis of a heterogeneous nuclear ribonucleoprotein A2 response element for RNA trafficking. *J. Biol. Chem*. 1999 274, 34389–34395.

Newberg, J.Y., Murphy, R.F. A framework for the automated analysis of subcellular patterns in human protein atlas images. *J Proteome Res.* 2008 7(6):2300–8.

Padrón, A., Iwasaki, S., Ingolia, N.T. Proximity RNA Labeling by APEX-Seq Reveals the Organization of Translation Initiation Complexes and Repressive RNA Granules. *Mol Cell.* 2019 Aug 22;75(4):875-887.e5.

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

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods*. 2008 Oct;5(10):877-9.

Ross, C.J., Rom, A., Spinrad, A., Gelbard-Solodkin D., Degani, N., Ulitsky, I. Uncovering deeply conserved motif combinations in rapidly evolving noncoding sequences. *Genome Biology.* 2021 22,29

Samacoits, A., Chouaib, R., Safieddine, A., Traboulsi, A-M., Ouyang, W., Zimmer, C., Peter, M., Bertrand, E., Walter, T., Mueller, F. A computational framework to study sub-cellular RNA localization. *Nat Commun.* 2018 9, 4584.

Savulescu, A.F, Brackin, R., Bouilhol, E., Dartigues, B., Warrell, J.H., Pimentel, M.R., Beaume, N., Cortunato, I.C., Dallongeville, S., Boulle, M., Soueidan, H., Agou, F., Schmoranzer, J., Olivo-Marin, J-C., Franco, C.L., Gomes, E.R., Nikolski, M., Mhlanga, M.M. Interrogating RNA and protein spatial subcellular distribution in smFISH data with DypFISH *Cell Reports Methods*. 2021 13 September, 100068

Savulescu, A.F., Jacobs, C., Negishi, Y., Davignon, L., Mhlanga, M.M. Pinpointing cell identity in time and space. *Front. Mol. Biosci*. 2020 Aug 14;7:209.

Shahbabian, K and Chartrand, P. Control of cytoplasmic mRNA localization. *Cell. Mol. Life Sci.* 2012 69:535–552

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

Shema, E., Bernstein, B.E., Buenrostro, J.D. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nat Genet.* 2019 Jan;51(1):19-25.

Schubert, P.J., Dorkenwald, S., Januszewski, M., Jain, V., Kornfeld, J. Learning cellular morphology with neural networks. *Nature Communications*. 2019 10, Article number: 2736

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

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

Sweeney, B.A., Hoksza, D.,Nawrocki, E.P. et al. R2DT is a framework for predicting and visualising RNA secondary structure using templates. *Nat Commun*. 2021 12, 3494

Su, Z.-D., Huang, Y., Zhang, Z-Y., Zhao, Y.-W., Wang, D., Chen, W., Chou, K.-C., and Lin, H. iLoc-lncRNA: Predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC, *Bioinformatics*. 2018 vol. 34, no. 24, pp. 4196–4204

Sullivan, D.P., Winsnes, C.F., Åkesson, L., Hjelmare, M., Wiking, M., Schutten, R., Campbell, L., Leifsson, H., Rhodes, S., Nordgren, A. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat Biotechnol.* 2018 36(9):820.

Suter, B. RNA localization and transport. *Biochim Biophys Acta Gene Regul Mech.* 2018 Oct;1861(10):938-951.

Tautz, D. & Pfeifle, C. A non-radioactive in situ hybridization method for the localization of specific RNAs in *Drosophilia* embryos reveals translational control of the segmentation gene hunchback. *Chromosoma*. 1989 *98*:81-85.

Trcek, T., Grosch, M., York, A., Shroff, H., Lionnet, T., Lehmann, R. Drosophila germ granules are structured and contain homotypic mRNA clusters. *Nat. Commun.* 2015 6, 7962.

Tzingounis, A.V. & Nicoll, R.A. Arc/Arg3.1: linking gene expression to synaptic plasticity and memory. *Neuron.* 2006 *52*:403.

Wan, S., Mak, M.-W. Machine Learning for Protein Subcellular Localization Prediction. *De Gruyter*. 2015 ISBN 978-1-5015-0150-0

Wang, F. *et al*. RNAscope A Novel *in Situ* RNA Analysis Platform for Formalin-Fixed, Paraffin-Embedded Tissues. *J Mol Diagn. 2012* 14, 22–29

Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., Nolan, G.P., Bava, F.A., Deisseroth, K. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science.* 2018 Jul 27;361(6400).

Weis, B.L., Schleiff, E. & Zerges, W. Protein targeting to subcellular organelles via mRNA localization. *Biochim. Biophys. Acta*. 2013 1833:260–273.

Wilbertz, J.H., Voigt, F., Horvathova, I., Roth, G., Zhan, Y., Chao, J.A. Single-Molecule Imaging of mRNA Localization and Regulation during the Integrated Stress Response. *Mol Cell.* 2019 Mar 7;73(5):946-958.e7.

Xijie Lu, A., Moses, A.M. An Unsupervised kNN Method to Systematically Detect Changes in Protein Localization in High-Throughput Microscopy Images. *PLoS One*. 2016 11(7): e0158712.

Yan, K., Arfat, Y., Li., D., Zhao, F., Chen, Z., Yin, C., Sun, Y., Hu, L., Yang, T., Qian, A. Structure prediction: New insights into decrypting long noncoding RNAs. *Int. J. Mol. Sci*. 2016 17 Jan 21;17(1):132.

Yan, Z., Lécuyer, E., Blanchette, M. Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics*. 2019 35(14):i333-i342

Yang, Y., Fu, X., Qu, W., Xiao, Y. & Shen, H. B. MiRGOFS: a GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA–disease association. *Bioinformatics.* 2018 34(20), 3547–3556

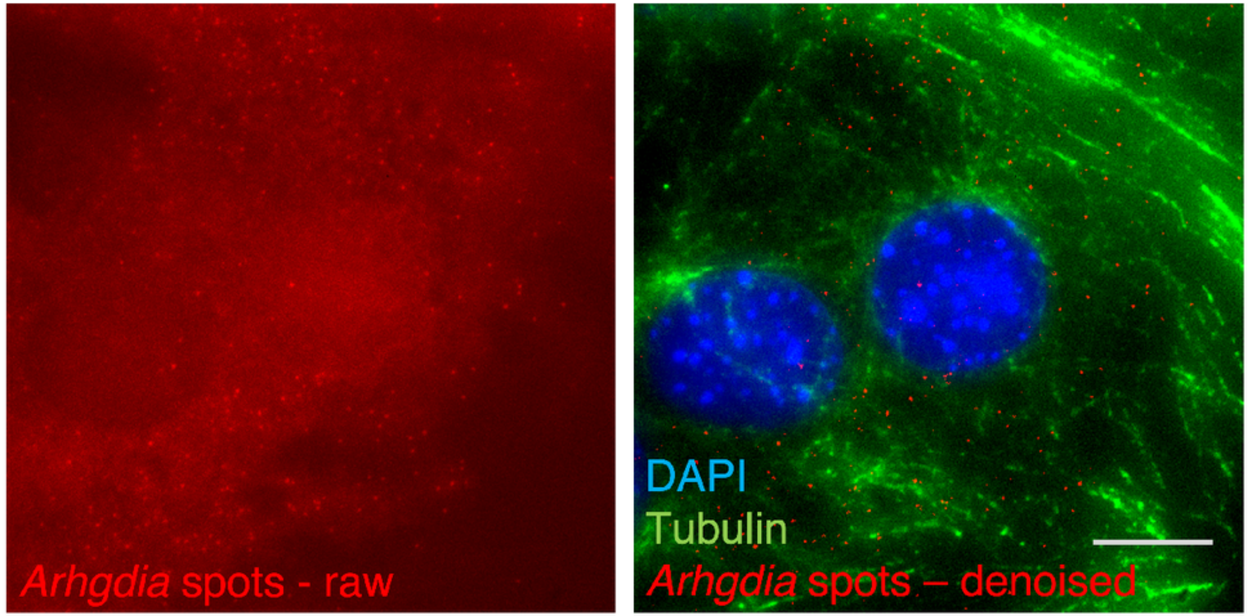
Yasuda, Y., Clatterbuck-Soper, S.F., Jackrel, M.E., Shorter, J., Mili, S. FUS inclusions disrupt RNA localization by sequestering kinesin-1 and inhibiting microtubule detyrosination. *J Cell Biol.* 2017 Apr 3; 216(4): 1015–1034.

Zhang, B., Gunawardane, L., Niazi, F., Jahanbani, F., Chen, X., Valadkhan, S. A Novel RNA Motif Mediates the Strict Nuclear Localization of a Long Noncoding RNA. *Mol Cell Biol*. 2014 Jun; 34(12): 2318–2329.

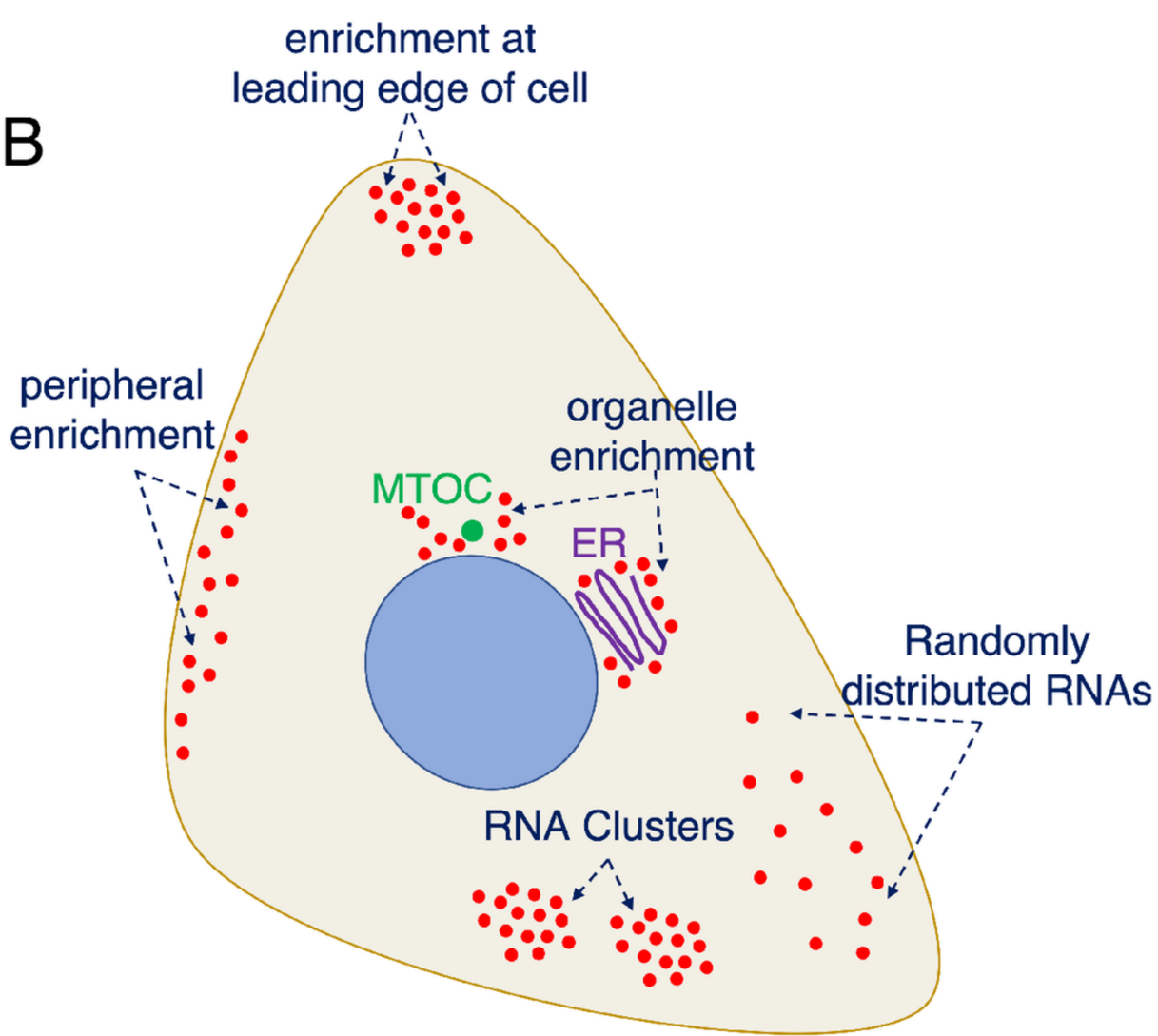
Zhang, T., Tan, P., Wang, L., Jin, N., Li, Y., Zhang, L., Yang, H., Hu, Z., Zhang, L., Hu, C., Li, C., Qian, K., Zhang, C., Huang, Y., Li, K., Lin, H., and Wang, D. Rnalocate: A resource for RNA subcellular localizations *Nucleic Acids Res*. 2016 vol. 45, pp. D135–D138,

Zappulo A., van den Bruck D., Ciolli Mattioli C., Franke V., Imami K., McShane E., Moreno-Estelles M., Calviello L., Filipchyk A., Peguero-Sanchez E., Müller T., Woehler A., Birchmeier C., Merino E., Rajewsky N., Ohler U., Mazzoni E.O., Selbach M., Akalin A., Chekulaeva, M. RNA localization is a key determinant of neurite-enriched proteome. *Nature Communications* 2017 Sep 19;8(1):583.

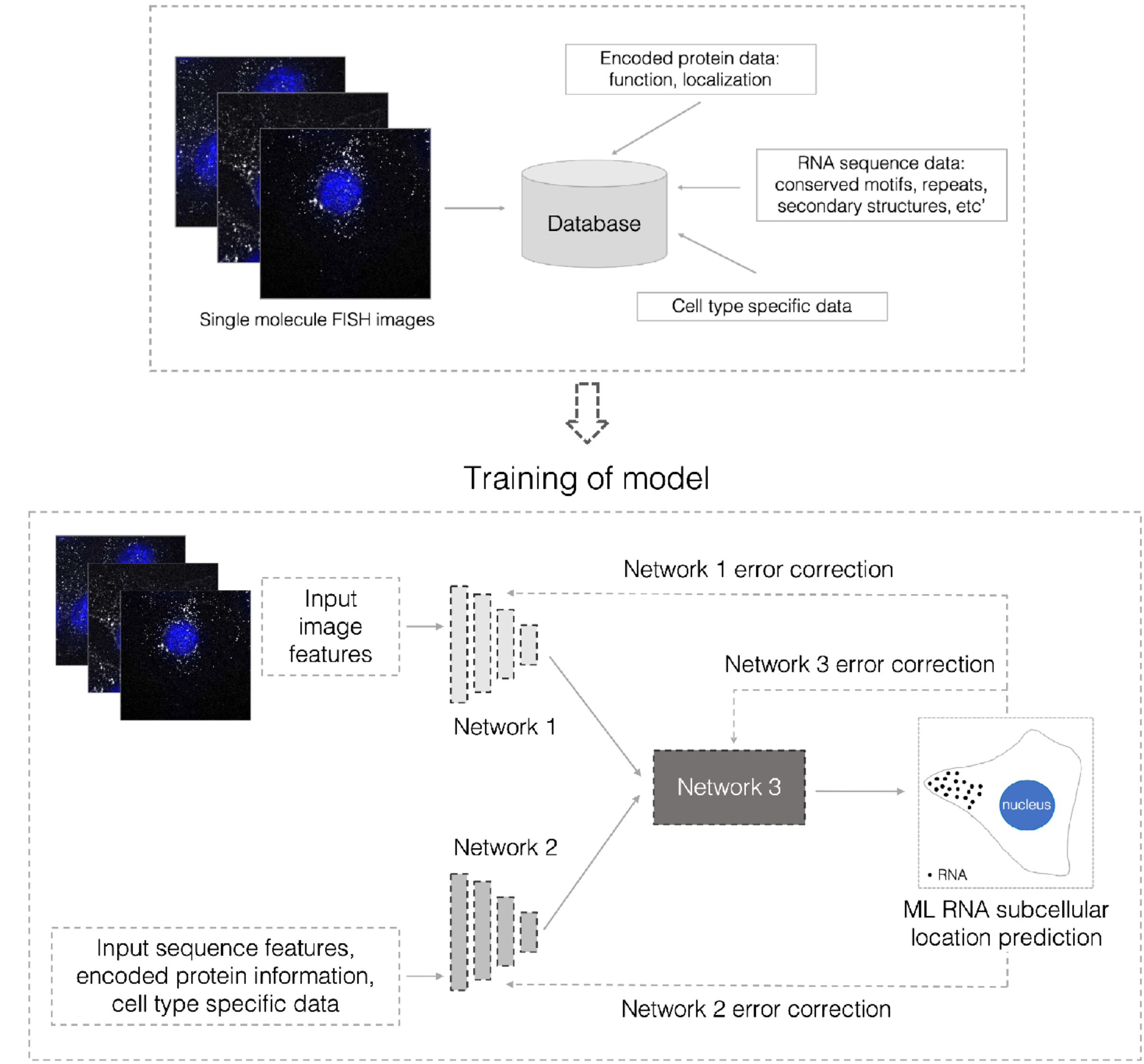
A


Arhgdia spots - raw
DAPI
Tubulin
Arhgdia spots – denoised

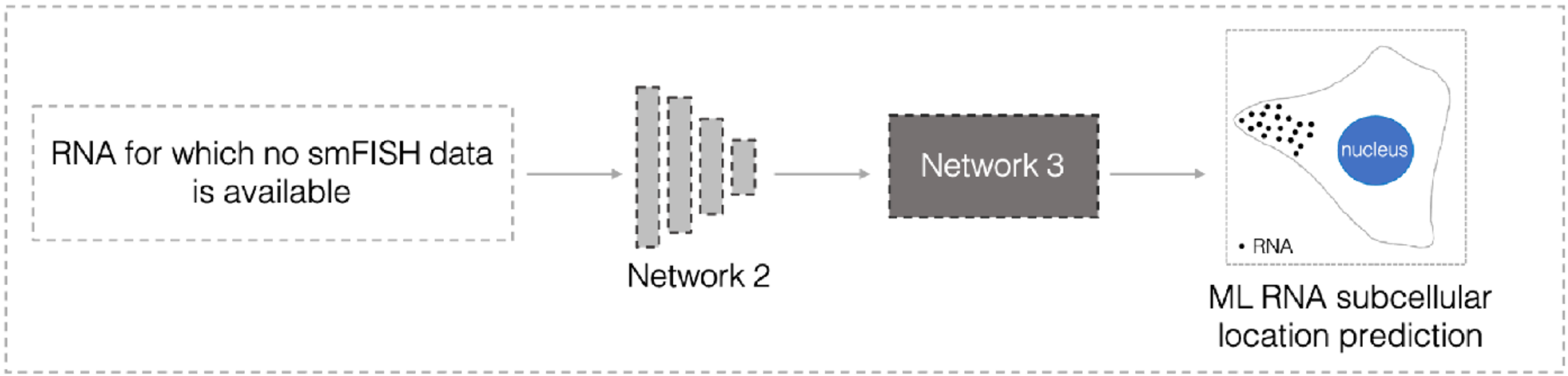
B

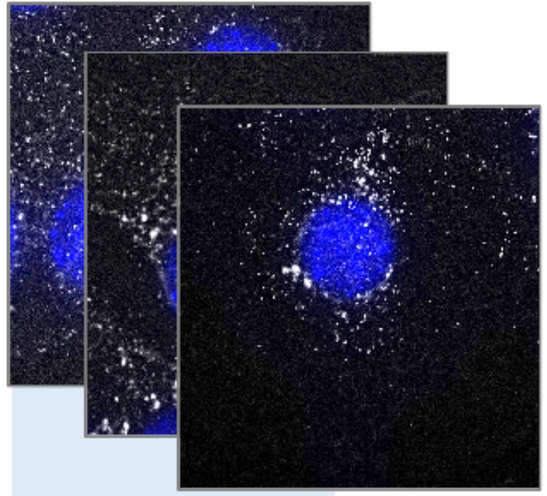

enrichment at
leading edge of cell
peripheral
enrichment
organelle
enrichment
MTOC
ER
Randomly
distributed RNAs
RNA Clusters

A


Creation of database
Encoded protein data:
function, localization
RNA sequence data:
conserved motifs, repeats,
secondary structures, etc'
Database
Cell type specific data
Single molecule FISH images
Training of model
Input
image
features
Network 1 error correction
Network 1
Network 3 error correction
Network 3
nucleus
• RNA
ML RNA subcellular
location prediction
Network 2
Input sequence features,
encoded protein information,
cell type specific data
Network 2 error correction

B


Prediction of new subcellular location based on trained model
RNA for which no smFISH data
is available
Network 2
Network 3
nucleus
• RNA
ML RNA subcellular
location prediction



Imaging data

Training phase

Prediction phase

Input sequence features, encoded protein information, cell type specific data, study model

Neural network

Neural network

Latent space

RNA location prediction network

Predicted subcellular location of RNA