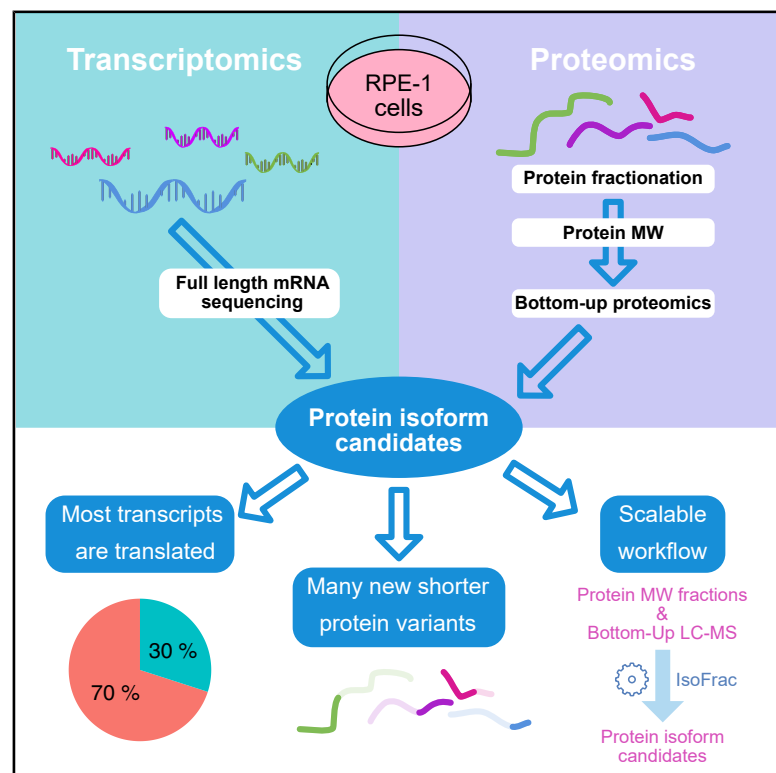


An integrated landscape of mRNA and protein isoforms

Graphical abstract



Authors

Amir Kedan, Henrik Zauber, Meng-Ran Wang, ..., Kathryn S. Lilley, Wei Chen, Matthias Selbach

Correspondence

chenw@sustech.edu.cn (W.C.), matthias.selbach@mdc-berlin.de (M.S.)

In brief

Kedan and Zauber et al. combine long-read transcriptomics with molecular-weight-resolved proteomics to generate an integrated landscape of mRNA and protein isoforms. Their IsoFrac framework enables scalable protein isoform discovery, revealing that numerous shorter protein variants contribute substantially to human proteome complexity.

Highlights

- IsoFrac enables scalable proteome-wide identification of protein isoform candidates
- Integrated long-read transcriptomics links mRNA and protein isoform diversity
- Most detectable alternative transcript isoforms are translated
- Shorter protein variants substantially expand proteome complexity



Resource

An integrated landscape of mRNA and protein isoforms

Amir Kedan,^{1,7} Henrik Zauber,^{1,7} Meng-Ran Wang,^{2,6} Suyeon Kim,³ Qionghua Zhu,^{2,4} Liang Fang,^{2,4} Kathryn S. Lilley,³ Wei Chen,^{2,4,*} and Matthias Selbach^{1,5,8,*}

¹Max Delbrück Center for Molecular Medicine, Berlin, Germany

²Department of Systems Biology, School of Life Science, Southern University of Science and Technology, Shenzhen, China

³Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge, UK

⁴Guangming Advanced Research Institute, Southern University of Science and Technology, Shenzhen, China

⁵Charité-Universitätsmedizin Berlin, Berlin, Germany

⁶Faculty of Synthetic Biology, Shenzhen University of Advanced Technology, Shenzhen, China

⁷These authors contributed equally

⁸Lead contact

*Correspondence: chenw@sustech.edu.cn (W.C.), matthias.selbach@mdc-berlin.de (M.S.)

<https://doi.org/10.1016/j.celrep.2026.117898>

SUMMARY

Alternative splicing and proteolytic processing expand proteome diversity by generating distinct protein isoforms from a single gene. However, the relationship between transcript isoforms and protein products remains poorly understood because of limitations in current proteomic workflows. Here, we combined full-length mRNA sequencing with protein fractionation and quantitative mass spectrometry to generate an integrated landscape of mRNA and protein isoforms in human RPE-1 cells. To overcome the ambiguity of bottom-up proteomics, we developed IsoFrac, a computational pipeline that resolves protein isoforms from molecular-weight-resolved peptide migration profiles. Using this approach, we identified ~45,000 full-length transcripts, ~32,000 open reading frames (ORFs), and ~14,000 protein isoform candidates. Comparative analyses revealed widespread translation of alternative transcripts and identified shorter protein variants, likely arising from proteolytic processing and/or alternative translation, as a major and underappreciated source of proteome complexity. Our results establish a scalable framework for isoform-resolved proteogenomics and provide a resource for studying protein isoform diversity.

INTRODUCTION

Gene expression is a complex and tightly regulated process that transforms the genetic information into functional proteins.^{1,2} This process begins with the transcription of genes into mRNAs and their subsequent translation into proteins and is further diversified by the generation of multiple mRNA and protein isoforms from individual genes. Posttranscriptional processes, such as alternative splicing and alternative polyadenylation, give rise to diverse mRNA species,^{3–6} while additional protein diversity arises through alternative translation initiation and post-translational modifications.^{7–9} It is estimated that an average cell line may contain over one million distinct proteoforms, though this number varies by cell type and tissue.

While long-read sequencing has provided a detailed view of transcriptome diversity, the global characterization of protein isoforms remains incomplete.^{10,11} A persistent discrepancy exists between the number of mRNA isoforms detected by RNA sequencing and the number of protein isoforms identified by proteomics. This raises two key questions: (1) to what extent do alternative transcripts yield functional protein products? and (2) how many additional protein isoforms arise through

mechanisms independent of mRNA variation, such as alternative translation initiation or proteolytic processing? One hypothesis suggests that many alternative transcripts are not translated and instead represent non-functional or regulatory RNA species,^{12,13} whereas another proposes that most splice variants contribute to proteome diversity.^{3,5} An alternative explanation is that technical limitations of current proteomic workflows prevent reliable detection of many protein isoforms. Resolving this discrepancy requires integrated transcriptomic and proteomic approaches.

Here, we refer to protein isoforms as distinct polypeptides arising from alternative transcripts, alternative translation, or proteolytic cleavage, distinguishing them from the broader class of proteoforms, which additionally includes post-translational modifications.^{7,8}

Current mass spectrometry (MS)-based approaches are broadly divided into top-down and bottom-up methods, each with distinct strengths and limitations. Top-down proteomics directly analyzes intact proteins and is therefore well suited for protein isoform identification,¹⁴ but still lacks the proteome coverage achieved by bottom-up techniques.^{15,16} In contrast, bottom-up proteomics routinely identifies more than 10,000



protein groups in a single experiment,¹⁷ but its ability to distinguish protein isoforms is limited because many peptides are shared among multiple isoforms.¹⁸

Most efforts to improve protein isoform identification have focused on deeper proteome sequencing and computational methods to extract isoform information from bottom-up datasets.^{19–23} For example, a recent deep bottom-up proteomics study employing six proteases and three tandem mass spectrometry fragmentation methods identified approximately one-third of alternative splice variants at the protein level and up to 64% among highly abundant transcripts.²⁴ Thermal proteome profiling has further enabled annotation-independent detection of functional proteoform groups based on peptide melting behavior.²⁵ However, systematic integration of long-read transcriptomics with proteome-wide protein isoform detection in the same biological sample has not yet been achieved.

Here, we present an integrated proteogenomic approach that combines SDS-PAGE-based protein separation with long-read transcriptome sequencing to systematically map mRNA and protein isoforms in the same cell type. By resolving proteins according to their intact molecular weight before mass spectrometry, we enhance the detection of protein isoforms while maintaining proteome-wide coverage. Although SDS-PAGE has traditionally been used to reduce sample complexity, its potential for protein isoform analysis is receiving renewed attention.^{26,27} Using this strategy in human RPE-1 cells together with full-length transcriptome profiling, we establish a matched landscape of transcript and protein isoforms. Our integrated proteogenomic analysis indicates that a large fraction of alternative transcripts are translated and highlights shorter protein variants generated by alternative translation and/or proteolytic processing as a major, underappreciated contributor to proteome complexity.

RESULTS

Peptide Correlation Profiling resolves protein-level information from peptide data

Correlating protein abundance across biochemical fractions under non-denaturing conditions, termed protein correlation profiling (PCP), is a well-established approach for mapping proteins to subcellular compartments and multiprotein complexes.²⁸ For example, fractionating proteins via sucrose density gradient centrifugation, gel filtration or non-denaturing gel electrophoresis yields protein abundance profiles across fractions. The distribution of correlated proteins across these fractions then provides information about their subcellular localization and/or multiprotein complexes of which they are members.^{29–36}

Building on the principles of PCP, we use peptide correlation profiling (PepCP) as a strategy to extract protein-level information from peptide-centric (bottom-up proteomic) data (Figure 1A). PepCP follows four key steps: (1) proteins are separated by SDS-PAGE based on their molecular weight into fractions.³⁷ (2) Each fraction is enzymatically digested into peptides. (3) Peptides are identified and quantified across all fractions using quantitative mass spectrometry-based proteomics. (4) Peptide-abundance profiles across fractions are analyzed to infer protein-level information.

We extracted proteins from the diploid human retinal pigmented epithelial cell line RPE-1 and fractionated them by SDS-PAGE using the GELFrEE system, which improves protein recovery.³⁸ To cover isoforms across molecular weight ranges we used gels with 10%, 8%, and 5% acrylamide. A total of 28 fractions per gel were collected, and proteins were digested using an automated SP3 workflow.³⁹ Peptides were labeled with tandem mass tags (TMT) to enable quantification across fractions.⁴⁰ To achieve this, the 28 fractions were combined into 2 TMT-16 plexes, each containing 14 fractions, along with an unfractionated whole-cell lysate (WCL) reference channel. This design allowed peptide abundance profiles across SDS-PAGE fractions to be quantified as ratios relative to the reference channel. Combined peptide mixtures were separated offline by high-pH reverse-phase high-performance liquid chromatography (HPLC) and subsequently analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS).

In total, we identified ~150,000 peptides (1% false discovery rate [FDR]) in ~13,000 protein groups corresponding to ~12,000 genes (Figure 1C). After restricting the analysis to uniquely mapping (genotypic) peptides, ~130,000 peptides representing 9,288 genes were retained for downstream analyses. Note that isoforms arising from alternative translation or proteolytic cleavage events reported in UniProt were included in the search (see STAR Methods).

To evaluate the resolving power of the SDS-PAGE separation, we projected the annotated molecular weights of identified proteins across all fractions (Figure 1B). As expected, protein molecular weight increased with fraction number, spanning approximately 15–250 kDa, while the accessible mass range increased with decreasing gel percentage.

The premise of PepCP is that peptides derived from the same protein isoform should display correlated abundance profiles across fractions. Accordingly, pairwise peptide correlations within genes showed a strong positive peak around $r = 0.9$, whereas randomly paired peptides were centered around zero (Figure 1D). Receiver operating characteristic (ROC) analysis confirmed high discrimination between true and random peptide pairs, with area under the curve (AUC) values of 0.93 and 0.94 for the 8% and 10% gels, respectively (Figure 1E). These estimates are conservative because peptides shared between multiple protein isoforms reduce within-gene correlations.

Having established that PepCP reliably groups peptides originating from the same protein species, we next assessed its ability to distinguish protein isoforms. Intronless genes, which are expected to encode a single protein isoform,⁴¹ predominantly produced a single peak (Figure 2A). In contrast, PepCP resolved protein isoforms generated by alternative splicing (CRK; Figure 2B), alternative translation initiation (CEBPB; Figure 2C), and proteolytic processing (NF- κ B1; Figure 2D), demonstrating its ability to detect diverse mechanisms contributing to proteome complexity.

Mapping the protein isoform landscape

The examples above demonstrate that PepCP resolves protein isoforms at the peptide-level. To analyze these data systematically, we developed IsoFrac, a computational pipeline that automatically assigns peptides to protein isoforms while remaining

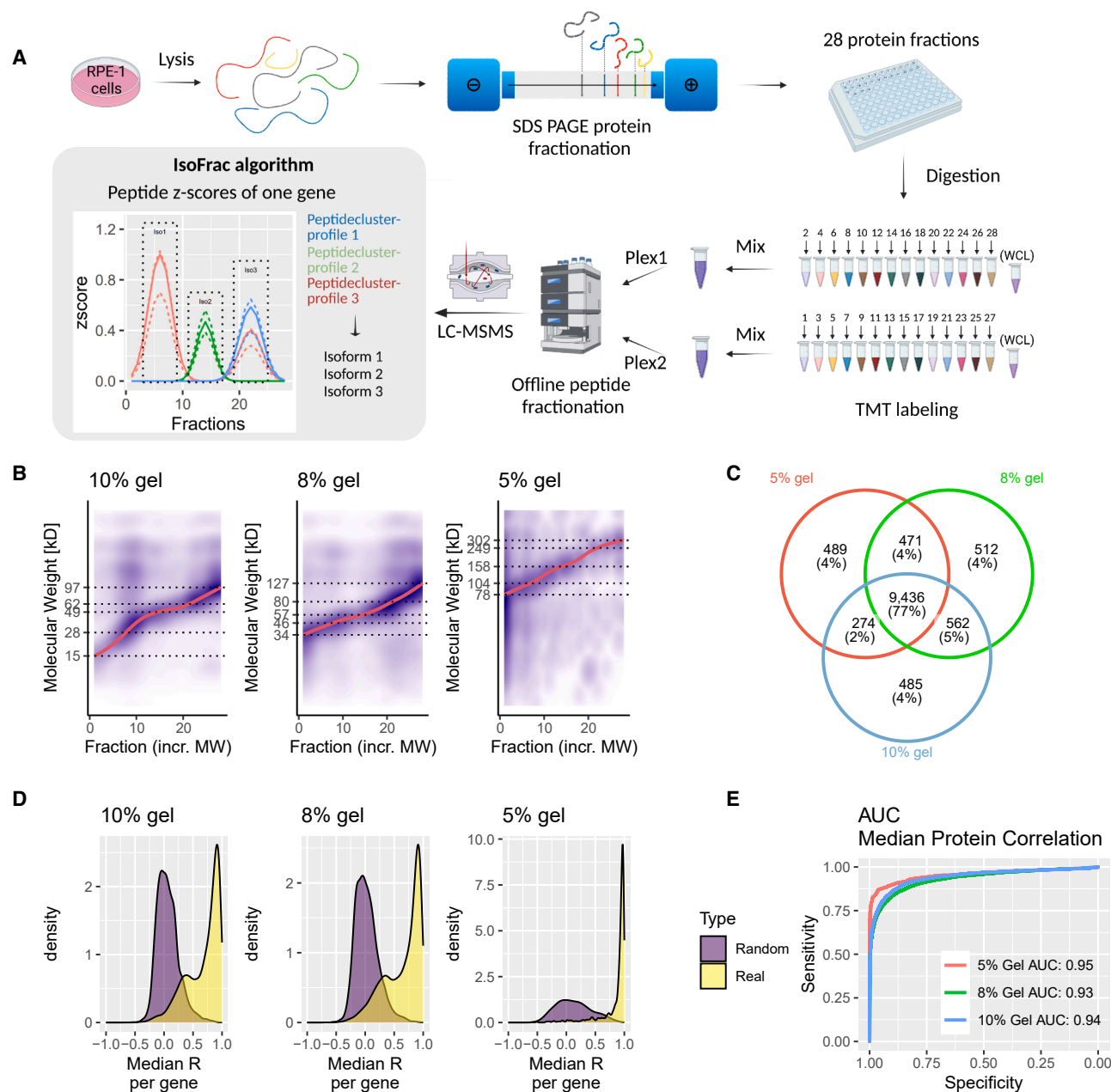


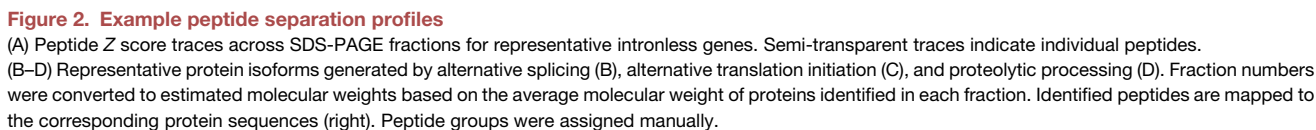
Figure 1. PepCP enables protein identifications from peptide data

(A) PepCP workflow.
 (B) Molecular weight distribution across fractions for the 10%, 8%, and 5% SDS-PAGE gels. Protein molecular weights were inferred from the major protein assignment of each identified peptide. Loess regression lines are shown.
 (C) Overlap of identified GENCODE gene IDs across the three gels.
 (D) Distribution of median Pearson correlations between peptide pairs from the same gene (real) or randomly paired peptides (random).
 (E) Receiver operating characteristic (ROC) curves derived from the distributions in (D). AUC values are indicated.

robust to noisy peptide fractionation profiles (Figure 3A and see STAR Methods).

Briefly, IsoFrac consists of three steps: (1) peptide traces are clustered using the density-based algorithm HDBSCAN to identify peptides with similar fractionation profiles,⁴² (2) peaks are detected from the average profile of each cluster, and (3) peptide

clusters with nearby peak positions are merged into protein isoforms. Isoforms are first identified separately for each gel (10%, 8%, and 5%) and subsequently merged across gels based on their predicted molecular weights. Applying this workflow together with additional quality filters (see STAR Methods) identified ~14,300 protein isoform candidates from ~8,100 genes,



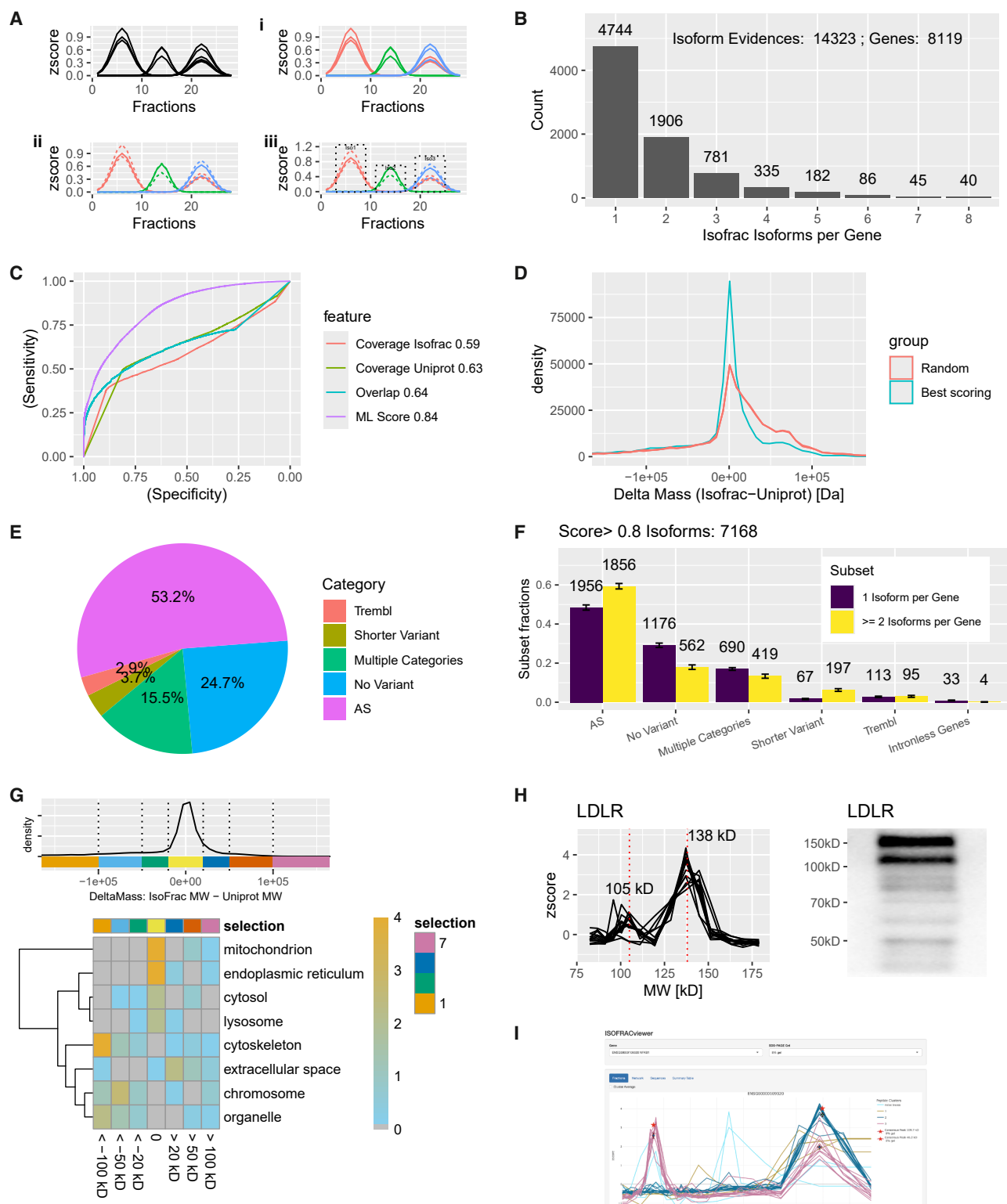


Figure 3. IsoFrac identifies protein isoforms on a global scale

(A) Schematic overview of the IsoFrac workflow. Peptide traces are clustered using HDBSCAN, cluster-average traces are generated, and peaks are identified and merged across clusters and gels.

(B) Distribution of identified protein isoforms per gene.

(legend continued on next page)

with roughly half of the genes producing more than one isoform (Figure 3B).

We next assessed whether IsoFrac reflects known protein isoform annotations. To this end, we compared each IsoFrac isoform to all annotated UniProt isoforms for the same gene using three features: (1) IsoFrac coverage—the percentage of IsoFrac peptides that match a given UniProt isoform, (2) UniProt coverage—the fraction of theoretical peptides of a given UniProt isoform that are detected in the corresponding IsoFrac peak, and (3) Overlap—a hypergeometric test to assess the significance of the overlap between UniProt and IsoFrac peptides. As a negative control, we randomly assigned peptides from the same gene to IsoFrac isoforms. A gradient-boosted classification model integrating these features substantially improved assignment accuracy over the individual metrics, achieving an AUC of 0.84 (Figure 3C). Because UniProt annotations do not fully capture protein isoform diversity, this value likely represents a conservative estimate.

As an independent validation, we compared the molecular weights of IsoFrac isoforms with their best-scoring UniProt assignments. Although randomized assignments also exhibited a peak around $\Delta\text{mass} = 0$, reflecting the limited search space of candidate isoforms within each gene, the best-scoring assignments were more strongly enriched around zero and displayed a narrower distribution (Figure 3D). Because molecular weight was not used during the scoring procedure, this provides orthogonal validation of the IsoFrac assignments.

Next, we classified high-confidence IsoFrac isoforms according to their best-matching UniProt annotation (Figure 3E). Alternative splicing represented the largest category, followed by proteins with a single reviewed UniProt entry (no variant). Consistent with this, genes annotated with multiple UniProt isoforms were more likely to yield multiple IsoFrac isoforms, whereas genes with a single annotated protein predominantly produced one IsoFrac isoform (Figure 3F). Likewise, 33 of 37 intronless genes in the high-confidence set (score > 0.8) produced only a single isoform. Together, these trends are consistent with known mechanisms of isoform generation and support the biological relevance of the IsoFrac-defined isoforms.

Gene Ontology analysis across the Δmass distribution revealed distinct functional trends (Figure 3G). Proteins localized to the endoplasmic reticulum (ER) and plasma membrane, which are frequently glycosylated, were enriched among positive Δmass values, consistent with increased apparent molecular weights caused by large post-translational modifications. For example, low-density lipoprotein receptor (LDLR) displayed two IsoFrac peaks with identical peptide composition, which were confirmed

by western blotting (Figure 3H). Conversely, shorter protein variants were enriched among cytoskeletal proteins.

Taken together, the IsoFrac dataset offers a comprehensive resource for global and gene-specific isoform analysis. To facilitate exploration of the dataset, we developed an interactive IsoFrac viewer that enables gene-specific inspection of peptide traces and protein isoforms (Figure 3I; <https://proteinisoforms.mdc-berlin.de>).

Long-read sequencing connects transcriptome complexity to protein isoform diversity

To complement our proteomic isoform analysis and capture potential RPE-1-specific splice variants, we characterized the RPE-1 transcriptome using an integrated long-read (PacBio) and short-read (next generation sequence, NGS) RNA-sequencing pipeline (Figure S1). After integration with GENCODE (Figure 4A), we identified 45,223 protein-coding transcripts comprising 32,612 open reading frames (ORFs) across 11,605 genes (Figure 4B). Of these, 20,420 transcripts (45%) were detected in both PacBio and GENCODE, 15,348 were supported only by GENCODE and NGS, and 7,652 (17%) were identified exclusively by PacBio, representing potential novel transcript isoforms. As an example, the integrated annotation of Thymopoietin (TMPO) resolved seven transcripts encoding four ORFs, corresponding to the four major TMPO protein isoforms identified by IsoFrac (Figures 4C and 4D).

Next, we repeated the IsoFrac assignment procedure using a custom protein database derived from the PacBio ORFs. Applying the same regression model used for UniProt assignments yielded an AUC of 0.85, slightly exceeding the performance obtained with UniProt (AUC = 0.84; Figure 4E).

The integrated transcriptome and proteome dataset enabled us to directly assess the contribution of alternative splicing to proteome complexity.^{5,12} Protein isoforms were exclusively assigned to their best-matching ORFs, yielding matches for approximately 43% of the 22,022 ORFs detected in both datasets (Figure 4F, left), which should be interpreted as a method-dependent mapping rate and not as definitive validation of all corresponding protein isoforms. At first glance, this might suggest that only a fraction of alternatively spliced transcripts are translated into stable protein products. However, a direct comparison between transcriptomic and proteomic datasets must account for technical differences between the two methods. In particular, SDS-PAGE cannot resolve protein isoforms with similar molecular weights. Collapsing ORFs differing by less than ± 5 kDa increased the overlap to more

(C) ROC curves for single metrics and combined regression score to assess the quality of isofrac uniprot pairs. Area under the curve values (AUC) for each metric are depicted in the legend.

(D) Δmass distributions for best-scoring and randomized UniProt assignments.

(E) Distribution of high-confidence IsoFrac protein isoforms (score > 0.8) across UniProt annotation categories. AS, alternative splicing; no variant, single reviewed UniProt entry; shorter variant, proteolytic cleavage or alternative translation; TrEMBL, unreviewed UniProt entries. Isoforms with ambiguous assignments are classified as combined categories.

(F) Comparison of protein fractions from subfigure E for genes with ≥ 2 isoforms in Uniprot (purple) against genes with a single entry (yellow) in Uniprot. Categories are based on the UniProt annotation status as used in (E) and expanded with the category of identified 'intron-less genes'. Error bars indicate the 5th and 95th percentiles derived from bootstrap resampling.

(G) Gene Ontology enrichment across Δmass bins for high-confidence (score > 0.8) IsoFrac isoforms. Color scale indicates $-\log_{10}p$ values.

(H) LDLR peptide traces showing two protein species (~ 105 and ~ 138 kDa) together with the corresponding western blot.

(I) Screenshot of the IsoFrac Viewer.

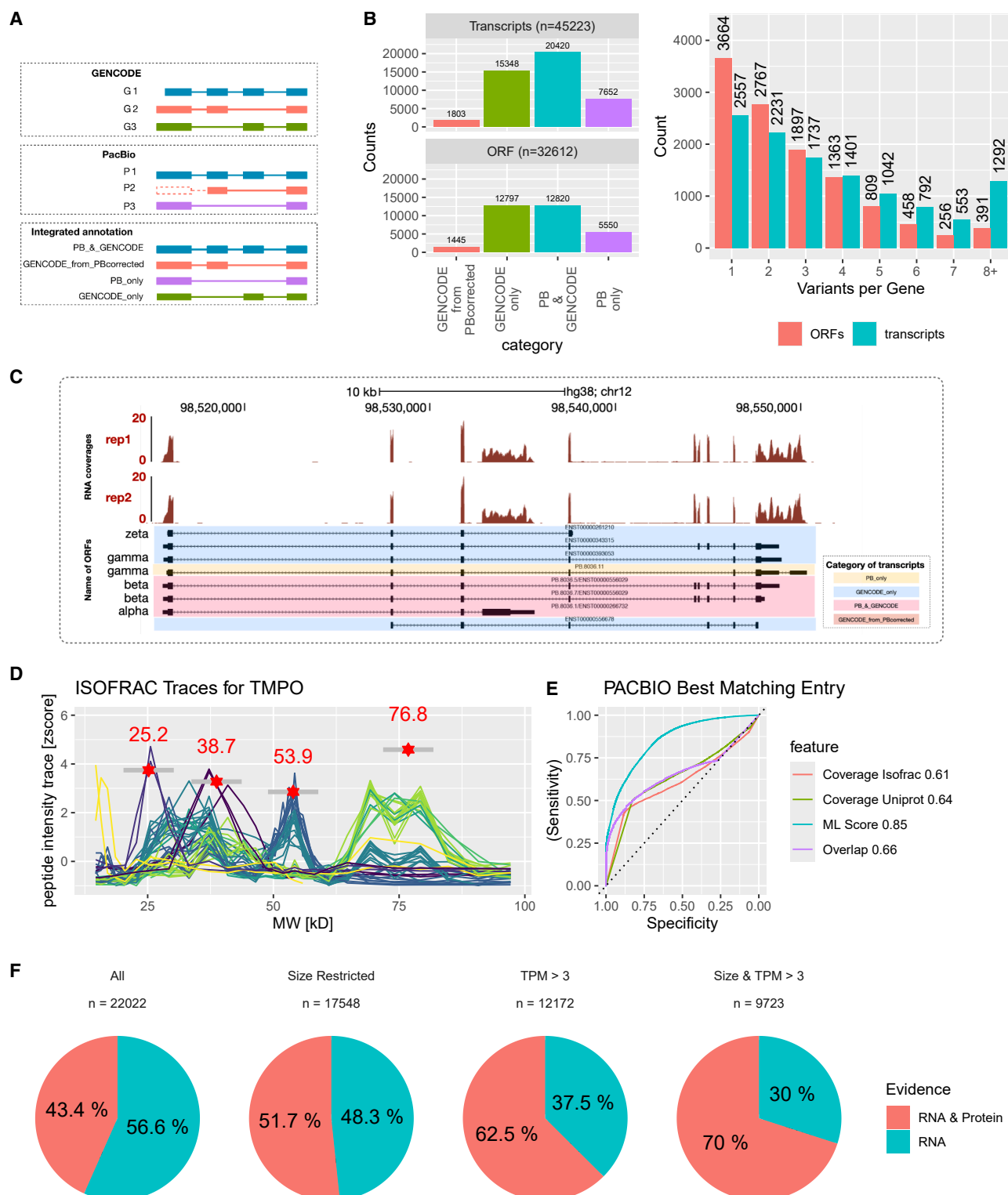


Figure 4. Integration of PacBio transcriptome sequencing

(A) Strategy for integrating PacBio-derived transcripts with GENCODE annotations.

(B) Number of transcripts and ORFs across annotation categories (left) and transcript variants per gene (right).

(C) Visualization of RNA-seq coverage and isoforms for TMPO gene. Tracks in red: RNA-seq coverages from NGS, with two replicates; blocks in black: transcripts in integrated annotation for TMPO genes, categories of each transcript are labeled with different colors. Names of ORFs are labeled on the left.

(legend continued on next page)

than 51% (Figures 4F and S2). Because low-abundance proteins are also more difficult to detect, restricting the analysis to transcripts with TPM (Transcripts Per Kilobase Million) > 3 further increased the overlap to over 60%, and combining both corrections increased the mapping rate to 70%. Together, these results indicate that most alternatively spliced transcripts detectable with our approach are translated into stable proteins, consistent with previous deep proteomic studies.²⁴

A substantial fraction of unannotated protein isoforms represent shorter variants

Our dataset provides evidence for previously unannotated protein isoforms, as several genes exhibited multiple IsoFrac isoforms despite having only a single annotated protein (Figure 3F). To determine whether these isoforms resulted from RPE-1-specific alternative splicing, we mapped each IsoFrac isoform to both UniProt- and the PacBio-derived database. Approximately 80% of isoforms showed identical assignment scores, indicating equivalent support from both databases (Figure 5A) and that most protein isoforms are already represented by existing transcript annotations. Only a few hundred isoforms aligned more strongly with PacBio, suggesting that RPE-1-specific alternative splicing accounts for only a small fraction of the observed protein isoforms. In contrast, isoforms preferentially assigned to UniProt were significantly enriched in the shorter variant category (Figure 5A), consistent with the PacBio database containing only transcript-derived ORFs and therefore lacking protein-level processing events.

We next compared the observed molecular weights of IsoFrac isoforms with their best-matching UniProt and PacBio entries ($\Delta\text{mass} = \text{IsoFrac MW} - \text{Database MW}$). For over 3,500 isoforms, the observed molecular weight closely matched the corresponding UniProt and PacBio annotations (Figure 5B). Isoforms were classified using an empirically selected tolerance of approximately ± 15 kDa around the expected molecular weight, chosen to balance specificity and sensitivity while applying a comparable tolerance to both the “correct mass” and “same offset” categories (see STAR Methods). Isoforms matching UniProt but not PacBio were enriched in the shorter variant category, demonstrating that IsoFrac efficiently detects known shorter protein variants. For most remaining isoforms, similar Δmass shifts relative to both databases suggested that alternative splicing alone cannot explain the majority of unannotated protein isoforms, implicating alternative translation and/or post-translational processing. Although Δmass provides a useful global benchmark, interpretation of individual isoforms requires orthogonal validation because assignments are probabilistic and molecular weights are estimated from gel migration.

To assess evidence for previously unannotated splice variants, we searched for peptides unique to the PacBio-derived transcriptome. We identified seven peptides absent from UniProt that were confidently identified with scores comparable to annotated peptides (Figure S3) and mapped to splice junctions absent from

GENCODE v40 (Figure S4). Notably, several of these variants have since been incorporated into newer GENCODE (e.g., FMNL2) and UniProt (e.g., CTBP1) releases. Thus, although alternative splicing explains only a minority of novel protein isoforms, IsoFrac successfully identifies previously unannotated splice variants.

Because many novel isoforms exhibited lower molecular weights than their best-matching UniProt entries, we next investigated post-translational processing. TEAD3, for example, was detected as three isoforms; including an additional C-terminal isoform (~26 kDa) lacking the DNA-binding domain (Figures 5E and 5F). As this isoform was absent from the PacBio dataset, it could not be explained by alternative splicing. Expression of codon-optimized N- and C-terminally tagged TEAD3 constructs confirmed both the full-length protein and the additional C-terminal isoform by western blot (Figure 5G). Similar agreement between PepCP and western blotting was observed for three additional proteins (Figure S5).

To assess shorter IsoFrac variants on a proteome-wide scale, we integrated peptide-level DIA-LOP (localization of organelle proteins by data-independent acquisition) spatial proteomics⁴³ since conventional protein-level summarization would obscure isoform-specific localization patterns. Unlike conventional protein-level analyses, peptide-level DIA-LOP preserves isoform-specific information. As a proof of principle, CTSB peptides correctly resolved the known localization difference between the ER/Golgi-localized procathepsin B (~37 kDa) and the mature lysosomal form (~26 kDa) (Figure 5H). More broadly, peptide centroids assigned to the same IsoFrac isoform clustered significantly more closely than peptides assigned to different isoforms of the same gene, whereas this separation was reduced after randomizing isoform assignments (Figure 5I). These findings provide independent proteome-wide support for the biological relevance of IsoFrac-defined isoforms and indicate that many shorter protein variants occupy distinct subcellular localizations.

Together, our integrated transcriptomic and proteomic analysis confirms that alternative splicing contributes substantially to proteome complexity while identifying shorter protein variants, generated by alternative translation and/or post-translational processing, as a major and underappreciated source of protein diversity.

DISCUSSION

Despite extensive efforts to characterize transcript diversity, systematic mapping of protein isoforms has remained elusive. Here, we show that PepCP enables automated identification of over 14,000 candidate protein isoforms, providing a comprehensive view of proteome diversity in RPE-1 cells. By separating intact proteins before digestion, PepCP preserves molecular-weight information throughout the bottom-up workflow. As defined throughout this study, these protein isoforms represent distinct polypeptides arising from alternative transcripts,

(D) Isofrac peptide traces across fractions for TMPO. Identified protein isoforms are depicted by a red star and the corresponding estimated MW in kD. Horizontal error bars indicate the 10th and 90th percentiles of the bootstrap-derived peak position estimates..

(E) ROC curves for assigning IsoFrac isoforms to PacBio-derived ORFs.

(F) Fraction of ORFs detected at the transcript (green) or transcript and protein (red) levels. Mapping rates are shown before and after accounting for molecular weight resolution (± 5 kDa) and transcript abundance (TPM >3).

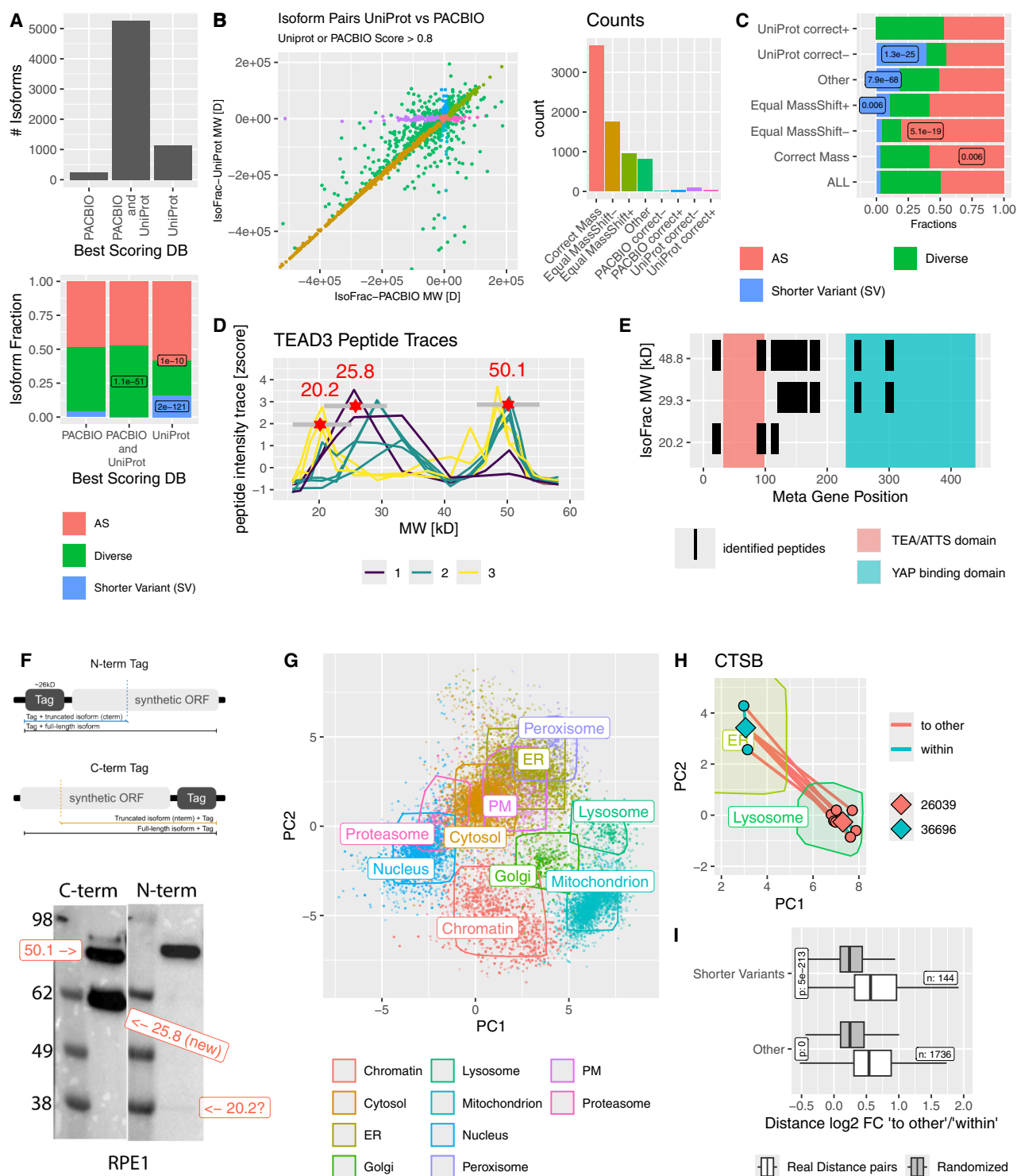


Figure 5. Validation of new protein isoforms

(A) Best-scoring IsoFrac assignments to PacBio or UniProt (top) and the corresponding distribution of UniProt isoform categories (bottom). Significant enrichments (hypergeometric test and $p < 0.01$) are indicated.

(B) Pairwise Δ mass comparison (IsoFrac – database molecular weight) between PacBio and UniProt assignments for high-confidence isoforms (score > 0.8). Isoforms are grouped according to agreement between IsoFrac, UniProt, and PacBio molecular weights (correct mass, same offset, UniProt correct, PacBio correct, and others); the corresponding counts are shown on the right.

(legend continued on next page)

alternative translation, or proteolytic cleavage and are supported by molecular weight-resolved peptide profiles and probabilistic assignment. While post-translational modifications can also alter apparent molecular weight, only larger modifications, such as glycosylation, are expected to substantially affect SDS-PAGE migration, as illustrated by LDLR (Figure 3H).

By integrating these data with long-read transcriptomics, we generated a matched landscape of mRNA and protein isoforms (<https://www.proteinisoforms.mdc-berlin.de>). Our analysis indicates that most alternatively spliced transcripts detectable with our technology are translated into proteins. In addition, we identify numerous previously unannotated protein isoforms, the majority of which are better explained by alternative translation and/or proteolytic processing than by alternative splicing. Together, these findings emphasize the importance of integrating transcriptomic and proteomic approaches to comprehensively characterize protein isoform diversity.

Protein isoform identification remains challenging because conventional bottom-up proteomics dissociates peptides from their protein of origin, giving rise to the protein inference problem.¹⁸ Ultradeep bottom-up proteomics improves sequence coverage through multiple proteases and extensive fractionation but does not fundamentally resolve ambiguity introduced by shared peptides.²⁴ Top-down proteomics directly analyzes intact proteins and can resolve protein isoforms,^{7,15,16,44} yet still lacks the proteome-wide depth of bottom-up workflows. PepCP complements these approaches by combining protein-level separation through SDS-PAGE with bottom-up proteomics, preserving molecular weight information while maintaining broad proteome coverage. Unlike methods that infer proteoforms from peptide co-variation across perturbations or thermal stability profiles,^{20,21} PepCP directly incorporates apparent protein size and enables integration with full-length transcript isoforms from the same sample.

The limited detection of protein isoforms in conventional proteomic studies has contributed to the view that many alternatively spliced transcripts do not produce stable proteins.^{12,13} After accounting for technical limitations, particularly molecular weight resolution and protein abundance, we find that the majority of detectable alternatively spliced transcripts do give rise to protein isoforms, consistent with recent ultradeep proteomic studies.²⁴

Our annotation-independent approach also provides an unbiased view of protein isoform diversity. While most annotated IsoFrac isoforms correspond to alternative splicing, many genes annotated with only a single protein nevertheless produced multiple IsoFrac isoforms, indicating that current reference an-

notations remain incomplete. Interestingly, RPE-1-specific alternative splicing accounts for only a small fraction of this unexplained diversity.

Instead, multiple lines of evidence indicate that alternative translation and/or proteolytic processing contribute substantially to proteome complexity. Most newly identified isoforms are shorter than their annotated counterparts and cannot be explained by alternatively spliced transcripts. Their well-defined molecular weights, together with independent peptide-level DIA-LOP localization data, argue against random degradation and support their biological relevance. Although we cannot globally distinguish alternative translation from proteolytic processing, the codon-optimized TEAD3 validation constructs make alternative splicing unlikely for the experimentally validated examples. More generally, current protein annotations are inherently biased toward transcript-derived isoforms because alternative splicing can be inferred from RNA sequencing, whereas alternative translation and proteolytic processing require direct protein-level evidence. Consequently, a substantial fraction of the human protein isoform landscape likely remains unannotated.

Together, our study provides an integrated view of transcript and protein isoform diversity. By combining full-length transcriptomics with PepCP, we demonstrate that most detectable transcript isoforms are translated and identify shorter protein variants as a major and underappreciated contributor to proteome complexity. More broadly, our work establishes a scalable framework for proteome-wide protein isoform discovery that complements existing transcriptomic and proteomic approaches and provides a resource for future studies of protein isoform biology.

Limitations of the study

Although IsoFrac enables proteome-wide protein isoform discovery, several limitations should be considered. First, the resolving power of SDS-PAGE is inherently limited, making it difficult to distinguish protein isoforms with similar molecular weights. Moreover, protein migration is influenced not only by molecular weight but also by charge, hydrophobicity, and residual secondary structure, which can cause deviations between apparent and theoretical molecular weights.^{45,46} Future integration of orthogonal separation strategies may therefore further improve isoform resolution.

Second, broad or overlapping peptide migration profiles, low signal-to-noise ratios, and low-abundance proteins can lead to ambiguous isoform assignments. Because IsoFrac prioritizes specificity, peptide traces that do not form coherent clusters are classified as noise and excluded from automated peak

(C) Enrichment of UniProt isoform categories across the Δ mass groups shown in (B). Significant enrichments are indicated (p values from hypergeometric tests).

(D) IsoFrac peptide traces for TEAD3 with identified protein isoforms (red stars) and the corresponding western blot.

(E) ROC of different metrics for predicting the best matching PACBIO entry for each IsoFrac isoform. Area under the curve values (AUC) for each metric are depicted in the legend.

(F) Western blot validation of N- and C-terminally tagged TEAD3 constructs.

(G) Peptide-level DIA-LOP projection of RPE-1 subcellular space based on the first two principal components. Hulls indicate the 90% quantile for each compartment.

(H) Peptide-to-centroid distance analysis of the known CTSB isoforms localized to the ER/Golgi (~37 kDa) and lysosome (~26 kDa).

(I) Subcellular localization distance ratios between other protein isoforms and within a protein isoform. Shorter Variants group and all other proteins are compared against a random protein isoform assignment of corresponding peptides (wilcoxon rank sum test, p -values depicted in the plot). Boxes represent the interquartile range (IQR), the center line indicates the median, and whiskers extend to the most extreme values within $1.5 \times$ IQR. Outlier points are not displayed.

detection. While this reduces false-positive assignments, it may also exclude genuine low-abundance protein isoforms or modified peptide species. Such borderline cases can be inspected using the accompanying IsoFrac web viewer.

Third, although IsoFrac efficiently detects shorter protein variants, it cannot distinguish whether these arise from proteolytic processing, alternative translation, or other mechanisms. Likewise, despite precautions to minimize post-lysis proteolysis, sample preparation artifacts cannot be completely excluded.

Finally, our study was performed in a single cell type (RPE-1), and the transcript reference was generated using the PacBio RSII platform, which has lower throughput than current long-read technologies. Consequently, the presented isoform landscape is cell-type specific, and rare transcript isoforms may remain underrepresented. Future studies using newer long-read platforms across diverse biological systems will further refine and expand the catalog of transcript and protein isoforms.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Matthias Selbach (matthias.selbach@mdc-berlin.de).

Materials availability

This study did not generate unique reagents. The GELFrEE 8100 system has been taken off the market. However, alternative systems can be considered to provide a similar performance.

Data and code availability

- The raw and processed RNA-seq data generated in this study are deposited in the Gene Expression Omnibus under the accession no. GSE261771.
- The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository^{47,48} with the dataset identifier PXD062269. R functions for processing the MaxQuant⁴⁹ output tables, Python code for PacBio processing, protein sequence databases and DIA-LOP marker proteins are available under <https://github.com/SelbachLab/IsoFrac> (Zenodo release: DOI: <https://doi.org/10.5281/zenodo.212607>). The dataset can be interactively explored on the website proteinisoforms.mdc-berlin.de. Protein isoform candidates as identified by IsoFrac are provided as a [supplemental information table \(Table S1\)](#).

ACKNOWLEDGMENTS

We thank Christian Sommer and Mohamad Haji for excellent technical assistance and Christopher Buccitelli (all MDC) for valuable input on peptide clustering algorithms. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, grant SE 1186/2-1), the National Natural Science Foundation of China (grant no. 31861133013), and the Shenzhen Science and Technology Program (grant no. KQTD20180411143432337).

AUTHOR CONTRIBUTIONS

M.S. and W.C. initiated the project and acquired funding. A.K., H.Z., and M.S. designed the proteomic experiments. A.K. performed all proteomic experiments. Q.Z. and M.-R.W. conducted all transcriptomic experiments and data analysis, respectively. L.F. designed and performed the western blot analysis of LDLR. H.Z. carried out the proteomic and the integrative proteogenomic data analysis, developed the IsoFrac pipeline, and created the online protein isoform shiny viewer. Q.Z., M.-R.W., and W.C. designed the transcrip-

tomic experiments. S.K. and K.S.L. designed, performed, preanalyzed, and discussed the DIA-LOP experiments. M.S. and H.Z. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

The authors declare no competing financial interest.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work the authors used ChatGPT-5.3 in order to improve wording and in a few cases optimize chunks of R code (e.g., peak-finding module). After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- [KEY RESOURCES TABLE](#)
- [EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS](#)
 - Human cell lines
- [METHOD DETAILS](#)
 - Cell culture
 - Sample preparation/GELFrEE/SP3/ TMT
 - MS measurements
 - Generation of FLAG-tagged genes
 - Transfections
 - Next-generation sequencing (NGS) for total RNA
 - PacBio library preparation and sequencing
 - Full-length transcriptome construction
 - DIA-LOP
- [QUANTIFICATION AND STATISTICAL ANALYSIS](#)
 - Peak detection and merging
 - Peptide trace generation and normalization
 - Peak merging across clusters and gels
 - IsoFrac backmapping/scoring
 - DIA-LOP data processing
 - Data integration
 - IsoFrac viewer
 - Nomenclature and terminology

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.celrep.2026.117898>.

Received: April 30, 2025

Revised: May 21, 2026

Accepted: August 6, 2026

REFERENCES

1. Bludau, I., and Aebersold, R. (2020). Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nat. Rev. Mol. Cell Biol.* 21, 327–340. <https://doi.org/10.1038/s41580-020-0231-2>.
2. Buccitelli, C., and Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* 21, 630–644. <https://doi.org/10.1038/s41576-020-0258-4>.
3. Ule, J., and Blencowe, B.J. (2019). Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Mol. Cell* 76, 329–345. <https://doi.org/10.1016/j.molcel.2019.09.017>.

4. Mitschka, S., and Mayr, C. (2022). Context-specific regulation and function of mRNA alternative polyadenylation. *Nat. Rev. Mol. Cell Biol.* 23, 779–796. <https://doi.org/10.1038/s41580-022-00507-5>.
5. Blencowe, B.J. (2017). The Relationship between Alternative Splicing and Proteomic Complexity. *Trends Biochem. Sci.* 42, 407–408. <https://doi.org/10.1016/j.tibs.2017.04.001>.
6. Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.* 102, 11–26. <https://doi.org/10.1016/j.ajhg.2017.11.002>.
7. Smith, L.M., Agar, J.N., Chamot-Rooke, J., Danis, P.O., Ge, Y., Loo, J.A., Paša-Tolić, L., Tsybin, Y.O., and Kelleher, N.L.; Consortium for Top-Down Proteomics (2021). The Human Proteoform Project: Defining the human proteome. *Sci. Adv.* 7, eabk0734. <https://doi.org/10.1126/sciadv.abk0734>.
8. Aebersold, R., Agar, J.N., Amster, I.J., Baker, M.S., Bertozzi, C.R., Boja, E.S., Costello, C.E., Cravatt, B.F., Fenselau, C., Garcia, B.A., et al. (2018). How many human proteoforms are there? *Nat. Chem. Biol.* 14, 206–214. <https://doi.org/10.1038/nchembio.2576>.
9. Hinnebusch, A.G., Ivanov, I.P., and Sonenberg, N. (2016). Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* 352, 1413–1416. <https://doi.org/10.1126/science.aad9868>.
10. Glinos, D.A., Garborcauskas, G., Hoffman, P., Ehsan, N., Jiang, L., Gokden, A., Dai, X., Aguet, F., Brown, K.L., Garimella, K., et al. (2022). Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* 608, 353–359. <https://doi.org/10.1038/s41586-022-05035-y>.
11. Wang, X., You, X., Langer, J.D., Hou, J., Rupprecht, F., Vlatkovic, I., Queidenau, C., Tushev, G., Epstein, I., Schaefer, B., et al. (2019). Full-length transcriptome reconstruction reveals a large diversity of RNA and protein isoforms in rat hippocampus. *Nat. Commun.* 10, 5009. <https://doi.org/10.1038/s41467-019-13037-0>.
12. 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>.
13. Ezkurdia, I., Rodríguez, J.M., Carrillo-de Santa Pau, E., Vázquez, J., Valencia, A., and Tress, M.L. (2015). Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.* 14, 1880–1887.
14. Toby, T.K., Fornelli, L., and Kelleher, N.L. (2016). Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem.* 9, 499–519. <https://doi.org/10.1146/annurev-anchem-071015-041550>.
15. Kaulich, P.T., Cassidy, L., and Tholey, A. (2024). Identification of proteoforms by top-down proteomics using two-dimensional low/low pH reversed-phase liquid chromatography-mass spectrometry. *Proteomics* 24, e2200542. <https://doi.org/10.1002/pmic.202200542>.
16. Tran, J.C., Zamdborg, L., Ahlf, D.R., Lee, J.E., Catherman, A.D., Durbin, K.R., Tipton, J.D., Vellaichamy, A., Kellie, J.F., Li, M., et al. (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 480, 254–258. <https://doi.org/10.1038/nature10575>.
17. Bekker-Jensen, D.B., Kelstrup, C.D., Batth, T.S., Larsen, S.C., Haldrup, C., Bramsen, J.B., Sørensen, K.D., Høyer, S., Ørntoft, T.F., Andersen, C.L., et al. (2017). An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst.* 4, 587–599.e4. <https://doi.org/10.1016/j.cels.2017.05.009>.
18. Nesvizhskii, A.I., and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* 4, 1419–1440. <https://doi.org/10.1074/mcp.R500012-MCP200>.
19. Umer, H.M., Audain, E., Zhu, Y., Pfeuffer, J., Sachsenberg, T., Lehtio, J., Branca, R.M., and Perez-Riverol, Y. (2022). Generation of ENSEMBL-based proteogenomics databases boosts the identification of non-canonical peptides. *Bioinformatics* 38, 1470–1472. <https://doi.org/10.1093/bioinformatics/btab838>.
20. Bludau, I., Frank, M., Dörig, C., Cai, Y., Heusel, M., Rosenberger, G., Picotti, P., Collins, B.C., Röst, H., and Aebersold, R. (2021). Systematic detection of functional proteoform groups from bottom-up proteomic datasets. *Nat. Commun.* 12, 3810. <https://doi.org/10.1038/s41467-021-24030-x>.
21. Dermit, M., Peters-Clarke, T.M., Shishkova, E., and Meyer, J.G. (2021). Peptide Correlation Analysis (PeCoRA) Reveals Differential Proteoform Regulation. *J. Proteome Res.* 20, 1972–1980. <https://doi.org/10.1021/acs.jproteome.0c00602>.
22. Miller, R.M., Jordan, B.T., Mehlferber, M.M., Jeffery, E.D., Chatzipantsiou, C., Kaur, S., Millikin, R.J., Dai, Y., Tiberi, S., Castaldi, P.J., et al. (2022). Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biol.* 23, 69. <https://doi.org/10.1186/s13059-022-02624-y>.
23. Dou, Y., Liu, Y., Yi, X., Olsen, L.K., Zhu, H., Gao, Q., Zhou, H., and Zhang, B. (2023). SEPeQuant enhances the detection of possible isoform regulations in shotgun proteomics. *Nat. Commun.* 14, 5809. <https://doi.org/10.1038/s41467-023-41558-2>.
24. 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>.
25. Kurzawa, N., Leo, I.R., Stahl, M., Kunold, E., Becher, I., Audrey, A., Mermelakas, G., Huber, W., Mateus, A., Savitski, M.M., and Jafari, R. (2023). Deep thermal profiling for detection of functional proteoform groups. *Nat. Chem. Biol.* 19, 962–971. <https://doi.org/10.1038/s41589-023-01284-8>.
26. Chen, C., Wen, M., and Jin, Y. (2022). 1DE-MS Profiling for Proteoform-Correlated Proteomic Analysis, by Combining SDS-PAGE, Whole-Gel Slicing, Quantitative LC-MS/MS, and Reconstruction of Gel Distributions of Several Thousands of Proteins. *J. Proteome Res.* 21, 2311–2330. <https://doi.org/10.1021/acs.jproteome.2c00180>.
27. Thiede, B., Treumann, A., Kretschmer, A., Söhlke, J., and Rudel, T. (2005). Shotgun proteome analysis of protein cleavage in apoptotic cells. *Proteomics* 5, 2123–2130. <https://doi.org/10.1002/pmic.200401110>.
28. Mann, M. (2020). The Origins of Organellar Mapping by Protein Correlation Profiling. *Proteomics* 20, e1900330. <https://doi.org/10.1002/pmic.20190330>.
29. Kristensen, A.R., Gsponer, J., and Foster, L.J. (2012). A high-throughput approach for measuring temporal changes in the interactome. *Nat. Methods* 9, 907–909. <https://doi.org/10.1038/nmeth.2131>.
30. Imami, K., Milek, M., Bogdanow, B., Yasuda, T., Kastelic, N., Zauber, H., Ishihama, Y., Landthaler, M., and Selbach, M. (2018). Phosphorylation of the Ribosomal Protein RPL12/uL11 Affects Translation during Mitosis. *Mol. Cell* 72, 84–98.e9. <https://doi.org/10.1016/j.molcel.2018.08.019>.
31. Dengjel, J., Høyer-Hansen, M., Nielsen, M.O., Eisenberg, T., Harder, L.M., Schandorff, S., Farkas, T., Kirkegaard, T., Becker, A.C., Schroeder, S., et al. (2012). Identification of Autophagosome-associated Proteins and Regulators by Quantitative Proteomic Analysis and Genetic Screens. *Mol. Cell. Proteomics* 11, M111.014035. <https://doi.org/10.1074/mcp.m111.014035>.
32. Christoforou, A., Mulvey, C.M., Breckels, L.M., Geladaki, A., Hurrell, T., Hayward, P.C., Naake, T., Gatto, L., Viner, R., Martinez Arias, A., and Lilley, K.S. (2016). A draft map of the mouse pluripotent stem cell spatial proteome. *Nat. Commun.* 7, 8992. <https://doi.org/10.1038/ncomms9992>.
33. Aryal, U.K., Xiong, Y., McBride, Z., Kihara, D., Xie, J., Hall, M.C., and Szymanski, D.B. (2014). A proteomic strategy for global analysis of plant protein complexes. *Plant Cell* 26, 3867–3882. <https://doi.org/10.1105/tpc.114.127563>.
34. Dunkley, T.P.J., Watson, R., Griffin, J.L., Dupree, P., and Lilley, K.S. (2004). Localization of Organelle Proteins by Isotope Tagging (LOPIT). *Mol. Cell. Proteomics* 3, 1128–1134. <https://doi.org/10.1074/mcp.t400009-mcp200>.
35. Andersen, J.S., Wilkinson, C.J., Mayor, T., Mortensen, P., Nigg, E.A., and Mann, M. (2003). Proteomic characterization of the human centrosome by

- protein correlation profiling. *Nature* 426, 570–574. <https://doi.org/10.1038/nature02166>.
36. Foster, L.J., de Hoog, C.L., Zhang, Y., Zhang, Y., Xie, X., Mootha, V.K., and Mann, M. (2006). A mammalian organelle map by protein correlation profiling. *Cell* 125, 187–199. <https://doi.org/10.1016/j.cell.2006.03.022>.
37. Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* 227, 680–685. <https://doi.org/10.1038/227680a0>.
38. Witkowski, C., and Harkins, J. (2009). Using the GELFREE 8100 Fractionation System for molecular weight-based fractionation with liquid phase recovery. *J. Vis. Exp.* 1842. <https://doi.org/10.3791/1842>.
39. Müller, T., Kalxdorf, M., Longuespée, R., Kazdal, D.N., Stenzinger, A., and Krijgsveld, J. (2020). Automated sample preparation with SP3 for low-input clinical proteomics. *Mol. Syst. Biol.* 16, e9111. <https://doi.org/10.1525/msb.20199111>.
40. Zecha, J., Satpathy, S., Kanashova, T., Avanesian, S.C., Kane, M.H., Clauser, K.R., Mertins, P., Carr, S.A., and Kuster, B. (2019). TMT Labeling for the Masses: A Robust and Cost-efficient, In-solution Labeling Approach. *Mol. Cell. Proteomics* 18, 1468–1478. <https://doi.org/10.1074/mcp.TIR119.001385>.
41. Grzybowska, E.A. (2012). Human intronless genes: functional groups, associated diseases, evolution, and mRNA processing in absence of splicing. *Biochem. Biophys. Res. Commun.* 424, 1–6. <https://doi.org/10.1016/j.bbrc.2012.06.092>.
42. Campello, R.J.G.B., Campello, R.J.G., Moulavi, D., and Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates (Springer Berlin Heidelberg). https://doi.org/10.1007/978-3-642-37456-2_14.
43. Geladaki, A., Kočevár Britovšek, N., Breckels, L.M., Smith, T.S., Vennard, O.L., Mulvey, C.M., Crook, O.M., Gatto, L., and Lilley, K.S. (2019). Combining LOPIIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat. Commun.* 10, 331. <https://doi.org/10.1038/s41467-018-08191-w>.
44. Melby, J.A., Brown, K.A., Gregorich, Z.R., Roberts, D.S., Chapman, E.A., Ehlers, L.E., Gao, Z., Larson, E.J., Jin, Y., Lopez, J.R., et al. (2023). High sensitivity top-down proteomics captures single muscle cell heterogeneity in large proteoforms. *Proc. Natl. Acad. Sci. USA* 120, e2222081120. <https://doi.org/10.1073/pnas.2222081120>.
45. Shi, Y., Mowery, R.A., Ashley, J., Hentz, M., Ramirez, A.J., Bilgic, B., Slunt-Brown, H., Borchelt, D.R., and Shaw, B.F. (2012). Abnormal SDS-PAGE migration of cytosolic proteins can identify domains and mechanisms that control surfactant binding. *Protein Sci.* 21, 1197–1209. <https://doi.org/10.1002/pro.2107>.
46. Burgess, R.R. (2009). Important but little known (or forgotten) artifacts in protein biochemistry. *Methods Enzymol.* 463, 813–820. [https://doi.org/10.1016/S0076-6879\(09\)63044-5](https://doi.org/10.1016/S0076-6879(09)63044-5).
47. Perez-Riverol, Y., Xu, Q.-W., Wang, R., Uszkoreit, J., Griss, J., Sanchez, A., Reisinger, F., Csordas, A., Ternent, T., Del-Toro, N., et al. (2016). PRIDE Inspector Toolsuite: Moving Toward a Universal Visualization Tool for Proteomics Data Standard Formats and Quality Assessment of ProteomeXchange Datasets. *Mol. Cell. Proteomics* 15, 305–317. <https://doi.org/10.1074/mcp.O115.050229>.
48. Perez-Riverol, Y., Bandla, C., Kundu, D.J., Kamatchinathan, S., Bai, J., Hewapathirana, S., John, N.S., Prakash, A., Walzer, M., Wang, S., and Vizcaino, J.A. (2025). The PRIDE database at 20 years: 2025 update. *Nucleic Acids Res.* 53, D543–D553. <https://doi.org/10.1093/nar/gkae1011>.
49. Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 26, 1367–1372. <https://doi.org/10.1038/nbt.1511>.
50. Pardo-Palacios, F.J., Arzalluz-Luque, A., Kondratova, L., Salguero, P., Mestre-Tomás, J., Amorín, R., Estevan-Morió, E., Liu, T., Nanni, A., McIntyre, L., et al. (2024). SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat. Methods* 21, 793–797. <https://doi.org/10.1038/s41592-024-02229-2>.
51. Hackl, T., Hedrich, R., Schultz, J., and Förster, F. (2014). proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* 30, 3004–3011. <https://doi.org/10.1093/bioinformatics/btu392>.
52. Salmela, L., and Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 30, 3506–3514. <https://doi.org/10.1093/bioinformatics/btu538>.
53. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
54. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. <https://doi.org/10.1038/nbt.3519>.
55. Demichev, V., Messner, C.B., Vernardis, S.I., Lilley, K.S., and Ralser, M. (2020). DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* 17, 41–44. <https://doi.org/10.1038/s41592-019-0638-x>.
56. Hahsler, M., Piekenbrock, M., and Doran, D. (2019). dbscan: Fast Density-Based Clustering with R. *J. Stat. Softw.* 91, 1–30. <https://doi.org/10.18637/jss.v091.i01>.
57. Gatto, L., Breckels, L.M., Wiczorek, S., Burger, T., and Lilley, K.S. (2014). Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics* 30, 1322–1324. <https://doi.org/10.1093/bioinformatics/btu013>.
58. Gatto, L., and Lilley, K.S. (2012). MSnbase—an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics* 28, 288–289. <https://doi.org/10.1093/bioinformatics/btr645>.
59. McCaskie, K., Hutchings, C., Feret, R., Kim, Y.-I., Breckels, L., Deery, M., and Lilley, K. (2025). Localization of Organelle Proteins Using Data-Independent Acquisition (DIA-LOP). *Mol. Cell. Proteomics* 24, 101047. <https://doi.org/10.1016/j.mcpro.2025.101047>.
60. Frankenfield, A.M., Ni, J., Ahmed, M., and Hao, L. (2022). Protein Contaminants Matter: Building Universal Protein Contaminant Libraries for DDA and DIA Proteomics. *J. Proteome Res.* 21, 2104–2113. <https://doi.org/10.1021/acs.jproteome.2c00145>.
61. Breckels, L.M., Mulvey, C.M., Lilley, K.S., and Gatto, L. (2016). A Bioconductor workflow for processing and analysing spatial proteomics data. *F1000Res.* 5, 2926. <https://doi.org/10.12688/f1000research.10411.2>.
62. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Duodoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5, R80. <https://doi.org/10.1186/gb-2004-5-10-r80>.
63. Hutchings, C., Krueger, T., Crook, O.M., Gatto, L., Lilley, K.S., and Breckels, L.M. (2025). An updated Bioconductor workflow for correlation profiling subcellular proteomics. *F1000Res.* 14, 714. <https://doi.org/10.12688/f1000research.165543.1>.
64. Crook, O.M., Mulvey, C.M., Kirk, P.D.W., Lilley, K.S., and Gatto, L. (2018). A Bayesian mixture modelling approach for spatial proteomics. *PLoS Comput. Biol.* 14, e1006516. <https://doi.org/10.1371/journal.pcbi.1006516>.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-DYKDDDDK Tag Antibody HRP Conjugate	Cell Signaling Technology	Cat#2044; RRID: AB_10707327
Biological samples		
Human retinal pigment epithelial cells (RPE-1)	ATCC	CRL-4000
Critical commercial assays		
Pierce BCA Protein Assay Kit	Thermo Fisher Scientific	Cat#23225
SMARTer PCR cDNA Synthesis Kit	Takara/Clontech	Cat.# 634926
TMTpro Reagent Set	Thermo Fisher Scientific	Cat#A44520
Deposited data		
RNA-seq data	This paper	GEO: GSE261771
Proteomics data	This paper	PRIDE: PXD062269
Recombinant DNA		
pTwist_ENTR Kozak constructs (TEAD3, NELFA, RRP1B, CTTNBP2NL)	Twist Bioscience	This paper
pDEST_pcDNA5_BirA_FLAG_Nterm	Couzens et al. 2013	N/A
pDEST_pcDNA5_BirA_FLAG_Cterm	Couzens et al. 2013	N/A
Software and algorithms		
MaxQuant v2.0.3.0	Cox & Mann ⁴⁹	https://www.maxquant.org
SQANTI3 v3.6.1	Pardo-Palacios et al. ⁵⁰	https://github.com/ConesaLab/SQANTI3
IsoSeq3 v3.4.0	PacBio	https://github.com/PacificBiosciences/IsoSeq
GMAP	Wu and Watanabe 2005	http://research-pub.gene.com/gmap
Proovread v2.14.1	Hackl et al. ⁵¹	https://github.com/Biolnf-Wuerzburg/proovread
LoRDEC v0.9	Salmela and Rivals ⁵²	https://gite.lirmm.fr/lordec/lordec
STAR v2.7.8a	Dobin et al. ⁵³	https://github.com/alexdobin/STAR
kallisto v0.46.0	Bray et al. ⁵⁴	https://pachterlab.github.io/kallisto
DIA-NN v1.9.2	Demichev et al. ⁵⁵	https://github.com/vdemichev/DiaNN
R 4.4.2	The R Foundation for Statistical Computing	https://cran.r-project.org/
Rstudio	Posit	https://posit.co/
Shiny v1.11.1	Winston Chang	https://shiny.posit.co/
ggplot2 R package 3.5.1	Hadley Wickham	–
dbscan R package v1.1-10	Hahsler et al. ⁵⁶	https://cran.r-project.org/package=dbscan
pracma R package v2.3.3	–	https://cran.r-project.org/package=pracma
h2o R package 3.46.0.7 rel-3.46.0	Tomas Fryda	https://cran.r-project.org/package=h2o
pRoloc	Gatto et al. ⁵⁷	https://github.com/lgatto/pRoloc
MSnbase	Gatto et al. ⁵⁸	https://github.com/lgatto/MSnbase
IsoFrac	This paper	https://doi.org/10.5281/zenodo.21260763
IsoFrac Viewer	This paper	https://proteinisoforms.mdc-berlin.de
Other		
UniProt Human Reference Proteome	UniProt Consortium	UP000005640
GENCODE annotation	GENCODE	Release v40
GELFrEE Fractionation System	Expedeon	GELFrEE 8100
Q Exactive HF-X Orbitrap	Thermo Fisher Scientific	N/A
Orbitrap Exploris 480	Thermo Fisher Scientific	N/A
timsTOF HT	Bruker Daltonics	N/A

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Agilent Bravo Liquid Handling Platform	Agilent Technologies	Bravo
nanoElute 2 UHPLC	Bruker Daltonics	nanoElute 2
Reagents		
Benzonase Nuclease	Merck Millipore	Cat#70664
Sera-Mag SpeedBeads A	Cytiva	Cat#45152105050250
Sera-Mag SpeedBeads B	Cytiva	Cat#45152105050250
Trypsin	Promega	https://www.promega.de/
LysC	Fuji	https://fujifilmbiosciences.fujifilm.com/us/lysyl-endopeptidase-mass-spectrometry-grade-lys-c.html
Lipofectamine 3000 Transfection Reagent	Thermo Fisher Scientific	–
TRIzol Reagent	Life Technologies	–
GELFrEE 5%, 8%, and 10% cartridges	Expedeon	N/A

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human cell lines

All experiments were performed using the immortalized human retinal pigment epithelial cell line hTERT RPE-1 (ATCC CRL-4000). Cells were maintained under standard culture conditions as described below.

METHOD DETAILS

Cell culture

hTERT RPE-1 cells were purchased from ATCC (CRL-4000) and maintained in DMEM/F-12 (1:1) + GlutaMAXTM medium (GIBCO 31331-093) containing 10% fetal calf serum (FCS) (PAN-Biotech). Cells were cultured at 37°C in a humidified incubator of 5% CO₂. Throughout the experiments the cells were kept under 15 passages.

Sample preparation/GELFrEE/SP3/ TMT

Lysis. hTERT RPE-1 cells were seeded on 15cm plates until ~80% confluency, then washed and scraped with ice-cold PBS, collected into 15mL falcons, washed again and centrifuged (500g, 3min, 4°C). Cell pellets were lysed at room temperature (20mM Tris-HCl pH 7.5, 100mM NaCl, 1% SDS, 1mM MgCl₂, Protease inhibitors) and immediately boiled (95°C, 5min). 1μL of Benzonase was added to digest DNA (37°C, 20min). Lysates were then centrifuged (14,000 rpm, 22°C, 15min) and the protein concentration in the supernatant was measured via BCA assay (Pierce BCA Protein Assay Kit). The proteins were then reduced (15mM DTT, RT, 30min), alkylated (20mM iodoacetamide, RT, 30min) and quenched (30mM DTT, RT, 30min). To desalt the protein lysate Acetone/Methanol precipitation was done. Namely, 6 volumes of –20°C Acetone/Methanol (8:1) was added to 1 volume of lysate, vortexed 20s and placed at –80°C overnight. Next, the samples were centrifuged (14,000rpm, 4°C, 30min), the supernatant was discarded and pellet was washed again with Acetone/Methanol and centrifuged (14,000rpm, 4°C, 10min). The pellet was dried and then solubilized in 2% SDS.

The proteins were fractionated according to their electrophoretic mobility employing the GELFrEE 8100 Fractionation system (Expedeon, Gel-Eluted Liquid Fraction Entrapment Electrophoresis) fractionation system with liquid phase recovery. We used the 5%- (mass range: 75–500kDa), 8%- (mass range: 35–150kDa) and 10%- (mass range: 15–100kDa) cartridges to generate 28 protein fractions (in total, 84 fractions per sample) according to the manufacturer instructions. Briefly, 150μL of protein sample (112μL de-salted protein, 8μL 1M DTT, 30μL Acetate Sample Buffer X5) was heated (95°C, 5min) and loaded onto the correspondent cartridge. The protein fractions were collected into 96 well-plates.

130μL of each protein fraction was transferred into a BRAVO-designated 96-well plate (28 fractions of each cartridge combined in one plate) along with two correspondent 96-well plates, one with magnetic beads (SeraMag beads A and B), and one with Trypsin-LysC (in 50mM HEPES pH 8). The digestion took place on the beads (1:25–50 protein:enzyme ratio).

We estimated the amount of peptides in the fractions to be 4μg–8μg. The digested fractions were then labeled with TMTpro reagent set (Thermo Fisher Scientific, Cat #: A44520, Lot #: UL296296) according to the manufacturer's instructions. Samples assignments: every 28 fraction set was divided into two separate plexes, where the "odd" fractions (fraction 1, fraction 3, ..., fraction 27) were randomly assigned to one plex and the "even" fractions (fraction 2, fraction 4, ..., fraction 28) were randomly assigned to a second plex. Assignments, plex #1: fraction 1–131N, fraction 3–127N, fraction 5–129N, fraction 7–132N, fraction 9–128C, fraction 11–127C, fraction 13–129C, fraction 15–131C, fraction 17–130C, fraction 19–126, fraction 21–130N, fraction 23–133C, fraction 25–132C, fraction

27–128N, Std1 - 133N, std2 - 134N. plex #2: fraction 2–129C, fraction 4–133C, fraction 6–127C, fraction 8–129N, fraction 10–132N, fraction 12–128N, fraction 14–126, fraction 16–130N, fraction 18–132C, fraction 20–130C, fraction 22–131C, fraction 24–128C, fraction 26–131N, fraction 28–127N, Std1 - 133N, std2 - 134N. Std1 and Std2 were whole cell lysates of hTERT-RPE1 cells that were not fractionated but went through the SP3-digestion (as described above). Std1 and Std2 contained 6μg and 60μg, respectively. The peptides from each plex were pooled, desalted with SepPak columns (Waters) and dried (speed-vac). We used the Dionex 3000 system (Thermo Fisher) to offline fractionate the desalted peptide mixture with high pH reverse phase fractionation (2p1mm C18 Xbridge Waters). Pooled peptides were resuspended in buffer A (0.0175% vol/vol NH₄OH; 0.01125% vol/vol FA + 2% MeCN), loaded, and fractionated (96 min gradient from 0% B to 40% B in 70 min, 44% B in 4 min, 60% B in 5 min and kept at 60% B for the remaining 17 min; Buffer B: 0.0175% vol/vol NH₄OH; 0.01125% vol/vol FA + 90% MeCN). Each 1 min a fraction was collected (in total 96 fractions). The automated collection procedure resulted in a total of 24 fractions, after recombination of selected fractions: (fraction i) + (fraction i + 24) + (fraction i + 48) + (fraction i + 72), i = 1 → 24. The fractions were dried and kept in –80°C.

MS measurements

Samples were resuspended in Buffer A (3% acetonitrile, 0.1 formic acid). Peptides were separated with an easy-nLC System (Thermo Scientific) on a selfpacked column (20cm, 75μm ID, DrMaisch 1.9μm AQ) with a reverse phase gradient of increasing buffer B concentrations (90% acetonitrile, 0.1 formic acid; 0 min 2% B, 1min 4% B, 67 min 20% B; 20 min 30% B; 10 min 60% B; 1 min 90% B). Peptides were subsequently measured over ESI spray on an Q-Exactive HFX MS System (Thermo Scientific) for the 8er and 10er gels or on an Exploris480 MS System (Thermo Scientific) for the 5er gel. Settings on the HFX were: MS1 Res. 60k; AGC target 3e6; MaxIT 10ms; scan range 350–1500 m/z; MS2 Res 45k; AGC target 1e5; MaxIT 86ms; TopN 20; IW 0.7 m/z; Scan Range 350–1500 m/z; NCE 30, Dynamic Exclusion 30s.

Settings on the Exploris were: MS1 Res. 60k; normalised AGC target 300%; MaxIT 50ms; scan range 375–1500 m/z; MS2 Res 45k; normalised AGC target 100%; MaxIT 86ms; Cycle Time 1s; IW 0.4 m/z; Scan Range 350–1500 m/z; HCD Collision Energy 31%, Dynamic Exclusion 20s.

Generation of FLAG-tagged genes

TEAD3, *NELFA*, *RRP1B* and *CTTNBP2NL* genes were ordered in a pTwist_ENTR Kozak vector (Twist Bioscience). Using the Gateway Clonase II system (Thermo Fisher Scientific) the genes were cloned into the vectors pDEST_pcDNA5_BirA_FLAG_Nterm and pDEST_pcDNA5_BirA_FLAG_Cterm (Couzens et al., 2013). Sequencing validated the resulting products. Gene sequences were optimized, don't contain introns and are therefore different from the human sequences.

Transfections

hTERT-RPE1 cells were seeded on 6-well plates and following the manufacturer's instructions transfected with 5μg of the DNA constructs using Lipofectamine 3000 (Thermo Fisher). The cell lysates were prepared following the same procedure described in Sample preparation. The samples were run on 4–12% Bis-Tris gradient gels (NuPAGE, Invitrogen) and then blotted on PVDF membrane (Immobilon-P). Ectopic expression was detected using anti-DYKDDDDK Tag Antibody HRP Conjugate (2044, Cell Signaling).

Next-generation sequencing (NGS) for total RNA

Total RNA from human retinal pigment epithelial-1 (RPE-1) cells was extracted following the manufacturer's protocol. Total RNA was extracted by TRIzol Reagent (Life Technologies). 1 μg total RNA was used to prepare the VAHTS Stranded mRNA-seq Library Prep Kit (Vazyme). After quality control, the libraries were sequenced in paired-end 2 × 150 nt manner on Illumina NovaSeq 6000 platform.

PacBio library preparation and sequencing

PacBio library preparation and sequencing were performed as described before.¹¹ Briefly, Total RNA from RPE-1 cells was extracted using TRIzol reagent (Life Technologies) following the manufacturer's protocol, and then was converted to cDNA using the Clontech SMARTer PCR cDNA Synthesis kit. Then we proceed to PCR cycle optimization, and use the cycle number to generate a sufficient amount of double stranded cDNA products. After purification of PCR products by using AMPure purification methods, SMRTbell™ template preparation was performed. The libraries were sequenced on PacBio RSII SMRT platform according to the manufacturer's instruction (Magbead mode for fragments >3 kb and Diffusion mode for all other fragment sizes; SMRT Cell v3 with 1200 min movie time). After sequencing, the raw reads were processed through the SMRT-Portal analysis suite (PacBio) to generate subread sequences for further analysis.

Full-length transcriptome construction

Pre-processing

The primer of raw PacBio sequencing reads was removed using lima (version 2.0.0, –isoseq –peek-guess). Then, the concatemer and poly(A) were removed by IsoSeq3 refine (version 3.4.0, –require-polya). For NGS sequencing, the clean reads were generated by fastp (version 0.21.0) with default parameters.

NGS-based error correction of PacBio sequencing reads

Proovread (version 2.14.1, with default parameters) and LoRDEC (version 0.9, with -k 17, -k 19 and -k 21, separately) were used to correct the sequencing errors in PacBio sequencing reads by taking advantage of the more accurate NGS reads. PacBio reads were corrected by Proovread⁵¹ and LoRDEC⁵² in parallel. After correction, GMAP (version 2021-02-22, Genome GRCh38, -sampling 1 -B 2 -n 0 -k 15 -min-intronlength 20) was used to align the corrected reads and reads with the best alignments (the longest alignment with the minimum number of mismatches) were kept for further analyses.

Clustering and collapsing

IsoSeq3 cluster (version 3.4.0) was applied to cluster and acquire consensus sequences from the corrected PacBio sequencing reads. Pbm2 align (version 1.9.0, with -preset ISOSEQ) was used to align the consensus sequences to genome GRCh38 and cDNA_Cupcake (collapse_isoforms_by_sam.py) was applied to collapse any identical isoforms based on the aligned exonic structure.

Annotating, quantification and filtering of full-length transcripts

SQANTI3⁵⁰ (version 3.6.1, with default parameters) was used to annotate the full-length transcripts with known GENCODE transcripts (gencode.v40.comprehensive.annotation.gtf downloaded from GENCODE) as well as discover novel transcripts. First, sqanti3_qc.py was used to annotate full-length transcripts by inputting transcript abundance information (parameter -e) and splicing junction information (parameter -c). The transcript abundances and splicing junction information was obtained from two NGS replicates, computed by kallisto⁵⁴ quant (version 0.46.0, with default parameters) and STAR⁵³ (version 2.7.8a, with -outFilterMultimapNmax 1). Second, sqanti3_RulesFilter.py (-a 0.9) was used to remove potential artifacts caused by RT-template switching, intra-priming, etc. ORFs of the full-length transcripts were annotated by SQANTI⁵⁰ at the same time.

Integration of full-length transcripts and ORFs with GENCODE annotation

Before integration, we first removed full-length transcripts (FLT) with potentially degraded 5' end. For each FLT, we used SQANTI⁵⁰ to compare its relationship with the annotated transcripts in GENCODE. For FLT that were not annotated as "Full-splice match" by SQANTI⁵⁰, we checked if they were artifacts resulting from degradation at the 5' end of the transcript due to technical limitations. Specifically, we compared the splice junction positions between FLT in pairs; for all FLT with the same "splice-junction-structure," we retained only the longest transcript for further analysis. Also, for transcripts annotated in GENCODE, we used kallisto⁵⁴ quant (version 0.46.0, with default parameters) to quantify each transcript and only retained transcripts with mean TPM > 1. Next, we integrated filtered FLT from PacBio with filtered transcripts from GENCODE by splice-junction structure, and obtained the union set. For each transcript in the union set, we tried to set a label for its source (Figure 4A). If the transcripts contain common splice-junction structure in both PacBio and GENCODE, they were labeled as "PB_&_GENCODE", indicating the PacBio FLT are totally "full-splice-match" compared to GENCODE transcripts; If the transcripts contain unique splice-junction structure in PacBio or GENCODE, they were labeled as "PB_only" or "GENCODE_only". It should be noted that some FLT are incomplete at 5'- or 3'- end likely due to technology limitation and they are partially "full-splice-match" compared to GENCODE transcripts. In this scenario, we hypothesized they are supported by both PacBio and GENCODE, and the PacBio FLT should be elongated based on splice-junction structure of GENCODE transcripts. We set the label "GENCODE_from_PBcorrected" for this group of transcripts.

After the integrated transcript annotation was obtained, the ORF for each transcript can be inferred and the source of ORFs can be defined. Similarly, we classified all ORFs into four categories with the following priority: 'PB_&_GENCODE' > 'GENCODE_from_PBcorrected' > 'PB_only' > 'GENCODE_only'. We also checked whether the ORF is from the canonical transcripts defined by Ensembl dataset and estimated the expression of ORF using the average TPM of all corresponding transcripts.

MaxQuant/peptide trace processing

Raw files from each gel preparation were analyzed separately using MaxQuant⁴⁹ v2.0.3.0 with the integrated Andromeda search engine, employing the TMT module and searching against the UniProt human reference proteome (UP000005640, 20220603), including annotated truncated variants, derived from alternative translation or proteolytic cleavage events. The annotated truncated variants were obtained from the available gff data format, as provided by UniProt. Trypsin/P was specified as the protease allowing up to 2 missed cleavages. In total, we extracted the following protein sequences per category: No Variant ~11,600; AS ~71,400; Shorter variant 35,400; TrEMBL ~18,800. Carbamidomethylation (C) was set as a fixed modification. Oxidation (M), protein N-terminal acetylation, and deamidation (NQ) were included as variable modifications. The first search precursor tolerance was set to 20 ppm and the main search tolerance to 4.5 ppm. Label-free quantification was disabled, and isobaric reporter ion quantification was performed using the TMT settings implemented in MaxQuant.⁴⁹ Match between runs was enabled (0.7 min matching window; 20 min alignment window), while matching of unidentified features was disabled. Protein quantification was restricted to unmodified peptides and those containing only oxidation (M) or protein N-terminal acetylation to reduce modification-driven variability. Peptide-spectrum matches and protein identifications were filtered at a 1% false discovery rate (FDR) at the peptide, protein, and site levels.

Further data analysis was conducted on the resulting *peptides.txt* table of the MaxQuant⁴⁹ output. This file contains peptide-level identification and quantification results after database searching and reporter ion quantification. The 28 fractions obtained from one gel separation were divided into two TMT experiments. For each plex 14 of the 16 available channels contained the quantitative information: 1. To correct for batch effects, peptide intensities were normalised within each TMT experiment to the respective reference channel intensity (channel 16), provided the peptide was quantified in the reference channel. In rare cases where a peptide was not identified in the reference channel, intensities were normalised to the average reference channel intensity within the respective TMT experiment. As a result we obtained peptide intensity traces across fraction for three different gels. Peptide intensities were

subsequently z-scored, and all peptide sequences were remapped to Ensembl Gene IDs using the UniProt.ws and biomaRt R packages. All downstream subsetting analyses were performed in a gene-centric manner.

Missing value imputation

Missing peptide intensity values across gel fractions were imputed to preserve continuity of fractionation traces. Missing values primarily arise from the multiplexed experimental design (two TMT plexes), where incomplete reporter quantification or stochastic sampling can result in peptide intensities being absent in individual fractions.

For a given peptide, missing data points flanked by quantified neighboring fractions were imputed by linear interpolation between the adjacent fraction intensities. That is, missing values were estimated by extrapolating in a linear fashion from the two closest observed fraction intensities for that peptide.

Imputation was performed only when adjacent fraction data were available; isolated missing values at the edges of traces without neighboring quantified fractions were not imputed. This approach preserves the overall shape of peptide elution profiles while minimizing artificial peak introduction.

DIA-LOP

DIA-LOP subcellular fractionation and differential centrifugation

RPE-1 cells were subjected to spatial proteomics using a DIA-LOP (Localization of Organelle Proteins by Data-Independent Acquisition) workflow adapted from the strategy described in.^{43,59} Briefly, cells were gently lysed in 1 mL of lysis buffer (250 mM sucrose, 10 mM HEPES pH7.4, 2 mM EDTA, 2 mM magnesium acetate, and cOmplete protease inhibitors from Sigma Aldrich) using a ball-bearing homogeniser (Isobiotec), passing the cells through the homogeniser 15 times with 10 μ m clearance on ice with 1 mL syringes. Biochemical fractionation of cells by differential centrifugation protocol was adapted from Geladaki et al.,⁴³ using a combination of a standard refrigerated benchtop centrifuge (Eppendorf 5804R) and a Beckman Optima MAX-XP ultracentrifuge with a TLA-55 fixed-angle rotor. Organellar enrichment of cells was achieved by spinning them at sequential spin speeds and times, transferring the supernatant of each fraction for the subsequent spin, resulting in 10 fractions. The final supernatant was precipitated with 6 volumes of acetone overnight in -20°C . The precipitated pellet and fractionated pellets were resolubilised in 8 M urea, 0.2% SDS, and 50 mM HEPES pH 8.5.

In addition to the 10 fractions, a separate nuclear enriched fraction was prepared to increase resolution for the nucleus. Briefly, cell pellets were resuspended in swelling buffer (SB; 10 mM Tris-HCl pH 7.5, 2 mM EDTA, with cOmplete protease inhibitors from Sigma Aldrich) at 10 x the pellet volume, vortexed at maximum speed for 15 s, and incubated on ice for 10 min. Plasma membrane lysis buffer (10 mM Tris-HCl pH 6.8, 2 mM MgCl_2 , 1% Triton X-100) was then added at 1/20th of the combined pellet and SB volume, followed by vortexing for 5 s and incubation on ice for 1 min. Samples were centrifuged at 16,000 RCF for 5 min at 4°C , and the supernatant was collected as the cytoplasmic fraction. The nuclear pellet was washed with 1 mL PBS, briefly vortexed, and centrifuged again at 16,000 RCF for 5 min at 4°C , after which the supernatant was discarded. The washed nuclear pellet was resuspended in 400 μ L RIPA buffer and incubated on ice for 30 min. Cells were sonicated at medium intensity for 1 min at 30 s on and off intervals using a Diagenode Bioruptor Plus with a water cooling system and spun at 20,000 RCF for 30 min at 4°C and the supernatant was collected.

DIA-LOP sample preparation and mass spectrometry

Proteins from each fraction were reduced by adding dithiothreitol (DTT) to a final concentration of 10 mM, incubating for 1 h at RT, then alkylated by adding iodoacetamide (IAA) to a final concentration of 25 mM. Samples were precipitated with 6 volumes of acetone at -20°C overnight.

Precipitated proteins were recovered by centrifugation at 3,200 RCF for 20 min at 10°C , and resuspended in triethylammonium bicarbonate (TEAB; Sigma-Aldrich) to a final protein concentration of 1 g/L. Each fraction was digested with trypsin, with trypsin to protein ratio of 1:20 (w/w) and incubated overnight at 37°C . The samples were desalted using Pierce Peptide Desalting Spin Columns (Thermo) according to the manufacturer's instructions. For analysis, the fractions were lyophilised and resuspended in 0.1% formic acid. All samples were analyzed on a timsTOF HT mass spectrometer (Bruker Daltonics), coupled to a nanoElute 2 UHPLC system (Bruker Daltonics). For each fraction, 750 ng of peptide was injected onto a PepMap Neo trap column (300 $\mu\text{m} \times 5 \text{ mm}$, 5 μm particle size; Thermo Scientific) and subsequently separated on a PepSep analytical column (150 $\mu\text{m} \times 150 \mu\text{m}$, 1.5 μm C18 particle size; Bruker Daltonics). Separation was performed at a flow rate of 800 nL/min using a binary mobile phase consisting of solvent A (0.1% formic acid in water) and solvent B (0.1% formic acid in acetonitrile). The elution gradient increased from 5% to 17% solvent B over 31 min, then to 26% over the following 7 min, and finally to 37% over the next 3 min, before a rapid increase to 85% solvent B over the final 1 min. The column temperature was maintained at 45°C throughout the separation. Peptides were ionised using a Captive Spray source equipped with a ZDV Sprayer emitter (20 μm ; Bruker Daltonics), and the mass spectrometer was operated in positive ion mode over an m/z range of 100–1,700. Ion mobility (IM) separation was set between 0.80 and 1.35 $1/\text{K}_0$ ($\text{V} \cdot \text{s} \cdot \text{cm}^{-2}$), with both accumulation and ramp times set to 100 ms. Collision energy was dynamically adjusted based on ion mobility, ranging from 20 eV at 0.60 $1/\text{K}_0$ to 59 eV at 1.6 $1/\text{K}_0$. The dia-PASEF window scheme was optimised for the RPE-1 cell proteome using py_diAID (v0.0.19) with a dda-PASEF spectral library as input. A total of 10 dia-PASEF scans were distributed across three ion mobility windows spanning 300–1,200 Da, resulting in an estimated cycle time of 1.17 s. MS1 data were acquired in single-shot acquisition mode for each fraction.

DIA-LOP database search and quantification

The dia-PASEF data obtained from the mass spectrometer were processed and analyzed using DIA-NN 1.9.2⁵⁵ to perform peptide sequence searches against the isofrac database alongside the Universal Protein Contaminants database (accessed in August 2024).^{55,60} A maximum of two missed cleavages was allowed, with peptide length restricted to 7 to 30 amino acids. Carbamidomethylation of cysteine residues was specified as a fixed modification, and variable modifications included methionine oxidation, protein N-terminal methionine excision and acetylation. To minimize missing identifications, the 'match between runs' option was enabled, using the QuantUMS (high precision) strategy. Mass tolerance was automatically optimized within the DIA-NN search. An FDR cutoff of 1% was used at the precursor level.⁵⁵

QUANTIFICATION AND STATISTICAL ANALYSIS

Peak detection and merging

The main basis for the subsequent analysis are peptide intensity data from across different fractions (peptide traces). We considered these fractionation profiles to be potentially noisy, particularly for low-intensity peptides. To reduce false-positive peak assignments driven by single peptides, we applied hdbscan⁵⁶ to identify density-based clusters of co-eluting peptides. Only peptide groups showing consistent fractionation trends are retained as clusters, whereas isolated or inconsistent peptide signals are classified as noise and excluded from automated peak definition.

Peptide trace generation and normalization

Peptide Z score traces from each gene were clustered using hdbscan from the dbscan R-package (1.1–10) using hdbscan function with the different settings and subsets: Option minPts set to 2 and 3 were used as different parameters. In order to improve the clustering depth, resulting Clusters with more than 30 peptide traces were subjected to an additional independent clustering round using hdbscan. Next, the best clustering setting was determined by calculating cluster silhouettes (cluster R-package 2.1.2). The clustering event with the highest silhouette was picked for further analysis. Hdbscan itself reports noisy traces as cluster number zero. Peptide traces assigned to 0 were excluded from the subsequent peak detection and further biological interpretation. Importantly, this classification reflects inconsistency in the fractionation profile rather than the validity of the peptide identification. Such traces may arise from low-abundance peptides, modified peptide forms with altered migration behavior, stochastic measurement variability, or low-abundance protein isoforms supported by only a few peptides.

To reduce artifacts from noisy but clusterable peptide traces and to decrease dataset complexity, median intensity profiles were computed for each cluster. These representative cluster traces were smoothed using a local moving average approach and subsequently Z score normalized. Peak detection was performed on the smoothed median profiles using the findpeaks function (pracma v2.3.3) with parameters nups = 1, ndowns = 1, and threshold = 0.1. This ensured that only consistent peak patterns spanning multiple fractions were retained, effectively excluding single-fraction spike events and reducing false positives.

To further refine peak positions and improve molecular weight (MW) resolution beyond the discrete gel fractions, peak apexes were re-estimated using local peptide-level information. Specifically, peak regions were re-evaluated within defined boundaries, and peak shapes were approximated using Gaussian models fitted via nonlinear least squares on the unsmoothed original data. This enabled extraction of peak width (σ), goodness-of-fit (pseudo R^2), and refined apex positions. Robustness and confidence of peak detection were assessed using bootstrap resampling. For each cluster, multiple bootstrap iterations were performed by resampling peptide traces, and peaks were re-detected within local windows. This provided empirical confidence intervals for peak positions and intensities.

Peak merging across clusters and gels

Redundant peak detections are expected, particularly when distinct peptide subsets (e.g., isoform-specific peptides) cluster independently or when partially overlapping fractionation patterns are observed across gels. Therefore, identified clusters are subjected to a merging step in which peaks with overlapping fraction centroids and consistent predicted molecular weights are combined. Merging is constrained by fraction proximity and expected mass range to preserve resolution between distinct proteoforms while consolidating redundant detections of the same underlying peak:

First candidate peptide clusters were filtered in two steps. Only non-noise clusters were retained (cluster2 != 0) and clusters were required to be supported by more than one peptide. For clusters with low z-scores (Height < 0.5), stricter support was applied, requiring more than two peptides per cluster; clusters with Height ≥ 0.5 were retained under the standard support criterion. Further, assignments required peptides to have limited fraction spread (Fraction_sd_80 ≤ 10), and were not part of gel-specific problematic or overlapping regions. These gel-specific exclusions removed low-mass peaks in the 5% and 8% gels (Peak < 8) and high-mass peaks in the 10% gel (Peak > 25).

Peaks were either merged within each gel separately (gel centric analysis) or merged across all gels (gel combined analysis) using the following procedure: Briefly, we selected the peak with the highest Z score as the "seed" of an isofrac isoform. We then tested the overlap of adjacent peaks to the same isofrac isoform if they fall within the peak boundaries ± 5000 Da. We therefore generated random distributions based on the Peak Apex and "start" and "end" of a peak. Then we sampled 20 data points for each peak and tested against the null hypothesis that the compared peaks are equal using the Wilcoxon Rank-Sum test, as implemented in

R. We did this in a bootstrap fashion, going for 50 rounds. Finally, we considered peaks to be similar if at least one resulting *p*-value is higher than 0.1.

IsoFrac backmapping/scoring

To assess our ability to detect known protein isoforms, we compared isoFrac isoforms with the Uniprot database, which includes information of splice variants and proteolytic cleavage variants. Specifically, we compared each isoFrac isoform to all annotated Uniprot isoforms of the same gene. For each pair of isoforms, we computed the fraction of IsoFrac peptides shared with Uniprot (IsoFrac coverage), the fraction of Uniprot peptides shared with IsoFrac (Uniprot coverage) and the statistical significance of the peptide overlap (hypergeometric test *p*-value). Since the ability to obtain significant *p*-values depends on the fraction of all peptides matching a given uniprot isoform, we also calculated this fraction (total uniprot fraction). This results in a feature table for each IsoFrac Uniprot pair (with: IsoFrac coverage, UniProt coverage, hypergeometric test *p*-value, total uniprot fraction, all peptides identified per gene). The entire procedure was then repeated after randomly assigning peptides to IsoFrac isoforms. We then built a regression model using gradient-boosted classification model (h2o.gbm, h2o package v3.46.0.7) using 10-fold cross validation (nfolds = 10). The obtained AUC from the ROC indicated that the prediction substantially improved with these scores in comparison of using each feature alone. Finally we used these scores to filter for the highest scoring isoform uniprot associations and calculated Δ mass (Uniprot MW - IsoFrac MW) and Quotient Mass (Uniprot MW/IsoFrac MW) values to independently evaluate the filtering efficiency on the resulting mass difference between Uniprot and IsoFrac isoforms. Finally, the best-matching protein database entry was selected based on the regression score. Notably, this score reflects peptide sequence information but does not incorporate protein molecular weight. Consequently, protein assignments are determined solely by peptide evidence, thereby avoiding biases in the observed delta mass distributions. In cases where ambiguity remained, multiple protein sequence annotations were retained.

For visualization and classification of Δ mass relationships between IsoFrac, UniProt, and PacBio assignments, isoforms were grouped using an empirical molecular-weight tolerance. “Correct Mass” assignments were defined as IsoFrac molecular weights lying within a circular distance of approximately 15 kDa from both database-derived molecular weights (Euclidean radius = 1.5×10 kDa). A comparable tolerance was applied when defining the “Same Offset” category to ensure consistent classification across Δ mass groups. The cutoff was chosen empirically to balance assignment specificity and robustness rather than representing a biological threshold. Importantly, the results were robust to the selected cutoff, for instance reducing the tolerance to 5 kDa still yielded more than 2,000 “Correct Mass” assignments, indicating that the overall conclusions are not dependent on the chosen threshold.

DIA-LOP data processing

Data processing was performed using the R Bioconductor packages MSnbase⁵⁸ and pRoloc, adapted from the established workflows^{57,59,61,62}. Contaminants and quality control filtering was based on a minimum intensity threshold, and global missing value filtering was performed with respect to missingness within a fraction (column) and across the fractionation profile (row). Fractions with missing data for over 70% of precursors, and precursors with fewer than two quantitative values across the remaining fractions were removed. Mixed imputation was used as established.^{59,63} Briefly, the missing at random values were imputed by k-NN, and the missing not at random values using the minimum value methodology (min). Precursors were sum normalized before aggregation to the peptide and protein-level by median summarization.

In contrast to classical protein-level DIA-LOP analyses, subcellular spatial profiles were generated at the peptide-level, including both proteotypic and non-proteotypic peptides. Inclusion of non-proteotypic peptides enabled isoform-resolved spatial analysis, allowing detection of protein isoform candidates that exhibit distinct subcellular distributions. Peptide localization was inferred based on protein group assignments derived from fractionation profiles, using a curated set of organelle marker proteins (Table S2). A manually curated reference set of marker proteins was used to define 10 subcellular compartments: chromatin, cytosol, endoplasmic reticulum, Golgi apparatus, plasma membrane, lysosome, mitochondrion, nucleus, peroxisome, proteasome. Peptides assigned to these marker protein groups were used to generate a corresponding marker peptide set, which served as the training input for peptide-level classification of unknown localisation using T-Augmented Gaussian Mixture model using Markov chain Monte Carlo (tagm-MCMC).⁶⁴ A collapsed Gibbs sampler was run in parallel across four independent chains, each for 3,000 iterations at the peptide-level. Convergence was assessed using the Gelman–Rubin diagnostic, and all chains were retained for downstream peptide analysis.

Data integration

To assess the biological relevance of shorter protein isoform candidates identified by IsoFrac, we leveraged independent spatial proteomics data obtained via DIA-LOP. The rationale was that functionally distinct isoforms should exhibit distinct subcellular localization patterns, reflected in separable DIA-LOP profiles.

Principal component analysis (PCA) was performed on the DIA-LOP dataset to capture the major spatial variance structure, derived from the fractionation profiles. From approximately 76,000 uniquely identified peptides in the DIA-LOP experiment, ~59,000 peptides corresponding to ~7,000 genes could be mapped to the IsoFrac dataset and were retained for further analysis.

The first nine principal components captured approximately 80% of the total variance, and only these components were used for downstream distance calculations to focus on biologically meaningful spatial structure while reducing noise.

For each gene with multiple IsoFrac protein isoform candidates, peptide-level DIA-LOP coordinates (PC1–PC9) were used to compute centroids for each isoform candidate. The Euclidean distance between isoform centroids within the same gene was calculated as a measure of spatial separation. To evaluate whether observed distances exceeded random expectation, we generated a randomized control by permuting IsoFrac isoform assignments within each gene while preserving peptide counts. Distances calculated from this randomized assignment were compared to the real distances. For the boxplot representation, peptide-level distances were summarized for each proteoform group by calculating the median log₂ ratio between the distance to the centroids of the other proteoform groups and the distance to the centroid of the respective proteoform group. Thus, each value in the boxplot represents the median relative distance of a proteoform group rather than an individual peptide-level distance.

An enrichment of larger centroid distances in the real dataset relative to the randomized control supports the biological relevance of IsoFrac-identified shorter protein isoform candidates, indicating that they occupy distinct spatial niches rather than representing stochastic fragmentation or noise.

IsoFrac viewer

Borderline or low-abundance cases that do not meet clustering thresholds are not automatically promoted to peaks but can be examined using the provided graphical user interface (GUI). The GUI enables visual inspection of peptide fractionation profiles, cluster assignments and peak detection results allowing expert-guided evaluation.

Nomenclature and terminology

In this study, an IsoFrac peak represents a molecular-weight-resolved protein species detected by IsoFrac. Such species may correspond to protein isoforms, which we define here as distinct polypeptides arising from alternative splicing, alternative translation initiation, or proteolytic processing. In rare cases, large post-translational modifications, such as glycosylation, can also produce molecular-weight shifts detectable by IsoFrac. These cases are best interpreted as PTM-dependent proteoform states rather than distinct protein isoforms.

Operationally, IsoFrac protein isoform candidates are defined as clusters of co-eluting peptide intensity traces across fractions. Individual peptides, identified by MaxQuant,⁴⁹ contribute their fraction-resolved profiles to density-based clustering and subsequent peak detection. The resulting clustered peptide traces represent a coherent molecular-weight-resolved protein species, which is then assigned to candidate database isoforms from UniProt or PacBio-derived entries based on peptide overlap.

UniProt Categories (column CleanCategory) were generated from a compiled UniProt FASTA/GFF resource (Human reference proteome UP000005640) and stored in the compiled UniProt GFF table (gff). Processing was performed in R using seqinr and data.table. UniProt consists of two different databases: UniProtKB/Swiss-Prot, a high quality and reviewed protein sequence database and TrEMBL, a database with unreviewed protein sequences. Based on the header information per sequence and gene we assigned each protein to one of the following categories.

- No Variant: Assigned when only one reviewed protein sequence per gene exists in the UniProt database or in the UniProtKB/Swiss-Prot database, independent of additional entries in the TrEMBL database.
- AS (Alternative Splicing): Assigned when isoform suffixes were present (e.g., “-2”) or when multiple splice variants were annotated for an UniProtKB/Swiss-Prot gene.
- Shorter Variant: Assigned when the UniProt feature type indicated Chain (e.g., post-translational maturation of a precursor protein) or alternative translation products.
- TrEMBL: Assigned to unreviewed TrEMBL entries when no reviewed Swiss-Prot representative was available and multiple TrEMBL variants existed per protein sequence.

We did not distinguish protein sequences from their N-terminal Methionine cleared version.