High-content imaging reveals microexons' ability to shape protein localisation

Peter Kjer-Hansen[1,2,3,8], Karina Pazaky[1,4], Daisy Kavanagh[1,3], Helen E. King[1,2], Savannah O'Connell[4], Gabriela Santos Rodriguez[3], Abigail K. Grootveld[4], Helaine Graziele Santos Vieira[1], Scott Berry[5,6,7], Javier Fernandez-Chamorro[1], Robert J. Weatheritt[1,3,8]

Affiliations & correspondence

1. EMBL Australia, Garvan Institute of Medical Research, Darlinghurst, New South Wales, Australia
2. St. Vincent's Healthcare Clinical Campus, School of Clinical Medicine, Faculty of Medicine and Health, UNSW Sydney, Darlinghurst, New South Wales, Australia
3. School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, Australia
4. Precision Immunology Program, Garvan Institute of Medical Research, Sydney, New South Wales, Australia
5. EMBL Australia Node in Single Molecule Science, University of New South Wales, Sydney, NSW, Australia
6. UNSW RNA Institute, University of New South Wales, Sydney, NSW, Australia
7. Department of Molecular Medicine, School of Biomedical Sciences, University of New South Wales, Sydney, NSW, Australia
8. Corresponding authors: r.weatheritt@garvan.org.au; p.hansen@garvan.org.au

Abstract

Alternative splicing massively expands the human proteome by generating multiple protein isoforms per gene. However, the functional properties of most protein isoforms are unknown, and it remains unclear whether most have distinct or redundant functions. The development of methods to address this question and characterise proteins at scale are needed. We hypothesized that high-content imaging combined with classifier-based analysis could provide a powerful and broadly applicable approach for characterizing subcellular localization and behavioural differences between protein isoforms. As a proof-of-concept, we examine 42 isoform pairs in HeLa cells, each pair consisting of two isoforms from the same gene that differ in the inclusion/exclusion of a tissue-regulated cassette exon, many differing by 10 or fewer amino acids. We identify localisation differences for ~38% (16/42) of isoform pairs, and features describing these differences. We perform follow-up examination of two sorting nexin 2 (SNX2) isoforms that differ by four amino acids inserted into an α-helix, which drastically altered subcellular localisation. We find the localisation differences are

caused by the presence of additional residues rather than their composition, which reduces vesicle association while enhancing interactions with the insulin receptor. Lastly, we generalise these findings by showing that microexons frequently overlap α-helices, are enriched in helix-stabilizing residues, have lengths that preserve helical register, and often modulate surface charge. Together, our findings support a broad role for alternative splicing in modulating protein subcellular localization and increasing functional proteomic diversity.

Introduction

Almost all human protein-coding genes are multi-exonic and their pre-mRNA must be spliced to form mature mRNAs that can be translated[1]. The RNA splicing process is flexible, however, and variations in which parts of the pre-mRNA that are included and excluded from the mature mRNA, called alternative splicing (AS), occurs for nearly all spliced mRNAs[2,3]. Since AS can change the coding region of mRNAs, it can change the encoded proteins, although the contribution of AS to proteome diversity has been debated[4,5]. Recent efforts, however, have shown most well-expressed alternative transcripts give rise to protein isoforms detectable by mass spectrometry[6,7], indicating the existence of vast AS-enabled protein diversity. While case studies have shown AS-derived proteins can impact practically every cellular process and are important for many biological processes[8–10], the function of most isoforms remain unexplored. This indicates there are substantial amounts of the proteome yet to be explored, and increasing the understanding of the AS-derived protein isoforms (hereafter referred to as protein isoforms) holds both fundamental and applied value, including to therapeutics[11].

Considering the regulatory context and conservation of AS events can help prioritize protein isoforms for examination. This is because conserved and regulated AS events, defined as AS occurring in a controlled and genetically encoded manner[12] e.g. via splicing factors acting on specific motifs, are considered more likely to be functional[13]. Using this framework, a particularly interesting group of AS events are neural microexons, which form the most highly conserved group of AS events across vertebrates and display strong tissue-specific inclusion in the nervous system[14], primarily through the activity of the splicing factor SRRM4[15]. These exons are small (3-27 nucleotides) and only minutely change the primary protein sequence but are enriched to occur within structured domains[14] and can have large effects on protein functions[16].

Protein isoforms arising through neural microexon splicing are among the more well-studied protein isoforms, but even for these protein isoforms, most remain poorly characterised[16]. Therefore, even when conservation and regulation are used to prioritise protein isoforms, the number of isoforms remain large, and high-throughput

methods are central in addressing whether most protein isoforms are functionally distinct or redundant.

Traditionally, high-throughput examination of protein isoforms has focused on protein-protein interactions (PPIs)[17–19], especially binary PPIs[17,18], which have shown AS frequently rewires protein interactomes, with most isoforms having distinct interactomes[17]. Notably, these analyses corroborated bioinformatical analyses of tissue-regulated AS's impact on proteins, which demonstrated AS was enriched in intrinsically disordered regions[20] containing short linear interaction motifs[18,21,22], which can mediate PPIs[23].

Recently, high-throughput protein isoform characterisation has expanded outside PPIs, and the impact of protein isoforms has been examined on cell fitness[24], transcription factor activity[25], transcriptional programmes during stem cell differentiation[26], and - at more modest throughput - zebrafish brain development[27,28]. These studies have expanded our understanding of isoform function, but their modest number still means only a limited number of isoforms have been examined. Additionally, only a limited number of readouts have been used. In this respect, microscopy-based examination of protein isoforms is largely absent. Indeed, we are not aware isoform imaging being done at scale until very recently, where it was used to assess transcription factor isoform localisation (nuclear or cytoplasmic), and condensate formation[25]. We hypothesize imaging is a powerful tool for protein isoform characterisation, especially high-content imaging[29], which offers a scalable and generic approach to measure numerous features that can capture different aspects of isoform-specific subcellular localisation, including - but not limited to - nuclear/cytoplasm localisation, for different types of proteins. Additionally, because it is generic, it can be part of multimodal descriptions of protein isoforms where different types of data contextualise each other. Note, we use a broad definition of 'subcellular localisation' to describe any change in the intracellular distribution of a protein, such as transitions from more diffuse to more granulated distribution in the cytosol. Intriguingly, high-content imaging was recently used in this way to examine the effects of missense variants at scale[30], suggesting it could work for protein isoform profiling.

To evaluate the use of high-content imaging for protein isoform characterisation and contribute to isoform characterisation at scale, we examined a modest-sized library of 42 protein isoform pairs where each pair consisted of two isoforms from the same gene. We focused on isoforms that differed in tissue-regulated cassette exons where the exons typically were included in neural tissues and excluded in non-neural tissues, with many exons being neural microexons. Many of these isoform pairs have been examined for PPI differences[18,31] but not subcellular localisation differences, which facilitated library generation and enabled comparison of localisation and PPI differences. We imaged the isoform pairs in HeLa cells and used machine-learning

classifiers to identify proteins with isoform-specific subcellular localisation. We found ~38% (16/42) of the isoform pairs contained differentially localising isoforms. This may suggest AS frequently alters the subcellular localisation of proteins, at least when acting on neural microexons. Notably, more pronounced localisation differences were associated with larger PPI differences. To showcase how imaging-based examination can inform more detailed isoform characterisation, we performed additional experiments on the previously uncharacterised sorting nexin 2 (SNX2) isoforms. Imaging, structural analysis, mutagenesis, and conservation-based examination suggested AS modulates SNX2's localisation and PPIs largely independent of the amino acids encoded by the alternative exon itself and potentially exemplifies a common way microexons affect helical structures.

Results

A high-content imaging approach for isoform characterisation

To generate a protein isoform library suitable for imaging analysis, we sub-cloned open reading frames (ORFs) encoding protein isoforms into plasmids where they were fused with mCherry fluorescent protein. For each gene, we cloned at least two isoforms that differed in the presence of a cassette exon, typically included in neural tissues. Many of the cassette exons were neural microexons, and the examined genes were chosen to represent gene ontologies that are enriched in genes containing neural microexons, namely vesicle-associated transport and cytoskeletal processes[14]. The library (**Supp. Table 1**) consisted of 68 isoforms across 32 genes with 91% (29/32) genes having 2 isoforms, and 9% (3/32) genes having >2 isoforms. This enabled 44 possible pairwise isoform comparisons of isoforms from the same gene. To our knowledge, only ~16% (7/44) of these isoform pairs have previously been examined for localisation differences using microscopy[32–38]. Conceptually, we aimed to examine if two isoforms from the same gene, visualised through fusion to mCherry, differed in subcellular localisation or behaviours that were quantifiable by measuring different features of the signal from the fluorescent protein (**Figure 1A)**. Potential differences could then be identified via classifier-based approaches. Comparison of cells expressing identical mCherry constructs and cells expressing mCherry constructs with clear localisation differences would serve as negative and positive controls, respectively.

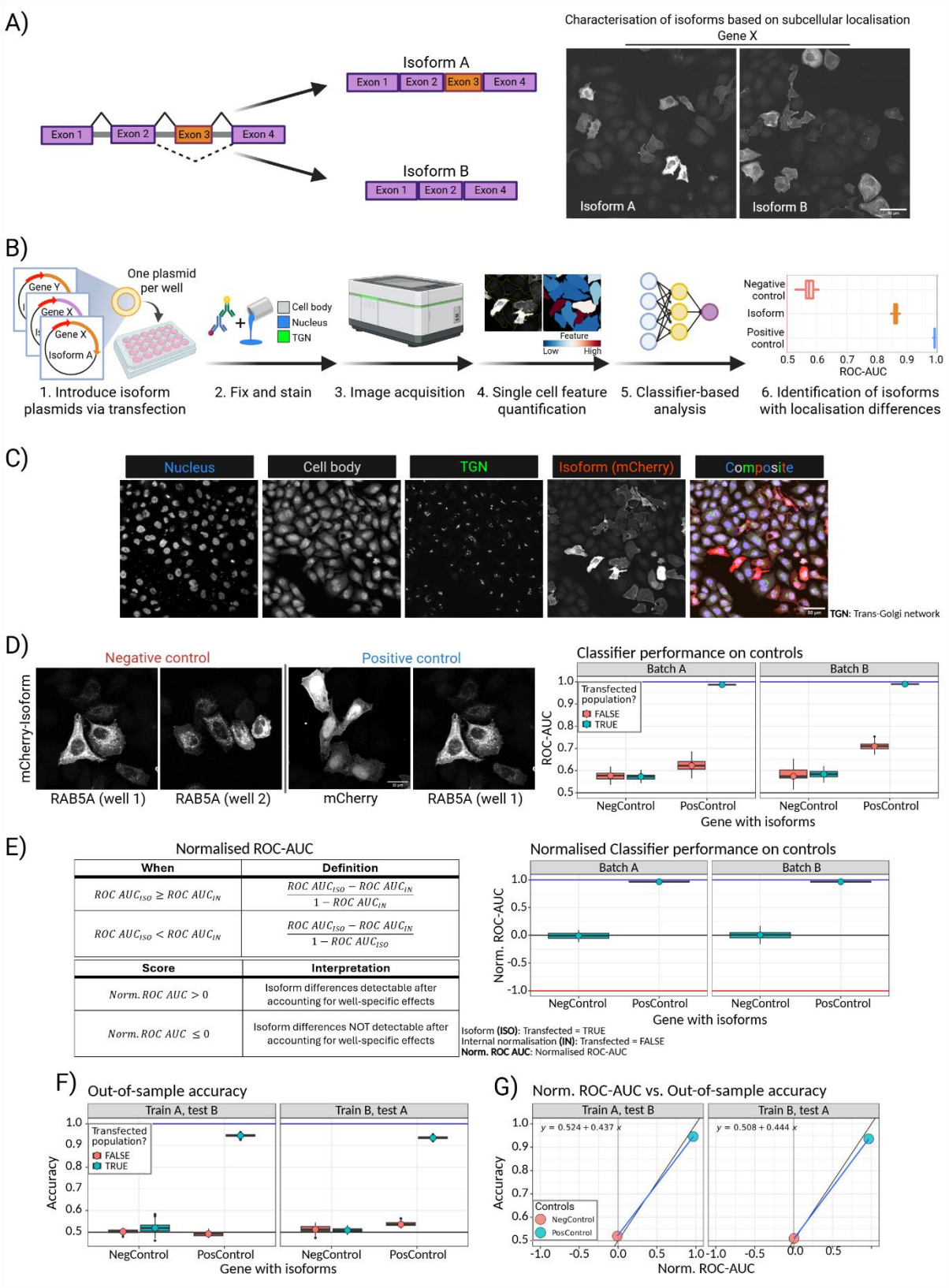


***Figure 1. Validation of high-content imaging and classifier-based approach as capable of identifying localisation differences. A) Left**, schematic illustrating inclusion or exclusion of cassette exon in spliced mRNA leads to different transcript isoforms that can encode different protein isoforms. **Right**, example images of fluorescent protein isoforms expressed in HeLa cells. **B)** Workflow of high-content imaging and classifier-based examination of protein isoforms (see Methods for details). TGN, Trans-Golgi network. **C)** Example of the four channels acquired for a field-of-view. The focus slice is shown after illumination correction. Scale bar is 50µm. **D) Left**, Example images of cells expressing the mCherry constructs used as positive and negative controls. The*

negative control consists of two separate wells transfected with the same mCherry construct (mCherry-RAB5A). The positive control consists of two separate wells transfected with mCherry and mCherry-RAB5A, respectively. Scale bar is 20µm. ***Right****, boxplot showing the ability of classifiers trained on mCherry features measured in cells expressing control plasmids to correctly assign the plasmid ID (which corresponds to well ID), quantified as ROC-AUC (Receiver Operating Characteristics (ROC), Area Under the Curve (AUC)). Cells used for the examination came from three experimental repeats that were pooled for the classifier-based analysis with n cells ≥ 500 for each comparison after pooling. The experiment with three repeats was performed twice, referred to as batch A and batch B, that were evaluated separately. The boxplots illustrate the ROC-AUCs from each of 50 iterations of cell sampling and classifier evaluation, and dot indicates median ROC-AUC (n = 50). Classifier-based analysis was performed separately on transfected (ISO) and non-transfected (internal normalisation) cells from the same wells to assess contribution of well-specific effects on classification. Coloured lines indicate perfect (blue) and worst performance (black), respectively.* ***E) Left****, definition and interpretation of the normalised ROC-AUC score, where the internal normalisation (IN) classifier (using non-transfected cells, Transfected = FALSE) is used to normalise the isoform classifier (consisting of transfected cells, Transfected = TRUE).* ***Right****, boxplot showing classifier performance measured as normalised ROC-AUC (Norm. ROC-AUC), for the controls, using ROC-AUCs plotted in Figure 1D, Right as input. Coloured lines indicate perfect (blue), neutral (black) or worst (red) performance, respectively.* ***E) Left****, definition and interpretation of the normalised ROC-AUC score, where the internal normalisation (IN) classifier (using non-transfected cells, Transfected = FALSE) is used to normalise the isoform classifier (consisting of transfected cells, Transfected = TRUE).* ***Right****, boxplot showing classifier performance measured as normalised ROC-AUC (Norm. ROC-AUC), for the controls, using ROC-AUCs plotted in Figure 1D, Right as input. Coloured lines indicate perfect (blue), neutral (black) or worst (red) performance, respectively.* ***F)*** *Boxplot showing performance of classifiers trained on mCherry features of cells transfected with control plasmids in batch A to predict plasmid identity in batch B and vice versa, as quantified by accuracy on the test batch (out-of-sample accuracy). The test batch (A or B) was down-sampled to have equal amounts of cells expressing either plasmid in a control pair before being used for prediction. The accuracy can therefore range from 0.5 to 1.0. 20 iterations of classifier training and testing was done, and the boxplots shows the mean accuracy from each iteration, and dot indicates median accuracy across iterations (n = 20). n cells ≥500 for each condition in both test and training sets.* ***G)*** *XY plot showing the correlation between the normalised ROC-AUC and the out-of-sample accuracy for the two controls. The normalised ROC-AUC for batch A is compared to the out-of-sample accuracy when trained on batch A and tested on B (vice versa for ROC-AUC for batch B). A linear regression was fitted to the data.*

To examine our library of AS-derived protein isoforms using high-content imaging (**Figure 1B**), protein isoform-encoding plasmids were transiently transfected into HeLa cells in an arrayed manner. The cells were then fixed, dye- and immuno-stained, and imaged using an automated confocal microscope setup where we acquired four channels (DAPI, NHS-ester, anti-GOLGA1, and mCherry-isoform), corresponding to the nucleus, cell body, trans-Golgi network (TGN) and mCherry-isoform, respectively (**Figure 1C**). From the images, we segmented single cells and extracted ~200 features for each cell describing the mCherry signal intensity and distribution (see **Supp. Figure 1** and Methods for workflow). We used mCherry intensity features to identify transfected cells. Cells with intensity values below the threshold defining transfected cells were labelled non-transfected. At least three experimental repeats were performed for each isoform pair.

Classifier-based approach to identify differences between fluorophore-tagged protein isoforms

We then used measurements of the mCherry signal in single, transfected cells to examine localisation differences between isoforms from the same gene in a pairwise-manner using a classifier-based approach (Methods, workflow illustrated in **Supp. Figure 2**). Here, we pooled the transfected cells expressing two isoforms from the same gene across replicates, down-sampled the cells in the more abundant isoform group to match the number of cells of the less abundant isoform in a replicate-aware manner,

and trained a machine-learning classifier (feedforward neural network) to assign the correct isoform identity (corresponding to the correct well identity) to the cells based on the measured mCherry features. The classifier was trained with 10-fold cross-validation, which allowed the entire dataset to be used for training and testing, and we used the average ROC-AUC (Receiver Operating Characteristic (ROC) Area Under the Curve (AUC)) across the test folds as the metric describing the degree to which isoforms were different from one another (ROC-AUC = 0.5 indicating random isoform identity assignment whereas ROC-AUC = 1.0 indicates perfect isoform identity assignment). These classifiers are referred to as isoform (ISO) classifiers. To consider for well-specific batch effects - such as slight variations in staining, focus, etc. - driving classification rather than true isoform differences, we also trained classifiers to assign well identities to the non-transfected cells from the same wells as the transfected cells (referred to as internal normalisation (IN) classifiers). The performance of IN classifiers can be directly compared to the ISO classifiers. Importantly, in addition to identifying whether isoforms show differences in subcellular localisation, the classifier-based analysis also assigned importance to features based on their contribution to classification. We could therefore identify key features that differed between isoform pairs based on their assigned importance.

To verify that the classifier-based approach worked, we tested its performance on clearly defined negative and positive controls (**Figure 1D**), which respectively consisted of two wells transfected with the same plasmid encoding mCherry-RAB5A and two wells transfected with plasmids encoding either mCherry-RAB5A or mCherry (**Figure 1D, left**). We performed the experiment twice (batch A and batch B), each time generating three biological replicates. For the negative control, we observed similar performance between classifiers trained on transfected (ISO) and non-transfected (IN) for both batches (**Figure 1D, right**). Conversely, we observed near perfect classification by the positive control's ISO classifier (ROC-AUC ~0.98 for both batches), clearly outperforming the associated IN classifiers. This suggested that the classifier-based approach worked as intended but also demonstrated that well-specific effects can affect classification. Moreover, the extent of these effects can vary between groups (IN classifiers of negative control and positive control) and across batches (IN classifiers for positive control across batches) (**Figure 1D, right**), which may confound comparison across isoform pairs. To address this, we defined a normalised ROC-AUC score that used both the ROC-AUC from the ISO classifier and associated IN classifier, so well-specific effects are considered (**Figure 1E, left**). The normalised ROC-AUC metric is symmetrical around 0 and ranges from -1 to +1. Scores >0 indicate isoform differences are detectable after accounting for well-specific effects with a score of 1 indicating perfect classification, which should translate to obvious differences. Conversely, scores ≤0 indicate isoform differences are not detected after accounting for well-specific effects and therefore suggests no isoform

differences. Notably, negative scores likely arise if the presence of the isoform adds variance to systematic differences in the red channel signal (mCherry channel) between the wells that are being compared (due to well-specific effects). Upon applying the normalised scores to the controls, the positive and negative controls had scores near 1 and 0, respectively, with nearly perfect correspondence between batches (**Figure 1E, right**). This indicated the normalised performance metric worked as intended and improved our ability to assess and compare classifier performances.

To further verify the classifier-based approach and using normalised ROC-AUC as readout, we utilised the fact we had performed the classifier-based examination of the controls in two separate batches. This allowed us to train the classifier on cells from one batch and test on the other batch and vice versa to perform out-of-sample testing (**Figure 1F**). In this setup, there is no overlap between the experiments that provide the cells for training and testing, and well-specific effects therefore ought to have a much smaller contribution to classifier performance. Correspondingly, we observed classifier performances (measured as accuracy) for the ISO classifier of the negative controls and IN classifiers of both controls around 0.5, while the ISO classifier accuracy for the positive control remained high (accuracy ~0.92-0.94). This verified that our approach also worked in a more classic setup, and suggested classification were driven by true mCherry construct-related differences. Importantly, the results from our out-of-sample test mirrored the results when using the normalised ROC-AUC score within batches (Figure 1E, right) (**Figure 1G**), with the correlation between the out-of-sample accuracy and normalised ROC-AUC being nearly perfect (slope ~0.44, with perfect slope = 0.5).

Collectively, these results demonstrated the ability of our combined high-content imaging and classifier-based approach to identify true mCherry construct-related signal differences between cells and supported using the normalised ROC-AUC score as a proxy for out-of-sample testing to identify differences.

High-content imaging reveals many isoforms differ in subcellular localisation

We next applied the high-content imaging and classifier-based approach to our isoform library. Due to the number of isoforms, the isoforms were processed in five batches (A, B, C, D, and E) regarding transfections, staining, and imaging, with A and B containing some of the same isoforms. Recurring isoforms in these batches were pooled for the subsequent classifier-based analysis to increase cell numbers, as pooling didn't negatively impact performance when we tested it for the controls (**Supp. Figure 3A-B**). Besides isoforms, all batches contained the negative and positive controls described above, and we examined these first (separately for each batch) to confirm similar performance across batches. We found the negative controls' performances were generally 0 and the positive controls' performances were close to 1 when quantified as normalised ROC-AUC (**Supp. Figure 4A-B**), although notably, the performance on the negative control in batch E was markedly below 0 (norm. ROC-AUC ~ -0.3), likely

because of a significantly brighter NHS-ester staining for this batch compared to the rest. While this may reduce our predictive power on Batch E isoforms, it shouldn't otherwise affect the analysis, and we therefore deemed all isoform batches suitable for analysis.

We therefore performed classifier-based analysis on isoforms from all batches, successfully analysing ~94% (64/68) of the protien isoforms of the library from ~94% of the genes (30/32). This resulted in 42 pairwise comparisons, where ~64% (27/42) was of isoforms from genes where we only examined those two isoforms. We quantified isoform differences using the normalised ROC-AUC (**Figure 2A**) (**Supp. Table 2**) (ROC-AUC_{ISO} and ROC-AUC_{IN} are reported in **Supp. Figure 5A**) and observed scores ranging from nearly 1.0 to -0.5 with most isoform pairs having normalised ROC-AUCs between 0.2 and -0.2. This suggested the examined isoforms contained both isoform pairs with and without localisation differences. We tested if the normalised ROC-AUC was significantly greater than 0 for the isoform pairs (99% bootstrap confidence interval lower bound did not include 0 (**Supp. Figure 5B**)), which would indicate isoform localisation differences. This was the case for 38% (16/42) of the isoform pairs as well as the positive control, but not the negative control (**Figure 2A**). Importantly, the indicated isoform differences were not associated with systematic differences in expression levels between the isoforms within pairs (**Supp. Figure 6A**) nor by how many cells that were used for classification came from a single experimental repeat (**Supp. Figure 6B**). This suggested these potential confounders didn't explain the indicated isoform differences, which instead may relate to true isoform differences.

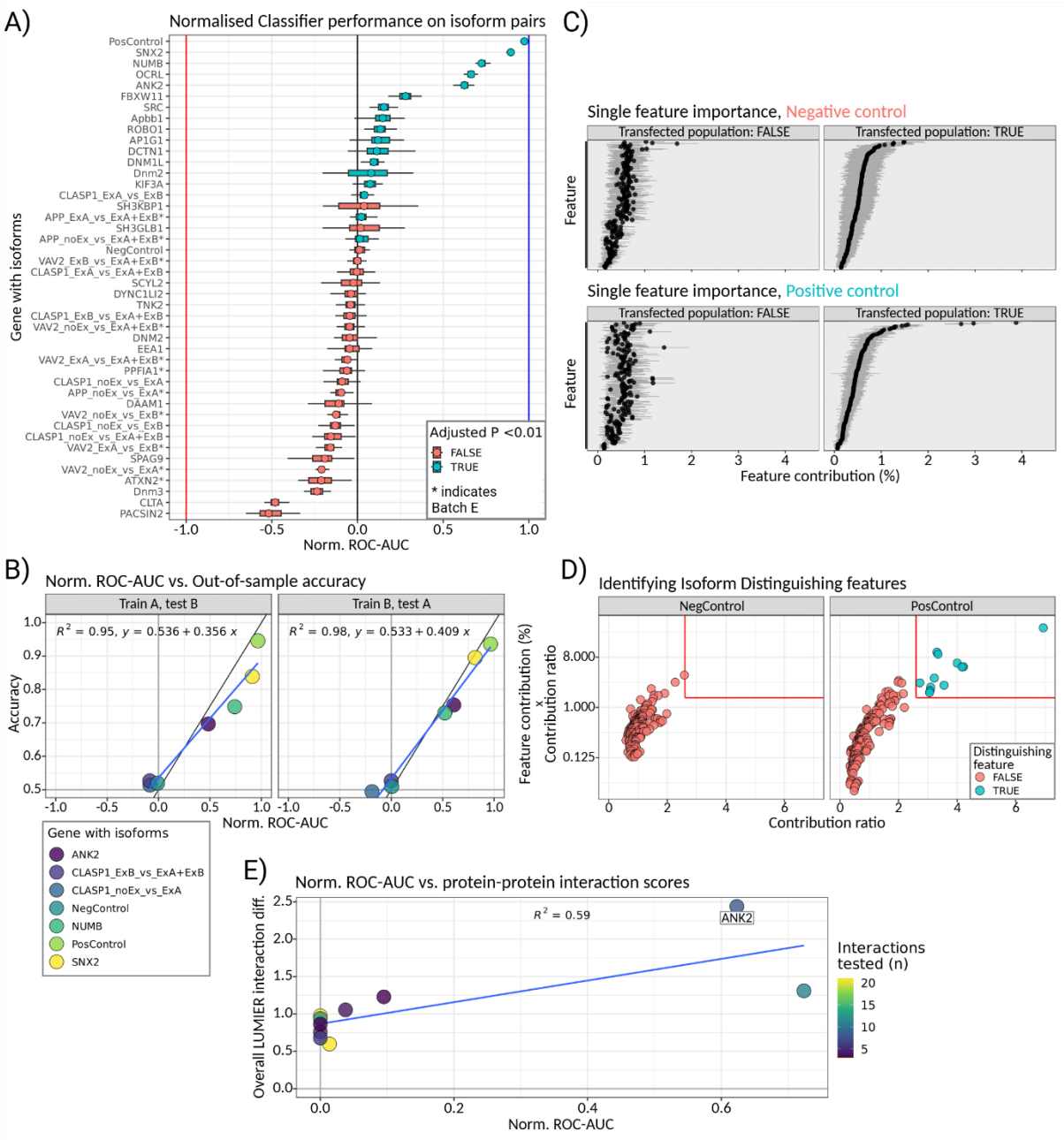


Figure 2. Many isoform pairs contain differentially localising isoforms and localisation differences correlate with interaction differences. A) *Boxplot showing the performance of classifiers trained on mCherry features of cells transfected with protein isoforms (or controls) making up isoform pairs, quantified as normalised ROC-AUC. All isoform pairs from batches A-E are shown. The cells used to examine a given isoform pair came from three experimental repeats from a given batch that were pooled for the classifier-based analysis. For isoform pairs present in multiple batches (batch A and B), cells expressing these were pooled prior to classifier-based analysis and so cells came from 6 experimental repeats. The shown controls also derive from pooling batch A and B. n cells ≥170 for each pair after pooling, with median n cells = 967, and numbers for each pair listed in Supp. Table 2. The boxplots illustrate the normalised ROC-AUCs from each of 50 iterations of cell sampling and classifier evaluation, and dot indicates median (n = 50). The isoform pairs are coloured according to whether the lower bound of a 99% bootstrap confidence interval included 0, after multiple testing correction (false discovery rate), with p-values < 0.01 being considered significant and indicating isoform differences. Batch E pairs are marked with *. Gene names in all caps indicate isoforms come from human, whereas genes only starting with upper case come from mouse.* ***B)*** *XY plot showing the correlation between the normalised ROC-AUC and the out-of-sample accuracy for the controls and five isoform pairs occurring in both batch A and B (n = 7). The normalised ROC-AUC for batch A is compared to the out-of-sample accuracy when trained on batch A and tested on B (vice versa for ROC-AUC for batch B). A linear regression was fitted to the data. After pooling cells across experimental repeats, ≥232 cells were present in both training and test set for a given comparison.* ***C)*** *XY plot showing the relative importance of features for classification, quantified as Feature contribution (%)*

(mean across classifier iterations is shown). Separate plots are shown for ISO (Transfected = TRUE) and IN (Transfected = FALSE) classifiers for the negative and positive controls. Lines indicate mean ± standard deviation. ***D)*** *XY plot of Contribution ratio and Contribution ratio multiplied with Feature contribution (%) for the ISO classifier for each feature for the positive and negative controls. The feature behaviours of the negative control are used to determine threshold values to define 'Isoform distinguishing' features (features important to the ISO classifier compared to the IN classifier). Features are coloured according to whether they are Distinguishing or not.* ***E)*** *XY plot examining the correlation between the normalised ROC-AUC and overall LUMIER interaction difference (Methods) for isoform pairs where ≥3 interactors have been found to interact with at least one of the isoforms in a pair, n = 10 isoform pairs. A linear regression was fitted to the data, and the correlation was tested using a correlation test (cor(method = 'Kendall')).*

To validate this, we performed out-of-sample testing on the isoforms that were recurring across batches A and B (**Figure 2B**). We found the accuracy and normalised ROC-AUC score was positively correlated for both batches (p-values = 0.01 (batch A) and 0.003 (batch B)) with slope values of ~0.36 (batch A) and ~0.41 (batch B) corresponding well with a value indicating perfect correlation (slope = 0.5). This supported the identified isoform differences were true mCherry construct differences and not well-specific effects. Importantly, our results were further supported by ~71% (5/7) of isoform pairs with previously described localisation differences being among the 16 isoform pairs with detected differences (**Supp. Table 2**), providing external support to our observations. Our results therefore suggest AS may frequently alter protein subcellular localisation, at least for proteins differentially including neural microexon.

Identification of known functional cassette exons confirms fidelity of high-content imaging approach

Having identified that a substantial fraction of isoforms localised differentially, we next examined what the differences were using the importance scores assigned to each signal feature during classification. We first calculated each feature's contribution percentages (%) in each isoform pair's ISO classifier, which reflected how many percent a single feature's importance score made up of the sum of feature importance scores (**Figure 2C**). We then normalised this feature contribution (%) to the feature contribution (%) in the associated IN classifier (referred to as contribution ratio) to remove features that described well-specific effects. To consider both feature enrichment (contribution ratio) and feature importance in the selection of features describing actual isoform differences, we finally plotted the contribution ratio against the feature contribution (%) (ISO) multiplied by the contribution ratio (**Figure 2D**) and used the negative control to set threshold values to define the features describing isoform differences, referred to as isoform distinguishing features (**Figure 2D**).

We validated this approach of identifying isoform distinguishing features by examining the NUMB and OCRL isoform pairs. Both were among the four best performing isoform classifiers (normalised ROC-AUC > 0.5) and had previously been examined using microscopy[33,34], which enabled comparison of our data and distinguishing features with previous reported localisation differences. For NUMB, the isoform distinguishing features related to Intensity_Edge (**Supp. Figure 7A-D**), suggesting differences in

plasma membrane localisation. This was apparent from the cell images (**Supp. Figure 7E**) and recapitulated the localisation differences previously reported in MDCK cells[34]. For OCRL, the top isoform distinguishing features described mCherry Granularity and intensity variance (**Supp. Figure 8A-D**), suggesting the isoforms differed in how diffusely they localised. Accordingly, we observed the mCherry-OCRL +EX isoform having a greater vesicular distribution (**Supp. Figure 8E**), replicating previous observations of OCRL isoforms in HeLa[33]. These observations supported the overall high-content approach and the classifier-based identification of isoform distinguishing features.

We next examined isoform distinguishing features simultaneously for the 16 isoform pairs with differentially localising isoforms (**Supp. Table 3**, **Supp. Figure 9**). However, our dataset likely wasn't big enough to identify trends in how AS changes protein localisation generally, as most isoform distinguishing features (69%, 29/42) were unique to single isoform pairs (see **Supplementary Results**).

Instead, we focused on the two remaining isoform pairs with highly scoring isoform classifiers (normalised ROC-AUC > 0.5), which were sorting nexin 2 (SNX2) and ankyrin-2 (ANK2) isoforms. To our knowledge, none of these had been examined previously using microscopy and the SNX2 isoforms had not been examined at all.

Isoform localisation differences correlate with larger protein-protein interaction differences

The ANK2 isoforms have previously been examined for binary PPIs using LUMIER, where the ANK2 +EX isoform interacted more strongly with various cytoskeletal and membrane proteins[18]. Interestingly, our imaging-based analysis highlighted mCherry intensity variance and RadialDistribution as isoform distinguishing features, which indicated an enrichment of ANK2 +EX at the plasma membrane (**Supp. Figure 10A-D**), which we confirmed in the images (**Supp. Figure 10E**). The correlation between increased membrane localisation and enhanced interactions with cytoskeleton and membrane proteins was notable, even if we can't determine causality, and hints at the complementary use of imaging and interaction data in isoform characterisation.

To further explore this, we utilised the fact that many isoform pairs in our library besides ANK2 had been examined for PPI differences via luminescence-based mammalian interactome (LUMIER)[18,31] to examine the correlation between localisation differences and overall PPI differences (**Figure 2E, Supp. Table 4**). We found a positive, although modest, correlation between the normalised ROC-AUC and interaction differences (R^2 = 0.58 linear regression. p = 0.027, cor.test), suggesting more readily detectable localisation differences were associated with more pronounced PPI changes. Similar to the ANK2 isoforms, we can't determine whether the examined PPIs differences drive the altered localisation or vice versa, but the two readouts help contextualise each

other and so likely aids interpretation, supporting them being complementary. Moreover, the positive association between different types of readouts further supported our analysis approach was detecting real isoform differences.

The sorting nexin 2 isoforms differ in their localisation to vesicular structures and propensity to form aggregates

We next examined the previously uncharacterised sorting nexin 2 (SNX2) isoforms, the last isoform pair with normalised ROC-AUC > 0.5, as a case study of how imaging-based characterisation facilitates more detailed isoform characterisation and how AS changes protein localisation. Briefly, SNX2 is a lipid binding protein involved in retrograde transport and recycling of internalised integral membrane proteins as well as regulating endosomal and vesicle-mediated processes[39]. It generally performs these functions as part of the retromer[40] and/or ESCPE-1[41,42] complexes. Structurally, SNX2 is part of the SNX-BAR subfamily of SNX proteins and therefore contains two structured domains: a Phox homology (PX) domain and a Bin/Amphiphysin/RVS (BAR) domain[43]. Both domains are involved in lipid binding, but the BAR domain also promotes dimerization with other SNX-BAR proteins through BAR-BAR interactions[43], resulting in SNX-BAR dimer complexes like ESCPE-1[41]. The BAR domain was particularly interesting as the SNX2 isoforms differed by 4 amino acids, EIEE, inserted in the BAR domain by a neural microexon (**Figure 3A**).

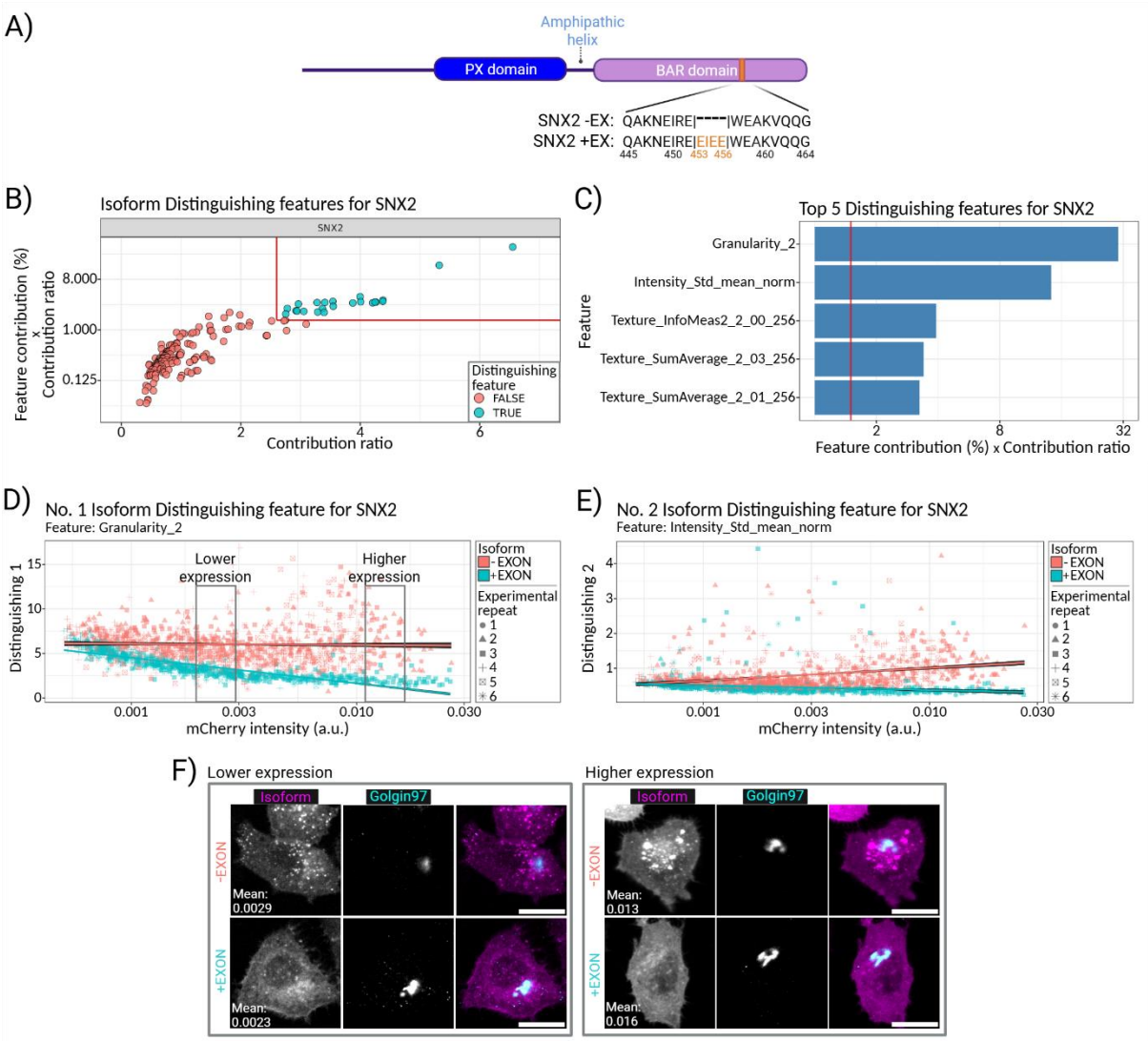


Figure 3. Two SNX2 isoforms differ in their localisation to vesicular structures and propensity to form aggregates. A) *Primary structure of human SNX2 with structured domains indicated as ovals and disordered regions as lines. The cassette exon (orange) is inserted into SNX2's BAR domain (purple), and the exon's residues as well as the residues immediately adjacent to the exon is indicated. Residue numbers prior to the exon corresponds to Uniprot ID: O60749. Residues after the exon have a position number that's +4 that of Uniprot ID: O60749.* ***B)*** *XY plot of Contribution ratio and Contribution ratio multiplied with Feature contribution (%) for the ISO classifier for each feature for the SNX2 isoform pair, used to identify Isoform Distinguishing features. Of ~200 features, ~20 was identified as Isoform distinguishing.* ***C)*** *Bar plot of the top 5 Isoform distinguishing features according to their Feature contribution (%) multiplied with Contribution ratio. The threshold value defining Isoform Distinguishing features is indicated as a red line.* ***D)*** *XY plot of Top 1 Isoform Distinguishing Feature for the SNX2 isoforms (mCherry Granularity) against cell mean mCherry intensity levels (proxy for expression level). Each data point corresponds to a cell with colour and shape indicating SNX2 isoform identity and experimental repeat ID, respectively. Cells come from n = 6 experimental repeats, with pooled n cells > 1300. Separate linear regression lines were fitted to SNX2 +EX and SNX2 -EX-expressing cells. Two rectangles (lower expression, higher expression) indicate where the representative cells (Figure 3F) fall within the overall population.* ***E)*** *XY plot of Top 2 Isoform Distinguishing Feature for the SNX2 isoforms (mCherry intensity standard deviation, normalised to mCherry mean intensity) against cell mean mCherry intensity levels (proxy for expression level). Each data point corresponds to a cell, same as Figure 3D.* ***F)*** *Representative images of HeLa cells expressing mCherry-SNX2 isoforms. Cells expressing the isoforms at two different levels (lower, higher) and their relation to the overall cell population is indicated by squares in Figure 3D, with the mean mCherry intensity of the cell stated in white. The mCherry isoform signal is in magenta and the Golgin97 signal, reflecting the trans-Golgi network, is shown in cyan. Scale bar: 15µm.*

Whereas the insertion resulted in minimal primary sequence changes, it drastically altered localisation with ~20 isoform distinguishing features (**Figure 3B**), where the top features were mCherry granularity and intensity variance (**Figure 3C**). This mirrored the OCRL isoforms' top isoform distinguishing features, suggesting differences in vesicular localisation patterns, which were apparent from the images (**Figure 3F**). Because our approach relied on over-expression, we next examined the relationship between the top features and isoform expression levels. This demonstrated both features differed between isoforms across all measured expression levels (**Figure 3D-E**), confirming general localisation differences. Interestingly, the relationship between the features and expression levels wasn't constant with differences exacerbated at higher expression levels due to altered behaviour of the SNX2 -EX isoform. We confirmed our population-based analysis by inspecting the cells (**Figure 3F**): At lower expression levels, SNX2 -EX had a more well-defined vesicular localisation pattern compared to SNX2 +EX, which had a more diffuse signal, although still localised in a granulated manner (**Figure 3F, left**). At higher expression levels, SNX2 -EX but not SNX2 +EX formed aggregates, which explained the observed increase in granularity and intensity variance (**Figure 3F, right**). Notably, SNX2 -EX has previously been reported to form aggregate-resembling swollen endosomes upon over-expression, and to localise to early endosomes upon less severe over-expression[44], which presumably is what we observe for both isoforms at lower expression levels. Collectively, this demonstrated pronounced localisation differences between the SNX2 isoforms and that this approach can be used to analyse isoform-specific localisation differences in an expression level-informed manner.

SNX2's microexon is highly conserved and does not disrupt folding of BAR domain

To investigate how the 4-residue microexon (**Figure 4A**) might alter localisation and aggregation behaviour of SNX2, we first assessed its functional, sequence, and structural conservation across vertebrate species. We inferred functional conservation by analysing the inclusion pattern of the microexon in bulk RNA-sequencing data from 13 tissues for four vertebrate species (human, rat, chicken, zebrafish) where the microexon was annotated in vastDB[45] (**Supp. Figure 11A**). This showed the exon was selectively expressed and highly included in neural tissues across the species despite >400 million years of evolution separating some of them, in line with previous observations of microexon conservation and tissue specificity[14]. This indicated a conserved biological function of the exon, that presumably related to the imaging phenotype.

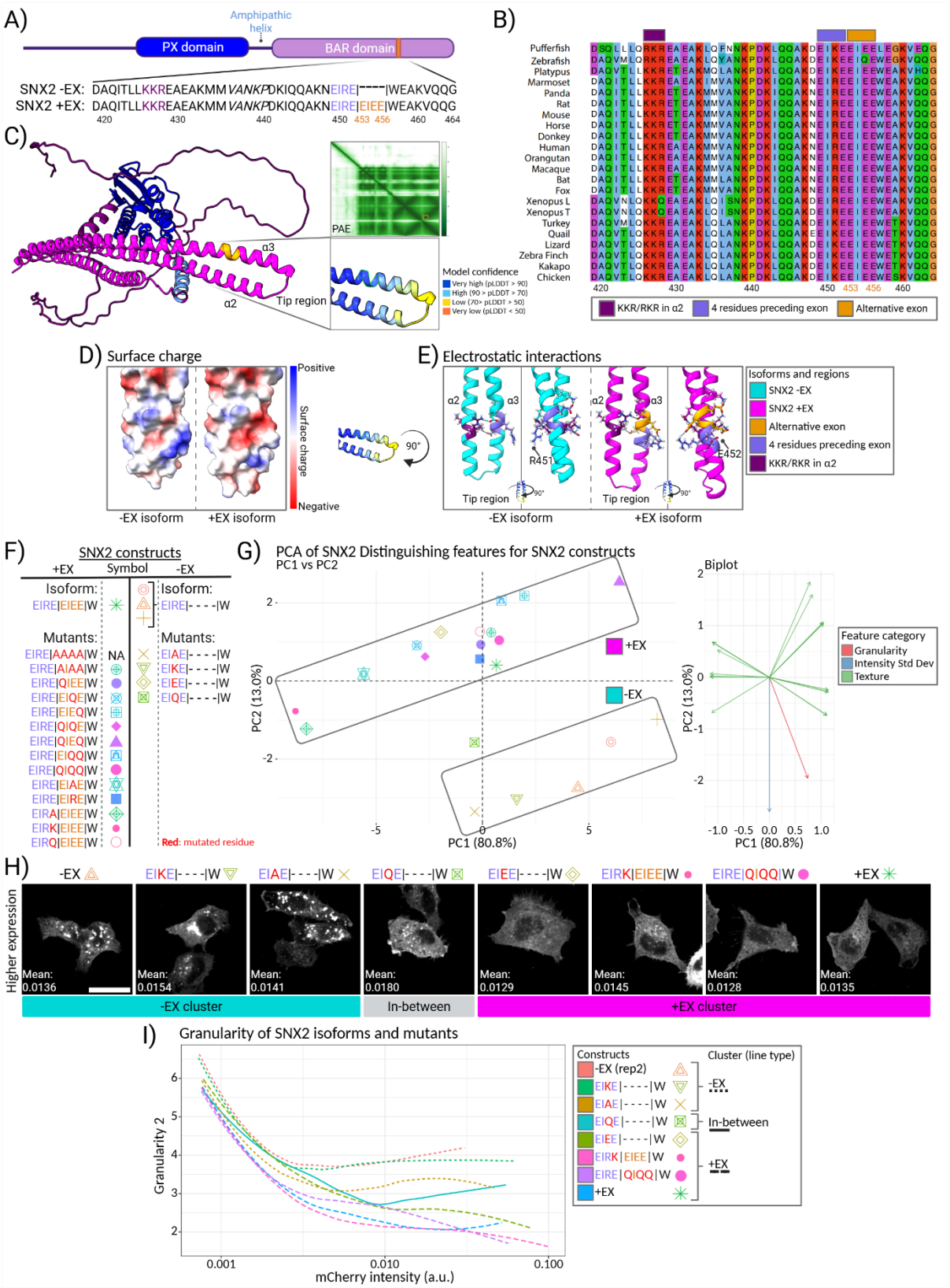


Figure 4. SNX2's subcellular localisation is altered by the insertion of the well-conserved microexon, not by the composition of its residues. A) *Primary structure of human SNX2 with structured domains indicated as ovals and disordered regions as lines. The cassette exon (orange) is inserted into SNX2's BAR domain (purple), and the exon's residues (orange) as well as the residues surrounding the exon are indicated (residues 419 - 464). The four residues preceding the exon are lavender, and the KKR motif (426-428) is dark purple. Residue numbers prior to the exon corresponds to Uniprot ID: O60749. Residues after the exon have a position number that's +4 that of Uniprot ID: O60749.* ***B)*** *Multiple sequence alignment of SNX2 region corresponding to human*

SNX2 residues 419 - 464 for 22 vertebrate species. The location of the cassette exon (orange), the 4 residues preceding the cassette exon (lavender), and the KKR motif (dark purple) are indicated by coloured blocks. The individual positions are coloured according to the residue properties that is conserved across the aligned sequences (default clustal colouring from JalView). ***C)*** *AlphaFold2 (AF2) structure of human SNX2 +EX. The BAR domain is coloured magenta, the amphipathic helix is light blue, and the PX domain is dark blue. The cassette exon is coloured orange and inserts as an additional helical turn in the alpha3 helix of the BAR domain. The α2 helix and tip region connecting α2 and α3 are indicated. The predicted aligned error (PAE) for the model (region with exon marked by orange square) and the predicted local distance difference test (pLDDT) for the BAR region with the microexon is shown. Exon highlighted by orange square in PAE plot and green outline in pLDDT model.* ***D)*** *Surface charge of AF2 models of human SNX2 isoforms in the BAR region containing residues 419 - 464. The BAR domain is viewed from the side that faces lipid membranes.* ***E)*** *Predicted electrostatic interactions between residues in the AF2 models of human SNX2 isoforms in the BAR region containing residues 419 - 464. Residues engaging in electrostatic interactions (salt bridges) are shown as well as residues that form electrostatic interactions in one isoform but not the other (R451, E452). The cassette exon, the four preceding residues. and the KKR motif is coloured. The left panel shows the BAR domain from the side that faces lipid membranes.* ***F)*** *List of SNX2 mutant constructs examined via high-content imaging. Residue colours correspond to colours in A). Mutated residues are indicated in red. Each construct has a symbol that indicates their position in the PCA. Note, three technical replicates for SNX2 -EX was included.* ***G) Left****, XY plot of PC1 versus PC2 for principal component analysis (PCA) of the mCherry-SNX2 construct imaging data using the population means of the SNX2 Isoform Distinguishing features from the transfected cells as input. Two clusters that are perpendicular to the granularity vector are shown, these contain -EX and +EX, respectively. Cells were pooled across three experimental repeats, with ≥185 cells expressing each construct (cell numbers for each construct are shown in Supp. Figure 12A).* ***Right****, A biplot of the Isoform Distinguishing features, where arrow directions show how variables contribute to each PC axis (PC1, PC2), and arrow length indicates their importance. Different feature classes are shown. Granularity is shown in red.* ***H)*** *Representative images of mCherry-SNX2 isoforms and select mutants in HeLa cells at higher expression levels. The mean mCherry intensity for the brightest cell in each image is indicated as white text. The PCA symbol and cluster identity for each construct is indicated. Scale bar: 25μm.* ***I)*** *Line plot of mCherry Granularity (Y-axis) against cell mean mCherry intensity levels (proxy for expression level) (X-axis) for mCherry-SNX2 isoforms and select mutants that demonstrate the transition in aggregation behaviour. The lines are fitted to all transfected cells expressing the specified constructs, pooled across three experimental repeats, with ≥185 cells expressing each construct (cell numbers for each construct are shown in Supp. Figure 12A). Line colours reflect construct identity and line type indicates PCA cluster.*

We then examined the exon's sequence conservation via multiple sequence alignment of the exon and surrounding region of SNX2's BAR domain for 22 species representing most of the major vertebrate clades to zebrafish (**Figure 4B**). We found no variation in the exon's residue composition (EIEE) across species except in zebrafish, which exhibited a conservative E to Q substitution (EIQE). Notably, the level of conservation was similar or higher than that of the surrounding residues in the BAR domain – which itself was highly conserved - including the upstream KKR residues which contribute to lipid binding for the closely related SNX1[46], as these were switched to KKQ in the Xenopus genus.

We next modelled the SNX2 isoforms from the four species where we had verified the splicing pattern (human, rat, chicken, and zebrafish) using AlphaFold2 (AF2)[47]. The microexon-containing region of the BAR was practically identical between organisms (**Supp. Figure 11B-C**) and suggested the microexon inserts into SNX2's BAR as a single additional helical turn in the α3 helix (**Figure 4C**). The insertion made the region's surface charge more negative (**Figure 4D**), which was less pronounced in zebrafish (**Supp. Figure 11D**), and resulted in changed electrostatic interactions for select residues immediately adjacent to the exon (R451/E452) (**Figure 5E**). Meanwhile, the exon's isoleucine (I454) faced the helix-helix interface of the coiled-coil structure

formed with the adjacent α2 helix (**Supp. Figure 11E**), which likely stabilised the BAR domain[48].

Collectively, these analyses supported a conserved function of SNX2's microexon across vertebrates, which presumably related to the highly conserved residues of the exon altering the surface charge and electrostatic properties of SNX2's BAR domain, while maintaining its fold.

SNX2's subcellular localisation is altered by the insertion of the microexon, not by the composition of its residues

To experimentally evaluate the importance of the compositional and electrostatic conservation of the microexon, we examined the SNX2 isoforms via directed mutagenesis. We first generated ~20 mCherry-SNX2 mutant constructs (**Figure 4F**), mutating the residues of the exon as well as the two flanking charged residues (R451, E452) with predicted electrostatic interaction changes (**Figure 4E**). Notably, we left I454 unchanged in all but one mutant (+EX: EIRE|AAAA|W, _ indicates changed residues) due to its hypothesized importance in stabilising the BAR domain. Next, we transfected the mCherry-SNX2 mutants into HeLa cells and examined them using high-content imaging. We initially verified that the constructs were expressed at comparable levels, finding that 18/19 constructs had similar cell population expression levels (**Supp. Figure 12A**). Conversely, +EX: EIRE|AAAA|W expressed at considerably lower levels, supporting I454 as potentially stabilising the BAR domain, and was excluded from subsequent analysis to avoid expression level differences impacting the analysis. For the remaining SNX2 mutants, we used the SNX2 isoform distinguishing features (**Figure 3B**) as input for principal component analysis (PCA), to identify connections between localisation and the specific mutations (**Figure 4G**). Plotting PC1 and PC2, which explained >93% of the variance (**Supp. Figure 12B**), we noticed the SNX2 constructs split into two parallel groups (**Figure 4G, left**) that were perpendicular to the granularity feature vector, as indicated by the biplot (**Figure 4G, right**). We focused on this split, as the granularity feature captured the visible isoform differences (**Figure 3F**) and was the top isoform distinguishing feature (**Figure 3C**). Importantly, the endogenous SNX2 isoforms (-EX/+EX) clustered in the separate groups while 3 replicates of SNX2 -EX clustered together (**Figure 4G**), whereas we didn't observe the same separation of SNX2 isoforms when performing PCA on non-transfected cells from the corresponding wells (**Supp. Figure 12C-E**), suggesting the clustering captured and were driven by construct-specific rather than well-specific effects.

Examining the mutants present in the two groups, we were surprised that the groups nearly perfectly reflected whether SNX2 had a 4-residue insertion or not. All +EX mutants grouped together with the +EX isoform, suggesting neither the exon's negative charge nor the unbonded adjacent E452 were driving the isoform differences. Conversely, the -EX mutants focused on R451 clustered with -EX upon R451K and

R451A but gradually moved toward the +EX cluster upon R451Q, and clustered with +EX upon R451E, which reflected the mutants' aggregation behaviour as observed in the images (**Figure 4H**) and granularity across expression levels (**Figure 4I**). Rather surprisingly, it therefore seemed that the insertion of 4 residues was causing the difference in isoform behaviour rather than the specific nature of the residues, except I454 that appeared important for coiled-coil stability. Upon re-examining the SNX2 isoform structures, we considered the effect might result from the insert displacing positively charged residues relative to other residues – as R451E caused -EX to resemble +EX - which might alter intra- or inter-molecular interactions with lipids or proteins, that the BAR engages in, that could then manifest in altered SNX2 localisation.

SNX2-dependent interactions are also disrupted by insertion of microexon not by its composition

To orthogonally validate the observations from the SNX2 mutagenesis and examine if inter-molecular interactions differed between isoforms, as speculated, we examined potential PPI differences between the SNX2 isoforms and select SNX2 mutants using LUMIER and affinity purification mass spectrometry (AP-MS).

We first performed AP-MS to identify interactors in a nontargeted manner. Here, we generated HeLa cell lines expressing either SNX2 isoform by knocking out endogenous SNX2 and then stably introducing the respective isoform via viral transduction, as verified by western blotting (**Supp. Figure 13A-B**) and end-point RT-PCR (**Supp. Figure 13C**). We subjected these cell lines to AP-MS and reliably identified canonical SNX2 interactors, including MPRI (*IGF2R*) and other SNX-BARs (SNX1, SNX5, and SNX6), for both isoforms (**Figure 5A**). This suggested both SNX2 isoforms formed SNX-BAR complexes, like ESCPE-1 (dimer between SNX1 or SNX2 with SNX5 or 6)[41], and could interact with cargo proteins like MPRI. We next wondered if the SNX2 isoforms differed in what cargos they preferentially interacted with, but because such interactions generally are transient, they are difficult to detect using AP-MS. To address this, we used LUMIER - an assay measuring co-immunoprecipitation of a luciferase-tagged bait protein with a FLAG-tagged prey protein[49] - which can detect transient interactions[49], to examine the interaction of the SNX2 isoforms with transmembrane cargo. We examined 6 transmembrane proteins that interact with and/or are regulated by ESCPE-1[41] or other SNX-BAR complexes[50]. We found both SNX2 isoforms interacted with ~66% of the examined proteins (4/6) (NLIR > 3[51]) (**Figure 5B**), including canonical ESCPE-1 cargo like the insulin-like growth factor 1 receptor (IGF1R). This supported the AP-MS data (SNX2 interacting with SNX5/6) and LUMIER's ability to measure cargo interactions. Interestingly, one of the transmembrane proteins, the insulin receptor (INSR), appeared to interact more strongly with SNX2 +EX than -EX, with the normalised interaction score (NtFLIR) for SNX2 +EX being two-fold that of SNX2 -EX (**Figure 5C**) and trending towards statistical significance ($p = 0.051$, one-sample t-test against $\mu = 0$, cohen's $D = 1.58$).

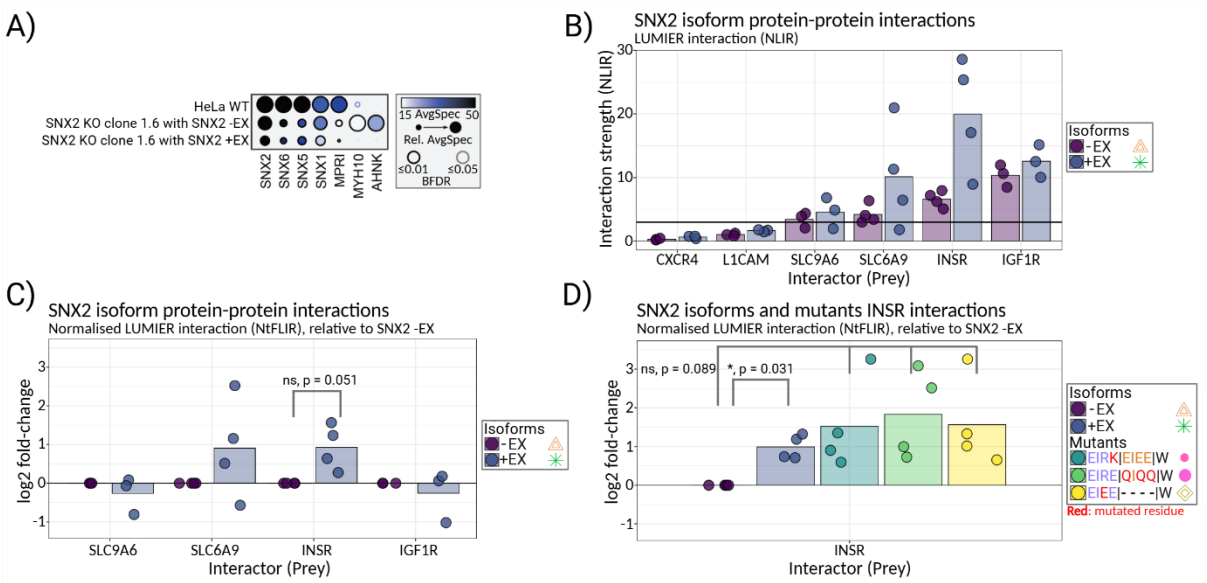


Figure 5. SNX2's microexon cause isoforms to differentially interact with the insulin receptor. A) *Bubble plot of proteins detected upon affinity purification mass spectrometry (AP-MS) analysis of HeLa wild type and HeLa SNX2 isoform lines. Only proteins with average spectral counts ≥15 are included. n = 3 independent experiments. Colours indicate average spectral counts, size indicates relative spectral counts, and line indicates Bayesian False Discovery Rate (BFDR).* ***B)*** *Bar plot showing LUMIER interaction scores for SNX2 isoforms with 6 transmembrane proteins that are suggested to interact with and/or be regulated by SNX-BAR complexes. Interaction strength is quantified as NLIR (normalised LUMIER interaction ratio), and an NLIR>3 (black line) is considered a positive interaction51. Column height indicates mean NLIR across experimental repeats (n = 3 or 4 independent experiments), and dots reflect NLIR from each experiment. Colours reflect SNX2 isoform identity.* ***C)*** *Bar plot of LUMIER interaction strength of SNX2 +EX with the 4 transmembrane proteins that were identified as positive interactors (Figure 5B), relative to the SNX2 -EX isoform's interaction with them (log2 fold-change). This is the same data used for Figure 5B but, the interaction score is normalised to both the level of prey and bait protein (NtFLIR). Column height indicates mean log2FC across experiments (n = 3 or 4 individual experiments), while dots reflect log2FC for individual experiments. Colours reflect SNX2 isoforms. Statistical test is one-sample t-test against μ = 0, p-value > 0.05, ns.* ***D)*** *LUMIER interaction strength of SNX2 +EX and three SNX2 mutants with INSR, relative to the SNX2 -EX isoform's interaction with INSR (log2 fold-change). Individual dots reflect log2FC for individual experiments, and column indicates mean across experiments (n = 4 independent experiments). Interaction differences were tested using one-sample t-tests against μ = 0. p-values were corrected for multiple testing using false discovery rate. P-value < 0.05, *. P-value > 0.05, ns.*

We then used the SNX2-INSR PPI to examine via LUMIER whether the likely interaction difference between the isoforms was caused by the microexon's residue composition or the insertion of 4 residues irrespective of their composition, which may displace residues, including R451, as we had observed for the localisation. We again found SNX2 +EX interacted twice as strongly with INSR as -EX (p-value = 0.03, one-sample t-test against μ = 0, cohen's D = 3.19) (**Figure 5D**), confirming the PPI difference, and that the two examined +EX mutants (EIRK|EIEE|W, EIRE|QIQQ|W) also interacted more strongly with INSR than -EX. This indicated that similarly to the imaging phenotype, the microexon wasn't driving PPI differences via its composition or the flanking E452. Relatedly, the -EX mutant (-EIEE|----|W) that resembled the +EX isoform during imaging (**Figure 4H**), also interacted more strongly with INSR than -EX. The INSR PPI result therefore mirrored our microscopy-based observation that the exon's residue composition was inconsequential, and its effect instead appeared to relate to displacing residues relative to one another within the BAR domain with the positively

charged R451 being important. Moreover, the differential interaction with INSR suggests SNX2's microexon potentially alters cargo specificity alongside subcellular localisation.

Microexon inclusion displaces charged residues within secondary structure elements

Considering the relatively generic way whereby the SNX2 microexon supposedly altered function, we wondered if microexons generally affected α-helices in this manner, focusing on neural microexons. Each helical turn in α-helices is ~3.6 residues, and we first examined if microexons were biased toward preserving this architecture by plotting the nucleotide length distribution of microexons in coding regions generally and microexons predicted to occur in α-helices (**Figure 6A**). We observed 12nt and 21nt long microexons, encoding 4 and 7 residues, respectively, were enriched in predicted helices (**Figure 6B**), suggesting microexons may frequently preserve helical structure, although enrichment of 24nt and 27nt microexons in helices suggest this observation is limited to shorter exons. To further explore this, we examined the relative position of microexons within predicted helices (**Figure 6C**). Interestingly, we observed microexon with helix preserving lengths (12nt, 21nt) were rather evenly distributed throughout helices, while microexons with supposed disruptive lengths were enriched near helix ends. This indicated microexons overlapping α-helices generally either preserve helical structure or occur near ends where they may induce or extend loops, exemplified by AlphaFold models of ERC1 (or SNX2) and TEAD1, respectively (**Figure 6D**).

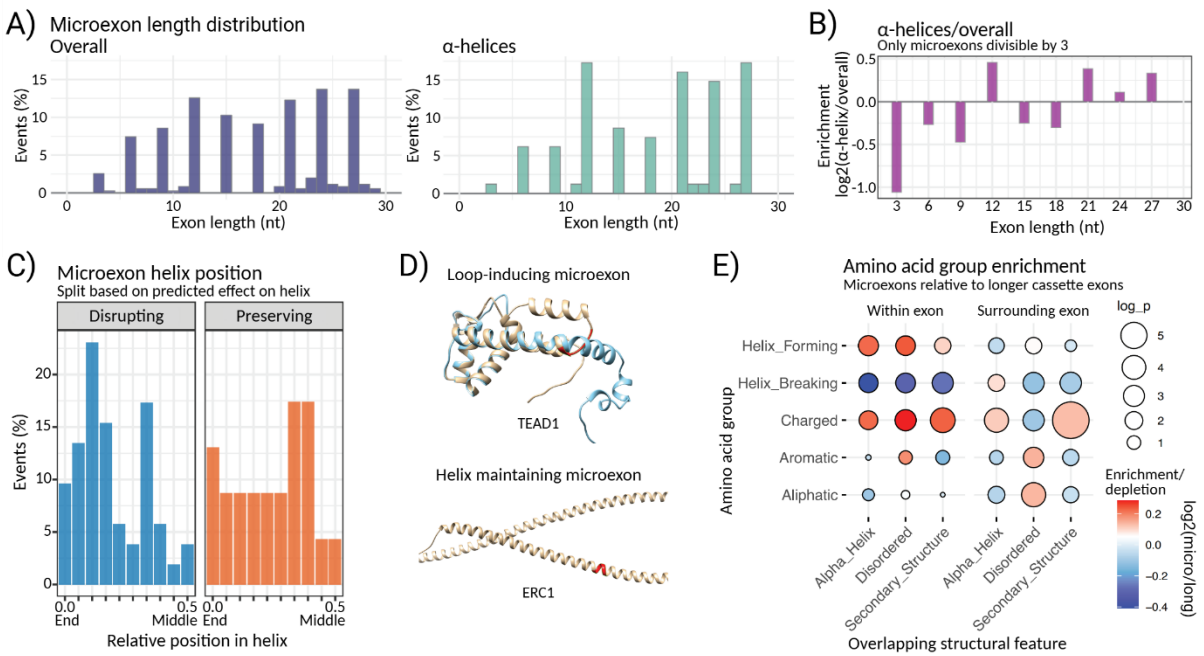


Figure 6. Microexons potentially affect α-helical structures in two common ways. A) *Relative frequency histogram of microexon nucleotide (nt) lengths for neural microexons, defined as exon <30nt and differentially spliced in neuronal tissues (Methods),* ***Left,*** *overall.* ***Right,*** *microexons predicted to overlap α-helices. n microexons = 353, n microexons in α-helices = 82.* ***B)*** *Bar plot showing the enrichment and depletion of neural microexons with specific lengths in α-helices, calculated as log2(Freq. α-helix / Freq. overall). Only microexon*

lengths divisible by 3 are included. ***C)*** *Relative frequency histogram of neural microexon positions within predicted α-helices, quantified as the relative distance of the point where the microexon is inserted to the nearest α-helix end (Methods). Microexons are grouped according to whether their length is predicted to preserve (12nt, 21nt) (Preserving) or disrupt (other lengths) the helical structure (Disrupting).* ***D)*** *AlphaFold structures of TEAD1 and ERC1 isoforms, showcasing microexons predicted to induce/extend loops in alpha helical structures and microexons predicted to preserve the helix. Microexons are coloured red. For TEAD1, isoforms with and without the microexon are shown in separate colours (blue and gold) and overlaid.* ***E)*** *Bubble plot showing amino acid enrichment within exons (exon) and flanking regions (surround), for neural microexons overlapping predicted α-helices, disordered regions, and secondary structure elements (α-helices, β-strands, loops), compared to longer neural cassette exons (≥30nt). Only exons divisible by three were included. Amino acids were grouped according to properties in a non-mutually exclusive manner: helix forming: L, I, F, E, Y, W, M, helix breaking: P, G , Charged: E, D, K, R , Aromatic: F, Y W, H, Aliphatic: L, I, V. Colours indicate degree of enrichment (calculated as log2(Freq. microexon / Freq. longer exon)), size indicates statistical significance calculated by fisher-exact test. Flanking residues were the first 30 residues upstream, and 30 residues downstream of the microexon. n microexons in α-helices = 77.*

To further investigate the consequences of these occurrences, we examined the amino acid compositions of neural microexons and the immediately flanking residues, compared to longer (>30nt) neurally-enriched cassette exons. We found microexons generally were enriched in helix promoting residues and charged residues, irrespective of whether they occurred in predicted α-helices or disordered regions (**Figure 6E**), although helix promoting residues were less enriched when other types of secondary structure was included (β-strands and loops considered alongside α-helices). Meanwhile, residues flanking microexons in α-helices and secondary structure generally were enriched in charged residues compared to longer exons, which wasn't observed for disordered regions (**Figure 6E**). These observations were in-line with our results for SNX2, which demonstrated microexon insertion displacing residues, including important charged residues, within a structured region, and suggests this mechanism potentially extends to other microexons inserted into secondary structure and may contribute to their effect.

Discussion

The function of most protein isoforms is unknown, despite protein isoforms being highly prevalent[6] and a deepened understanding of protein isoforms can have both fundamental and applied value[11]. Characterisation of protein isoforms, especially at scale, is therefore warranted. To complement existing approaches for high-throughput examination of protein isoforms, we combined high-content imaging with classifier-based analysis to characterise protein isoform subcellular localisation in a pairwise-manner. To our knowledge, this is the first time high-content imaging has been used to directly study AS-derived protein isoforms. We applied the approach to study tissue-regulated cassette exons, most of which were microexons included in neural tissues through SRRM4 activity and successfully performed 42 protein isoform pair comparisons of 64 protein isoforms from 30 genes, where we analysed only two protein isoforms for 90% (27/30) of genes. We found the protein isoforms to be different in 38% (16/42) of the comparisons, with most examples being from genes where we only

compared two protein isoforms (**Figure 2A**). The differentially localising protein isoforms included ~71% (5/7) of protein isoform pairs with previously reported localisation differences (Dnm2[37,38], DNM1L, NUMB[34], OCRL[33], SRC[32]), and we verified the localisation patterns for NUMB and OCRL replicated previous findings[33,34], which validated our approach's ability to identify isoform differences.

Alternative splicing as a frequent regulator of protein subcellular localisation

While our dataset is not genome-wide, the relatively high fraction of differentially localising isoforms suggests AS is a prevalent regulator of protein subcellular localisation. The universality of this may however depend on how representative our library is of the overall pool of protein isoforms. We focused on neural AS events, and AS is more prevalent neural tissues than in other tissues[13,52]. Also, many of our events occur in vesicle- and cytoskeleton-associated genes, and these gene types frequently undergo AS[14,53]. Both indicate our library captures some common characteristics of AS. On the other hand, many of our events are neural microexons. Whilst microexons only minimally change primary structure, which may result in less dramatic effects than larger changes, they are enriched in structured regions and are well-conserved[14] (which is associated with biological function)[13,54]. While we can only speculate to what extent these factors cancel each other out, a recent study on AS-derived transcription factor isoforms reported comparable numbers to our results, which supports our estimate[25]. Here, the authors examined whether transcription factor isoforms differed in nuclear versus cytoplasmic localisation and condensate formation - both of which we would report as subcellular localisation differences - and found ~50% of isoforms differed in one or both aspects[25]. Notably, their approach to selecting isoforms for analysis differed substantially to ours, which further supports our estimate not arising through biased isoform selection.

Association between localisation and interactions

Altered protein subcellular localisation relies on altered intermolecular interactions with, for instance, proteins and lipids. Altered localisation can then subsequently result in more interaction differences. We identified a modest, positive correlation between classifier performance (using localisation features) and overall PPI differences based on LUMIER (**Figure 2E**), which supported this notion, and speculate the correlation may be stronger if more interactions are profiled. It also suggested the potential complementarity of imaging and interaction profiling in isoform characterisation, where different data types can contextualise each other. This was recently demonstrated in the characterisation of transcription factor isoforms where isoforms that maintain PPIs with self or co-factors but don't localise to the nucleus may serve as negative regulators[25].

Considering most of the examined AS events differed between neural and non-neural contexts, we were intrigued that nearly 40% showed localisation differences. Here, AS supposedly regulates interactions with universal interactors (i.e., found in HeLa) and they are the ones influencing localisation, as seen for NUMB and OCRL where AS affects localisation by modulating interactions with phospholipids[34] and clathrin[33], respectively. This doesn't mean cell type-specific interactors are not affected, however, and it remains unclear whether AS generally changes localisation through altered interactions with ubiquitous or cell type-specific interactors, as both supposedly occur. Relatedly, our SH3GLB1 isoform pair was one of two pairs with previously reported localisation differences, that we didn't identify as different, which may be due to cell-type specific interactors driving the localisation changes as the difference was reported in neurons[35]. While HeLa or other exogenous models won't detect such isoform behaviours, the relatively high proportion of differentially localising isoforms suggests initial profiling in exogenous models is feasible and can likely guide characterisation in more challenging models that more closely resemble the endogenous setting.

Assessing how alternative splicing-based changes to proteins affects subcellular localisation

The combination of high-content imaging with classifier-based analysis could highlight the main features describing isoform differences alongside identifying isoforms as different (**Figure 3, Supp. Figures 7,8,10**). We found this resulted in features that were robustly different between isoforms even though we analysed populations of cells with varying expression levels. It also facilitated selection of representative cell images, which can otherwise be challenging for heterogeneous populations. As these are common issues, our approach is likely broadly useful for analysing cell imaging data.

More excitingly, features describing isoform difference can potentially be used to unravel whether different types of AS-enabled protein changes (e.g. microexons in BAR domains) alter localisation in specific ways, by analysing isoform distinguishing features across isoform pairs. Whereas our dataset was too small for such analysis (**Supp. Figure 9**) - most of our isoform distinguishing features were unique - the general principle was recently demonstrated. Here, machine-learning models were trained on Human Protein Atlas (HPA) data to predict the importance of specific amino acid sequence features for protein subcellular localisation in different cell line settings[55]. Notably, because HPA relies on antibody-based detection of proteins[56] and generally won't have isoform-level information, the model's applicability to protein isoforms is unclear, but the concept of unravelling such localisation grammar for protein isoforms is exciting, although will require larger datasets than ours.

SNX2's microexon affects protein behaviour largely independent of composition

From the imaged isoform pairs, we selected the SNX2 isoforms for further examination as a case study of how imaging-based analysis can contribute to more detailed isoform characterisation, and how AS may affect protein localisation. The SNX2 isoforms differed by 4 amino acid residues, EIEE, that inserts into the α3 helix of SNX2's BAR domain. The insertion resulted in a more diffuse localisation phenotype, supposedly due to reduced association with early endosomes, which SNX2 has been described to localise to[44], and the abrogation of swollen endosomes at high expression levels[44]. Interestingly, in the closely related SNX1, the specific region of the BAR domain that contains the microexon in SNX2 is involved in both lipid binding[46] and formation of higher order SNX-BAR oligomers that assemble as protein lattices on lipid membranes[57,58], and disruption of either can cause reduced vesicle association[46,57].

The localisation effects were largely independent of the residues of the exon itself except I455, which we suggest stabilises the BAR domain. We speculate the effect may stem from the residue insertion shifting flanking residues, including R451, relative to other residues. How this manifests as localisation changes is unclear, but if residues on either side of the microexon in α3 form an interaction interface or both residues in α2 and α3 form interaction interfaces, the residue insertion could disrupt these.

We profiled PPIs alongside localisation and found both SNX2 isoforms interacted with other SNX-BAR proteins and supposedly formed functional ESCPE-1 (heterodimer of SNX1 or 2 with SNX5 or 6) complexes, as known ESCPE-1 cargo proteins interacted with both isoforms. Interestingly, SNX2 +EX and three SNX2 mutants (including one -EX mutant) resembling +EX during imaging, interacted more strongly with INSR than the SNX2 -EX exon, suggesting the mechanism by which the exon alters localisation also affects PPIs. Notably, because LUMIER relies on over-expression, the detected PPI differences could be associated with different aggregation propensities between SNX2 constructs, but it seems unlikely since the PPI differences were not consistent across all tested cargos.

Whereas the microexon's ability to impact vesicular localisation is unsurprising based on its localisation in SNX2, the differential interaction with INSR is somewhat puzzling as cargo canonically is bound by the cargo binding SNX-BAR (SNX5 or 6) in ESCPE-1, while SNX2 acts as the lipid binding subunit[39]. Lopez-Robles *et al.* suggested a connection between the lipid binding subunit's tip/α3 region and the cargo binding subunit's amphipathic helix and PX domain, which may enable the lipid binding subunit to influence cargo binding[57], in a manner that may be modulated by the microexon. The same study also suggested cooperativity between lipid and cargo binding[57], which creates a potential link between cargo PPIs and localisation, although surprisingly, it is the isoform with the well-defined vesicular localisation that interacts less strongly with INSR. Our observations are only based on a single cargo, however, and it is possible SNX2 -EX interacts more strongly with other cargos. It is also noteworthy that it is INSR

that the isoforms differentially interact with. Unlike IGF1R, INSR is not a canonical ESCPE-1 cargo[41,42,59] and its association with ESCPE-1 may be indirectly mediated through INSR-IGF1R heterodimers[42], which could suggest the SNX2 isoforms either may impact cargo selection or work independently of ESCPE-1. Either way, the microexon may potentially differentially impact cargo selection and recycling.

Microexons as modulators of secondary structure elements

Interestingly, the SNX2 microexon's effect on the helical makeup (preserving and extending the helical structure but disrupting hypothetical interaction interfaces, potentially through displacing charged residues), may represent a common way microexons affect helices. Our analysis also indicates another common mechanism may be microexons disrupting helices, typically when overlapping helix ends, resulting in extended loop regions. We are not familiar with studies detailing how these types of microexon events affect protein function, but similar behaviours have been documented in case studies of microexons in β-strands.

Here, extended loop regions can directly affect ligand interaction sites[60,61] - in some cases through displacing charged residues relative to binding sites[60] - but can also alter inter-molecular interactions through altered flexibility of loops connecting the regions that make up the interaction face[62]. Meanwhile, the few reports on microexons extending β-strands typically describe complex, indirect ways whereby interactions are modulated, including changes to the likelihood of autoinhibitory conformations[63] or protein rigidity, which sterically limits which PPIs other parts of the protein can form[64]. Whereas these studies may indicate an overall pattern with microexons in loops more likely to directly affect inter-molecular interactions while microexons within structured elements indirectly affect interactions – which potentially extends to α-helices - the effect is heavily dependent on the protein, demonstrating a remarkable complexity of how microexons affect protein behaviours. More studies of microexons in various secondary structure contexts are therefore needed to identify more general principles of how they shape protein behaviour.

Collectively, we demonstrate the use of high-content imaging in the direct characterisation of protein isoforms, exemplify how it can be an initial step in isoform characterisation, and propose AS often alters subcellular localisation of proteins.

Limitations of this study

We introduced protein isoforms via transient plasmid transfections, which results in variations in expression levels between transfected cells with the introduced protein often being over-expressed compared to endogenous levels[65], while the endogenous protein is still present. This was the easiest way of introducing protein isoforms that could be selectively visualised but may result in non-physiological behaviours. We verified, however, that classifier performances didn't correlate with differences in

expression levels between isoforms, that the top isoform distinguishing features for the top performing classifiers differed consistently across expression levels, and the SNX2 isoforms showcased that the varied expression levels enable some degree of dose-response analysis. Nonetheless, a more sophisticated method could perhaps use the recently described CRAFT trans-splicing system[66].

We used HeLa cells as the model system as they are easy to manipulate, flat, and relatively large, which makes them well-suited for imaging. However, HeLa are of non-neural origin and therefore shouldn't express neural-specific interactors that may be required to dissect isoform-specific localisation behaviours.

Our study was of relatively small-scale. This limited our ability to identify structural features that were associated with specific types of localisation behaviour and the estimate of how frequently AS alters protein localisation, although our estimate is supported by Lambourne *et al.*[25]. While large-scale examination can be achieved via arrayed screening, pooled screening relying on introducing a pool of protein isoforms and then identifying isoform identity via, for instance, *in situ* sequencing of bar codes[67–70] is likely more feasible for genome-wide studies. A pooled approach would presumably also greatly reduce well-specific effects which reduce sensitivity, although comes at the cost of being technically challenging.

We used a feedforward neural network (FFN) for classification, but note many other options are available and may offer better performance. Moreover, we only used features that described cell shape or the mCherry-isoform signal but additional signals from landmark stains can be included to improve classification and interpretability. This may also enable 'guilt by association' analysis where isoform effects can be interpreted based on how landmark stains are impacted, as are often used in drug or genetic perturbation-based high-content imaging experiments[67,70,71]. In our study, however, the inclusion of trans-Golgi network features didn't improve classification, perhaps due to well-specific effects and/or the increased number of features, and we therefore didn't use them for the final classifiers.

Acknowledgements

We would like to thank Benjamin J. Blencowe and Jonathan D. Ellis for sharing the protein isoform library with us, the Katarina Gaus Light Microscopy Facility at University of New South Wales for helpful advice on image acquisition and analysis, the Bioanalytical Mass Spectrometry Facility at University of New South Wales for performing mass spectrometry and Jonathan Roth for assisting with the analysis, and the members of the Weatheritt lab for discussions on the project and the manuscript.

S.B. is the recipient of an Australian Research Council Discovery Early Career Award (DE230100271).

Data availability

Datasets containing cell measurements from all cells transfected with the protein isoform library, classifier performance metrics across all iterations, and cell measurements from all cells transfected with SNX2 mutant constructs will be made available upon publication via Figshare.

Materials and methods

See tables at the end of this section for reagent (**Table 3**), plasmid (**Table 4**), and oligonucleotide details (**Table 5**). Antibody and dye dilutions are included in the text when their use is described. Figures were created using Biorender.com.

Molecular Cloning

Gateway recombination

Most subcloning of protein isoform open reading frames (ORFs) was carried out using the Gateway system (Thermo-Scientific), using BP and LR clonase II enzyme mixes (donor: pDONR303, destination: mCherry-Nterminus, Cterminus-mCherry). Gateway cloning was similarly used to generate the SNX2 isoform lentiviral plasmids (Destination: pLX303[72]), and the bait and prey plasmids used for LUMIER (destinations: 3xFLAG-Nterminus or Cterminus-3xFLAG, RenillaLuciferase-Nterminus) gifted by Benjamin J. Blencowe's group. Plasmids gifted by researchers through Addgene are listed in **Table 4**.

The Gateway BP reactions were set up according to manufacturer's protocol. If the insert-of-interest was not in a plasmid flanked by attB1/attB2 sites, these sites were introduced by PCR with primers tailed by the attB1/attB2 recombination sites. The PCR product was then used for BP reaction. If the insert-of-interest was already in a plasmid flanked by attB1/B2 sites, the plasmid was linearised by restriction enzyme digestion prior to the BP reaction. The BP reactions were generally incubated 2-3h at +25°C before proteinase K-based inactivation, and 30% of the reaction was then used for bacterial transformation. The Gateway LR reactions were set up in a reaction volume half of that described by the manufacturer with every component halved, including DNA amounts, but where otherwise performed according to the manufacturer's protocol. The LR reactions were generally incubated 1-2h at +25°C before proteinase K-based inactivation, and 50% of the reaction was then used for bacterial transformation. If either BP or LR reactions weren't used immediately for bacterial transformation, the reaction was frozen at -20°C and used within a week.

Site-directed mutagenesis

Protein isoform ORFs not obtained from the Blencowe lab were generated by site-directed mutagenesis. Similarly, site-directed mutagenesis was used to generate the SNX2 mutants, and to introduce stop codons in a few ORFs. All site-directed mutagenesis was done using the Q5 Site-Directed mutagenesis Kit (New England Biolabs) according to the manufacturer's protocol.

All plasmids (generated via Gateway recombination or site-directed mutagenesis) were verified via Sanger sequencing at the Garvan Molecular Genetics facility.

Bacterial transformation and plasmid purification

Home-made DH5α *E. coli* was used for most transformation reactions. Briefly, 50µL cells in a 2mL Eppendorf tubes were used per transformation. The cells were mixed with 1-10ng plasmid DNA, 30% of an BP reaction, 50% of an LR reaction, or 50% of a site-directed mutagenesis reaction, and then incubated on ice for 30min. The cells were then heat shocked at +42°C for 1min and then incubated on ice for 2min. 900µL room temperature LB media was then added, and the tubes were incubated in a shaking incubator (+37C, 200-250rpm) for 1h. The bacterial suspension was then plated on LB agar plates with the appropriate antibiotic for the plasmid in question. Ampicillin was used at 100µg/mL, kanamycin was used at 50µg/mL, and spectinomycin was used at 75µg/mL. Transformations of constructs that transformed less efficiently (for instance, BP reactions of large inserts), was done using commercial DH5α *E. coli* (New England Biolabs, C2987H) according to manufacturer's protocol, using the same DNA input amounts as listed for Home-made DH5α *E. coli*. Transformation of plasmids used for lentivirus production was done using commercial OneShot Stbl3 *E. coli* (Thermo-Scientific, C737303), according to manufacturer's protocol. The Stbl3 cells were incubated at +30°C rather than +37°C.

Home-made DH5α *E. coli* was from commercial DH5α (New England Biolabs, C2987H) as previously described[73].

Plasmids were purified from liquid cultures of bacterial transformants expressing the plasmid-of-interest, using various plasmid purification kits (Promega's Wizard® Plus SV Miniprep kit (A1330), Macherey-Nagel's NucleoBond Xtra Midi kit (740410), Promega's PureYield kit (A2392)) according to manufacturer's instructions Liquid bacterial cultures were grown at +37°C and 200-250rpm for non-lentiviral plasmids, and at +30°C and 200-250rpm for lentiviral plasmids.

Routine cell culture

HeLa cells (both wild type and stable and knockout lines) and HEK293T wild type cells were grown in non-coated TC-treated polystyrene plastic ware (Corning) in DMEM (High glucose, sodium pyruvate, Thermo-Scientific#11995065) supplemented with 10% FBS, with or without Pen/Strep (penicillin: 100 units/mL, streptomycin: 100µg/mL), at +37C, 5% CO_2 in a humidified cell incubator. Both cell lines were passed by washing the cells

2x in heated 1xPBS (137mM NaCl, 10mM Na_2HPO_4, 1.8mM KH_2PO_4, 2.7mM KCl) then treating the cells with trypsin-EDTA 0.25% for 4min at +37°C to detach them. The cells were then resuspended in supplemented DMEM, spun at 200 x g for 4min at room temperature, and the supernatant was discarded. The cell pellet was resuspended in fresh supplemented DMEM, counted using a Countess II Automated Cell Counter, diluted further in supplemented DMEM to the desired cell density and used for downstream assays or continued culture.

HeLa and HEK293T wild type cells were gifted by Ruth Lyons (Garvan Tissue Culture Facility manager).

Cell line generation

CRISPR-Cas9-based gene knockout

First, we cloned gRNAs targeting SNX2, either previously described[74] or predicted using the CRISPRscan online tool[75], into the pSpCas9(BB)-2A-GFP (PX458)[76] (Addgene #48138) plasmid under a U6 promoter. The gRNAs generally targeted exon junctions of exon 2 and were predicted to cause gene knockout by disrupting exon 2 usage The gRNA introduction into the plasmids was done as described by Ran *et. al.* for synthetic oligonucleotides[76]. Briefly, the gRNA oligo and the complementary sequence were ordered separately and then annealed *in vitro*. Both the gRNA and complementary strand were ordered with overhanging 5'-ends that are complementary to the BbsI-generated sticky ends in the pSpCas9(BB)-2A-GFP (PX458) plasmid. The annealed gRNA/complementary strand could therefore be inserted via into the plasmid via DNA ligation.

To generate the SNX2 KO HeLa cells, 3µg gRNA-containing pSpCas9(BB)-2A-GFP (PX458) plasmid was transiently transfected into HeLa cells in p60 dishes using ViaFect transfection reagent (4:1 reagent/DNA ratio), according to the manufacturer's protocol. Notably, the two plasmids with different SNX2-targeting gRNAs were transfected into different p60 dishes to generate independent SNX2 KO HeLa cell lines. The cells were incubated for 48h after transfection at normal culture condition. The cells were then detached and spun as for normal cell passage, then resuspended in 1xPBS w. 2% FBS and Pen/Strep and passed through a 35µm nylon mesh to remove cell clumps. The cell suspension was then fluorescence-activated cell sorted (FACS'ed) on a BD FACS Aria III (nozzle diameter: 100µm) to isolate EGFP-positive cells positive which were sorted into 96-well plates with one cell being seeded per well. The wells contained 50µL fresh supplemented DMEM with Pen/Strep and 50µL conditioned supplemented DMEM. The single clones were then allowed to grow into individual clonal lines. Clonal lines were subsequently tested for successful gene knockout by western blotting.

Conditioned supplemented DMEM was generated by aspirating the media from HeLa cells after the cells had been incubated with the media for 24-48h. The isolated media was then spun at 3220 g x 10min at +4°C to pellet potential cell debris, and the supernatant was filtered through a 0.22µm filter. The filtered supernatant was used as conditioned supplemented DMEM.

Stable cell lines

We generated cell lines that stably expressed the gene-of-interest via lentiviral transduction. We generated the lentiviruses by co-transfecting 60-80% confluent HEK293T in T-25 flasks with 3 plasmids: 4µg psPAX2 (packing vector, Addgene (Plasmid #12260)), 2µg pCAG-VSVG (envelope vector, Addgene (Plasmid #35616)), and 4µg pLX303 lentiviral plasmid containing the ORF for one of the two SNX2 isoforms. The transfection was performed using Lipofectamine 3000 reagent (Thermo-Scientific) (ratios: 1.5:1 reagent/DNA, 2:1 P3000/DNA), according to manufacturer's protocol. 24h after the transfection, the media was changed to supplemented DMEM. After another 24h, the media was isolated from the HEK293T and spun at 2500 x g for 10min at +4°C to pellet cell debris. The supernatant was then filtered through at 0.45µm filter, and this filtered supernatant was used for viral transductions (virus-containing supplemented DMEM not used immediately was frozen at -70°C for later use).

To transduce the target cells, the virus-containing supplemented DMEM was added to the media of HeLa cells (either HeLa wild type or SNX2 knockout clones) in 6-well plates. The cells were incubated 48h with the virus and the media was then removed. The cells were washed 10x in 1xPBS, and fresh supplemented DMEM was then added, and the cells were removed from the PC2 facility. The cells were subsequently passed (after 30min of rest in a cell incubator) and seeded into new 6-well plate wells in supplemented DMEM with 2µg/mL blasticidin to select for transduced cells. The cells were treated for 4 days with blasticidin at this concentration (media changed every second day) and passed as normal when they reached confluency. After 4 days, the blasticidin concentration was increased to 10µg/mL to make sure surviving cells were transduced. This concentration was used for 6 days after which the drug was removed, and the cells were cultured as per normal to grow cells for freezing. Upon reviving frozen stocks of the stable cells, they were cultured at 3µg/mL blasticidin when not used for experiments to keep selecting for cells expressing the transgene. The drug was removed when cells were used for AP-MS. Expression of the transgenic SNX2 (either neural or non-neural isoform) was verified by western blotting and end-point RT-PCR.

Generation of stable cells by viral transduction was done in a Garvan PC2 laboratory designated for viral work (the point at which the cells were removed from the facility is specified). This was done in accordance with Garvan's PC2 guidelines.

Cell preparation for imaging

Protein isoform library

HeLa wild type cells were detached, counted, and diluted in supplemented DMEM (as above) to make a cell suspension with a density of 90k/mL. 500µL of the cell suspension was then added into each well in a glass-bottom 24-well plate (Cellvis) (45k cells/well). Immediately after adding the cell suspension to the wells, the cells were reverse transfected by dripping in 50µL transfection mix consisting of 400ng plasmid encoding mCherry-tagged construct mixed with ViaFect (4:1 reagent/DNA ratio) in OPTI-MEM as per manufacturer's protocol. Each well was transfected with a different construct (except the negative controls). One well was left untransfected, and instead 50µL OPTI-MEM was added. Three experimental replicates for a given protein isoform batch were set up sequentially on the same day with cell suspensions and transfection mixes being made separately for each replicate.

After seeding and reverse transfection, the cells were incubated for 24h in the cell incubator. After the incubation, all wells were washed 2x in 500µL heated 1xPBS and then fixed with 200µL 4% paraformaldehyde for 10min at room temperature. The wells were subsequently washed 3x in 500µL 1xPBS, and 1mL 1xPBS was then added to each well. The well plate was then wrapped in parafilm and kept at +4°C until used for immunofluorescence staining (within 7 days of fixation).

SNX2 mutants

HeLa wild type cells were detached, counted, and diluted in supplemented DMEM (as above) to make a cell suspension with a density of 100k/mL. 450µL of the cell suspension was then added into each well in a poly-D-lysine coated glass-bottom 24-well plate (Cellvis) (45k cells/well). Cells were incubated for 16h in a cell incubator and were then forward transfected with plasmids encoding mCherry-SNX2 constructs (both endogenous isoforms and mutants) or mCherry itself by dripping in 50µL transfection mix consisting of 375ng plasmid DNA mixed with jetPRIME (2:1 reagent/DNA ratio) in jetPRIME buffer as per manufacturer's protocol. Each well was transfected with a different construct (except 3 wells that were all transfected with mCherry-SNX2 -EX). One well was left untransfected, and instead 50µL jetPRIME buffer was added. Three experimental replicates were set up sequentially on the same day with cell suspensions and transfection mixes being made separately for each replicate.

The cells were fixed 24h after transfection by removing the media and adding 200µL 4% paraformaldehyde for 10min at room temperature. The fixed sample was washed and stored as for the protein isoform library samples.

Immunofluorescence staining of protein isoform pairs and SNX2 mutants

The well plates with fixed, transfected cells were stained for imaging within 7 days of fixation. All replicates for a given protein isoform batch or SNX2 mutants were stained at the same time. The staining was done as follows: (note, between each step, the wells were washed 2-4x with 500µL 1xPBS per wash. During incubations, the well plates were covered with alu foil). 1) Permeabilisation. Cells were permeabilised in 200µL 1xPBS with 0.2% Triton-X100 for 12min at room temperature. 2) Blocking. The cells were then blocked in 200µL 1xPBS with 3% BSA for 60min at room temperature. 3) Primary antibody staining. The cells were then stained in 200µL 1xPBS with 3% BSA and mouse anti-Golgin97 (0.4µg/mL) for 120min at +37°C. 4) Secondary antibody staining. The cells were then stained in 200µL 1x PBS with 3% BSA and goat anti-mouse-Alexa488 (3µg/mL) for 60min at room temperature. 5) DAPI staining. The cells were then stained in 200µL 1xPBS with DAPI (1µg/mL) for 15min at room temperature. 6) NHS-ester staining. The cells were then stained in 200µL 1xPBS with NHS ester-Alexa647 (0.125µg/mL) for 30min at room temperature (note, the NHS ester was made up just before being added to the cells). 7) Final wash. The cells were washed 4x in 500µL 1x PBS and a final 1mL of 1xPBS was added to each well to prevent the samples from drying out. The stained well plates were wrapped in parafilm and kept at +4°C until used for imaging (typically less than a week).

Imaging

Imaging of protein isoform library

Fixed and immunofluorescence-stained cells were imaged using a PerkinElmer Operetta CLS system using its spinning disk confocal setting. The cells were imaged with a 40x water immersion 1.1NA objective, resulting in fields-of-view of 0.32mm x 0.32mm (2160 x 2160 pixels), with a pixel size of 0.148µm x 0.148µm. We acquired ~50 fields-of-view (FOV) per sample per experimental repeat with 14 slices being acquired per FOV with 0.5µm between each slice (z-dimension). Three biological repeats were generated for each isoform batch.

Four channels (referred to as DAPI, Golgin97-Alexa488, mCherry, and NHS Ester-Alexa647) were acquired for each slice using the following excitation, emission, and exposure settings (**Table 1**). Note, the microscope uses LEDs with bandpass filters (BPF) rather than lasers for excitation.

Table 1. Excitation, emission, and exposure for the four acquired channels

Channel	LED wavelength	Excitation BPF	Emission BPF	Power	Exposure Time
DAPI	365 nm	355-385	430-500	130 mW	100 ms
Golgin97-Alexa488	475 nm	460-490	500-550	110 mW	500 ms
mCherry	550 nm	530-560	570-650	170 mW	2500 ms

Channel	LED wavelength	Excitation BPF	Emission BPF	Power	Exposure Time
NHS Ester-Alexa647	630 nm	615-645	655-760	120 mW	700 ms

Imaging SNX2 mutants

The SNX2 mutant samples were imaged similarly to the protein isoform library except only 8 slices were acquired per FOV with a step size of 1.0µm. 41 FOVs were acquired per sample per repeat. Three biological repeats were generated.

Image processing, cell measurements, cell filtering, and transfection status

Protein isoform screen and SNX2 mutants

We performed analysis of single slices from the acquired confocal stacks, using the slice corresponding to where the cell adheres to the glass surface.

The acquired confocal stacks were processed as outlined in **Supp. Figure 1**. This was done by: **1)** identifying the focus slice in each confocal stack using ImageJ v1.54f[77] and the plugin Select focus slice v1.0 (unpublished, from: https://sites.google.com/site/qingzongtseng/find-focus, downloaded June 2022) with a user-defined variance threshold on the NHS-Alexa647 channel. The same threshold was applied to all images for a given repeat. The selected focus slice was then used for all channels for the given FOV. **2)** We illumination corrected all channels in ImageJ using the BaSiC v1.0 plugin[78] where all focus images from the experimental repeat for the given channel were used to calculate the correction profiles (both flat and dark field) using default settings. The correction profiles were then used to illumination correct the images. **3)** We segmented individual cells using CellPose 2.0[79], using the illumination corrected NHS-Alexa647 and DAPI as input to define the cell body, and nucleus, respectively. We used a custom segmentation model for segmentation generated via transfer learning using the default 'cyto' model with default settings as the starting model. Cell diameter was set to 150 pixels. **4)** We then trained a custom classifier to label cells as either interphase or mitotic using Ilastik v1.4[80], based on cell morphology and NHS staining, to focus analysis on interphase cells. **5)** We then loaded the illumination corrected channel images and cell segmentation maps into CellProfiler v4.1[81] and first subtracted the median background (defined as the part of the image not covered by cells) from all channel images. **6)** We then measured features describing cell morphology, the mCherry channel, and the Golgin97 channel, as well as correlation between Golgin97 and mCherry. **7)** The dataset was then filtered to remove mitotic cells, cells on image borders, non-mononucleated cells, cells with touching nuclei (typically a result of mis-segmentation of multinucleated cell into two cells), and cells where >85% of their cell border contacts other cells as this typically indicated tight packing of cells, which impaired segmentation accuracy. **8)** We then assigned transfection status to the cells based on their cell body mean mCherry intensity and

75% quantile mCherry intensity, where both values had to surpass user-defined thresholds based on the non-transfected cell sample. **9)** For transfected cells, we removed the brightest cells from the transfected populations on a pairwise-basis that reflected which samples would be compared. We defined an outlier threshold for each sample as: $Outlier = mCherry_{MeanIntensity} > Population\ mCherry_{75\%QuantileIntensity} + 1.5 \cdot Population\ mCherry_{InterQuantileRange}$

We then applied the lower threshold for a given isoform to both samples in that pair. Notably, this filtering step generally removed very few cells and so only removed the brightest cells.

Step 9) was not performed for the SNX2 mutant experiment as we didn't perform pairwise comparisons. Instead, we removed the outliers for each population using the outlier threshold calculated (as above) for the same population.

The non-transfected and transfected cells made up the cells used for downstream analysis.

Representative images

Representative images were selected based on our analyses of the overall population behaviour of the given mCherry constructs. These images were prepared in ImageJ v 1.54 using the illumination corrected channel images. All images that are directly compared have the same contrast settings. Mean mCherry intensities for representative cells were determined by manually segmenting the cells and measuring their mean mCherry intensity, then subtracting the mean background mCherry intensity measured adjacent to each representative cell. The pixel values were then scaled to reflect the input image being 16-bit by: $scaled\ mean = \frac{mean_{pixel}}{65535}$, which directly correspond to the plotted intensity values in the population XY plots.

Analysis of cell measurements

Classifier-based analysis of protein isoform pairs

We used the single cell feature measurements to compare localisation of mCherry constructs (such as two isoforms from the same gene) in a pairwise-manner using a classifier-based approach as outlined in **Supp. Figure 2**. We performed separate analyses for transfected and non-transfected cells from the same wells. These made up the 'isoform' (ISO) and 'internal normalisation' (IN) groups, respectively.

The analysis consisted of 1) generating an appropriate dataset through sampling, and 2) analysing the dataset through a classifier-based approach. For 1), we pooled transfected cells expressing the two isoforms to be compared across replicates, then

down-sampled the cells in the more abundant isoform group to match the number of cells of the less abundant isoform in a replicate-aware manner. Accordingly, replicates could contribute different numbers of cells compared to each other, but the cells contributed by a replicate was always 50% isoform A and 50% isoform B. Cell numbers used for classifier assessment after down-sampling and pooling was ≥170, although most pairs had substantially more cells (median cell number = 967). Cell numbers for each comparison are listed in **Supp. Table 2**. The dataset from 1) was then used for classifier-based analysis 2), where the classifier was trained to assign the correct isoform identity (corresponding to well identity) based on the measured features.

The classifier-based analysis was performed in R (4.3.3) using Caret v6.0-94[82] to control classifier training, testing, and feature assessment, with the classifier itself being a feedforward neural network, from nnet[83] v7.3-19. The initial input features for the classifier were all features describing the mCherry signal (although intensity features were normalised to mean mCherry intensity) and cell morphology. Zero variance features were then removed, and the remaining features (around 200 for all classifiers) were centred, then scaled, before being used for classifier assessment. For the assessment, the classifier was trained with 10-fold cross-validation, and we used the average ROC-AUC (Receiver Operating Characteristic (ROC) Area Under the Curve (AUC)) at the parameters resulting in the best performance, across the test folds as the metric describing the degree to which isoforms were different from one another. The 1) sampling and 2) classifier assessment was repeated 50 times.

To normalise the ROC-AUC metric for potential well-specific effects driving classification, we performed the same analysis on non-transfected cells from the same wells as the two mCherry constructs being compared and used these for internal normalisation. Notably, the internal normalisation cell population was sampled to best mirror the isoform pair population in question, and the non-transfected cells were therefore sampled to match the proportion of cells contributed by each replicate for the transfected cells, while also contributing the same number of cells from the two wells.

Normalised ROC-AUCs were calculated using the ROC-AUC from the 'Isoform' (ISO) and 'Internal normalisation' (IN) classifiers as outlined in **Table 2** (same as **Figure 1E**):

Table 2. Definition and interpretation of normalised ROC-AUC performance metric

When	Definition	Score	Interpretation
$ROC\ AUC_{ISO} \geq ROC\ AUC_{IN}$	$\frac{ROC\ AUC_{ISO} - ROC\ AUC_{IN}}{1 - ROC\ AUC_{IN}}$	$ROC\ AUC_{normalised} > 0$	Isoform differences detectable after accounting for well-specific effects
$ROC\ AUC_{ISO} < ROC\ AUC_{IN}$	$\frac{ROC\ AUC_{ISO} - ROC\ AUC_{IN}}{1 - ROC\ AUC_{ISO}}$	$ROC\ AUC_{normalised} \leq 0$	Isoform differences NOT detectable after accounting for well-specific effects

The normalised ROC-AUC ranges from -1 and 1 and is symmetric around 0. Scores above 0 indicates mCherry construct differences can be detected after accounting for well-specific effects.

We statistically tested whether the normalised ROC-AUC for a given isoform comparison was significantly greater than zero using a one-sided bootstrap approach in R using boot v1.3-31[84]. For each pairwise comparison, we performed 2000 bootstrap iterations by resampling with replacement from the normalised ROC-AUC values, calculating bootstrap confidence intervals for the mean performance. If zero was excluded from the lower bound of the 99% one-sided confidence interval, the ISO classifier was determined to significantly outperform the IN classifier, indicating isoform differences. We corrected for multiple testing using the Benjamini-Hochberg method to control the false discovery rate.

Out-of-sample testing

We assessed out-of-sample performance of classifiers trained on controls and select isoform pairs by using cells from one batch (batch A or B) for the training set, and cells from the other batch for testing. For the training set, we sampled and trained the classifier as described in 'Classifier-based analysis of protein isoform pairs'. For the test set, we sampled the cells as described in 'Classifier-based analysis of protein isoform pairs', to have the same number of cells of the two groups and then predicted isoform identity for the cells using Caret predict(). The performance was assessed as accuracy. Notably, we sampled the test set to have equal classes to both mirror the training set and to make the accuracy comparable between conditions (as accuracy is sensitive to class imbalance). The accuracy can therefore range from 0.5 to 1.0. 20 iterations of classifier training and testing was done, and the performance values were the mean accuracy from each iteration. For each iteration, a classifier was trained on the training batch and was then used to predict isoform identities for the down-sampled test batch, where the test batch down-sampling and prediction was performed 50 times. The mean accuracy across the 50 times was used as the performance.

To examine the association between the accuracy of the out-of-sample tests and the normalised ROC-AUC approach, we performed linear regression, Accuracy ~ normalised ROC-AUC. We statistically tested if the accuracy and normalised ROC-AUC were correlated in R using a correlation test, cor(method = 'kendall').

Replicate sampling imbalance

We quantified replicate sampling imbalance as the maximum fraction of cells used in given isoform pair comparison that came from a single experimental repeat over the total number of cells used for the comparison (cells pooled across all repeats).

mCherry intensity imbalance

We quantified mCherry intensity (expression) imbalance between isoforms making up each pair by calculating the cell population mean of single cell mean mCherry intensities for each isoform within each pair on a replicate basis. We defined mean intensity differences between the isoforms as: $\Delta Intensity_{repeat} = |\log_2\left(\frac{Isoform\ +EXON_{intensity}}{Isoform\ -EXON_{intensity}}\right)|$, which was calculated for each repeat. The repeat-level intensity differences were subsequentially scaled to reflect the fraction of total cells within the given isoform pair that was contributed by the repeat. The sum of the scaled repeat-level intensity differences for all repeats was used as the expression imbalance metric. $Imbalance_{intensity} = \frac{cells_{rep1}}{cells_{total}} \cdot \Delta Intensity_{rep1} + \frac{cells_{rep2}}{cells_{total}} \cdot \Delta Intensity_{rep2} + \frac{cells_{rep3}}{cells_{total}} \cdot \Delta Intensity_{rep3}$

Identifying Isoform distinguishing features

Using the default classifier importance scores generated by Caret for the nnet classifiers, we first calculated the Feature contribution percentage (%), that specified how large a percentage of the sum of feature importance scores were contributed by a given feature:

$$Feature\ contribution\ (\%) = \frac{Importance\ Score_{Feature_i}}{\sum_{i=1}^{n_{features}} Importance\ Score_{Feature_i}} \cdot 100$$

(n features: number of features used for the given classifier, ~200 for all classifiers)

We did this for both the ISO classifier and associated IN classifier, which allowed us to calculate contribution ratios, specifying the relative contribution of a feature to the ISO classifier compared to the IN classifier:

$$Contribution\ ratio = \frac{Feature\ contribution\ (\%)_{ISO}}{Feature\ contribution\ (\%)_{IN}}$$

(ISO: isoform (Transfected = TRUE). IN: internal normalisation (Transfected = FALSE)).

We then defined "Isoform Distinguishing" features (features that were notably more important for ISO classifier compared to the associated IN classifier), as features with a contribution ratio and a contribution ratio multiplied with the feature contribution percentage for the ISO classifier above two user-defined threshold values.

$$Isoform\ Distinguishing\ feature = \ Contribution\ ratio > threshold_{ratio}$$

$$\&$$

$$Contribution\ ratio \cdot feature\ contriubtion\ (\%)_{ISO} > threshold_{ratio\ x\ contribution}$$

The user-defined threshold values were defined so the negative control had no Isoform Distinguishing features. The contribution ratio threshold was 2.6, and the ratio x contribution threshold was 1.5.

Principal component analysis of SNX2 mutants

We analysed the SNX2 mutant constructs by principal component analysis (PCA) in R. As input, we used the mean values of the SNX2 isoform distinguishing features (about 20), calculated across the entire cell population (pooled across the three repeats) for the given construct, with the lowest number of cells for a construct being n = 185, while most constructs had n > 300 (cell numbers for each construct listed in **Supp. Figure 12A**). The feature means were centred and scaled before performing PCA using prcomp(). The plots of PCA scores, biplots, and the scree plots were created using the factoextra[85] and tidyverse packages[86]. We performed this analysis separately for transfected and non-transfected cells.

Affinity purification mass spectrometry

We examined the protein interactomes of the SNX2 isoforms using affinity purification mass spectrometry (AP-MS) on HeLa wild type and HeLa SNX2 cell lines. The immunoprecipitation protocol was based on[87] with minor modifications. To isolate SNX2 and its interactors, we grew HeLa wild type and SNX2 rescue cells lines to 80-90% confluency in P150 dishes in supplemented DMEM with Pen/Strep. The cells were then washed 2x in ice cold 1xPBS, after which 1mL 1xPBS was added to each dish and the cells were detached by cell scraping. The detached cells were transferred to Eppendorf tubes and spun at 800 x g for 5min at +4°C to pellet the cells. The supernatant was then removed, and the cell mass was weighed and then snap frozen in liquid nitrogen. The cell pellets were stored at -70°C until the immunoprecipitation was continued (within a week of harvest). To continue the immunoprecipitation, the cell pellets were thawed on ice and then resuspended in immunoprecipitation buffer (20 mM HEPES pH 7.2, 100 mM K-acetate, 2.5 mM Mg-acetate, 0.1% Nonidet-P40 (NP40) (v/v), 1mM DTT, 10% glycerol (v/v)) at a ratio of 4µL/mg pellet by gentle pipetting. The cell suspensions were then frozen on dry ice and thawed by holding the tubes to promote cell lysis (freeze-thaw). The lysates were then transferred to a tube spinner and spun at 12rpm for 30min at +4°C. The cell lysates were then transferred to a 2mL Dounce homogenizer for mechanical shearing (25 strokes). Homogenised lysates were then transferred to Eppendorf tubes and spun at 17000 x g for 20min at +4°C to pellet nuclei. The supernatant was recovered and transferred to a new Eppendorf tube. 15µL of the isolated supernatant was transferred to a separate Eppendorf tube and served as a measure of the input fraction for western blotting (kept at -20°C until use). The remaining supernatant was mixed with 5µL anti-SNX2 antibody (corresponding to 1.25µg antibody) and incubated on a tube spinner at 7rpm overnight at +4°C. The next day, the antibody-lysate mix was mixed with 50µL immunoprecipitation buffer-

equilibrated Protein G Dynabeads or Protein A Dynabeads (Thermo-Scientific), and the bead lysate suspension was then incubated on a tube spinner at 7rpm for 2h at +4°C. After this incubation, the suspension was put on a magnetic rack to pull down the beads and the antigen-antibody complexes. The supernatant was removed and transferred to a separate tube that was frozen at -20°C (it served as the unbound fraction for western blotting). The beads were then resuspended in 1mL immunoprecipitation buffer and transferred to a new tube. The tubes were then placed on a magnet again and the supernatant was removed. The beads where then resuspended in 1mL wash buffer (20 mM HEPES pH 7.2, 10mM $MgCl_2$) by gentle pipetting, and placed on the magnet a final time and the supernatant was removed. The sample was then eluted from the beads by adding 30µL 4xLDS sample buffer diluted to 1x in 1xPBS with 20mM DTT. The sample was then incubated at +95°C for 5min, after which it was put back on magnet so the eluate could be isolated from the beads. 5µL of the eluate was kept as the immunoprecipitated fraction for western blotting and the remaining 25µL was used for mass spectrometry analysis. Eluted samples were kept at -20°C until use (within a week). Samples were subjected to tryptic digest (bottom-up) prior to mass spectrometry analysis on a Thermo-Scientific LTQ Orbitrap XL instrument. Downstream sample processing, including tryptic digest, and label-free quantification mass spectrometry analysis was performed by the Bioanalytical Mass Spectrometry Facility at University of New South Wales. The mass spectrometry data was analysed using SAINT[88] using default settings. The experiment was repeated three times.

LUMIER

Luminescence-based mammalian interactome mapping (LUMIER) was performed to examine binary protein-protein interactions as described[51], with slight modifications. Specifically, we doubled the input cells (by seeding HEK293T into 48-well plates (TC-treated, non-coated polystyrene) at 40k cells/well) and scaled the plasmid transfections accordingly but kept the rest of the LUMIER assay (from cell lysis and onwards) consistent with previous reports in terms of volumes[51]. This resulted in more concentrated samples and therefore better signal-to-noise.

LUMIER experiments were repeated 3-4 times for all examined interactions.

LUMIER data analysis

The LUMIER interaction scores, NLIR and NtFLIR, were calculated using luminescence measurements of the TOTALs and IP plate and IP ELISA as described[51]. NLIR describes luminescence signal in the IP plate divided by luminescence in the TOTAL plate. NLIR > 3 were considered positive interactions[51]. NtFLIR was calculated as the NLIR score divided by the absorbance measured in the ELISA against FLAG (which quantifies Prey amount) and served as an interaction score (NLIR) normalised to the amount of prey protein present.

For each experimental repeat, we used the mean NLIR and NtFLIR across the technical replicates. We calculated the interaction differences between SNX2 isoforms with a given prey protein as: $\log_2\left(\frac{NtFLIR_{SNX2+EXON}}{NtFLIR_{SNX2-EXON}}\right)$, which we calculated for each experimental repeat. The SNX2 -EX isoform was used as the reference for the SNX2 +EX isoform as well as for all SNX2 mutants, where interaction differences were calculated in the same way. We statistically tested if the SNX2 +EX isoform or SNX2 mutants interacted differently with prey proteins compared to SNX2 -EX using two-sided one-sample t-tests on the interaction difference score, where the interaction differences for the different SNX2 constructs were compared to zero (μ = 0). We corrected for multiple testing using the Benjamini-Hochberg method to control the false discovery rate.

Anti-FLAG plate coating

Lumitrac 96-well plates were coated with anti-FLAG antibody by incubating each well with 100µL 1x PBS with anti-FLAG antibody (10µg/mL) overnight at +4°C. The next day, the antibody solution was removed and 250µL LUMIER blocking buffer (1xPBS, 3% BSA (w/v), 5% sucrose (w/v), and 0.5% Tween 20 (v/v)) was added to each well and incubated at 60min at RT at gentle agitation on orbital shaker. The blocking buffer was then removed, and the plate was sealed and kept at +4°C until use.

Western blots

Western blots were performed using standard protocols. Briefly, all protein samples were run on Bolt Bis-Tris 8% gels (Invitrogen) using 10-30µg input protein per sample (determined through a Bradford assay according to manufacturer's protocol) or, for immunoprecipitated samples, a specific percentage of the isolated fractions for (1/150 input/unbound and 1/3 immunoprecipitated sample). After gel electrophoresis, the proteins were transferred to methanol-activated PVDF membranes (Immobilon). The membranes were then blocked with TBS-T (1x TBS, 0.1% Tween 20 (v/v)) and 6% skim milk powder (w/v) for 1h under constant, gentle agitation on orbital shaker. The membranes were then briefly washed in TBS-T, and the membranes were then transferred to small plastic bags and covered with primary antibody solution, consisting of TBS-T, 1% BSA (w/v), and the primary antibody (see below for concentrations). Note, blots were typically cut into pieces to probe different parts of the membrane for specific proteins (based on size) after blocking but prior to primary antibody incubation. The membranes were incubated with the primary antibody solution at +4°C overnight under constant agitation. The next day, the membranes were washed 3x for 10min at room temperature in TBS-T under constant gentle agitation. Then the membranes were incubated with HRP-conjugated secondary antibody solution, consisting of TBS-T, 1% skim milk powder, and secondary antibody-HRP raised against the host species of the primary antibody (see specifics below), for 1h at room temperature under constant agitation. This was also done in plastic bags. The membranes were then washed again

for 3x10min at room temperature with TBS-T under constant gentle agitation. The membranes were then developed with Clarity Western ECL Substrate (BioRad) according to the manufacturer's instructions and imaged on a ChemiDoc.

To quantify the amount of SNX2 protein, the intensity of all bands-of-interest were measured in ImageJ v1.54, after which the background was subtracted. The amount was then normalised to the amount of GAPDH (reference gene) from the same sample, measured in the same way SNX2.

SNX2 cell line verification blots

To verify knockout or transgene expression, cells from HeLa wild type, SNX2 KO, and SNX2 rescue cell lines were washed in 1xPBS then lysed in RIPA buffer (150mM NaCl, 1% Nonylphenol ethoxylated (v/v), 0.5% Sodium deoxycholate (w/v), 0.1% SDS (w/v), 17.1mM EDTA, 50mM Tris(HCl), pH 8) on ice for 5min. The lysates where then transferred to Eppendorf tubes, spun at 17000 x g for 10min at +4C, and the supernatant was isolated and used for western blots. Cell lysates were mixed with LDS sample buffer (Thermo-Scientific) and DTT for a final concentration of 1x, and 20mM, respectively, and examined via western blotting as described above. The primary antibodies were mouse anti-SNX2 used at 0.25µg/mL (1/1000), and mouse anti-GAPDH used at 0.04µg/mL (1/5000). The secondary antibody was goat anti-mouse-HRP, used at 0.16µg/mL (1/5000). The experiment was performed once.

SNX2 immunoprecipitation verification blots

The three isolated fractions (input, unbound, and immunoprecipitated) were diluted in 1x PBS; for the input fraction, and the unbound fraction this was 1/150, and for the immunoprecipitated fraction, this was 1/3. The diluted samples were then prepared in 4x LDS sample buffer and 20mM DTT and examined via western blotting as described above. The primary antibodies were mouse anti-SNX2 used at 0.25µg/mL (1/1000), and mouse anti-GAPDH used at 0.04µg/mL (1/5000). The secondary antibody was goat anti-mouse-HRP, used at 0.16µg/mL (1/5000). The experiment was performed once.

End-point RT-PCR

End-point reverse transcription polymerase chain reaction (RT-PCR) was carried out to examine inclusion of SNX2's microexon to verify correct isoform expression in HeLa SNX2 stable cell lines. We isolated RNA for the analysis from the HeLa SNX2 cell lines by washing adherent cells 2x in 1xPBS, then lysing them in TRIzol reagent (Thermo-Scientific). Total RNA was subsequently purified according to the manufacturer's protocol for TRIzol-based RNA extraction. The precipitated RNA was redissolved in RNase-free H_2O. The reverse transcription and subsequent PCR amplification was done as a single step using QIAGEN OneStep RT-PCR kit. The reactions were set up in 10µL total volume, as described[89]. Briefly, each 10µL reaction consisted of: 3.7µL RNase-free

H_2O, 2mL 5x QIAGEN OneStep RT-PCR Buffer, 1.2µL 5µM FORWARD PRIMER, 1.2µL 5µm REVERSE PRIMER, 0.4µL 10mM dNTP Mix, 0.5µL enzyme mix, 1µL RNA template (10-30ng/µL) in RNase-free H_2O. The reactions were run in a Thermocycler with the following steps. 1) Reverse transcription: 30min at +50C, 2) Initial PCR activation: 15min at +95C, 3) 3-step cycling (40 cycles): Denaturation, 1min at +94°C. Annealing, 1min at +54°C. Extension, 1min at +72°C. 4) Final extension: 10min at +72°C. The reactions were loaded on 2% agarose gels containing GelRed (Biotium), separated via gel electrophoresis, and imaged on a ChemiDoc to examine RT-PCR product sizes. Size differences corresponding to the size of the microexon indicated which isoform was expressed. The primer sequences for the primers used to examine SNX2's microexon are listed in **Table 5**. The experiment was performed once.

Correlation between normalised ROC-AUC and LUMIER interactions

We examined the correlation between normalised ROC-AUC (localisation readout) and the overall PPI differences as quantified by LUMIER for isoform pairs where ≤3 interactions had been tested across ≤2 experimental repeats and resulted in positive interactions with at least one isoform of the pair. The LUMIER data originated from several publications by the same authors[18,31] and can be found in **Supp. Table 5**.

We calculated overall PPI differences between two isoforms from the LUMIER data by calculating the fold-change in NLIR between isoforms for a given interactor for each repeat: $NLIR_{FC} = \left(\frac{NLIR_{Isoform+EXON}}{NLIR_{Isoform-EXON}}\right)$. Then averaging the $NLIR_{FC}$ for all experimental repeats and calculating the absolute log2 fold-change for the averaged value: $|\log_2(mean(NLIR_{FC})|$. This score was calculated for each interactor that had been examined for an isoform pair, and we used the mean of these interaction difference scores as a measure of the overall PPI difference for an isoform pair. For the localisation readout, we used the median normalised ROC-AUC. We statistically tested the correlation between the median normalised ROC-AUC and collective LUMIER interaction score in R using a correlation test, cor(method = 'kendall'). A p-value ≤ 0.05 was considered significant.

Protein modelling

SNX2 isoforms from human, rat, chicken, and zebrafish were modelled using AlphaFold2 v2.1 with the default input settings through the application that is set up via Google Colaboratory[90]. The isoform models were visualised in ChimeraX v1.8rc202405240417[91]. ChimeraX was also used to align, examine surface charge, and examine electrostatic interactions (only salt bridges) of the SNX2 region corresponding to 417 – 460 of human SNX2 (Uniprot ID: O60749) with default input settings.

Sequences for TEAD1 and ERC1 protein isoforms were extracted from UniProt database[92] (human TEAD, Uniprot ID: P28347-1. Human ERC1, Uniprot ID: Q8IUD2-1)

and amino acid sequences corresponding to the microexons-of-interest were added/removed manually. Protein structures were modelled using AlphaFold3 on the AlphaFold server[93] (https://alphafoldserver.com) with default settings. Models were downloaded and visualized using ChimeraX.

Conservation analysis of SNX2 microexon

Amino acid conservation analysis

All isoforms of SNX2 annotated on the UniProt server[92] were identified and isoforms containing the microexon and of canonical size (~500 amino acids) were extracted. Isoforms were further manually filtered to ensure no evolutionary clade was over-represented. Sequences were aligned using Clustal Omega (v1.2.4) using default parameters[94]. JalView was then used to visualize sequences[95]. Secondary structure was predicted using JPred secondary structure prediction (Lupas_14) and JNetPRED, buried residues were predicted using Jnet burial prediction of solvent accessibility, and conservation using JNETConf[95].

Inclusion pattern analysis of SNX2 microexon

We analysed the inclusion pattern of SNX2 microexon using the percent-spliced in (PSI) values in vastDB v11[45] for the four species where the exon was annotated: human (HsaEX0060835), rat (RnoEX0084548), chicken (GgaEX0007383), and zebrafish (DreEX0008261). We compared the PSI values across tissue samples that were shared between all or most organisms. For tissue types with multiple samples, we used the mean PSI across the sample replicates. We plotted the PSI values in R (4.3.3) using pheatmap v1.0.12[96] with column-based clustering where we for the clustering inferred missing values by using the column mean of the available samples.

Analysis of microexons in α-helices

We extracted microexons (nucleotide (nt) length <30 nt) from vastDB[45] v11 and filtered them to only include alternatively spliced microexons expressed in neural tissues, defined as microexons differentially included in neural tissues compared to non-neural tissues (using vast-tools compare), and with a percent spliced in (PSI) greater >15 in neurally annotated adult tissues in vastDB resource.

The localization of microexons within secondary protein structures was assessed using DSSP (database of secondary structure) assignments based on protein entities in the Protein Data Bank (PDB)[97]. Here, we defined α-helices as annotated by DSSP as H (α-helix) or G (3_{10}-helix), β-strands as B (isolated β-bridge) or E (extended β-strand), loops as S (bend) or T (hydrogen-bonded turn). Microexons were considered within a secondary structure element if any of its residues (or the immediately preceding or proceeding residue) was annotated as a secondary structure element by DSSP. The

relative position of microexon within α-helices was defined based on the minimum distance from microexon to the beginning or end of the helix as defined by DSSP.

We analysed the residue composition of microexons, and flanking residues compared to those of larger cassette exons (≥30nt), that were alternatively included in neural tissues (inclusion defined similarly to neural microexons). Only exon lengths divisible by 3 were included. Residues were grouped as follows: Aliphatic residues: Ile (I), Leu (L) or Val (V). Charged residues: Glu (E), Lys (K), Asp (D) or Arg (K). Helix breaking as Pro (P) or Gly (G). Helix forming: Leu (L), Ile (I), Phe (F), Glu (E), Tyr (Y), Trp (W), Met (M). Aromatic: Phe (F), Tyr (Y), Trp (W), His (H). Residue proportions within exons and in the flanking regions were defined by counts. The flanking regions were defined as residues 30 amino acids upstream and 30 amino acids downstream of the inserted exon. Enrichment was calculated as $\log_2\left(\frac{Freq.\ microexon}{Freq.\ longer\ exon}\right)$, and statistically examined using fisher-exact test. Residue analysis of microexon overall as well as microexon subsets (microexons in helices, microexons in secondary structure) was compared to longer neural cassette overall.

Statistics

Data was analysed and statistically tested as indicated in the sections relevant to said data. When multiple tests were performed, p-values were corrected for false discovery rate using the Benjamini–Hochberg Method. The exact number of experimental repeats are stated in the Figure Legends and/or Methods relating to the specific experiment. P-values ≤0.05 were considered significant for LUMIER and correlation tests, while P-values ≤0.01 were considered significant for the examination of normalised classifier performances.

Reagents, plasmids, and oligonucleotides

Table 3. List of used reagents

Reagents	Provider	Catalogue #
MOLECULAR CLONING		
Gateway™ LR Clonase™ II Enzyme mix	ThermoFisher	11791020
Gateway™ BP Clonase™ II Enzyme mix	ThermoFisher	11789020
Q5® Site-Directed Mutagenesis Kit	New England Biolabs	E0554S
NEB® 5-alpha Competent E. coli (High Efficiency)	New England Biolabs	C2987H
One Shot™ Stbl3™ Chemically Competent E. coli	ThermoFisher	C737303
Wizard® Plus SV Minipreps DNA Purification Systems	Promega	A1330
NucleoBond Xtra Midi kit for transfection-grade plasmid DNA	Macherey-Nagel	740410
PureYield™ Plasmid Maxiprep System	Promega	A2392
RT-PCR		
TRIzol reagent	ThermoFisher	15596026
Chloroform	SigmaAldrich	C2432
QIAGEN OneStep RT-PCR kit	QIAGEN	210212
GelRed® Nucleic Acid Gel Stain	Biotium	#41003
MAMMALIAN CELL CULTURE		
DMEM, high glucose, pyruvate	ThermoFisher	11995065

Fetal Bovine Serum (FBS)	ThermoFisher	
Penicillin-Streptomycin (10,000 U/mL)	ThermoFisher	15140122
Trypsin-EDTA (0.25%), phenol red	ThermoFisher	25200056
Blasticidin S HCl (10 mg/mL)	ThermoFisher	A1113903
Lipofectamine™ 3000 Transfection Reagent	ThermoFisher	L3000001
ViaFect™ Transfection Reagent	Promega	E4981
IMMUNOFLUORESCENCE		
Precision cover glasses thickness No. 1.5H (tol. ± 5 µm)	Marienfeld	0117580
24 Well glass bottom plate with high performance #1.5 cover glass	Cellvis	P24-1.5H-N
DAPI (4',6-diamidino-2-phenylindole, dihydrochloride)	ThermoFisher	D1306
Golgin-97 Monoclonal Antibody (CDF4), eBioscience™	ThermoFisher	14-9767-82
Alexa Fluor® 488 AffiniPure™ Fab Fragment Goat Anti-Mouse IgG (H+L)	Jackson ImmunoResearch	115-547-003
Alexa Fluor™ 647 NHS Ester (Succinimidyl Ester)	ThermoFisher	A20006
Phalloidin-iFluor 647 Reagent	Abcam	ab176759
Fluoromount-G™ Mounting Medium	ThermoFisher	00-4958-02
CELL LINE GENERATION AND IMMUNOPRECIPITATION		
BD Transduction Laboratories™ Purified Mouse Anti-SNX2	BD Biosciences	611308
GAPDH Antibody (0411): sc-47724	Santa Cruz Biotechnology	sc-47724
Goat anti Mouse IgG (H+L) Secondary Antibody, HRP	ThermoFisher	31430
DTT (dithiothreitol)	ThermoFisher	R0861
NuPAGE™ LDS Sample Buffer (4X)	ThermoFisher	NP0007
Bolt™ Bis-Tris Plus Mini Protein Gels, 8%, 1.0 mm, WedgeWell™ format	ThermoFisher	NW00085BOX
Immobilon®-P PVDF Membrane	SigmaAldrich	IPVH00010
Clarity Western ECL Substrate	BioRad	#1705061
Dynabeads™ Protein A for Immunoprecipitation	ThermoFisher	10001D
Dynabeads™ Protein G for Immunoprecipitation	ThermoFisher	10003D
Halt™ Protease and Phosphatase Inhibitor Single-Use Cocktail (100X)	ThermoFisher	78442
LUMIER		
Monoclonal ANTI-FLAG® M2 antibody produced in mouse	SigmaAldrich	F1804-50UG
Monoclonal ANTI-FLAG® M2-Peroxidase (HRP) antibody produced in mouse	SigmaAldrich	A8592
Renilla-Glo® Luciferase Assay System	Promega	E2710
SuperSignal™ ELISA Pico Chemiluminescent Substrate	ThermoFisher	37070
Microplate, 96 Well, Ps, F-Bottom, (Chimney Well), White, Lumitrac, High Binding, Sterile	Greiner	655074

Table 4. Used Addgene plasmids

Plasmid	Addgene catalogue #	Reference
pSpCas9(BB)-2A-GFP (PX458)	Addgene (Plasmid #48138)	[76]
pCAG-VSVG	Addgene (Plasmid #35616)	Unpublished
psPAX2	Addgene (Plasmid #12260)	Unpublished
pLX303	Addgene (Plasmid #25897)	[72]
pLenti CMV Puro DEST (w118-1) IGF1R	Addgene (Plasmid #107499)	[98]
R777-E106 Hs.INSR-nostop	Addgene (Plasmid #70390)	Unpublished
pcDNA3-hL1	Addgene (Plasmid #89411)	[99]
pDONR223_CXCR4_WT	Addgene (Plasmid #81957)	[99]
pDONR221_SLC6A9	Addgene (Plasmid #132220)	[100]
pDONR221_SLC9A6	Addgene (Plasmid #132175)	[100]

Table 5. Key oligonucleotides

Oligo purpose	Sequence (5'-3')	Source
hSNX2 knockout gRNA, 1 (sense)	CACCGCAGCACTGTCTCCACCCTAG	[74]
hSNX2 knockout gRNA, 1 (antisense)	CTAGGGTGGAGACAGTGCTGC	[74]
hSNX2 knockout gRNA, 2 (sense)	GGGACGCGGTCTCACCTCTA	This study
hSNX2 knockout gRNA, 2 (antisense)	TAGAGGTGAGACCGCGTCCC	This study
human SNX2 RT-PCR primer, forward	GCTGAAGCAAAAATGATGGTTGC	This study

Oligo purpose	Sequence (5'-3')	Source
human SNX2 RT-PCR primer, reverse	TTCCCCTTGTTGCACTTTCGC	This study

References

1. Rogalska, M.E., Vivori, C., and Valcárcel, J. (2023). Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. Nat Rev Genet *24*, 251–269. https://doi.org/10.1038/s41576-022-00556-8.

2. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476. https://doi.org/10.1038/nature07509.

3. Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40*, 1413–1415. https://doi.org/10.1038/ng.259.

4. Tress, M.L., Abascal, F., and Valencia, A. (2017). Alternative Splicing May Not Be the Key to Proteome Complexity. Trends Biochem Sci *42*, 98–110. https://doi.org/10.1016/j.tibs.2016.08.008.

5. Blencowe, B.J. (2017). The Relationship between Alternative Splicing and Proteomic Complexity. Trends in Biochemical Sciences *42*, 407–408. https://doi.org/10.1016/j.tibs.2017.04.001.

6. Sinitcyn, P., Richards, A.L., Weatheritt, R.J., Brademan, D.R., Marx, H., Shishkova, E., Meyer, J.G., Hebert, A.S., Westphall, M.S., Blencowe, B.J., et al. (2023). Global detection of human variants and isoforms by deep proteome sequencing. Nat Biotechnol *41*, 1776–1786. https://doi.org/10.1038/s41587-023-01714-x.

7. Kedan, A., Zauber, H., Wang, M.-R., Zhu, Q., Fang, L., Chen, W., and Selbach, M. (2025). An integrated landscape of mRNA and protein isoforms. Preprint at bioRxiv, https://doi.org/10.1101/2025.04.16.648713 https://doi.org/10.1101/2025.04.16.648713.

8. Kjer-Hansen, P., and Weatheritt, R.J. (2023). The function of alternative splicing in the proteome: rewiring protein interactomes to put old functions into new contexts. Nat Struct Mol Biol, 1–13. https://doi.org/10.1038/s41594-023-01155-9.

9. Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S. (2013). Function of alternative splicing. Gene *514*, 1–30. https://doi.org/10.1016/j.gene.2012.07.083.

10. Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol *18*, 437–451. https://doi.org/10.1038/nrm.2017.27.

11. Kjer-Hansen, P., Phan, T.G., and Weatheritt, R.J. (2024). Protein isoform-centric therapeutics: expanding targets and increasing specificity. Nat Rev Drug Discov, 1–21. https://doi.org/10.1038/s41573-024-01025-z.

12. Mantica, F., and Irimia, M. (2022). The 3D-Evo Space: Evolution of Gene Expression and Alternative Splicing Regulation. Annual Review of Genetics *56*, null. https://doi.org/10.1146/annurev-genet-071719-020653.

13. Ule, J., and Blencowe, B.J. (2019). Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. Molecular Cell *76*, 329–345. https://doi.org/10.1016/j.molcel.2019.09.017.

14. Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., et al. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell *159*, 1511–1523. https://doi.org/10.1016/j.cell.2014.11.035.

15. Calarco, J.A., Superina, S., O'Hanlon, D., Gabut, M., Raj, B., Pan, Q., Skalska, U., Clarke, L., Gelinas, D., van der Kooy, D., et al. (2009). Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. Cell *138*, 898–910. https://doi.org/10.1016/j.cell.2009.06.012.

16. Gonatopoulos-Pournatzis, T., and Blencowe, B.J. (2020). Microexons: at the nexus of nervous system development, behaviour and autism spectrum disorder. Curr Opin Genet Dev *65*, 22–33. https://doi.org/10.1016/j.gde.2020.03.007.

17. Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G.M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y.A., Murray, R.R., et al. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell *164*, 805–817. https://doi.org/10.1016/j.cell.2016.01.029.

18. Ellis, J.D., Barrios-Rodiles, M., Çolak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., et al. (2012). Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. Molecular Cell *46*, 884–892. https://doi.org/10.1016/j.molcel.2012.05.037.

19. Roth, J.F., Braunschweig, U., Wu, M., Li, J.D., Lin, Z.-Y., Larsen, B., Weatheritt, R.J., Gingras, A.-C., and Blencowe, B.J. (2023). Systematic analysis of alternative exon-dependent interactome remodeling reveals multitasking functions of gene regulatory factors. Molecular Cell *83*, 4222-4238.e10. https://doi.org/10.1016/j.molcel.2023.10.034.

20. Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z., et al. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc Natl Acad Sci U S A *103*, 8390–8395. https://doi.org/10.1073/pnas.0507916103.

21. Weatheritt, R.J., Davey, N.E., and Gibson, T.J. (2012). Linear motifs confer functional diversity onto splice variants. Nucleic Acids Res *40*, 7123–7131. https://doi.org/10.1093/nar/gks442.

22. Buljan, M., Chalancon, G., Eustermann, S., Wagner, G.P., Fuxreiter, M., Bateman, A., and Babu, M.M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol Cell *46*, 871–883. https://doi.org/10.1016/j.molcel.2012.05.039.

23. Davey, N.E., Van Roey, K., Weatheritt, R.J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., and Gibson, T.J. (2012). Attributes of short linear motifs. Mol Biosyst *8*, 268–281. https://doi.org/10.1039/c1mb05231d.

24. Xiao, M.-S., Damodaran, A.P., Kumari, B., Dickson, E., Xing, K., On, T.A., Parab, N., King, H.E., Perez, A.R., Guiblet, W.M., et al. (2024). Genome-scale exon perturbation screens uncover exons critical for cell fitness. Molecular Cell *84*, 2553-2572.e19. https://doi.org/10.1016/j.molcel.2024.05.024.

25. Lambourne, L., Mattioli, K., Santoso, C., Sheynkman, G., Inukai, S., Kaundal, B., Berenson, A., Spirohn-Fitzgerald, K., Bhattacharjee, A., Rothman, E., et al. (2025). Widespread variation in molecular interactions and regulatory properties among transcription factor isoforms. Molecular Cell. https://doi.org/10.1016/j.molcel.2025.03.004.

26. Joung, J., Ma, S., Tay, T., Geiger-Schuller, K.R., Kirchgatterer, P.C., Verdine, V.K., Guo, B., Arias-Garcia, M.A., Allen, W.E., Singh, A., et al. (2023). A transcription factor atlas of directed differentiation. Cell *186*, 209-229.e26. https://doi.org/10.1016/j.cell.2022.11.026.

27. Lopez-Blanch, L., Rodríguez-Marin, C., Mantica, F., Iñiguez, L.P., Permanyer, J., Kita, E.M., Mackensen, T., Codina-Tobias, M., Romero-Ferrero, F., Fernandez-Albert, J., et al. (2024). Phenotypic impact of individual conserved neuronal microexons and their master regulators in zebrafish. eLife *13*. https://doi.org/10.7554/eLife.104275.1.

28. Calhoun, C.C., Capps, M.E., Muya, K., Gannaway, W.C., Martina, V., Conklin, C.L., Klein, M.C., Webster, J.M., Torija-Olson, E.G., and Thyme, S.B. (2024). Removal of developmentally regulated microexons has a minimal impact on larval zebrafish brain morphology and function. eLife *13*. https://doi.org/10.7554/eLife.101790.1.

29. Buchser, W., Collins, M., Garyantes, T., Guha, R., Haney, S., Lemmon, V., Li, Z., and Trask, O.J. (2004). Assay Development Guidelines for Image-Based High Content Screening, High Content Analysis and High Content Imaging. In Assay Guidance Manual, S. Markossian, A. Grossman, M. Arkin, D. Auld, C. Austin, J. Baell, K. Brimacombe, T. D. Y. Chung, N. P. Coussens, J. L. Dahlin, et al., eds. (Eli Lilly & Company and the National Center for Advancing Translational Sciences).

30. Lacoste, J., Haghighi, M., Haider, S., Reno, C., Lin, Z.-Y., Segal, D., Qian, W.W., Xiong, X., Teelucksingh, T., Miglietta, E., et al. (2024). Pervasive mislocalization of pathogenic coding variants underlying human disorders. Cell *0*. https://doi.org/10.1016/j.cell.2024.09.003.

31. Jonathan D. Ellis, Ulrich Braunschweig, Dave O'Hanlon, Miriam Barrios-Rodiles, Thomas Gonatopoulos-Pournatzis, Lim Caden Comsa, Jeffrey L. Wrana, Robert J. Weatheritt, and Benjamin J. Blencowe Activity-Dependent Splicing Rewires Protein-Protein Interaction Networks. In preparation.

32. Lewis, P.A., Bradley, I.C., Pizzey, A.R., Isaacs, H.V., and Evans, G.J.O. (2017). N1-Src Kinase Is Required for Primary Neurogenesis in Xenopus tropicalis. J Neurosci *37*, 8477–8485. https://doi.org/10.1523/JNEUROSCI.3881-16.2017.

33. Choudhury, R., Noakes, C.J., McKenzie, E., Kox, C., and Lowe, M. (2009). Differential clathrin binding and subcellular localization of OCRL1 splice isoforms. J Biol Chem *284*, 9965–9973. https://doi.org/10.1074/jbc.M807442200.

34. Dho, S.E., French, M.B., Woods, S.A., and McGlade, C.J. (1999). Characterization of four mammalian numb protein isoforms. Identification of cytoplasmic and membrane-associated variants of the phosphotyrosine binding domain. J Biol Chem *274*, 33097–33104. https://doi.org/10.1074/jbc.274.46.33097.

35. Wang, D.B., Uo, T., Kinoshita, C., Sopher, B.L., Lee, R.J., Murphy, S.P., Kinoshita, Y., Garden, G.A., Wang, H.-G., and Morrison, R.S. (2014). Bax interacting factor-1 promotes survival and mitochondrial elongation in neurons. J Neurosci *34*, 2674–2683. https://doi.org/10.1523/JNEUROSCI.4074-13.2014.

36. Strack, S., Wilson, T.J., and Cribbs, J.T. (2013). Cyclin-dependent kinases regulate splice-specific targeting of dynamin-related protein 1 to microtubules. J Cell Biol *201*, 1037–1051. https://doi.org/10.1083/jcb.201210045.

37. Cao, H., Garcia, F., and McNiven, M.A. (1998). Differential Distribution of Dynamin Isoforms in Mammalian Cells. Mol Biol Cell *9*, 2595–2609. https://doi.org/10.1091/mbc.9.9.2595.

38. Liu, Y.-W., Surka, M.C., Schroeter, T., Lukiyanchuk, V., and Schmid, S.L. (2008). Isoform and splice-variant specific functions of dynamin-2 revealed by analysis of conditional knock-out cells. Mol Biol Cell *19*, 5347–5359. https://doi.org/10.1091/mbc.e08-08-0890.

39. Gopaldass, N., Chen, K.-E., Collins, B., and Mayer, A. (2024). Assembly and fission of tubular carriers mediating protein sorting in endosomes. Nat Rev Mol Cell Biol. https://doi.org/10.1038/s41580-024-00746-8.

40. Rojas, R., Kametaka, S., Haft, C.R., and Bonifacino, J.S. (2007). Interchangeable but essential functions of SNX1 and SNX2 in the association of retromer with endosomes and the trafficking of mannose 6-phosphate receptors. Mol Cell Biol *27*, 1112–1124. https://doi.org/10.1128/MCB.00156-06.

41. Simonetti, B., Paul, B., Chaudhari, K., Weeratunga, S., Steinberg, F., Gorla, M., Heesom, K.J., Bashaw, G.J., Collins, B.M., and Cullen, P.J. (2019). Molecular identification of a BAR domain-containing coat complex for endosomal recycling of transmembrane proteins. Nat Cell Biol *21*, 1219–1233. https://doi.org/10.1038/s41556-019-0393-3.

42. Yong, X., Zhao, L., Deng, W., Sun, H., Zhou, X., Mao, L., Hu, W., Shen, X., Sun, Q., Billadeau, D.D., et al. (2020). Mechanism of cargo recognition by retromer-linked SNX-BAR proteins. PLoS Biol *18*, e3000631. https://doi.org/10.1371/journal.pbio.3000631.

43. Shortill, S.P., Frier, M.S., and Conibear, E. (2022). You can go your own way: SNX-BAR coat complexes direct traffic at late endosomes. Curr Opin Cell Biol *76*, 102087. https://doi.org/10.1016/j.ceb.2022.102087.

44. Carlton, J.G., Bujny, M.V., Peter, B.J., Oorschot, V.M.J., Rutherford, A., Arkell, R.S., Klumperman, J., McMahon, H.T., and Cullen, P.J. (2005). Sorting nexin-2 is associated with tubular elements of the early endosome, but is not essential for retromer-mediated endosome-to-TGN transport. J Cell Sci *118*, 4527–4539. https://doi.org/10.1242/jcs.02568.

45. Tapial, J., Ha, K.C.H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., Quesnel-Vallières, M., Permanyer, J., Sodaei, R., Marquez, Y., et al. (2017). An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. *27*, 1759–1768. https://doi.org/10.1101/gr.220962.117.

46. Carlton, J., Bujny, M., Peter, B.J., Oorschot, V.M.J., Rutherford, A., Mellor, H., Klumperman, J., McMahon, H.T., and Cullen, P.J. (2004). Sorting nexin-1 mediates tubular endosome-to-TGN transport through coincidence sensing of high- curvature membranes and 3-phosphoinositides. Curr Biol *14*, 1791–1800. https://doi.org/10.1016/j.cub.2004.09.077.

47. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

48. Coiled coil-based therapeutics and drug delivery systems - PubMed https://pubmed.ncbi.nlm.nih.gov/33378707/.

49. Barrios-Rodiles, M., Brown, K.R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R.S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I.W., et al. (2005). High-throughput mapping of a dynamic signaling network in mammalian cells. Science *307*, 1621–1625. https://doi.org/10.1126/science.1105776.

50. Haft, C.R., de la Luz Sierra, M., Barr, V.A., Haft, D.H., and Taylor, S.I. (1998). Identification of a Family of Sorting Nexin Molecules and Characterization of Their Association with Receptors. Molecular and Cellular Biology *18*, 7278–7287. https://doi.org/10.1128/MCB.18.12.7278.

51. Barrios-Rodiles, M., Ellis, J.D., Blencowe, B.J., and Wrana, J.L. (2017). LUMIER: A Discovery Tool for Mammalian Protein Interaction Networks. Methods Mol Biol *1550*, 137–148. https://doi.org/10.1007/978-1-4939-6747-6_11.

52. Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science *338*, 1587–1593. https://doi.org/10.1126/science.1230612.

53. Blue, R.E., Curry, E.G., Engels, N.M., Lee, E.Y., and Giudice, J. (2018). How alternative splicing affects membrane-trafficking dynamics. J Cell Sci *131*, jcs216465. https://doi.org/10.1242/jcs.216465.

54. Wright, C.J., Smith, C.W.J., and Jiggins, C.D. (2022). Alternative splicing as a source of phenotypic diversity. Nat Rev Genet. https://doi.org/10.1038/s41576-022-00514-4.

55. Zhang, X., Tseo, Y., Bai, Y., Chen, F., and Uhler, C. (2025). Prediction of protein subcellular localization in single cells. Nat Methods, 1–11. https://doi.org/10.1038/s41592-025-02696-1.

56. Thul, P.J., and Lindskog, C. (2018). The human protein atlas: A spatial map of the human proteome. Protein Sci *27*, 233–244. https://doi.org/10.1002/pro.3307.

57. Lopez-Robles, C., Scaramuzza, S., Astorga-Simon, E.N., Ishida, M., Williamson, C.D., Baños-Mateos, S., Gil-Carton, D., Romero-Durana, M., Vidaurrazaga, A., Fernandez-Recio, J., et al. (2023). Architecture of the ESCPE-1 membrane coat. Nat Struct Mol Biol *30*, 958–969. https://doi.org/10.1038/s41594-023-01014-7.

58. Zhang, Y., Pang, X., Li, J., Xu, J., Hsu, V.W., and Sun, F. (2021). Structural insights into membrane remodeling by SNX1. Proc Natl Acad Sci U S A *118*, e2022614118. https://doi.org/10.1073/pnas.2022614118.

59. Kvainickas, A., Jimenez-Orgaz, A., Nägele, H., Hu, Z., Dengjel, J., and Steinberg, F. (2017). Cargo-selective SNX-BAR proteins mediate retromer trimer independent retrograde transport. Journal of Cell Biology *216*, 3677–3693. https://doi.org/10.1083/jcb.201702137.

60. Dergai, M., Tsyba, L., Dergai, O., Zlatskii, I., Skrypkina, I., Kovalenko, V., and Rynditch, A. (2010). Microexon-based regulation of ITSN1 and Src SH3 domains specificity relies on introduction of charged amino acids into the interaction interface. Biochem Biophys Res Commun *399*, 307–312. https://doi.org/10.1016/j.bbrc.2010.07.080.

61. Wilson, S.C., White, K.I., Zhou, Q., Pfuetzner, R.A., Choi, U.B., Südhof, T.C., and Brunger, A.T. (2019). Structures of neurexophilin-neurexin complexes reveal a regulatory mechanism of alternative splicing. EMBO J *38*, e101603. https://doi.org/10.15252/embj.2019101603.

62. Garcia, J., Gerber, S.H., Sugita, S., Südhof, T.C., and Rizo, J. (2004). A conformational switch in the Piccolo C2A domain regulated by alternative splicing. Nat Struct Mol Biol *11*, 45–53. https://doi.org/10.1038/nsmb707.

63. Gerth, F., Jäpel, M., Sticht, J., Kuropka, B., Schmitt, X.J., Driller, J.H., Loll, B., Wahl, M.C., Pagel, K., Haucke, V., et al. (2019). Exon Inclusion Modulates Conformational Plasticity and Autoinhibition of the Intersectin 1 SH3A Domain. Structure *27*, 977-987.e5. https://doi.org/10.1016/j.str.2019.03.020.

64. Gan, Y., Li, Y., Long, Z., Lee, A.R., Xie, N., Lovnicki, J.M., Tang, Y., Chen, X., Huang, J., and Dong, X. (2018). Roles of Alternative RNA Splicing of the Bif-1 Gene by SRRM4 During the Development of Treatment-induced Neuroendocrine Prostate Cancer. EBioMedicine *31*, 267–275. https://doi.org/10.1016/j.ebiom.2018.05.002.

65. Fus-Kujawa, A., Prus, P., Bajdak-Rusinek, K., Teper, P., Gawron, K., Kowalczuk, A., and Sieron, A.L. (2021). An Overview of Methods and Tools for Transfection of Eukaryotic Cells in vitro. Front Bioeng Biotechnol *9*, 701031. https://doi.org/10.3389/fbioe.2021.701031.

66. Fiflis, D.N., Rey, N.A., Venugopal-Lavanya, H., Sewell, B., Mitchell-Dick, A., Clements, K.N., Milo, S., Benkert, A.R., Rosales, A., Fergione, S., et al. (2024). Repurposing CRISPR-Cas13 systems for robust mRNA trans-splicing. Nat Commun *15*, 2325. https://doi.org/10.1038/s41467-024-46172-4.

67. Funk, L., Su, K.-C., Ly, J., Feldman, D., Singh, A., Moodie, B., Blainey, P.C., and Cheeseman, I.M. (2022). The phenotypic landscape of essential human genes. Cell. https://doi.org/10.1016/j.cell.2022.10.017.

68. Kappel, K., Strebinger, D., Edmonds, K.K., Chau-Duy-Tam VO, S., Vockley, C.M., Biswas, T., Farhi, S.L., Macrae, R., Zhang, F., and Regev, A. (2025). Characterizing protein sequence determinants of nuclear condensates by high-throughput pooled imaging with CondenSeq. Nat Methods, 1–1. https://doi.org/10.1038/s41592-025-02726-y.

69. Feldman, D., Singh, A., Schmid-Burgk, J.L., Carlson, R.J., Mezger, A., Garrity, A.J., Zhang, F., and Blainey, P.C. (2019). Optical Pooled Screens in Human Cells. Cell *179*, 787-799.e17. https://doi.org/10.1016/j.cell.2019.09.016.

70. Ramezani, M., Weisbart, E., Bauman, J., Singh, A., Yong, J., Lozada, M., Way, G.P., Kavari, S.L., Diaz, C., Leardini, E., et al. (2025). A genome-wide atlas of human cell morphology. Nat Methods, 1–13. https://doi.org/10.1038/s41592-024-02537-7.

71. Berry, S., Müller, M., Rai, A., and Pelkmans, L. (2022). Feedback from nuclear RNA on transcription promotes robust RNA concentration homeostasis in human cells. Cell Syst *13*, 454-470.e15. https://doi.org/10.1016/j.cels.2022.04.005.

72. Yang, X., Boehm, J.S., Yang, X., Salehi-Ashtiani, K., Hao, T., Shen, Y., Lubonja, R., Thomas, S.R., Alkan, O., Bhimdi, T., et al. (2011). A public genome-scale lentiviral expression library of human ORFs. Nat Methods *8*, 659–661. https://doi.org/10.1038/nmeth.1638.

73. Maniatis, T. (1982). Molecular cloning. A laboratory manual.

74. Simonetti, B., Danson, C.M., Heesom, K.J., and Cullen, P.J. (2017). Sequence-dependent cargo recognition by SNX-BARs mediates retromer-independent transport of CI-MPR. J Cell Biol *216*, 3695–3712. https://doi.org/10.1083/jcb.201703015.

75. Moreno-Mateos, M.A., Vejnar, C.E., Beaudoin, J.-D., Fernandez, J.P., Mis, E.K., Khokha, M.K., and Giraldez, A.J. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat Methods *12*, 982–988. https://doi.org/10.1038/nmeth.3543.

76. Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat Protoc *8*, 2281–2308. https://doi.org/10.1038/nprot.2013.143.

77. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat Methods *9*, 676–682. https://doi.org/10.1038/nmeth.2019.

78. Peng, T., Thorn, K., Schroeder, T., Wang, L., Theis, F.J., Marr, C., and Navab, N. (2017). A BaSiC tool for background and shading correction of optical microscopy images. Nat Commun *8*, 14836. https://doi.org/10.1038/ncomms14836.

79. Pachitariu, M., and Stringer, C. (2022). Cellpose 2.0: how to train your own model. Nat Methods *19*, 1634–1641. https://doi.org/10.1038/s41592-022-01663-4.

80. Berg, S., Kutra, D., Kroeger, T., Straehle, C.N., Kausler, B.X., Haubold, C., Schiegg, M., Ales, J., Beier, T., Rudy, M., et al. (2019). ilastik: interactive machine learning for (bio)image analysis. Nat Methods *16*, 1226–1232. https://doi.org/10.1038/s41592-019-0582-9.

81. Stirling, D.R., Swain-Bowden, M.J., Lucas, A.M., Carpenter, A.E., Cimini, B.A., and Goodman, A. (2021). CellProfiler 4: improvements in speed, utility and usability. BMC Bioinformatics *22*, 433. https://doi.org/10.1186/s12859-021-04344-9.

82. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software *28*, 1–26. https://doi.org/10.18637/jss.v028.i05.

83. Venables, W.N., and Ripley, B.D. (2002). Modern Applied Statistics with S (Springer) https://doi.org/10.1007/978-0-387-21706-2.

84. Angelo Canty and B. D. Ripley (2024). boot: Bootstrap R (S-Plus) Functions.

85. Kassambara A and Mundt F (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses.

86. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the Tidyverse. Journal of Open Source Software *4*, 1686. https://doi.org/10.21105/joss.01686.

87. Wassmer, T., Attar, N., Harterink, M., van Weering, J.R.T., Traer, C.J., Oakley, J., Goud, B., Stephens, D.J., Verkade, P., Korswagen, H.C., et al. (2009). The retromer coat complex coordinates endosomal sorting and dynein-mediated transport, with carrier recognition by the trans-Golgi network. Dev Cell *17*, 110–122. https://doi.org/10.1016/j.devcel.2009.04.016.

88. Choi, H., Larsen, B., Lin, Z.-Y., Breitkreutz, A., Mellacheruvu, D., Fermin, D., Qin, Z.S., Tyers, M., Gingras, A.-C., and Nesvizhskii, A.I. (2011). SAINT: Probabilistic Scoring of Affinity Purification - Mass Spectrometry Data. Nature methods *8*, 70. https://doi.org/10.1038/nmeth.1541.

89. Gueroussov, S., Weatheritt, R.J., O'Hanlon, D., Lin, Z.-Y., Narula, A., Gingras, A.-C., and Blencowe, B.J. (2017). Regulatory Expansion in Mammals of Multivalent hnRNP Assemblies that Globally Control Alternative Splicing. Cell *170*, 324-339.e23. https://doi.org/10.1016/j.cell.2017.06.037.

90. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat Methods *19*, 679–682. https://doi.org/10.1038/s41592-022-01488-1.

91. Meng, E.C., Goddard, T.D., Pettersen, E.F., Couch, G.S., Pearson, Z.J., Morris, J.H., and Ferrin, T.E. (2023). UCSF ChimeraX: Tools for structure building and analysis. Protein Sci *32*, e4792. https://doi.org/10.1002/pro.4792.

92. UniProt Consortium (2025). UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res *53*, D609–D617. https://doi.org/10.1093/nar/gkae1010.

93. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature *630*, 493–500. https://doi.org/10.1038/s41586-024-07487-w.

94. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol *7*, 539. https://doi.org/10.1038/msb.2011.75.

95. Procter, J.B., Carstairs, G.M., Soares, B., Mourão, K., Ofoegbu, T.C., Barton, D., Lui, L., Menard, A., Sherstnev, N., Roldan-Martinez, D., et al. (2021). Alignment of Biological Sequences with Jalview. Methods Mol Biol *2231*, 203–224. https://doi.org/10.1007/978-1-0716-1036-7_13.

96. Kolde, R. (2019). Pheatmap: pretty heatmaps. R package version *1*, 726.

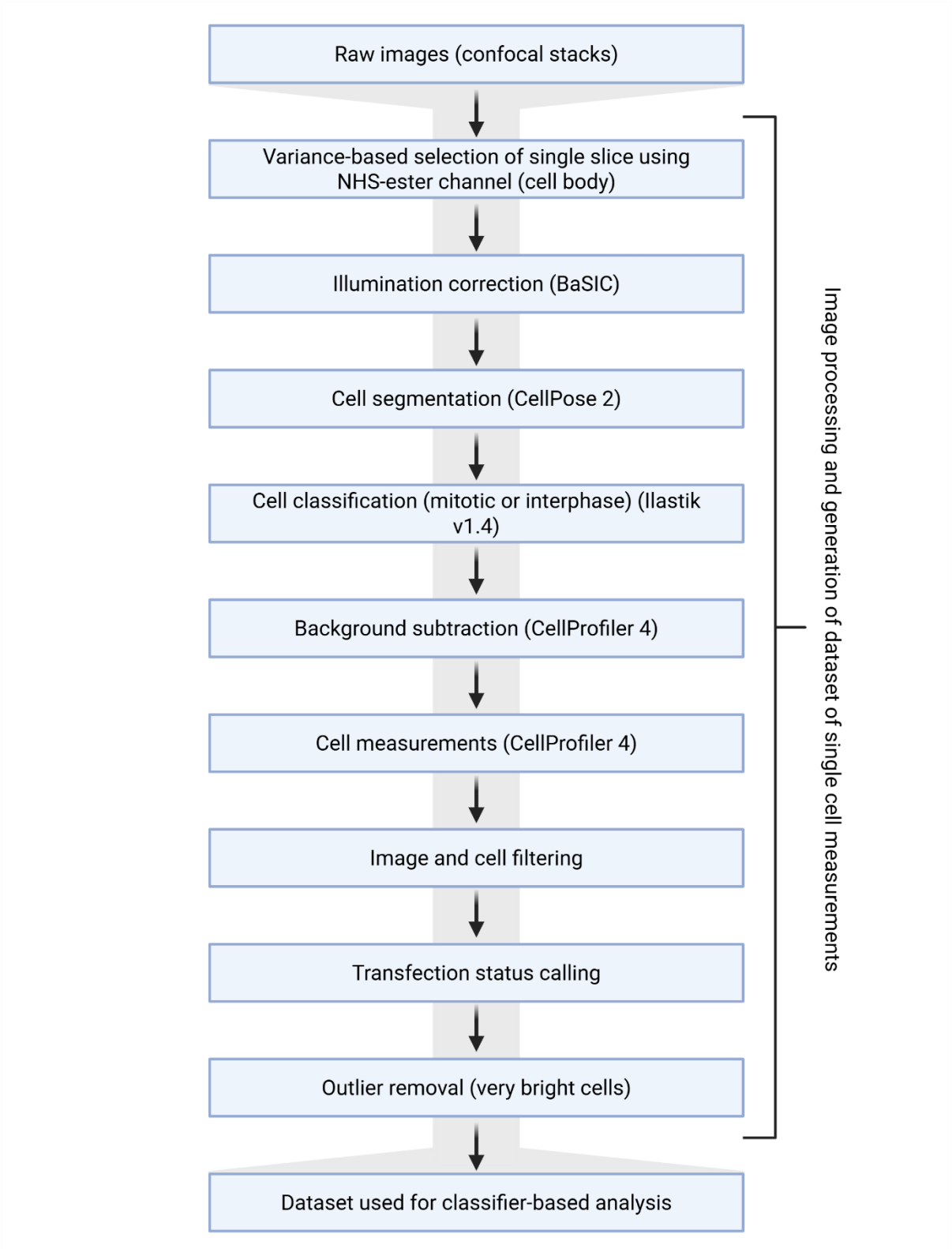
97. Joosten, R.P., te Beek, T.A.H., Krieger, E., Hekkelman, M.L., Hooft, R.W.W., Schneider, R., Sander, C., and Vriend, G. (2011). A series of PDB related databases for everyday needs. Nucleic Acids Res *39*, D411-419. https://doi.org/10.1093/nar/gkq1105.

98. Ariës, I.M., Bodaar, K., Karim, S.A., Chonghaile, T.N., Hinze, L., Burns, M.A., Pfirrmann, M., Degar, J., Landrigan, J.T., Balbach, S., et al. (2018). PRC2 loss induces chemoresistance by repressing apoptosis in T cell acute lymphoblastic leukemia. J Exp Med *215*, 3094–3114. https://doi.org/10.1084/jem.20180570.

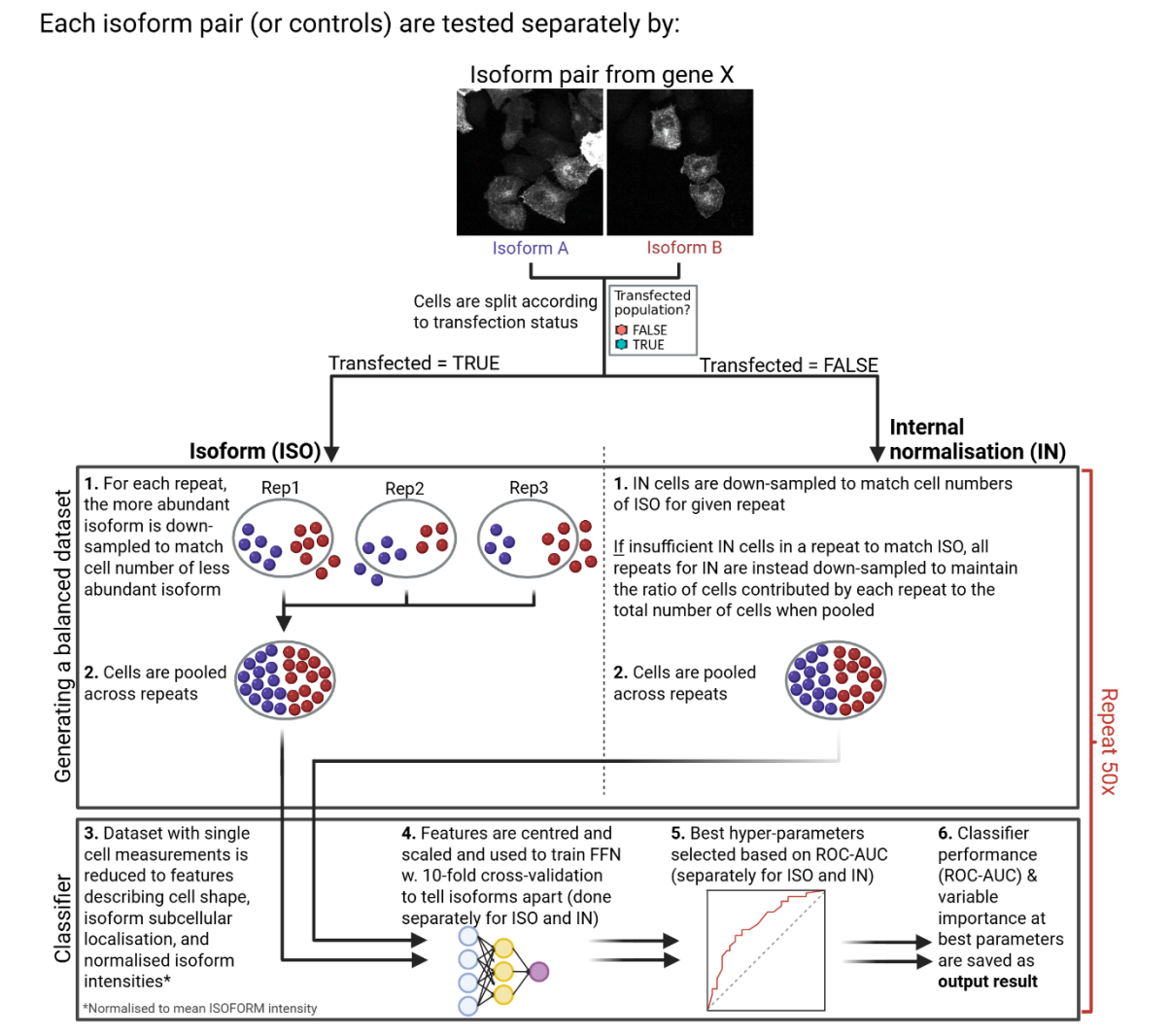
99. De Angelis, E., MacFarlane, J., Du, J.S., Yeo, G., Hicks, R., Rathjen, F.G., Kenwrick, S., and Brümmendorf, T. (1999). Pathological missense mutations of neural cell adhesion molecule L1 affect homophilic and heterophilic binding activities. EMBO J *18*, 4744–4753. https://doi.org/10.1093/emboj/18.17.4744.

100. Superti-Furga, G., Lackner, D., Wiedmer, T., Ingles-Prieto, A., Barbosa, B., Girardi, E., Goldmann, U., Gürtl, B., Klavins, K., Klimek, C., et al. (2020). The RESOLUTE consortium: unlocking SLC transporters for drug discovery. Nat Rev Drug Discov *19*, 429–430. https://doi.org/10.1038/d41573-020-00056-6.

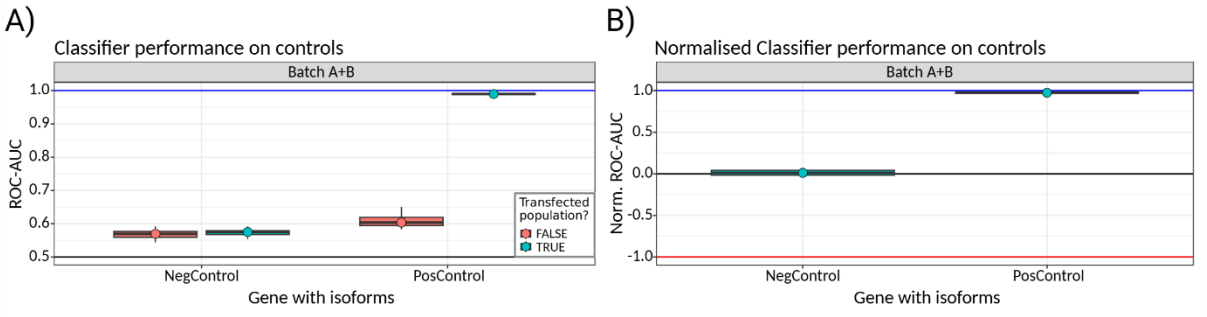
Supplementary Materials



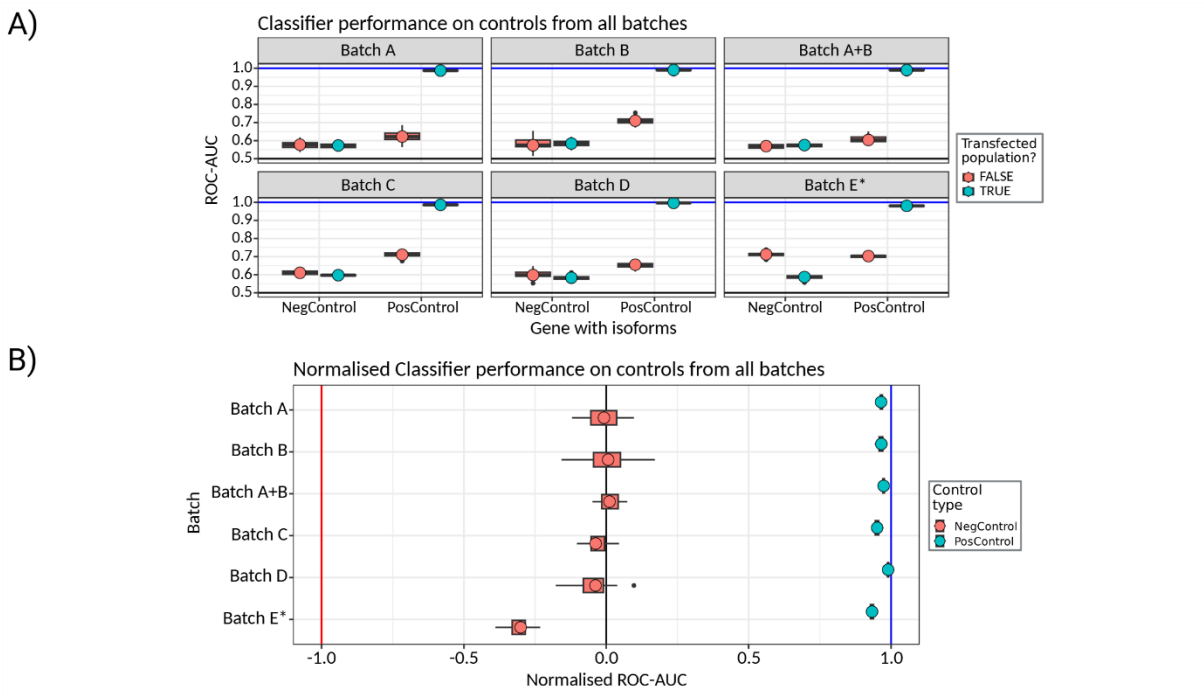
Supp. Figure 1. Generation of a single cell isoform measurement dataset from the acquired images. *Overview of image processing to generate the cell datasets used for the classifier-based analysis (Supp. Figure 2). The selected focus slice from each confocal stack is used for all four channels for all subsequent processing. See Methods for details.*



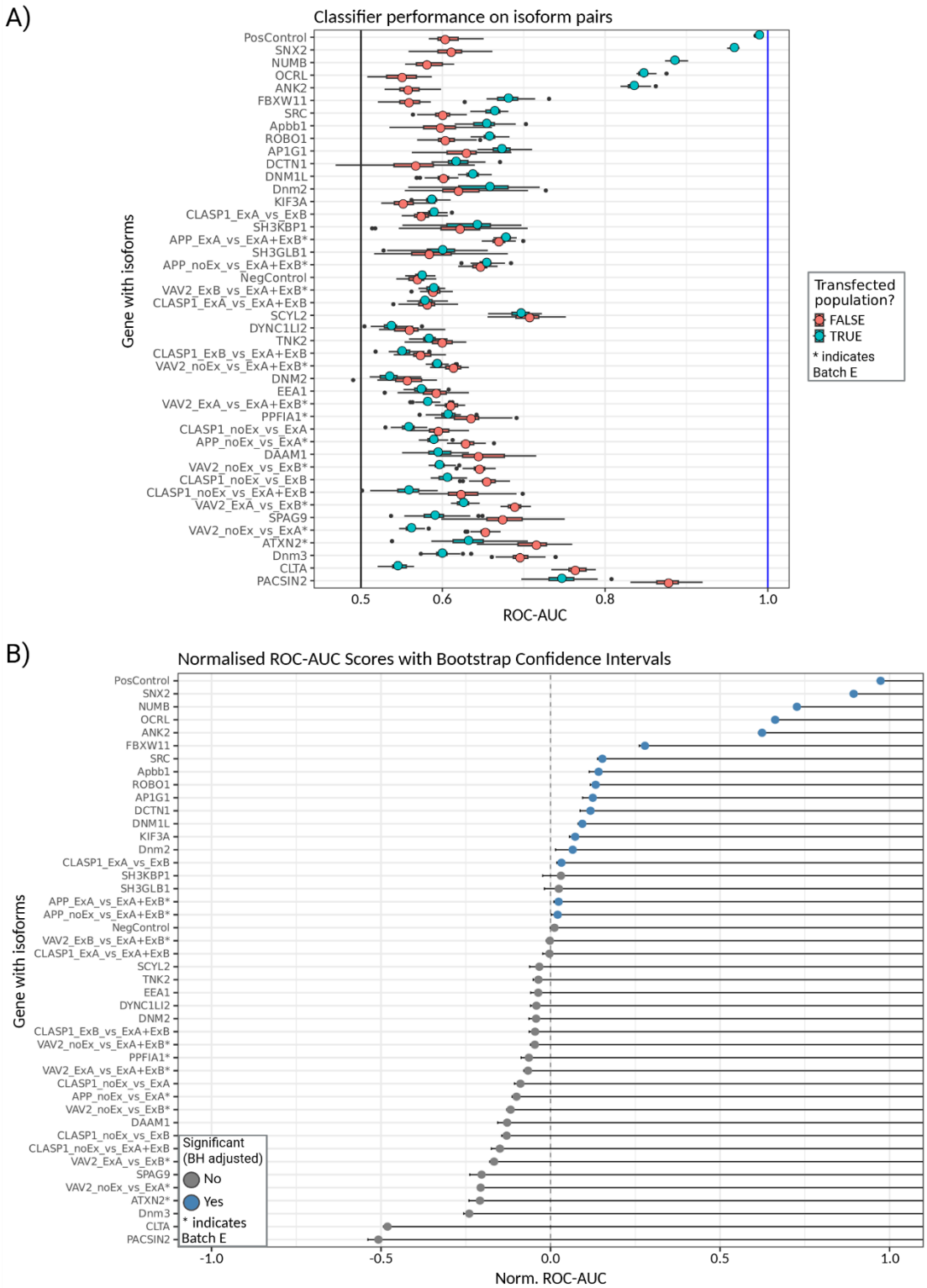
Supp. Figure 2. Isoform differences are detected using down-sampling and classifier-based analysis of single cell measurement dataset. *Overview of classifier-based analysis. Isoforms (or the positive or negative control) are compared in a pairwise-manner, where isoform-expressing cells (transfected, referred to as ISO) and internal normalisation cells (non-transfected, referred to as IN) initially are split. For the ISO cells, the more abundant isoform is down-sampled to match the less abundant one in terms of cell numbers on a replicate basis (1). The cells are then pooled across replicate (2), which means replicates can contribute different number of cells, but always 50% of each isoform. This corrects for replicate-specific effects driving classification. The IN is sampled to match the ISO. The down-sampled datasets are then used for classifier-based analysis where features describing cell morphology, mCherry localisation, and normalised intensity values are used as input for a feedforward neural network (FFN) (3). The classifier is trained with 10-fold cross-validation where the average performance on the test folds at optimal parameters is used as a readout for classifier performance (4, 5). Feature importance scores are also saved to identify features driving classification (6). The classifier evaluation is done separately for the ISO and IN datasets. The combined down-sampling and classifier evaluation is done 50 times, where the performance for each iteration is used as the readout. See Methods for details.*



Supp. Figure 3. Pooling cells across experiment batches doesn't alter performance for the positive and negative controls. A) *Boxplots showing the ability of classifiers trained on mCherry features measured in cells expressing control plasmids to correctly assign the plasmid ID (which corresponds to well ID), quantified as ROC-AUC. Cells were pooled from two separate experiments (batch A and B), each with three experimental repeats, before the classifier-based evaluation (n cells ≥ 1500 after pooling). The boxplots illustrate the ROC-AUCs from each of 50 iterations of cell sampling and classifier evaluation, and dots indicate median ROC-AUC. Classifier-based analysis was performed separately on transfected (isoform) and non-transfected (internal normalisation) cells from the same, indicated by colour. Coloured lines indicate perfect (blue) and worst performance (black), respectively.* ***B)*** *Same as Supp. Figure 3A, but the boxplots display classifier performance quantified as normalised ROC-AUC. Coloured lines indicate perfect (blue), neutral (black) or worst (red) performance, respectively.*

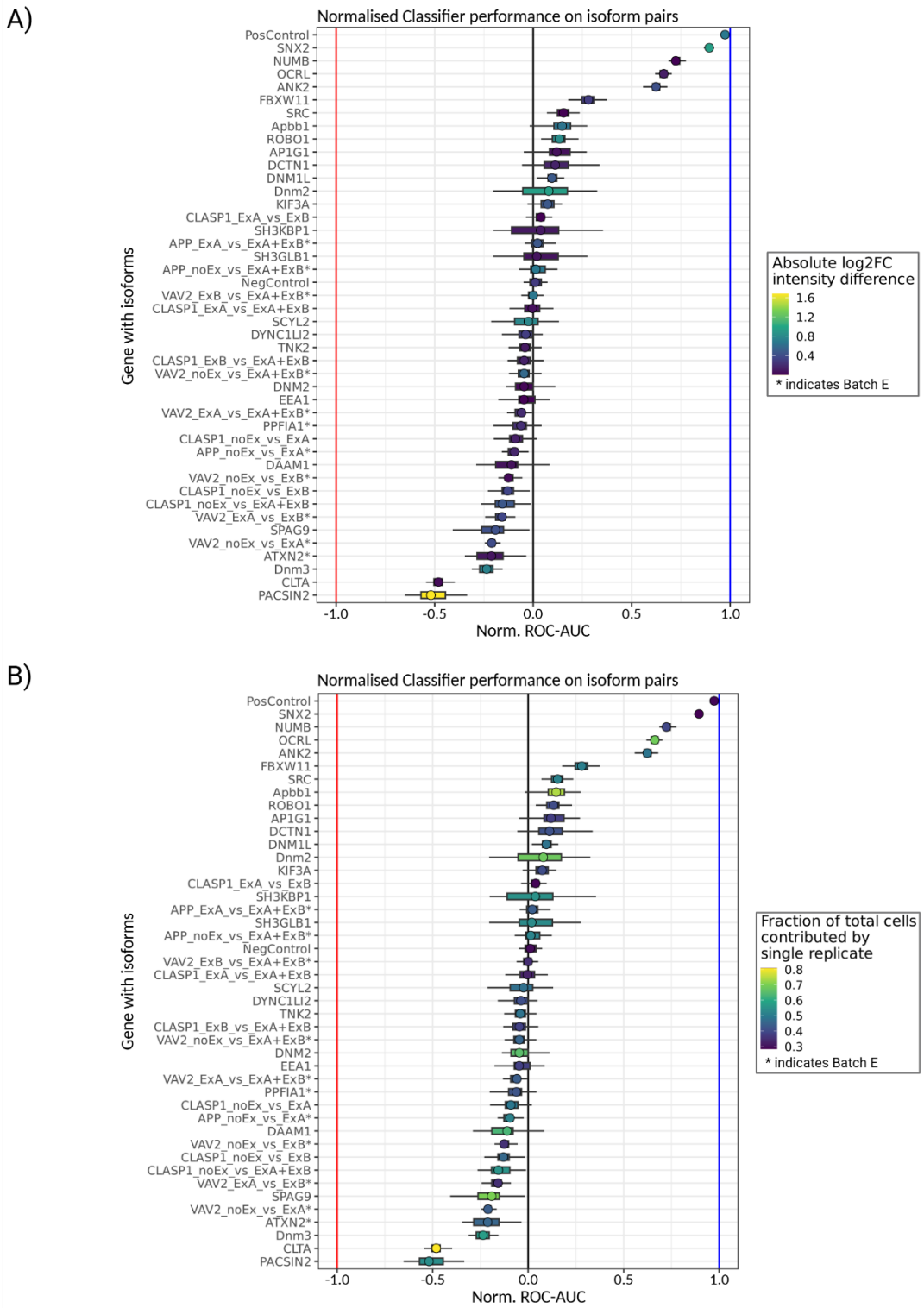


Supp. Figure 4. The positive and negative controls are largely consistent across batches. A) *Boxplots showing the ability of classifiers trained on mCherry features measured in cells expressing control plasmids to correctly assign the plasmid ID (which corresponds to well ID), quantified as ROC-AUC. The evaluation was performed separately on all experimental batches as well as the pooled batch A + B. Each batch consisted of three experimental repeats. n cells ≥ 500 for each comparison after pooling. The boxplots illustrate the ROC-AUCs from each of 50 iterations of cell sampling and classifier evaluation, and dots indicate median ROC-AUC across iterations. Classifier-based analysis was performed separately on transfected (isoform) and non-transfected (internal normalisation) cells from the same wells, indicated by colour. Coloured lines indicate perfect (blue) and worst performance (black), respectively. Batch E is indicated *.* ***B)*** *Same as Supp. Figure 4A, but the boxplots display classifier performance quantified as normalised ROC-AUC. Coloured lines indicate perfect (blue), and neutral (black) performance, respectively.*

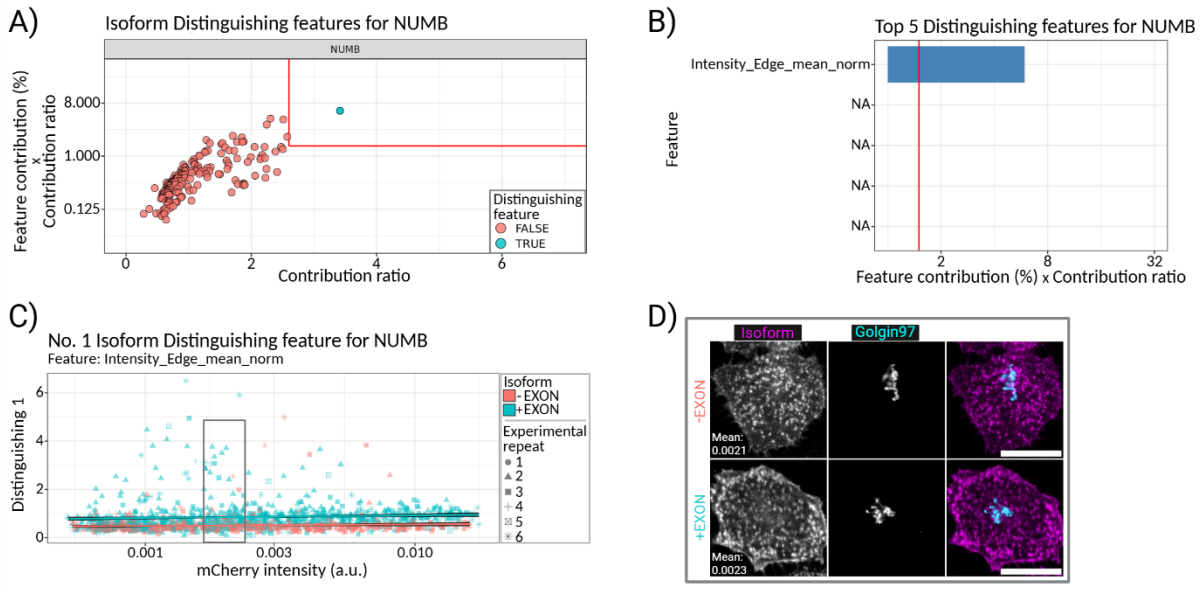


Supp. Figure 5. Performance (ROC-AUC) for isoform and internal normalisation classifiers, and bootstrap analysis of normalised ROC-AUCs. A) *Boxplots showing the performance of classifiers trained on mCherry features of cells transfected with protein isoforms (or controls) making up isoform pairs, quantified as ROC-AUC. All isoform pairs from batches A-E are shown. The cells used to examine a given isoform pair came from three experimental repeats from a given batch that were pooled for the classifier-based analysis. For isoform pairs present in multiple batches (batch A and B), cells expressing these were pooled prior to classifier-based analysis and so cells came from 6 experimental repeats for these isoform pairs. n cells ≥170 for each pair after pooling, with median n cells = 967, and numbers for each pair listed in Supp. Table 2. The shown controls also derive from pooling batch A and B. Classifier-based analysis was performed separately on transfected (isoform) and non-transfected (internal control) cells from the same wells, indicated by colour. Gene names in all caps indicate*

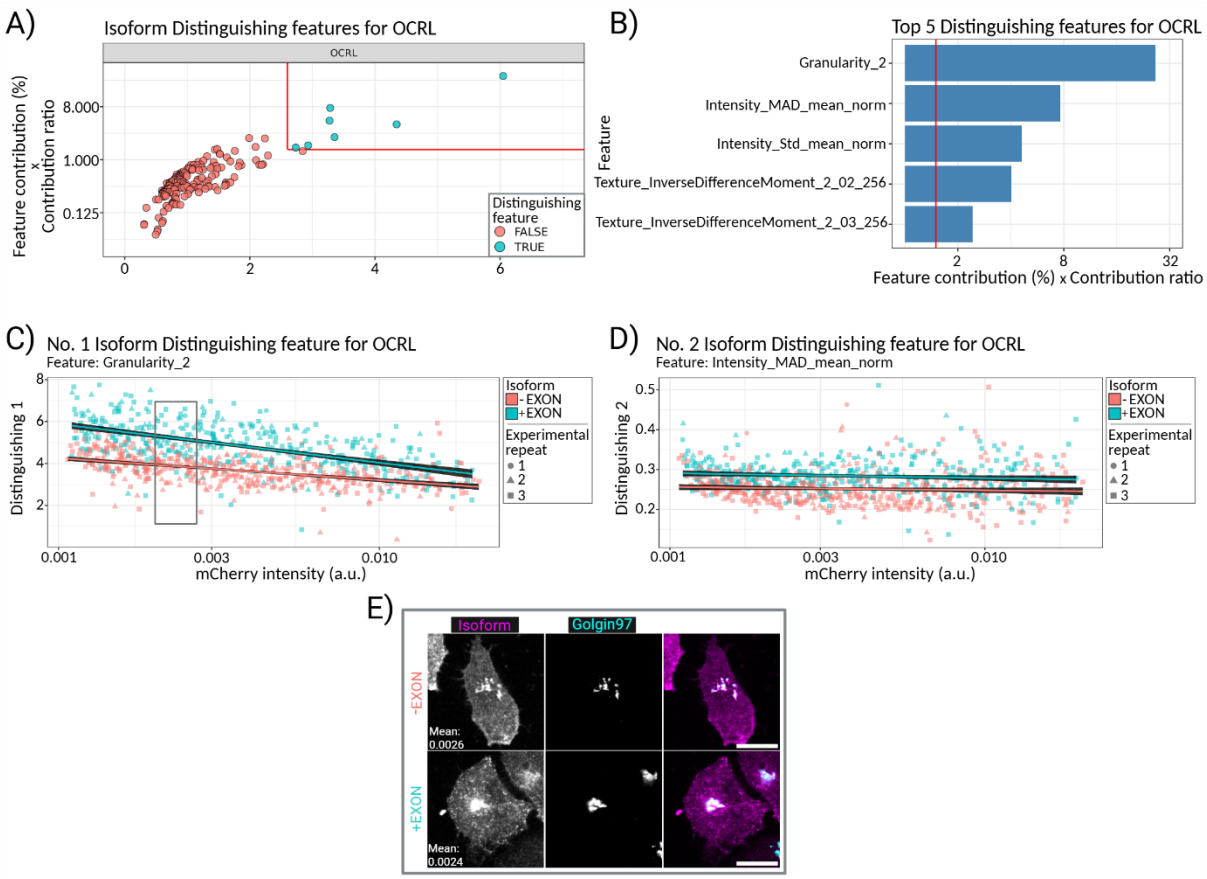
isoforms come from human, whereas genes only starting with upper case come from mouse. ***B)*** *XY plots showing bootstrap statistical testing of the normalised ROC-AUC scores for the isoform pairs from Figure 2A. The normalised ROC-AUC scores for each pair were analysed via bootstrapping, and the displayed values are the bootstrap mean and 99% confidence intervals (CI). Isoform pairs that didn't contain 0 in the lower bound of the CI after false discovery rate correction were identified as having ISO classifiers significantly outperforming the IN classifier and are coloured blue. Isoform pairs with 0 in the CI were found as non-significant, coloured grey. n bootstrap iterations = 2000, one-sided test, $p < 0.01$, *.*



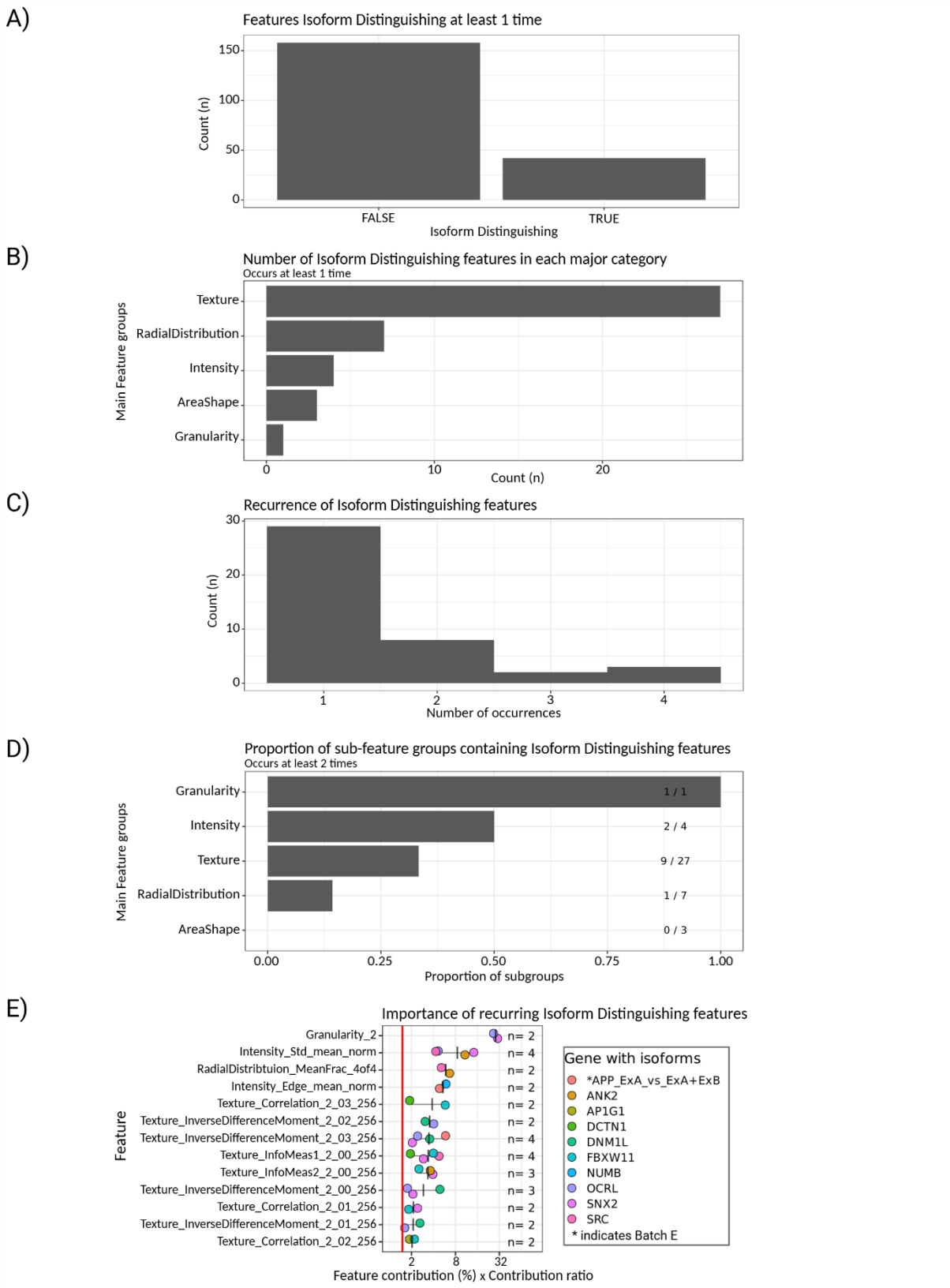
Supp. Figure 6. Classifier performance on isoform pairs doesn't appear to be driven by imbalanced isoform expression nor replicate sampling. A) *Boxplots showing the performance of classifiers trained on mCherry features of cells transfected with protein isoforms (or controls) making up isoform pairs, quantified as normalised ROC-AUC. Same as Figure 2A, but isoform pairs are coloured according to population-level difference in mCherry mean intensity between the isoforms. The difference is quantified as the absolute log2 fold-change between isoforms, calculated for each experimental replicate (n = 3 or 6), scaled according to the fraction of total cells that comes from that replicate, and then summed together (Methods).* ***B)*** *Boxplots showing the performance of classifiers trained on mCherry features of cells transfected with protein isoforms (or controls) making up isoform pairs, quantified as normalised ROC-AUC. Same plot as Figure 2A, but isoform pairs are coloured according to the how large a fraction of total cells is contributed by a single replicate, n = 3 or 6 experimental replicates.*



Supp. Figure 7. Two NUMB isoforms differ in their localisation to the cell periphery/membrane, replicating previous reports. A) *XY plot of Contribution ratio and Contribution ratio multiplied with Feature contribution (%) for the ISO classifier for each feature for the NUMB isoform pair, used to identify Isoform Distinguishing features. Of ~200 features, 1 was identified as 'Isoform distinguishing'.* ***B)*** *Bar plot of top 5 Isoform Distinguishing features for the NUMB isoform pair, according to their Feature contribution (%) multiplied with Contribution ratio. The threshold value defining Isoform Distinguishing features is indicated as a red line.* ***C)*** *XY plot of Top 1 Isoform Distinguishing Feature for the NUMB isoforms (mean mCherry intensity at cell edge, normalised to mean mCherry intensity for whole cell) against cell mean mCherry intensity levels (proxy for expression level). Each data point corresponds to a cell with colour and shape indicating NUMB isoform identity and experimental repeat ID, respectively. Cells come from n = 6 experimental repeats, with pooled n cells > 1000. Separate linear regression lines were fitted to NUMB +EX and NUMB -EX-expressing cells. Rectangle indicates where the representative cells (Figure Supp. Figure 7D) fall within the overall population.* ***D)*** *Representative images of HeLa cells expressing mCherry-NUMB isoforms. The expression level (white text) and its relation to the whole cell population is indicated (rectangle, Supp. Figure 7C). The mCherry isoform signal is in magenta and the Golgin97 signal, reflecting the trans-Golgi network, is shown in cyan. Scale bar: 15µm.*

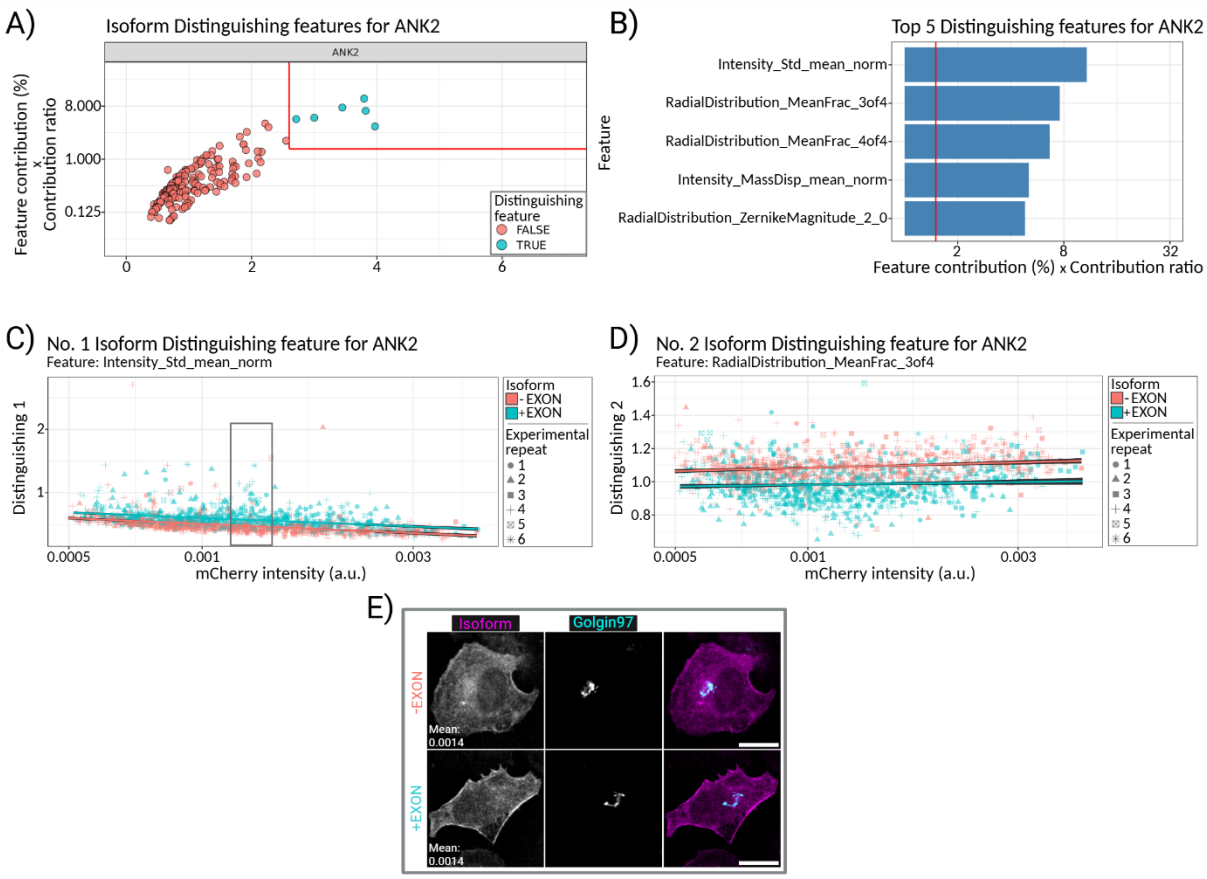


Supp. Figure 8. Two OCRL isoforms differ in their localisation to vesicular structures, replicating previous reports. A) *XY plot of Contribution ratio and Contribution ratio multiplied with Feature contribution (%) for the ISO classifier for each feature for the OCRL isoform pair, used to identify Isoform Distinguishing features. Of ~200 features, 7 were identified as Isoform Distinguishing.* ***B)*** *Top 5 Isoform Distinguishing features according to their Feature contribution (%) multiplied with Contribution ratio for the OCRL isoform pair. The threshold value defining Isoform Distinguishing features is indicated as a red line.* ***C)*** *XY plot of Top 1 Isoform Distinguishing Feature for the OCRL isoforms (mCherry Granularity) against cell mean mCherry intensity levels (proxy for expression level). Each data point corresponds to a cell with colour and shape indicating OCRL isoform identity and experimental repeat ID, respectively. Cells come from n = 3 experimental repeats, with pooled n cells > 700. Separate linear regression lines were fitted to OCRL +EX and OCRL -EX-expressing cells. Rectangle indicates where the representative cells (Figure Supp. Figure 8E) fall within the overall population.* ***D)*** *XY plot of Top 2 Isoform Distinguishing Feature for the OCRL isoforms (mCherry intensity MAD, normalised to mean mCherry intensity) against cell mean mCherry intensity levels (proxy for expression level). Each data point corresponds to a cell, same as Supp. Figure 8C.* ***E)*** *Representative images of HeLa cells expressing mCherry-OCRL isoforms. The expression level (white text) and its relation to the whole cell population is indicated (rectangle, Supp. Figure 8D). The mCherry isoform signal is in magenta and the Golgin97 signal, reflecting the trans-Golgi network, is shown in cyan. Scale bar: 15µm.*

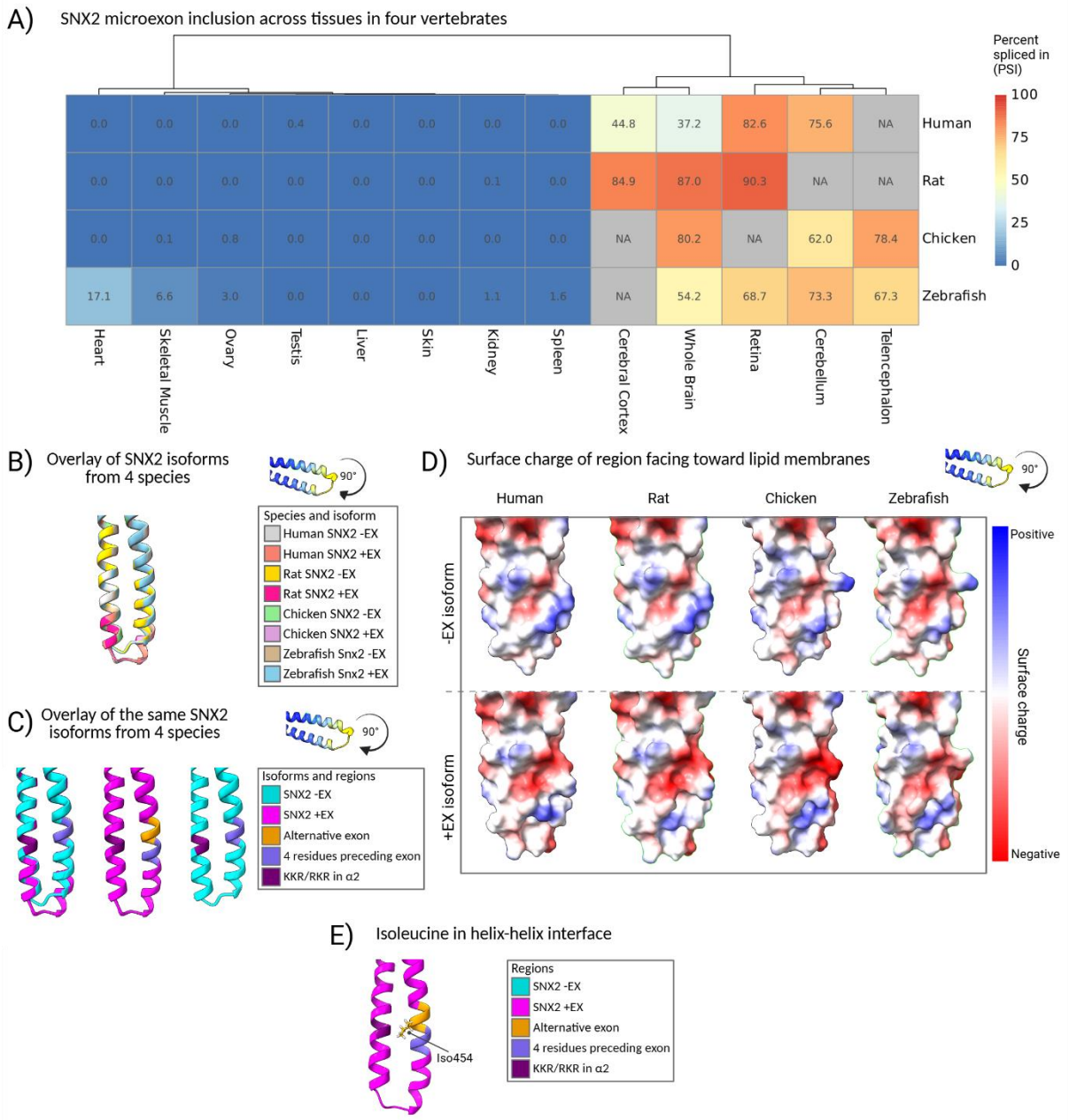


Supp. Figure 9. Most Isoform Distinguishing features are unique to single isoform pairs across the 16 isoform pairs with detected isoform differences. A) *Bar plot showing the number of features used for classifier evaluation found 'Distinguishing' ≥1 for the 16 isoform pairs where the classifier-based analysis indicated isoform differences.* ***B)*** *Histogram showing the number of different Isoform Distinguishing feature types from each of the major feature categories for the 16 isoform pairs where the classifier-based analysis indicated isoform differences.* ***C)*** *Histogram of how many times a feature is detected as Distinguishing across the 16 isoform pairs where the classifier-based analysis indicated isoform differences. Only features Distinguishing ≥1*

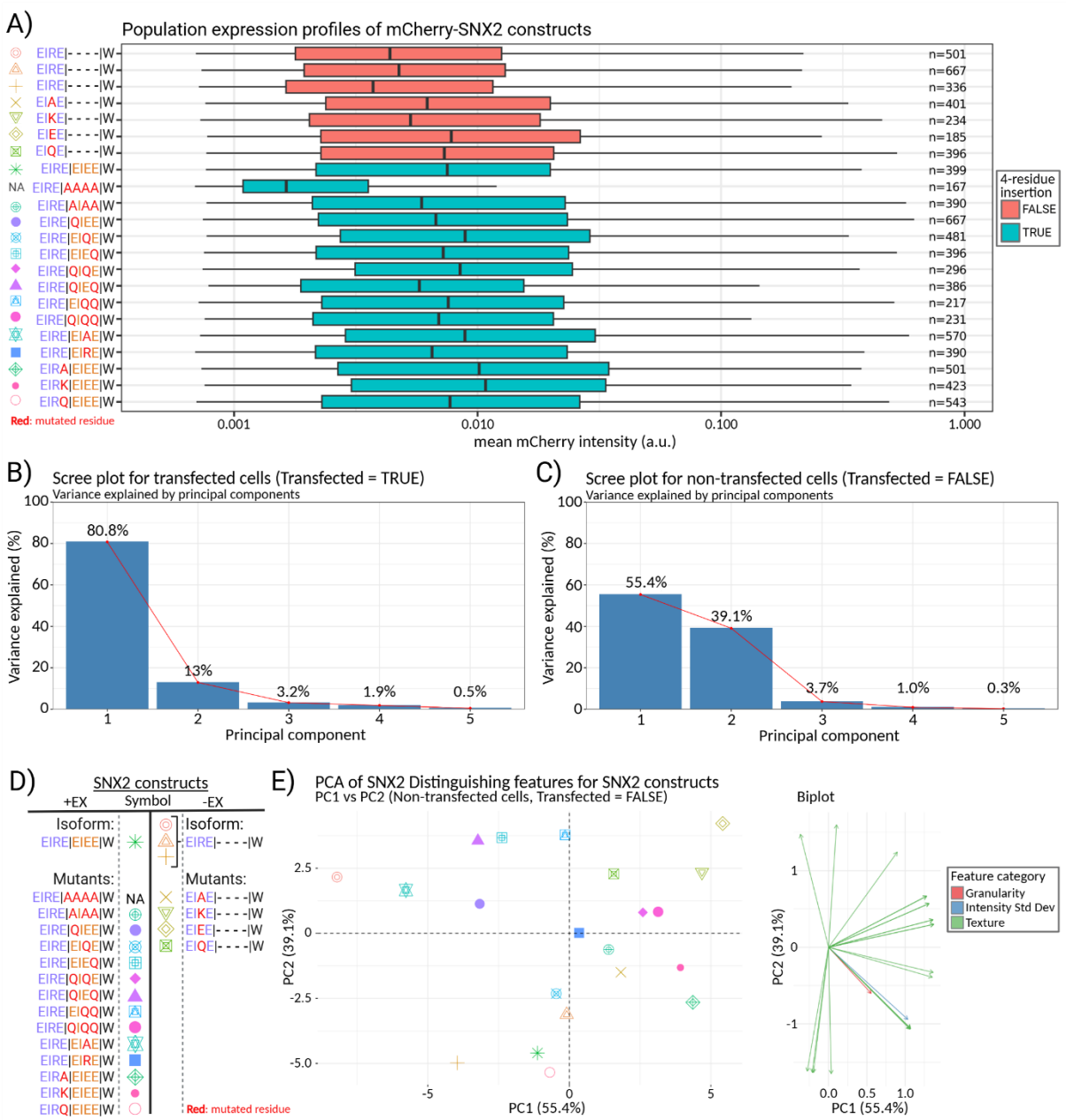
are shown. ***D)*** *Bar plot showing the proportion of minor feature groups in each major feature category (Y-axis) that are recurring Distinguishing feature (Distinguishing ≥2).* ***E)*** *Importance of the recurrent Isoform Distinguishing features (Distinguishing ≥2) for the 16 isoform pairs where the classifier-based analysis indicated isoform differences and the respective feature (Y-axis) was Isoform Distinguishing. Importance is quantified as the Feature contribution (%) (ISO) multiplied with the Contribution ratio (X-axis). Each data point represents the importance for a specific isoform pair, indicated by colour. n indicates the number of isoform pairs the given feature is Distinguishing for.*



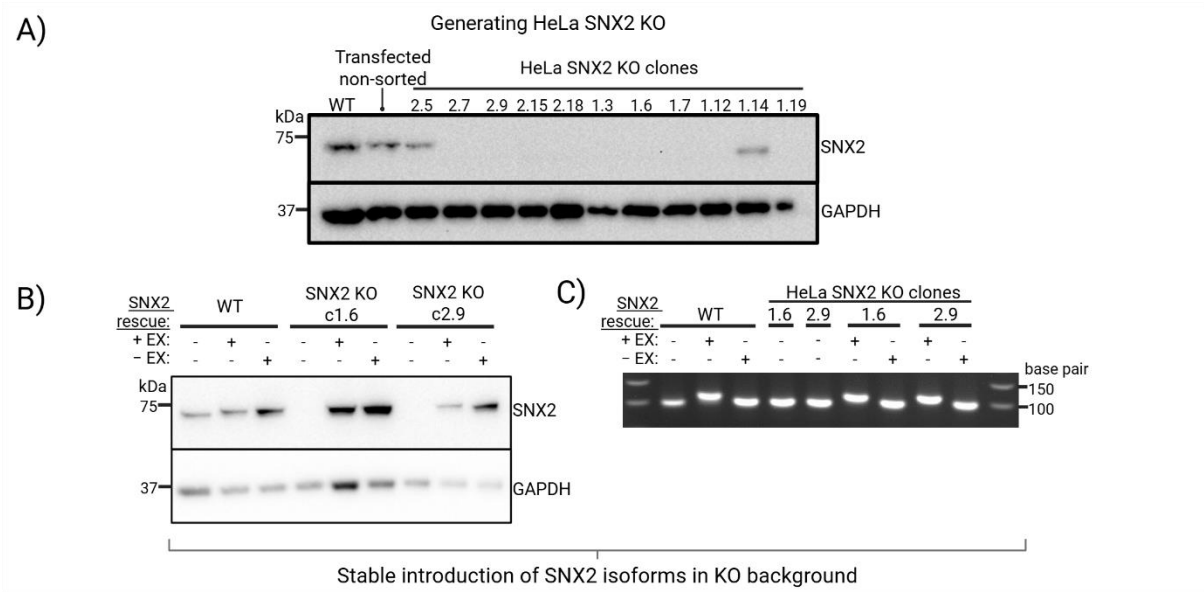
Supp. Figure 10. Two ANK2 isoforms differ in their localisation to the cell periphery/membrane. A) *XY plot of Contribution ratio and Contribution ratio multiplied with Feature contribution (%) for the ISO classifier for each feature for the ANK2 isoform pair, used to identify Isoform Distinguishing features. Of ~200 features, 6 were identified as Isoform Distinguishing.* ***B)*** *Bar plot of top 5 Distinguishing features according to their Feature contribution (%) x Contribution ratio for the ANK2 isoform pair. The threshold value defining Isoform Distinguishing features is indicated as a red line.* ***C)*** *XY plot of Top 1 Isoform Distinguishing feature for the ANK2 isoforms (mCherry intensity standard deviation, normalised to mean mCherry intensity) against cell mean mCherry intensity levels (proxy for expression level). Each data point corresponds to a cell with colour and shape indicating ANK2 isoform identity and experimental repeat ID, respectively. Cells come from n = 6 experimental repeats, with pooled n cells > 900. Separate linear regression lines were fitted to ANK2 +EX and ANK2 -EX-expressing cells. Rectangle indicates where the representative cells (Figure Supp. Figure 10E) fall within the overall population.* ***D)*** *XY plot of Top 2 Isoform Distinguishing feature for the ANK2 isoforms (mCherry radial distribution, signal in fraction 3 of 4) against cell mean mCherry intensity levels (proxy for expression level). Each data point corresponds to a cell, same as Supp. Figure 10C.* ***E)*** *Representative images of HeLa cells expressing mCherry-ANK2 isoforms. The expression level (white text) and its relation to the whole cell population is indicated (rectangle, Supp. Figure 10C). The mCherry isoform signal is in magenta and the Golgin97 signal, reflecting the trans-Golgi network, is shown in cyan. Scale bar: 15µm.*



Supp. Figure 11. SNX2's microexon show similar usage across human, rat, chicken, and zebrafish, and is predicted to affect structure in a highly similar manner. A) *Heatmap displaying usage of the SNX2 microexon in 13 tissues across 4 species (human, rat, chicken, zebrafish), quantified as percent spliced-in (PSI). The PSI values were extracted from vastDB, where bulk RNA-sequencing data has been used to determine PSI for the SNX2 microexons.* ***B)*** *Overlay of AlphaFold2 models of the SNX2 BAR region containing residues 419 - 464 (human SNX2 residue numbering) of both SNX2 isoforms from the four species where inclusion was examined. The BAR domain is viewed from the side that faces lipid membranes. Each species isoform combination has a distinct colour.* ***C)*** *Same as Supp. Figure 11C), but in addition to have both SNX2 isoforms for each of the four species overlaid, the isoforms from the four species are also separated to reflect isoform identity. Colours reflect isoform (-EX cyan, +EX magenta), microexon (orange), the four residues preceding the microexon (lavender), and the KKR motif (dark purple).* ***D)*** *Surface charge of AlphaFold2 models of SNX2 isoforms from the four species, in the BAR region containing residues 419 - 464 (human SNX2 residue numbering). The BAR domain is viewed from the side that faces lipid membranes. Colour indicates charge (blue, positive. Red, negative).* ***E)*** *AlphaFold2 model showing the position of I454's sidechain, encoded by SNX2's microexon, for human SNX2 +EX isoform in the BAR region containing residues 419 - 464. The BAR domain is viewed from the side that faces lipid membranes. The cassette exon is orange, the four residues preceding the exon are lavender, and the KKR motif (426-428) is dark purple.*



Supp. Figure 12. All but one mCherry-SNX2 constructs express at comparable levels and the PCA clustering driven by well-specific effects. A) *Boxplots showing single cell mCherry mean intensity values in transfected cells expressing the various mCherry-SNX2 isoforms and mutants. Cells were pooled from three experimental repeats (n = 3 independent experiments). The number of cells after pooling is indicated for each construct (n).* ***B)*** *Scree plot showing principal component (PC) variance in PCA of SNX2 isoform Distinguishing Features from transfected cells expressing SNX2 isoforms and mutants, relates to Figure 4G.* ***C)*** *Scree plot showing PC variance in PCA of SNX2 isoform Distinguishing Features from non-transfected cells in the same wells as the SNX2 isoform/mutant-expressing cells, relates to Supp. Figure 12E.* ***D)*** *List of SNX2 isoform and mutant constructs examined via high-content imaging. Same as Figure 4F.* ***E)*** *Left, XY plot showing PC1 versus PC2 for principal component analysis (PCA) of the mCherry-SNX2 construct imaging data using the population means of the SNX2 isoform Distinguishing features from the non-transfected cells as input. Right, biplot showing the rotation of the Distinguishing features. Different feature classes are shown. Granularity is shown in red.*



Supp. Figure 13. Generation of HeLa cell lines expressing select SNX2 isoforms. A) *Western blot against human SNX2 and human GAPDH in lysates from HeLa cell clones isolated after SNX2 knockout in parental HeLa population. Experiment was performed once.* ***B)*** *Western blot against human SNX2 and human GAPDH in lysates from wild type HeLa cells or HeLa SNX2 knockout clones, where SNX2 -EX or SNX2 +EX had subsequently been re-introduced via lentiviral transduction. Experiment was performed once.* ***C)*** *End-point RT-PCR amplification of SNX2 region that flanks the SNX2 microexon in RNA from wild type HeLa cells or HeLa SNX2 lines. The PCR products were run on agarose gels, and a slight reduction in band mobility was visible upon microexon inclusion (12nt). The presence of PCR product in KO cells may stem from low amounts of SNX2 mRNA that contains the amplified region, as KO was directed at locus upstream of the amplified region. Experiment was performed once.*