UNIVERSITÀ DEGLI STUDI DI MILANO

PhD degree in Systems Medicine

Dept. Of Oncology and Emato-Oncology

A *de novo* computational discovery platform from RNA to Protein
BIOS07/A

Roberto Albanese
Matr. R13460
ORCID n. 0009-0008-5256-1682

Tutor name and surname: Dr. Lorenzo Calviello

PhD Program Coordinator: DIEGO PASINI

A.A. 2024-2025

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Lorenzo Calviello, for his constant support, guidance, encouragement, and availability. He conveyed to me not only his scientific knowledge but also enthusiasm and passion for research.

Special thanks also to my advisors, Dr. Francesco Nicassio and Prof. Jeroen Krijgsveld, for their comments and suggestions during my PhD journey.

I would like to thank our collaborators from the Jagannathan Lab, especially Sujatha Jagannathan and Amy E. Campbell, who shared with us beautiful datasets to work with and who were always available and supportive.

I am also grateful to Andrea Graziadei, Syed Azmal Ali, Alberto Riva and Matteo Bonfanti for sharing their knowledge and expertise.

Heartfelt thanks to all members of the Calviello Lab and to all HT people for creating a stimulating and collaborative environment. Thank you also to Francesca Fiore and Veronica Viscardi for their help and guidance with administrative aspects of SEMM and University of Milan.

Finally, I would like to thank my family, who has always believed in me and encouraged me, and my friends, who have made my life brighter and enjoyable during these years.

Thank you to all of you!

Table of Contents

List of Abbreviations

BAM	Binary Alignment Map
cDNA	complementary DNA
CDS	Coding Sequence
CFC-seq	Cap-trap full-length cDNA sequencing
ChIP-seq	Chromatin immunoprecipitation sequencing
CPM	Counts Per Million
CPU	Central Processing Unit
DDA	Data-Dependent Acquisition
DIA	Data-Independent Acquisition
DMSO	Dimethyl sulfoxide
DNA	DeoxyriboNucleic Acid
FDR	False Discovery Rate
FSHD	Facioscapulohumeral muscular dystrophy
GTF	Gene Transfer Format
ID	Identifier
IGV	Integrative Genomics Viewer
iPSCs	induced Pluripotent Stem Cells
LC	Liquid Chromatography
LC-MS/MS	Liquid Chromatography-Mass Spectrometry/Mass Spectrometry
LFQ	Label-Free Quantification
log2FC	log2 Fold Change
LR	Long-Read (data)
NGS	Next-Generation Sequencing
NMD	Nonsense-Mediated Decay
NTC	Normal Termination Codon
ORF	Open Reading Frame
PSM	Peptide Spectrum Match
PTC	Premature Termination Codon
PTM	Post-Translational Modification
RNA	RiboNucleic Acid
SAM	Sequence Alignment Map
SR	short-read (data)
TES	Transcription End Site
TMT	Tandem Mass Tag
TPM	Transcripts Per Million
TSS	Transcription Start Site
UMI	Unique Molecular Identifier
UTR	Untranslated region

Figures Index

Tables Index

Abstract

Data-driven identification and functional characterization of human transcripts and proteins remain challenging tasks in the post-genomics era. Transcriptional and post-transcriptional regulation mechanisms hugely increase RNA isoform diversity, while their contribution to protein synthesis remains vastly unexplored. Moreover, the transcriptome composition changes in different human cell types, tissues, and conditions. Therefore, there is great need for unbiased, dataset-specific annotation efforts. In this regard, transcriptomic and proteomic methods can help elucidate transcript functions and the detection of actively translated Open Reading Frames (ORFs). The main goal of this project is the development of methods for *de novo* identifications of RNAs, ORFs, and proteins directly from the data. We implement a pipeline which couples *de novo* transcriptome assembly, *de novo* ORF detection, and proteome characterization using proteogenomic approaches. Furthermore, we devise computational strategies for the evaluation of *de novo* detection from RNA to protein.

By using our pipeline, we characterize the effects of *DUX4* activation in human skeletal muscle cells as a model for facioscapulohumeral muscular dystrophy (FSHD). Our results show that misexpression of *DUX4*, which encodes an embryonic transcription factor, impairs RNA metabolism by inhibiting Nonsense-Mediated Decay, thus leading to the accumulation of incomplete transcripts and truncated proteins. *De novo* transcriptome assembly allows detection of several unannotated genes and transcripts, including potential novel DUX4 targets, whereas *de novo* ORF finding reveals the presence of translated ORFs within novel transcripts. By using a custom protein database and a deep Tandem Mass Tag (TMT)-labeling proteomics dataset, we identify upregulated novel proteins with evidence for RNA expression in patient-derived data. When further extending the transcriptome annotation by adding long-read-derived transcripts and genes, we find a modest increase in the number of detected changes, showcasing the power of *de novo* approaches even with short read data. Moreover, we analyze how transcript and ORF expression levels as well as the choice of annotation and protein database influence downstream analysis.

In conclusion, our data analysis strategy allows an improved characterization of the functions of the transcribed genome. We characterize the effects of a gene misexpression on RNA metabolism and on the proteome, we identify novel targets in a rare disease, and we investigate the factors influencing results of our unbiased analyses from RNA to Protein.

Introduction

1. Background

This section contains the background of this PhD project. After providing an overview on biological diversity from RNA to Protein, we focus on technologies and platforms for identification and quantification. The section concludes with a summary of the main aims of the project.

1.1 Functional heterogeneity from RNA to Protein

Through transcription of DNA regions, our genome produces RNA molecules, which are collectively referred to as the human transcriptome [1]. However, the transcriptome composition is not static, as it changes in different tissues, cell types, and diseases [2], [3], [4]. Moreover, transcript diversity and functional heterogeneity of the transcribed genome are greatly increased through co-transcriptional and post-transcriptional regulation mechanisms [5].

Alternative splicing, consisting in removal of introns and combinatorial joining of exons, largely contributes to this diversity, with around 95 % of multiexon genes being alternatively spliced [6], [7]. As a consequence of this process, multiple RNA isoforms are produced, differing not only in terms of exon composition, coding sequences, and functions [6], [8] , but also in terms of stability and translational levels [9].

Usage of alternative transcription start sites (TSSs) also increases transcriptomic heterogeneity [5], [10]. It has been estimated that in mammalian genomes more than half of genes have alternative TSSs [5]. Through alternative TSS usage, transcripts with different first exons or different lengths of the 5' untranslated region (UTR) are generated [10]. While using alternative first exons can alter translated open reading frames (ORFs) and lead to the production of proteins with different N-termini, changes in 5' UTR lengths can influence post-transcriptional regulation [10].

In addition, usage of alternative transcription end sites (TESs) and polyadenylation sites contributes to RNA diversity [5], [10]. At least 70 % of genes use alternative polyadenylation sites [5]. Transcripts with different 3' ends can have different coding sequences or 3' UTR lengths, with potential consequences in terms of produced proteins or RNA stability [10].

Other mechanisms contributing to transcriptomic complexity include antisense transcription, RNA cleavage events, and RNA editing [11], [12], [13]. Antisense transcription consists in the production of transcripts from the strand that is opposite to that of protein-coding or non-coding genes [11]. While antisense transcripts do not generally code for proteins, they are known to be involved in gene expression regulation [11].

Such heterogeneity is accompanied by a diversity of translated transcript regions, with different canonical ORFs, upstream ORFs (uORFs), downstream ORFs (dORFs), alternative ORFs, ORFs on alternative RNA isoforms, and translated regions on transcripts which were previously annotated as non-coding [14], [15], [16]. In general, the function of translation is not only the synthesis of stable proteins since this process can also have regulatory roles such as translation regulation, as often reported for uORFs [17], or mRNA decay [18]. Moreover, different co-translation RNA surveillance mechanisms exist, including Nonsense-Mediated Decay (NMD), which removes transcripts containing premature termination codons (PTCs), thus preventing the production of truncated proteins [19]. Additional sources of diversity at the level of translation are alternative translation initiation, ribosomal frameshifting, and stop codon readthrough [20], [21], [22]. Moreover, recent studies have suggested that translation errors, consisting in the incorporation of not encoded amino acids, are more frequent than previously hypothesized [23], [24].

An even larger degree of heterogeneity is present at the level of the proteome, with production of a wide variety of species, differing in terms of structure, function, subcellular localization, and molecular interactions [25], [26], [27]. Proteome diversity is further enhanced by post-translational modifications (PTMs) and proteolytic cleavage events [26], [28], as well as by RNA editing, which can lead to non-synonymous substitutions [13].

To describe isoform-level diversity of the proteome, a single-term, proteoform, has been proposed [26]. This term, which has become widely used, refers to any form of protein that a single gene can produce, taking into account all the different types of heterogeneity, including protein isoforms differing because of genetic variations, alternative RNA splicing, PTMs, and RNA editing [26], [29].

Altogether, mechanisms generating diversity at the level of RNAs, translated ORFs, and proteins can act independently or in combination, thus making the human transcriptome and proteome highly flexible and complex systems. Notably, the extent to which specific RNA isoforms contribute to the proteome needs to be elucidated, and it is still a matter of debate [30], [31]. Therefore, when integrating between transcriptomes and proteomes it is essential to consider the aforementioned mechanisms creating molecular heterogeneity, to get a complete and detailed understanding of the gene expression regulatory cascade (Figure 1).

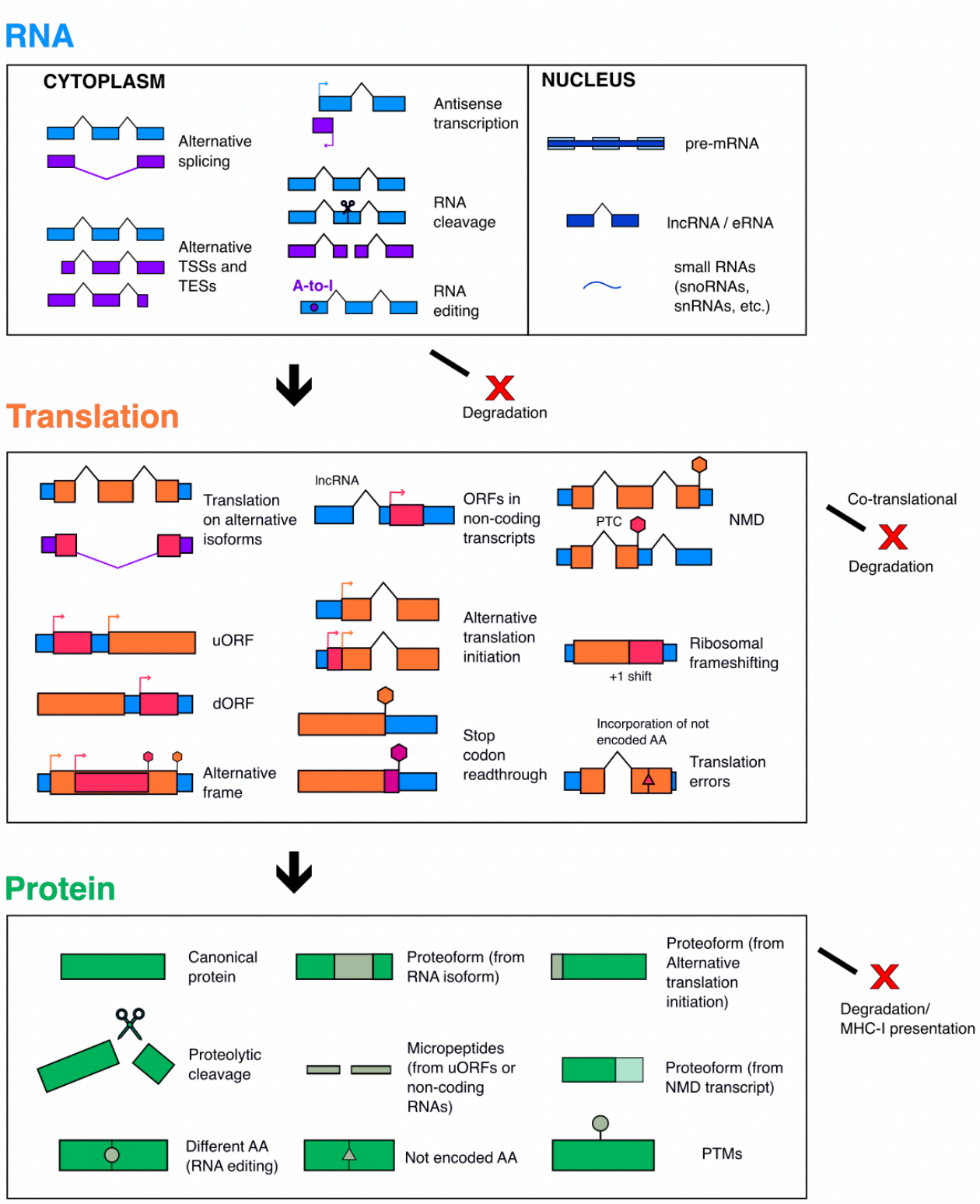


Figure 1 - Overview of the sources of molecular heterogeneity from RNA to Protein.

1.1.1 Post-transcriptional regulation and RNA-Protein correlations

As outlined in the previous section, gene expression is a complex process that happens in multiple steps [32]. Imperfect correlations between mRNA and protein levels often arise, with each step providing its contribution to gene expression [32]. In particular, not only transcription rates but also mRNA stability, protein synthesis rates, and protein stability represent important factors determining protein abundance [32]. However, it is also worth mentioning that discrepancies between detected mRNA and protein levels might arise not only because of biological reasons, but also because of technical limitations, measurement errors, and different contextual confounders [32].

Such discrepancy between RNA and protein is often the focus of integrative approaches in human disease, while often simplifying the molecular heterogeneity of the species involved.

1.1.2 Molecular heterogeneity in disease: the FSHD case

Alterations of processes contributing to molecular heterogeneity are often involved in the emergence and progression of pathological conditions, including cancer, neurodegenerative diseases, and neurodevelopmental disorders [33], [34], [35], [36], [37], [38], [39]. Examples of alterations comprise not only splicing and RNA processing aberrations but also changes in transcription start and termination site usage, which can lead to the presence of different transcript isoforms, translated regions, and protein isoforms.

An interesting example of a disease with altered gene expression is represented by facioscapulohumeral muscular dystrophy (FSHD), which has been linked to activation of the gene *DUX4* in human skeletal muscle [40]. Misexpression of the gene leads to multiple pathological changes, including alterations of RNA metabolism, inflammation, and oxidative stress [41]. Consequences also comprise expression of DUX4 target genes as well as activation of germline genes, retroelements, immune mediators, and previously unannotated intergenic loci [41],

[42], [43]. Studies on FSHD alterations have also focused on chromatin architecture, highlighting the role of 3D nuclear contacts in modulating the expression of muscle atrophy genes [44]. However, while transcriptomic changes in FSHD have been investigated, disease-specific alterations in the proteome are largely ignored, with recent studies suggesting NMD inhibition and consequent production of truncated RNA-binding proteins [45] (Results – Section 2.1), but also proteomic dysregulation of multiple molecular pathways and post-transcriptional buffering of the protein levels of stress-response genes [46].

1.2 From transcriptomes to proteomes: -omics technologies for identification and quantification

In this section, we describe technologies and platforms for the high-throughput detection and quantification from RNA to Protein.

Section 1.2.1 discusses RNA sequencing technologies and *de novo* transcriptome assembly. Section 1.2.2 deals with the experimental protocol Ribo-seq and computational methods for the *de novo* detection of translated regions. Section 1.2.3 focuses on shotgun mass spectrometry and protein database searches. Finally, section 1.2.4 pertains to the emerging and integrative field of proteogenomics, with a particular emphasis on proteogenomic approaches and pipelines.

1.2.1 Transcriptomics and *de novo* transcriptome assembly

Transcriptomics is a fundamental field of research in molecular biology [47]. The profiling of human transcriptomes, which are complex and heterogeneous, allows to shed light on the ways RNA composition and abundances change in different cell types, diseases, and tissues [2], [3], [4]. Such detailed views on transcriptomes have been possible thanks to the emergence of RNA sequencing (or RNA-seq) technologies [47].

Second-generation or next-generation sequencing (NGS), in particular Illumina sequencing, has represented a revolution for the study and analysis of RNAs as it

has enabled the fast production of millions of short sequences (or reads) [47]. The Illumina platforms are based on sequencing by synthesis: fluorescently labelled nucleotides are incorporated and imaged permitting the determination of the sequence of each deoxyribonucleic acid (DNA) fragment [47]. The generated sequences are useful not only to quantify gene expression and identify differential expression patterns, but also to detect splice junctions [47], [48]. Nevertheless, sequences are short, and this is an obstacle for full-length isoform identification [49], [50].

Third-generation sequencing platforms, including PacBio and Oxford Nanopore technologies, overcome some of the NGS limitations [47]. The PacBio IsoSeq protocol produces full-length complementary DNA (cDNA) reads for transcripts that are up to 15 kb long [47]. Oxford Nanopore sequencing is based on a different approach: individual DNA or RNA molecules pass through nanopores, and changes in ionic current are measured; then, basecalling is performed, i.e., electrical current signals are converted to long nucleotide sequences [51]. Notably, this technology offers the possibility of directly sequencing RNAs without a step of reverse transcription [47]. Significant improvements in data quality are continuously reported thanks to the development of new technologies and basecalling methods [51]. Third generation sequencing technologies have been extensively used [2], as they can produce longer sequences when compared with Illumina sequencers, thus helping full-length isoform identification and detection of splicing aberrations [52]. Despite these advantages, long-read technologies suffer from three main limitations when compared with Illumina platforms; the higher cost, the lower throughput, and the lower basecalling accuracy [50].

Independently from the used sequencing technology, aligning reads to a reference genome is an important step in the analysis of RNA-seq data [47]. For short-read data, a widely used tool is the *STAR (Spliced Transcripts Alignment to a Reference)* aligner, which was designed for RNA-seq data and was optimized for speed and the accurate identification of exon-exon junctions [48]. Transcriptome annotations can be provided to *STAR* to improve junction detection [48].

After aligning reads to the genome, transcripts which are present in the analysed samples can be reconstructed starting from the alignment results, and this step is called transcriptome assembly [47]. A widely used transcriptome assembler for

short-read and long-read RNA-seq data is *StringTie3*, a recent and improved version of the original *StringTie* algorithm [53], [54]. After grouping reads into clusters, the *StringTie* assembler constructs a splicing graph for each cluster [53]. In this graph exons or exonic parts are represented as nodes, while candidate splicing junctions as edges [53]. Starting from the splicing graph the assembler identifies transcripts and it uses a maximum flow algorithm to estimate the expression level of each transcript [53].

The development of long-read technologies was later accompanied by the development of long-read-based assemblers, including *IsoQuant* [55]. The *Isoquant* algorithm comprises multiple steps: first, mapped reads are assigned to known transcript isoforms; then, transcripts are quantified and alignments are corrected; finally, transcript models are built [55]. If known transcript isoforms are not provided, transcript discovery is directly performed [55].

Thanks to the joint application of sequencing technologies and computational tools, transcriptomics has become a central, powerful field of research. While next-generation sequencing is useful for high-precision quantification, long-read technologies are useful to directly study entire transcripts and to uncover previously unappreciated layers of transcriptomic complexity. Tools such as *STAR*, *StringTie3*, and *IsoQuant* are often fundamental pillars of many pipelines for discovery-based transcriptomics, a research field which will remain fundamental for molecular biology, even thanks to constant improvements and refinements in technologies and computational methods.

1.2.2 Translation and isoform-aware *de novo* ORF finding

At the interface between the transcriptome and the proteome, translation is a fundamental step of gene expression [18]. A revolutionary technique for the study of translation at a genome-wide scale is ribosome profiling (Ribo-seq), which has particularly high precision and resolution [56]. In the Ribo-seq protocol, a nuclease is added to a cell lysate, and it degrades transcript parts that are not protected by ribosomes [56]. This allows the profiling of ribosome positions on transcripts with single-nucleotide resolution. By detecting 3-nucleotide periodicity in the Ribo-seq

signal it is possible to identify active translation directly from the data [57]. Several studies have used ribosome profiling to identify translated regions [58], [59], [60], [61]. Importantly, Ribo-seq data can also be used not only to quantify translational levels but also to study translation-coupled regulatory mechanisms, such as Nonsense-Mediated Decay (NMD) [18].

The detection of actively translated ORFs represents an important aspect of functional genomics [16]. The same genomic region can contain different coding sequences depending on which transcript isoforms are produced [62]. Methods for isoform-aware *de novo* ORF finding address this challenge by jointly using transcriptome annotations and ribosome profiling data [16]. One of the tools which were developed for isoform-aware ORF detection is the R package *ORFquant* [16]. Given ribosome profiling data and transcriptome annotations, the package identifies translated ORFs using an Occam's razor strategy to identify a subset of transcripts which can fully explain the Ribo-seq data [16]. Importantly, *ORFquant* produces output files containing not only the detected ORFs, but also quantitative estimates of their translation and information on their associated transcript isoform(s) [16].

By jointly using ribosome profiling and isoform-aware ORF finding methods, it is possible to obtain a detailed overview of the translational landscape. Tools such as *ORFquant* exploit ribosome profiling data for the accurate annotation of translated sequences, while also considering the heterogeneity of transcript isoforms and the intricate mechanisms of translational regulation.

1.2.3 Shotgun mass spectrometry and proteomic searches

Modern proteomics heavily relies on mass spectrometry, which allows to identify and quantify the abundance of peptides and proteins. A mass spectrometer is an instrument to measure the mass-to-charge ratio of ions, and its main components are the ion source, the mass analyser, and the detector. By using techniques such as electrospray ionisation or matrix-assisted laser desorption/ionisation, the ion source converts sample molecules into ions, which are then separated according to their mass-to-charge ratios by the mass analyser. Finally, the number of ions at each

mass-to-charge value are recorded by the detector, and mass spectra are generated.

Comprehensive protein identification can be achieved with shotgun mass spectrometry, which is also known as bottom-up proteomics [63]. When using this approach, proteins are extracted from biological samples, and they are digested into peptides using proteolytic enzymes [63]. Peptides are separated, often through liquid chromatography (LC), and ionised [63]. Then, after a first mass spectrometry step and fragmentation, they go through a second mass spectrometry step [63]. Finally, through this procedure (liquid chromatography-mass spectrometry/mass spectrometry, or LC-MS/MS), tandem mass spectra, which are signals corresponding to the injected peptides, are detected [63]. Such spectra are used to identify the peptides and to infer the proteins from which they derive [64].

When using the approach of shotgun proteomics, data can be obtained by using two main strategies: Data-Dependent Acquisition (DDA) and Data-Independent Acquisition (DIA) [65]. In DDA specific precursor ions are first selected for subsequent fragmentation [65]. A limitation of this approach is that lower-abundance peptides are not selected, leading to significant portions of missing data [65]. On the other hand, when using DIA all ions within predefined mass-to-charge windows are systematically fragmented [66]. In this way, highly complex and challenging to resolve spectra are produced, requiring advanced computational methods for data analysis [66]. DIA data analysis can be performed with a spectral library built with DDA data, but now neural networks can predict such a library without DDA [67]. Even if DIA is characterized by higher reproducibility and improved data completeness [65], both acquisition strategies are important and widely used in proteomic studies [65].

Peptides and proteins can be identified by performing database searches [64]. In conventional workflows, experimentally derived spectra are compared against theoretical spectra deriving from a reference protein database [64]. Peptides are identified by peptide spectrum matches and reporting a subset of the matches as identifications [68]. Proteomic searches can also be performed by providing custom databases, which can include predicted protein sequences [64]. By using custom protein databases, it is possible not only to increase the likelihood of correct peptide identifications but also to detect unannotated peptides [64]. Such discovery-oriented

proteomic analyses can also be complemented with orthogonal datasets, including transcriptomic and ribosome profiling data, thus strengthening confidence in novel identifications (Introduction 1.2.4).

The size, quality, and content of the protein database can heavily influence the data analysis results [69], as the database is meant to contain sequences of proteins that are present in the samples of interest, plus contaminants. To control for false identifications, the database also contains decoys, obtained by reverting protein sequences in the database: the scores of decoys are assumed to be distributed as scores of incorrect identifications [70]. However, when this does not hold true, there can be loss of control of False Discovery Rate (FDR) [70]. The risk of false positives can increase when using very large databases [64], and often complicated by the presence of numerous PTMs [71].

Inferring proteins from peptide-level evidence is not a trivial task, as identification errors propagate when aggregating peptide-level evidence, and as peptides are often mapping to multiple proteins, especially when considering the presence of protein isoforms from the same gene [72]. To address this issue, which is referred to as the protein inference problem, protein groups containing multiple database entries can be defined with different algorithms and used in proteomic analyses.

By performing proteomic searches, proteins and peptides can be not only identified but also quantified with labeling and label-free approaches [73]. For example, when using Tandem Mass Tag (TMT) labeling, chemical tags are added to peptides, with each tab being applied to a specific sample [74]. By performing demultiplexing of the tags, peptides are then associated to their samples of origin. The number of tags in TMT labeling kits has increased over time [75], [76], allowing for more efficient and cost-effective proteomic analysis.

Altogether, shotgun proteomics and protein database searches are instrumental for the characterization of complex proteomes, and they provide powerful insights into proteome composition and regulation.

1.2.4 Proteogenomic applications and workflows

The interaction between the fields of genomics, transcriptomics, and proteomics is referred to as proteogenomics [64]. Proteogenomic applications are based on the idea that genomic variation and transcript-level events strongly affect the final proteome. It is possible to detect mutations with a possible impact on already existing protein sequences. By combining multiple -omic datasets, proteogenomic approaches can be used to quantify evidence at protein-level for genomic and transcriptomic variation. A few large-scale initiatives which jointly analyse RNA-seq and proteomic data exist, with the Clinical Proteomic Tumor Analysis Consortium (CPTAC) representing an example of collaboration in which proteogenomic methods are applied to study different cancers, albeit without isoform-level analyses or strong focus on novel ORFs [77].

Typically, the first step of a proteogenomic workflow consists in obtaining genomic or transcriptomic data [64]. Genomic and transcript sequences are often analysed to detect variants, insertions or deletions that can affect protein sequences. RNA-sequencing and ribosome profiling data are also used.

The next step consists in building a customized database of protein sequences [64]. In standard proteomic workflows, a reference protein database is used. In contrast, proteogenomic workflows generate and use databases that include not only canonical proteins, but also other protein sequences, including variant sequences, alternative isoforms, and translation products of unannotated transcripts. The control of database size is important (Introduction 1.2.3).

An important characteristic of proteogenomic approaches is the possibility to improve and filter protein databases based on data deriving from multiple -omics. First, only proteins from transcripts with evidence for RNA expression in specific samples might be included in the database [72]. Moreover, analysis of ribosome profiling data is often performed to expand protein databases [78]. Finally, integration of long-read RNA-seq and proteomic data can improve the detection of protein isoforms [69].

Proteogenomic approaches are useful for their wide variety of possible applications. First, they can be used to detect protein variants as well as potential neoantigens

[79]. Recent progress in modern, personalized medicine greatly benefitted from precise identification of disease or patient-specific transcriptomes and proteomes, as exemplified by targeted immunotherapy strategies making use of cancer-specific neoantigens [80]. Moreover, proteogenomic approaches allow the discovery of novel translation products, including peptides that are produced from unannotated translated regions [64]. Recently, proteogenomic workflows that are focused on the detection of microproteins also started to become more popular [81]. Proteogenomics is also important for the study of translation products of fusion transcripts [82]. The analysis of RNA-seq data can provide evidence for the presence of fusion transcripts, and this information can be used when building protein databases.

Overall, proteogenomic applications and workflows are fundamental and useful methods to understand and characterize how heterogeneity at the level of RNAs and translation affect proteomic diversity at the level of single isoforms.

2. Aims

As highlighted in the previous section, there is great need for integrating between RNA and protein, with isoform-level resolution and including dataset-specific functional annotations. For this reason, the main goal of this project is the development of computational methods for the *de novo* identifications and integration of transcripts, ORFs, and proteins directly from the data.

To achieve our goal, we defined and pursued different aims, which are briefly outlined below.

- **A computational discovery platform from RNA to Protein** (Results – Section 1): We devised and implemented a pipeline that combines *de novo* transcriptome assembly, *de novo* ORF finding, and proteome-wide discovery using a proteogenomic approach.
- **Investigation of the consequences of *DUX4* activation in a model of FSHD** (Results – Section 2): We applied our pipeline to analyse RNA-seq, Ribo-seq, and TMT proteomic data of human skeletal muscle cells in which the gene *DUX4*, which encodes an embryonic transcription factor, was activated. These cells are used as a model of FSHD (Introduction – Section 1.1.2).
- **Benchmarking methods from RNA to Protein** (Results – Section 3): We developed and used computational strategies for the evaluation of results of our *de novo* pipeline. In the context of the international FANTOM consortium [83], we also analysed and compared long-read-derived transcriptome annotations.

Overall, this work shows that data-driven approaches allow to obtain a more accurate characterization of human transcriptomes and proteomes, revealing the presence of transcripts, ORFs, and proteins that would otherwise be undetected. It also presents methods for the assessment of the obtained results at each step.

Results

1. Accelerating *de novo* discovery from RNA to Protein with R2T2P

This section of the thesis describes our computational pipeline, R2T2P (RNA to Translation to Protein)., which links *de novo* discovery of transcripts and translated transcript regions and the identification and quantification at protein level. Section 1.1 contains a general, high-level overview of the entire workflow, while the sections 1.2, 1.3, and 1.4 provide information on its three modules, with a particular focus on the rationale, data processing, and computational strategies for the main steps. Section 1.5 concerns the Nextflow implementation of R2T2P, with a comprehensive representation of the entire pipeline, followed by details on computational requirements and performances.

1.1 General overview of the pipeline

The pipeline R2T2P comprises three main steps, i.e., *de novo* transcriptome assembly, isoform-level *de novo* ORF finding, and proteome characterization (Figure 2).

First, transcripts are reconstructed with short-read RNA-seq data and the transcriptome assembler *StringTie3* (RNA module). Then, translated regions on transcripts are detected with Ribo-seq data and the R package *ORFquant* (Translation module). Finally, peptides and proteins are identified and quantified with the computational platform *FragPipe* by running custom protein database searches (Protein module). R2T2P also includes quality control checks and differential expression analyses with all the used data types, i.e., RNA-seq, Ribo-seq and LC-MS/MS. Their output files are integrated into a final report summarizing results of the entire workflow.

Figure 2 - High-level representation of the computational pipeline R2T2P.

1.2 The RNA module

The RNA module comprises steps for *de novo* transcriptome assembly (Figure 3). Specific tools and parameter settings were chosen to improve the reliability of the assembled transcripts (more details in the section Materials and Methods – RNA-seq and Ribo-seq alignments; Materials and Methods – Transcriptome assembly).

Before transcriptome assembly, two genome alignments are performed. In the first alignment, short-read RNA-seq data and a reference transcriptome annotation file, such as a GENCODE annotation file [84], are provided to the aligner *STAR* [48] (Introduction – Section 1.2.1), which is used in two-pass mode. This approach enables *STAR* to detect and quantify on-the-fly exon-exon junctions that are not present in the annotation. All the novel junctions discovered during this first alignment, as well as the ones in the reference annotation, are subsequently provided to *STAR* for a second genome alignment. *De novo* splice junction discovery is disabled during this second alignment, ensuring that only annotated and previously identified junctions are considered and thereby reducing the risk of including spurious junctions.

The alignments produced in the second step are used for transcriptome assembly with *StringTie3*, the most recent major update of the widely used *StringTie* [53], [54] (Introduction – Section 1.2.1). Alignments of all replicates are pooled to increase sensitivity for transcript detection, while later in the pipeline counts from individual replicates are used to robustly identify regulated novel species (Results – Section 1.3). Even if R2T2P does not produce replicate-specific transcriptome annotations,

a custom annotation can be specified and included in the analysis (following paragraph and Results – Section 2.4). The reference annotation is provided during assembly to guide transcript reconstruction, while still allowing the identification of novel isoforms. *StringTie3* was selected and included in the pipeline as it is well suited for large datasets and for the discovery of novel transcripts in complex transcriptomes. Moreover, it can process RNA-seq data from different library types. *StringTie3* generates an annotation file containing the reconstructed transcripts.

In the final stage of the RNA module, the *StringTie3* annotation is compared to the reference annotation using *GFFCompare*, which is a specialized tool for the evaluation and comparison of transcriptome annotations [85]. *GFFCompare* is used to classify the reconstructed transcripts based on their positions relative to reference transcripts. It also returns performance metrics such as sensitivity and precision at various levels, including exon, intron, intron chain, and transcript levels. Then, the two annotations are integrated using a custom R script (Materials and Methods – Combining GTF files). Original transcript biotypes (e.g. protein-coding, retained_intron, etc..) and *GFFCompare* transcript classes are harmonized when merging the two annotations. *StringTie3*-assembled transcripts which are already present in the reference annotation are not included in the output file of the merging: transcripts with exact intron-chain matches to reference transcripts, as well as those fully contained within reference transcripts and intron compatible (as defined by *GFFCompare*), are filtered out to avoid redundancy. Optionally, an additional, user-provided annotation can also be included in the final step of the RNA module. In this case, *GFFCompare* and R are used to compare and merge the *StringTie3* and user-provided annotations (Materials and Methods – Combining GTF files). When merging the *StringTie3* and user-provided annotations, new gene entities are defined whenever transcripts from the user-provided annotation are associated with *StringTie3* genes. The resulting file is then merged with the reference annotation.

The final output of the RNA module is therefore a transcriptome annotation containing annotated and novel transcripts. Alignment results and the *StringTie3* annotation are subjected to quality checks: mapping statistics, including the number and percentage of mapped reads, are obtained; moreover, the *StringTie3* annotation is evaluated with *GFFCompare* performance metrics.

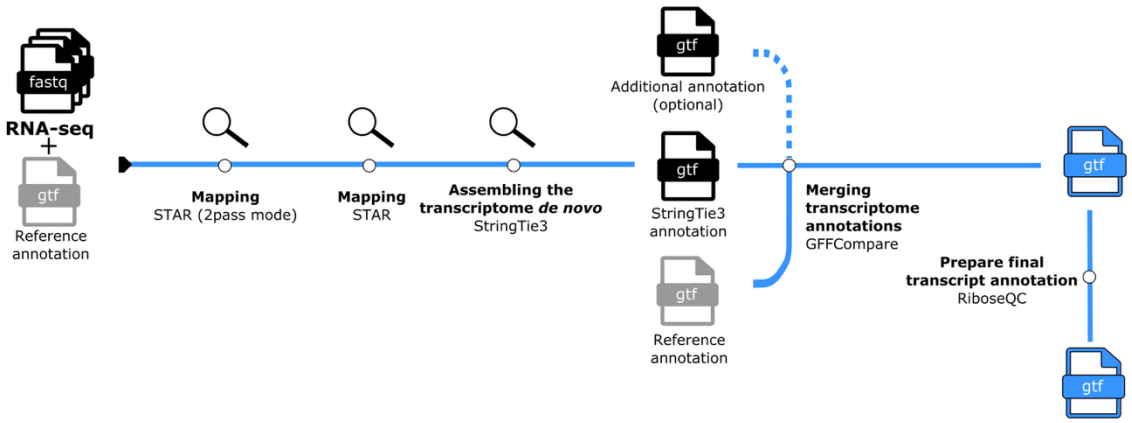


Figure 3 - The R2T2P RNA module.

Main steps are shown in bold; tools and files, including their formats, are also indicated. Quality checks are represented by magnifying glass icons.

Quality checks:

- Mapping statistics in *STAR* log output files for both alignment steps
- *GFFCompare* performance statistics for transcriptome assembly

1.3 The Translation module

The Translation module comprises steps for *de novo* open reading frame (ORF) identification at the isoform level (Figure 4). This module uses RNA-seq and ribosome profiling (Ribo-seq) data as well as exon-exon junctions of the transcriptome annotation generated by the RNA module (Results – Section 1.2), enabling a comprehensive characterization of transcript-specific translated regions (Materials and Methods – RNA-seq and Ribo-seq alignments; Materials and Methods – ORF finding).

First, RNA-seq and Ribo-seq reads are mapped to the reference genome using the *STAR* aligner [48], which is provided with exon-exon junctions of the augmented transcriptome annotation to guide splice-aware alignment. The resulting alignments are subsequently analysed using the R package *RiboseQC* [86], which returns read counts and statistics on a wide range of annotation features, including genes, coding regions, and untranslated regions. The package also performs automatic strandedness and paired-end status detection.

Then, Ribo-seq alignments and *RiboseQC* output files are provided to the R package *ORFquant* to identify actively translated regions within the augmented transcriptome [16]. Unlike many ORF detection tools available, *ORFquant* performs isoform-aware analyses (Introduction – Section 1.2.2). This is particularly useful when analysing complex transcriptomes that are characterized by alternative splicing, as different isoforms of the same gene may have different translational levels (Introduction – Section 1.1). The package allows the systematic investigation of translation events and the discovery of unannotated translated regions in the human transcriptome. RNA-seq and Ribo-seq differential analyses are also performed at various levels, including at the level of genes and individual ORFs, using the R packages *DESeq2* and *DEXSeq* [87], [88], [89].

Finally, three protein databases are built, i.e., a database of annotated proteins, a database of protein sequences corresponding to *ORFquant*-detected ORFs (hereafter, the *ORFquant* database), and a database comprising both sets of protein sequences. Annotated protein sequences are extracted from the coordinates of coding sequences (CDSs) of the reference annotation file. The three databases are used for multiple proteomic searches in the Protein module (Results – Section 1.4). The annotated protein database allows identifications relying on the already existing annotation, while the *ORFquant* database permits the detection of novel translation products. The combined database can be used to run a comprehensive search leveraging the strengths of both sets of protein sequences.

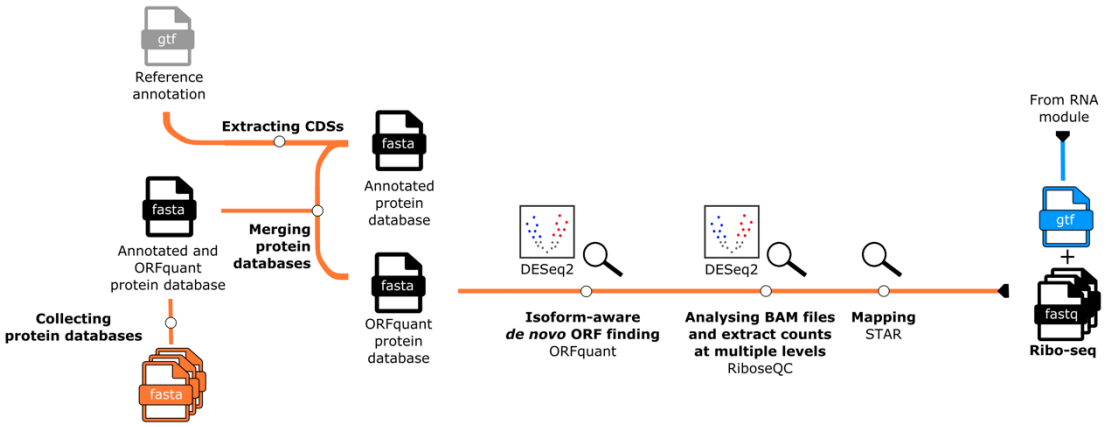


Figure 4 - The R2T2P Translation module.

Main steps are shown in bold; tools and files, including their formats, are also indicated. Differential expression analyses are represented by volcano plot icons.

Quality checks:

- Mapping statistics in *STAR* log output files for the alignment step
- Read counts, read lengths, and statistics in *RiboseQC* output files
- Numbers, lengths, annotation, and quantification estimates of *ORFquant*-detected ORFs

1.4 The Protein module

The Protein module contains various steps for mass spectrometry data analysis (Figure 5; Materials and Methods – Proteomic searches). The module takes as input experimental data, i.e., DDA TMT, DDA label-free quantification (LFQ), or DIA data, as well as the three protein databases that are generated from module Translation (Results – Section 1.3).

In the first step of the module, decoys and contaminants (Introduction – Section 1.2.3) are added to protein databases with the software *Philosopher* [90].

Then, proteomic searches are performed by providing experimental data and the protein databases to the computational platform *FragPipe* [91]. *FragPipe* is a powerful platform for proteomic data analysis as it combines different well-established tools within a single framework. Depending on the type of mass spectrometry protocol, a combination of the following data analysis tools is employed.

MSFragger is a fast and efficient database search engine which is used for peptide identification [92]. *MSFragger* also includes the module *MSFragger-DIA*, which allows to directly identify peptides from DIA data [91]. *MSBooster* improves *MSFragger* results by rescoring peptide-to-spectrum matches (PSMs) using additional features from the mass spec run within a deep learning analysis framework [93]. PSM-level results are improved with the post-processing tool *Percolator* [94], which uses support vector machines to accurately distinguish between targets and decoys. *IonQuant* is a tool for quantification with label-free and labelling data [95], while *TMT-Integrator* provides quantification reports starting from data of isobaric labelling samples [96].

Other tools are specifically designed for the analysis of DIA data: while the Python package *EasyPQP* is useful for library generation, *DIA-Umpire* and *DIA-NN* support DIA data analysis without the presence of a spectral library. [97], [67].

Finally, the *Philosopher* toolkit comprises multiple proteomic data analysis tools and performs different steps, including FDR scoring and generation of reports at the levels of PSMs, peptides, and proteins [90].

A key characteristic of *FragPipe* is the use of workflow files, which contain user-specified search settings and allow to specify the combinations of tools to use, depending on the data type (DDA, DIA, or TMT) and on user choices. *FragPipe* log files then provide confirmation of the successful execution of all data analysis steps. Identification and quantification results can be generated at multiple levels, enabling downstream analyses at the levels of peptides, protein groups, and genes. Finally, in R2T2P differential expression analyses at the peptide level are conducted in R using *limma* [98]

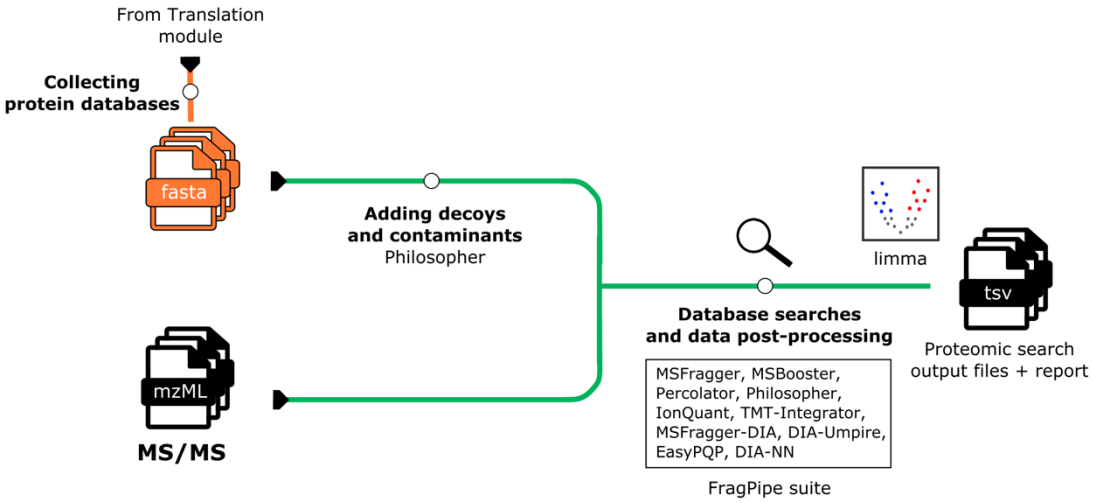


Figure 5 - The R2T2P Protein module.

Main steps are shown in bold; tools and files, including their formats, are also indicated.

Quality checks for proteomic data analysis:

- Detected PSMs across samples
- Number of detected and quantified peptides, protein groups, and genes
- FDR filtering algorithm and FDR thresholds for PSMs, peptides, ions, and proteins
- FragPipe log files, with details on execution of all the analysis steps

1.5 Nextflow implementation of R2T2P

The pipeline R2T2P is implemented in Nextflow [99]. Using the workflow manager Nextflow offers several advantages. First, Nextflow implements the pipeline with a modular architecture. This modularity facilitates not only code reusability, but also maintenance, update, extension, and replacement of specific components. Nextlow also makes the pipeline portable as it supports containerization, ensuring the usage of consistent software versions and environment configurations. Moreover, it improves reproducibility of results across different operating systems.

Figure 6 contains a comprehensive representation of R2T2P. The Nextflow pipeline was applied to analyse an example dataset including RNA-seq, Ribo-seq, and proteomic data.

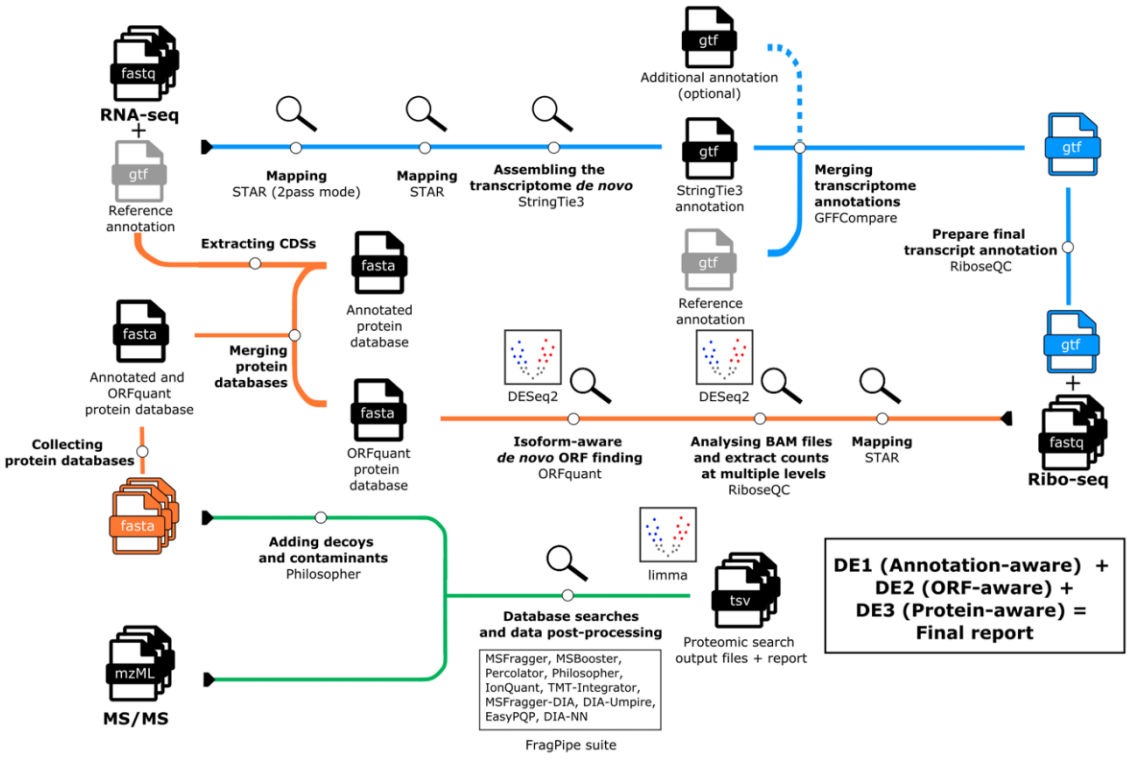


Figure 6 - The entire pipeline R2T2P.

Main steps are shown in bold; tools and files, including their formats, are also indicated. Details on the differential expression analyses and the final report are also shown in the bottom right.

The example dataset is described in Results – Section 2. Figure 7 shows details about running times and required computational resources for each step of the pipeline in this analysis.

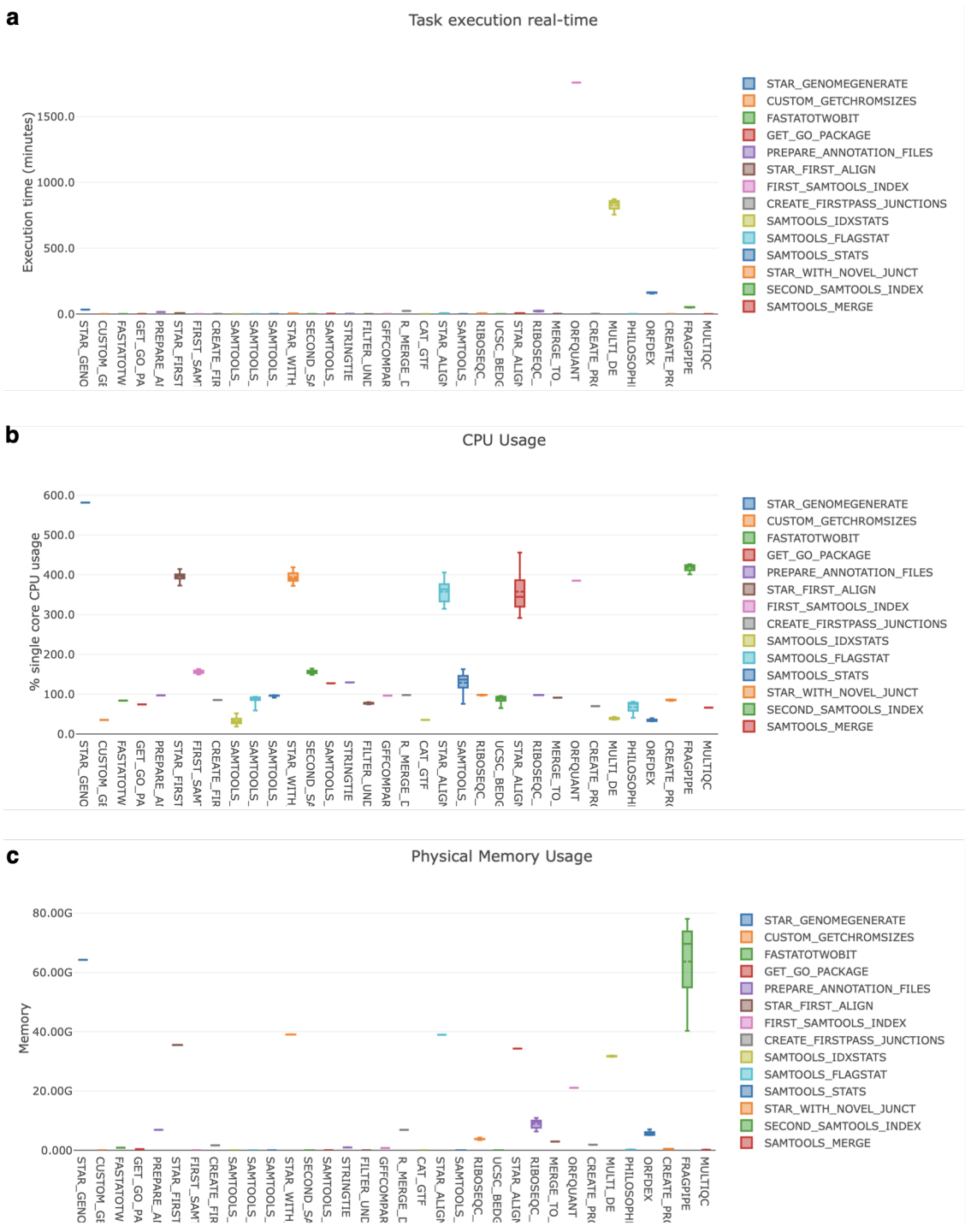


Figure 7 - Execution times and required computational resources for all the steps of the Nextflow implementation of R2T2P.

Execution times (a), percentage of single core CPU (Central Processing Unit) usage (b), and memory usage (c) for all the steps of the Nextflow pipeline, when applying it to an example dataset which is described in Results – Section 2. The plots are included in the Nextflow execution report of the pipeline.

Isoform-aware *de novo* ORF finding with *ORFquant* and the differential expression analyses using read counts of the alignments were the most time-consuming steps of the pipeline. Different steps, including the *STAR* genome index generation and the *FragPipe* proteomic searches, used more than a single CPU. Genome index generation and *FragPipe* searches were also the steps which required the highest amount of memory.

Optimization of the execution times and fine-tuning of the required computational resources will be possible after running the pipeline with multiple datasets and collecting data on the performance of each step. In this optimization and refinement process it will be essential to use datasets exhibiting different characteristics, including different numbers of sequencing reads, RNA-seq library types, and proteomic data types (DDA and DIA).

2. Unbiased analysis from RNA to Protein in FSHD models

This section of the thesis concerns the analysis of RNA-seq, Ribo-seq, and TMT proteomics of a model of facioscapulohumeral muscular dystrophy (FSHD). This myopathy has been linked to misexpression of the gene *DUX4*, which encodes an embryonic transcription factor, in human skeletal muscle (Introduction – Section 1.1.2).

The FSHD model consists of human skeletal muscle cells in which the gene *DUX4* was activated. The dataset comprises RNA-seq and Ribo-seq data from three replicates at 0, 4, 8, and 14h after *DUX4* activation, as well as proteomic data from three replicates at matching time points (excluding 0h) after gene induction or DMSO treatment (Figure 8). The proteomic dataset was obtained with the TMT 16plex, and it included a reference channel, consisting of a mix of all samples and to be used for normalization (Materials and Methods – Proteomic searches).

The entire dataset (RNA-seq, Ribo-seq, and proteomic data) was provided by our collaborators of the Jagannathan Lab from University of Colorado (Materials and Methods – Data description and availability). However, while the RNA-seq and Ribo-seq data were used in a *Cell Reports* paper to which we contributed and are publicly available [45], the proteomic dataset is unpublished.

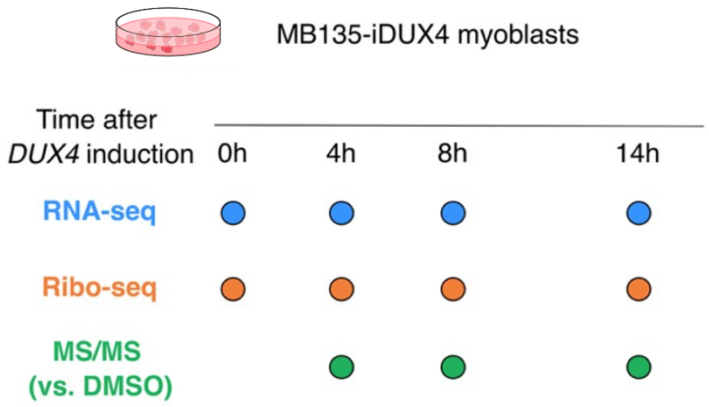


Figure 8 - Experimental procedures used to generate the FSHD model dataset.

Section 2.1 contains details about the RNA-seq and Ribo-seq data, with a particular emphasis on their depth and resolution, as well as our previously published results obtained by analysing the same data [45].

Section 2.2 shows results of the RNA and Translation modules of R2T2P. It reports the numbers of annotated and novel transcripts; then, it focuses on upregulated novel genes, as well as evidence for direct DUX4 binding. Finally, this section concludes with statistics on the detected translated ORFs, including their categories, numbers, lengths, and estimates of their translational levels.

Section 2.3 focuses on results obtained with the Protein module of R2T2P: after providing the numbers of peptides which were detected in the combined database search (Results – Sections 1.3-1.4), this section considers the effects of *DUX4* activation on the proteome, highlighting upregulation of novel peptides following the gene induction, showcasing some examples. Moreover, we provide evidence for novel genes in two RNA-seq datasets from patient-derived myotubes and patient biopsies (RNA-seq datasets from [42], [100]; Materials and Methods – Data description and availability).

Finally, section 2.4 shows results of the entire pipeline when adding a long-read-derived annotation in the final step of the RNA module (Results – Section 1.2): this section examines how including long-read-derived transcripts improved identification and quantification results at the transcriptome and proteome levels.

2.1 Inhibition of nonsense-mediated decay following *DUX4* activation results in production of truncated proteins

RNA-seq and Ribo-seq data of the FSHD model were inspected to ensure adequate quality before proceeding with the computational analyses of this study. Numbers of raw, cleaned, mapped, and uniquely mapped reads are shown in Table 1.

	Raw reads	Cleaned reads	Mapped reads	Uniquely mapped reads
RNA_00h_rep1	8,431,849	8,362,109	8,171,358	7,086,062
RNA_00h_rep2	9,252,046	9,172,024	8,985,280	7,630,053
RNA_00h_rep3	8,768,030	8,692,011	8,515,409	7,197,980
RNA_04h_rep1	8,676,700	8,583,778	8,418,310	7,157,571
RNA_04h_rep2	9,109,763	8,697,890	8,525,774	7,182,759
RNA_04h_rep3	8,758,875	8,620,353	8,460,169	7,074,396
RNA_08h_rep1	9,284,255	9,167,494	9,035,998	7,341,264
RNA_08h_rep2	9,923,741	9,844,099	9,705,049	7,729,796
RNA_08h_rep3	9,087,630	8,966,607	8,836,825	7,047,713
RNA_14h_rep1	9,698,512	9,575,893	9,465,309	7,836,756
RNA_14h_rep2	7,623,289	7,534,328	7,441,749	6,021,528
RNA_14h_rep3	8,974,779	8,854,789	8,742,631	6,929,292
Ribo_00h_rep1	19,839,044	10,351,767	8,052,975	6,042,578
Ribo_00h_rep2	14,307,922	7,188,514	5,613,384	4,102,999
Ribo_00h_rep3	15,862,982	7,513,450	3,153,463	2,078,784
Ribo_04h_rep1	61,429,483	10,415,691	7,642,914	3,944,600
Ribo_04h_rep2	9,447,570	4,442,700	1,771,088	1,206,994
Ribo_04h_rep3	18,133,986	6,646,372	5,409,141	3,865,159
Ribo_08h_rep1	19,492,323	10,048,359	8,008,834	5,656,782
Ribo_08h_rep2	62,739,718	9,619,659	5,995,074	1,817,877
Ribo_08h_rep3	19,485,233	10,841,797	8,725,064	6,282,900
Ribo_14h_rep1	19,953,228	7,701,944	7,335,647	4,885,360
Ribo_14h_rep2	12,254,080	4,301,836	3,764,756	2,436,575
Ribo_14h_rep3	51,262,280	11,369,944	5,521,166	1,928,998

Table 1 - Mapping statistics of RNA-seq and Ribo-seq data from iDUX4 cells

Albeit not very deep, the Ribo-seq data exhibited excellent resolution and the expected three-nucleotide periodicity (Figure 9).

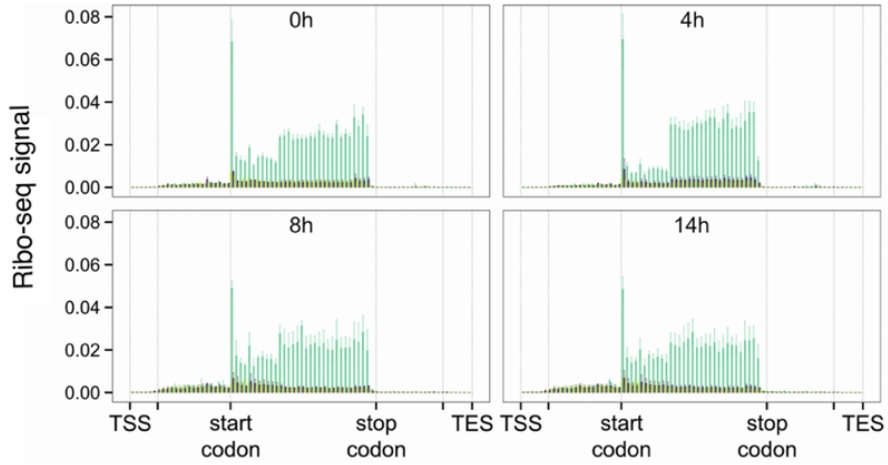


Figure 9 - Quality control of Ribo-seq data (Adapted from Campbell et al., Cell Reports, 2023). Aggregate profiles of P-sites are shown on the y axis, and different colours are used for the frames.

We then analysed the RNA-seq and Ribo-seq data to investigate the consequences of *DUX4* activation on nonsense-mediated decay (NMD). By using the R package *ORFquant* we detected unannotated, NMD-inducing ORFs on NMD target transcripts.

We used those ORF coordinates as well as canonical ORFs to quantify and visualize changes: we visualized RNA-seq and Ribo-seq profiles around stop codon pairs, premature termination codons (PTCs) and normal termination codons (NTCs), from the same genes (Figure 10; Materials and Methods – Metaplots of RNA-seq and Ribo-seq profiles around stop codons).

As expected, ribosome footprints exhibited 3nt periodicity also at PTCs, and signal coverage dropping after the stop codons, thus showing active translation of NMD-sensitive transcripts.

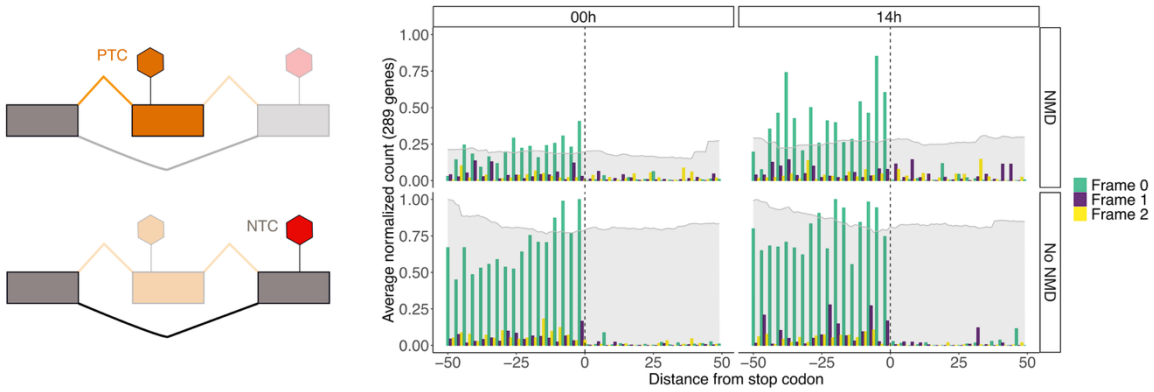


Figure 10 - RNA-seq and Ribo-seq profiles (metaplots) around the stop codons (Adapted from Campbell et al., Cell Reports, 2023).

Windows around stop codons of NMD-inducing ORFs (top) and canonical ORFs (bottom) at 0h and 14h after activation of the gene DUX4. The x-axis represents positions inside 100 nt windows around genomic coordinates of stop codons, whereas the average normalized count is on the y axis. A barplot and a grey area are used for Ribo-seq and RNA-seq, respectively. A conceptual schematic of NMD and no-NMD transcripts is shown on the left.

RNA-seq and Ribo-seq signals around PTCs increased over time (Figure 11), with increase of translation and RNA expression already visible at early time points, with signal at NTCs not increasing. These findings suggest that *DUX4* activation leads to NMD inhibition and stabilization of NMD-targeted transcripts. Since these transcripts are not degraded, their RNA-seq levels increase over time. Moreover, NMD target transcripts are actively translated, leading to the production of different, truncated proteoforms.

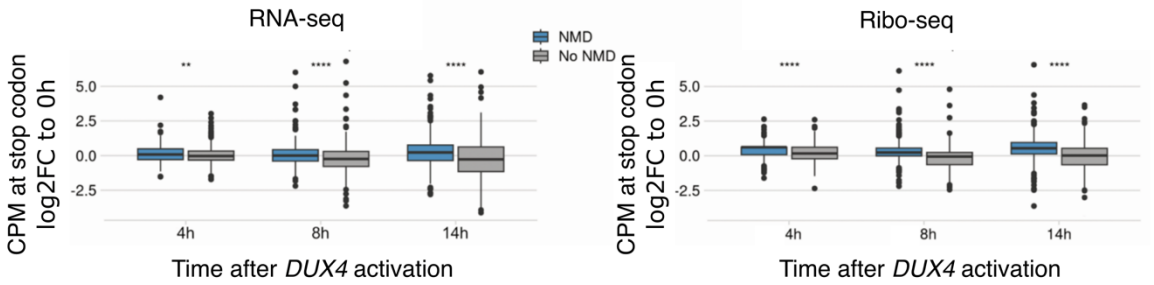


Figure 11 - Quantification around stop codons (Adapted from Campbell et al., Cell Reports, 2023).

RNA-seq (left) and Ribo-seq (right) log2 fold changes relative to 0h using Counts Per Million (CPM) values are on the y axis.

2.2 Novel genes among the top upregulated DUX4 targets in human muscle cells

We were interested in further characterizing the consequences of *DUX4* activation in our model of FSHD. For this reason, we used R2T2P to analyse the data of the FSHD model, and we obtained a comprehensive annotation of transcripts, genes, and open reading frames (ORFs): this approach produced results comprising both known features (in GENCODE v47 annotation) as well as unannotated genes and transcripts, obtained from the RNA module using the RNA-seq data (Table 2).

Transcript type	**Number of transcripts (genes)**
Annotated	385640 (78705)
Novel antisense	407 (357)
Novel antisense intronic	573 (540)
Novel intergenic	624 (585)
Novel intronic	1034 (890)
Novel isoform	8127 (5308)

Table 2 - Numbers of annotated and novel transcripts in the output annotation of the R2T2P RNA module.

Number of genes are in parentheses.

Given the low depth of our sequencing data, we detect a few hundred novel transcripts and genes. However, perhaps more importantly, the pipeline allowed the identification of differentially (novel and known) expressed genes at each time point after gene activation, relative to the control condition. Collectively, these analyses provided detailed insights into changes in this novel transcriptome, revealing its heterogeneity and dynamic regulation over time following *DUX4* activation.

2.2.1 Regulated expression of novel genes following *DUX4* activation

To gain deeper insights into the effects of *DUX4* activation on transcriptome composition, we used the transcriptome annotation generated by the pipeline to perform gene-level differential expression analyses using RNA-seq and Ribo-seq data (Materials and Methods – Gene-level differential expression analyses). This provided a time-resolved view of gene regulation, allowing the identification of novel genes exhibiting significant changes in RNA-seq and Ribo-seq levels. Notably, we observed a progressive increase over time in the number of upregulated novel genes detected by RNA-seq and Ribo-seq analyses, reaching ~500 upregulated novel genes at the level of RNA, and 136 genes upregulated in their translation (Figure 12). This trend suggests the upregulation of previously uncharacterized transcriptional units as part of the transcriptome remodelling following gene activation. We classified the upregulated novel genes according to their biotype, enabling the distinction between sense and antisense genes. In this analysis, known DUX4 target genes were used as positive controls. As expected, the number of upregulated genes (both novel and known targets) increased over time, albeit lower for Ribo-seq.

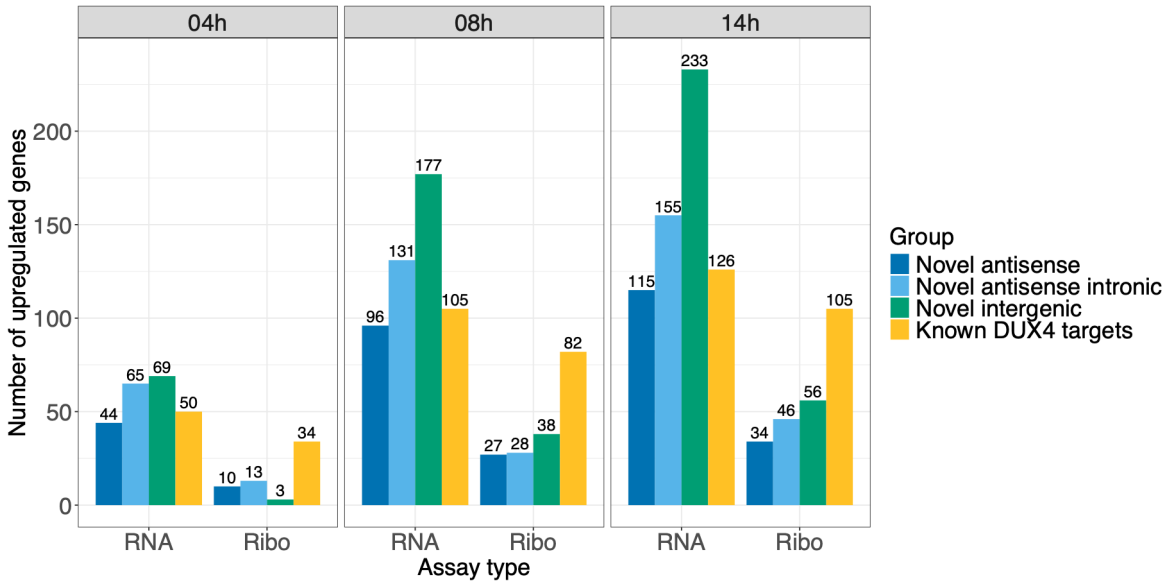


Figure 12 - Gene-level RNA-seq and Ribo-seq differential expression analyses (numbers of upregulated genes).

The assay type is on the x axis, whereas the number of upregulated genes is on the y axis.

We then focused on the comparison of gene expression levels at 14 h after *DUX4* activation relative to 0 h. We visualized log2 fold changes and adjusted p-values of differentially expressed genes to assess both the extent of expression changes and their statistical significance (Figure 13).

This visualization provided a clear overview of which genes were most strongly affected by the gene activation, revealing the presence of upregulated novel genes whose log2 fold changes were comparable to those observed for known DUX4 targets. This similarity in the magnitude of expression changes suggests that these novel genes may be subject to similar regulatory mechanisms and could represent previously uncharacterized targets.

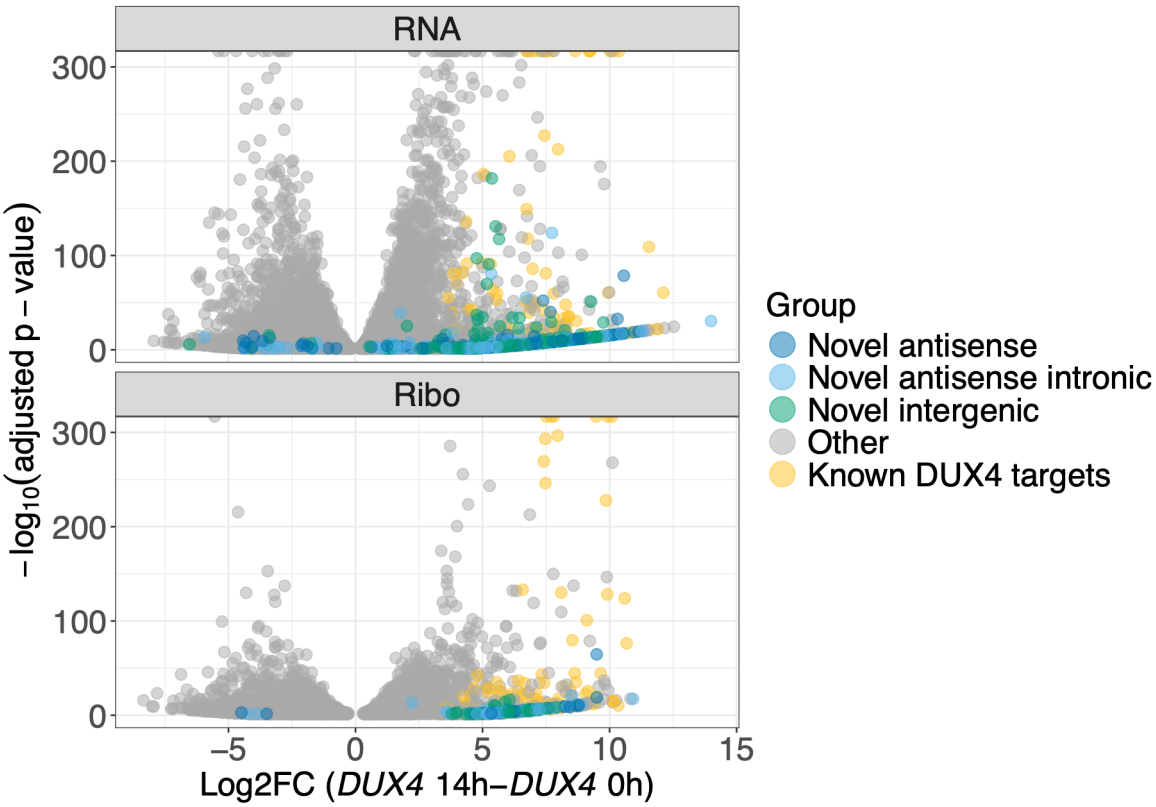


Figure 13 - RNA-seq and Ribo-seq volcano plots of the DUX4 14h-DUX4 0h comparison. Adjusted p-values are on the y axis, whereas log2 fold changes are on the x axis.

We further investigated the upregulated novel genes identified in the comparison by examining their expression dynamics over time (Figure 14). To provide appropriate benchmarks for the interpretation of the results, we included two additional sets of genes in the analysis: the log2FCs of known DUX4 targets, which served as positive control, and log2FCs of randomly selected genes, which served as a baseline.

The analysis revealed that the log2 fold changes of the upregulated novel genes followed a trend over time that was remarkably similar to that of the known targets, with log2 fold changes gradually increasing relative to the 0-hour baseline. This observation held true also for changes in Ribo-seq, pointing to the presence of a subset of novel translated genes upon DUX4 induction.

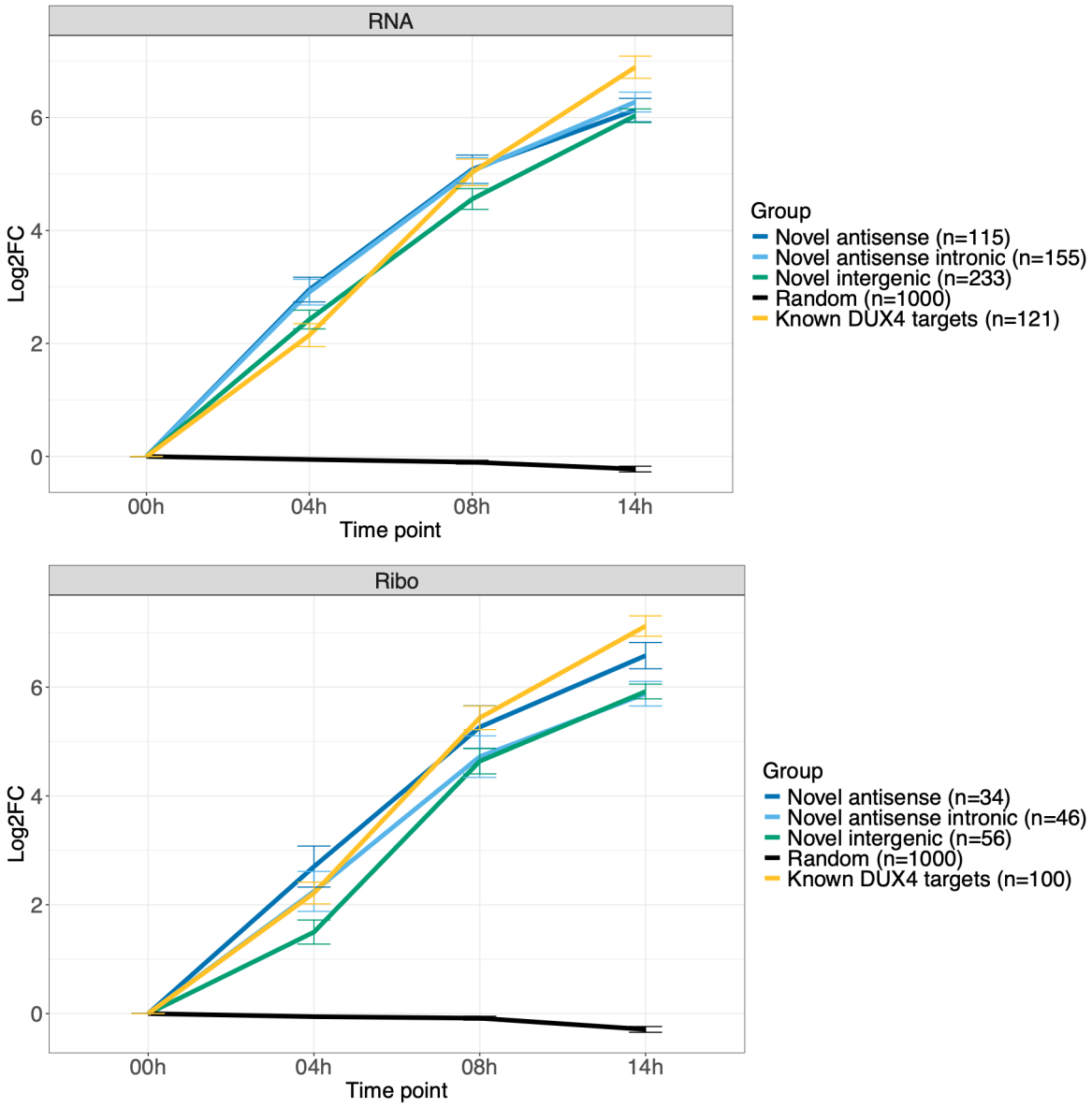


Figure 14 - Gene-level RNA-seq and Ribo-seq log2 fold changes relative to DUX4 0h.

Log2 fold changes are on the y axis, and time points on the x axis. Log2 fold changes of novel genes, known DUX4 target genes, and randomly selected genes are shown.

Overall, these analyses identified a set of novel genes exhibiting expression changes that are comparable to those of known DUX4 target genes. This suggests that they may constitute uncharacterized targets of the transcription factor.

2.2.2 Assessing DUX4 binding within transcript promoters of the upregulated novel genes

We were interested in investigating whether the previously identified set of upregulated novel genes represented direct targets of the transcription factor DUX4. For this reason, we used previously published chromatin immunoprecipitation sequencing (ChIP-seq) data from [43] (Materials and Methods – Data description and availability) to assess whether there was evidence of DUX4 binding within transcript promoters of the upregulated novel genes. We focused on the same gene groups as in the previously described analysis about RNA-seq log2 fold changes over time. For each gene group, we obtained the number of genes with ChIP-seq support and divided it by the total number of genes in the group. In this way, we obtained the proportions of ChIP-seq-supported genes. Then, we computed pairwise comparisons of proportions (Materials and Methods – Assessment of DUX4 binding with chromatin immunoprecipitation sequencing data). We found strong statistical evidence that the proportion of ChIP-seq-supported genes was significantly higher in known DUX4 target genes and in upregulated novel genes, especially in novel antisense intronic genes, compared to randomly selected genes (Table 3). Interestingly, we found no statistical evidence that the proportion of genes with ChIP-seq support was either higher or lower in upregulated novel genes than in known DUX4 target genes (Table 3). These results are consistent with the conclusion that the upregulated novel genes might represent new direct DUX4 targets.

Group1	Group2	Adjusted p-value (alternative = "greater")	Adjusted p-value (alternative = "less")
Known DUX4 target genes (n=121)	Random genes (1000)	2.47e-06	1
Novel intergenic genes (n=233)	Random genes (1000)	0.000358	1
Novel antisense genes (n=115)	Random genes (1000)	0.0187	1
Novel antisense intronic genes (n=155)	Random genes (1000)	1.85e-07	1
Novel intergenic genes (n=233)	Known DUX4 target genes (n=121)	0.913	0.483
Novel antisense genes (n=115)	Known DUX4 target genes (n=121)	0.913	0.483
Novel antisense intronic genes (n=155)	Known DUX4 target genes (n=121)	0.666	1

Table 3 - Testing DUX4 binding evidence in known and novel genes

An interesting example emerging from our analyses on gene expression levels is represented by the gene *DDX10*. In this locus, we identified two massively upregulated antisense transcripts that are produced from a novel gene and that overlap the first exon of the annotated gene (Figure 15). Moreover, a ChIP-seq peak overlapped the promoters of the two novel transcripts. Notably, these novel antisense transcripts would not be detected using standard analysis pipelines that rely exclusively on reference transcriptome annotations. This example highlights the added resolution provided by *de novo* transcriptome assembly.

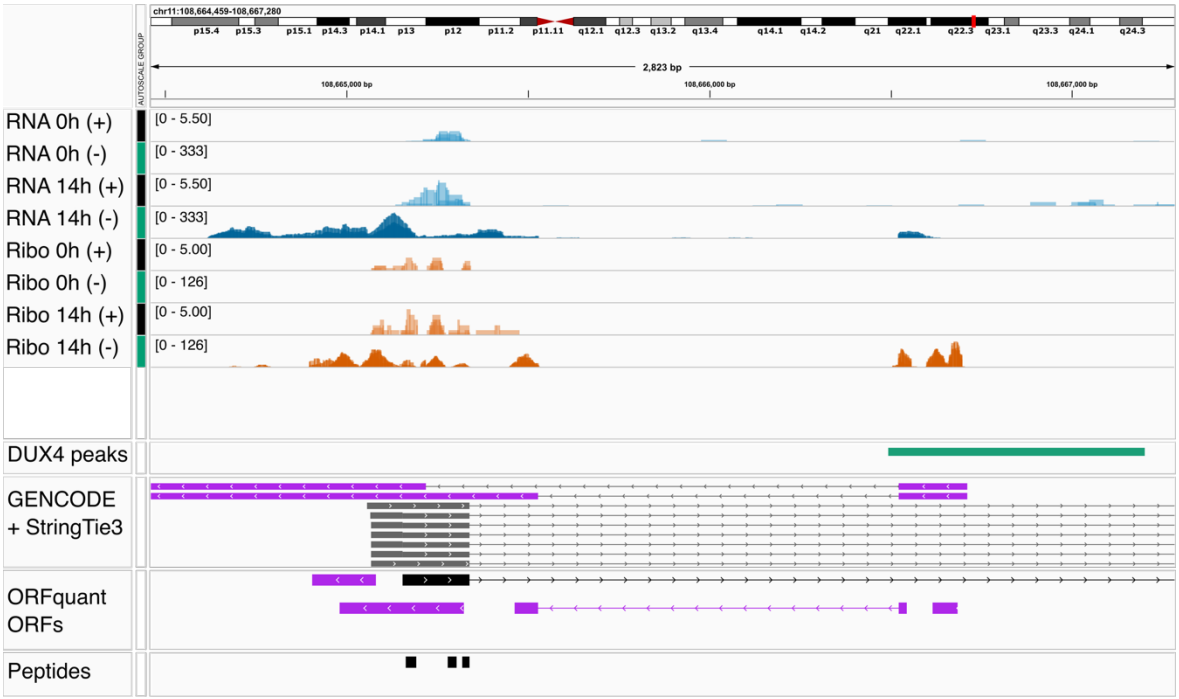


Figure 15 - IGV visualization of DDX10.

RNA-seq and Ribo-seq data of the three replicates at DUX4 0h and DUX4 14h (grouped by strand) are on top. RNA-seq data are shown in blue, whereas Ribo-seq data in orange. GENCODE v47 transcripts are in grey, whereas novel StringTie3 transcripts and unannotated ORFquant-detected ORFs are in violet. An annotated ORF and three peptides mapping to its corresponding protein sequence are in black. DUX4 ChIP-seq peaks from Geng et al., Developmental Cell, 2012 are shown in green.

2.2.3 *De novo* ORF finding with an augmented transcriptome

We used the translation module to analyse Ribo-seq data of the FSHD model, and detect translated ORFs at isoform-resolution, thus obtaining a detailed overview of the DUX4-induced translational landscape.

Different types of detected ORFs were obtained (Figure 16). Notably, we identified several unannotated ORFs, including 1218 ORFs from novel isoforms of annotated protein-coding genes, as well as 243 ORFs located on transcripts of novel genes.

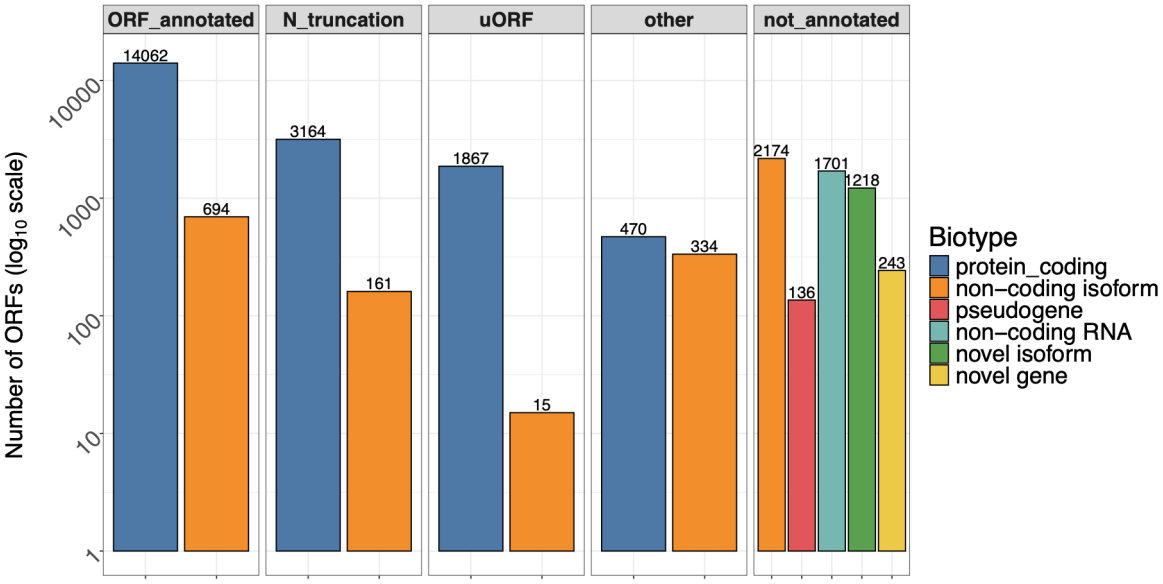


Figure 16 - Numbers of the ORFquant-detected translated ORFs.

Length distributions of the different types of ORFs were visualized using violin plots (Figure 17). This allowed the comparison of median lengths and overall distribution shapes across categories. ORFs on transcripts of novel genes had a length similar to uORFs or ORFs on non-coding RNAs, thus falling overall in the category of small ORFs.

Moreover, the length distribution of ORFs from novel isoforms of annotated protein-coding genes appeared slightly lower than canonical ORFs, likely representing translation on novel NMD isoforms.

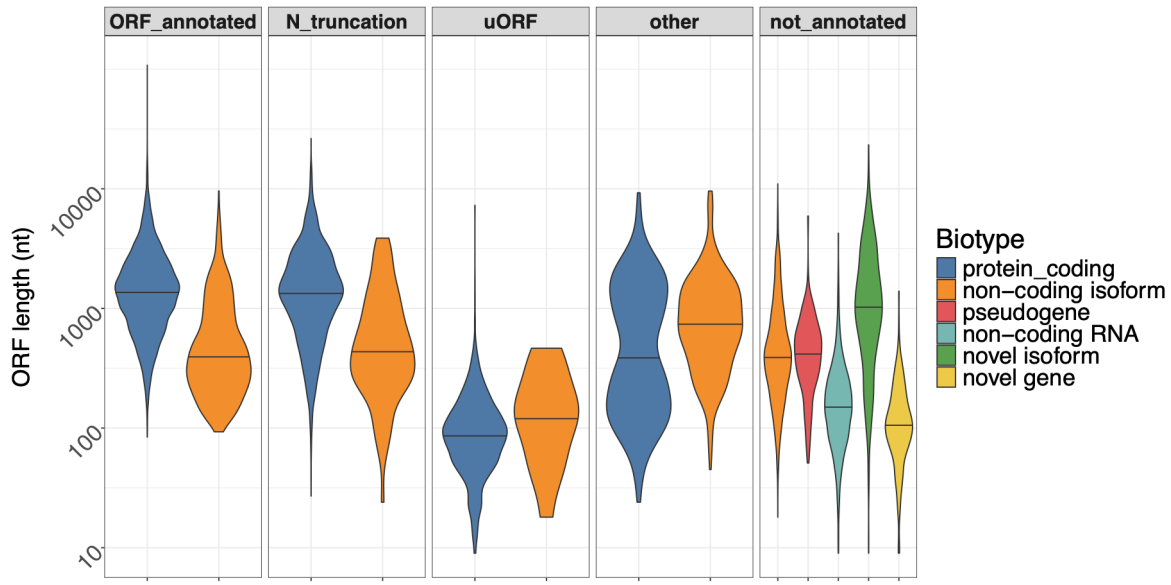


Figure 17 - Length distributions of the ORFquant-detected translated ORFs.

Distributions of ORFs per million values (ORFs_pM), which represent length-normalized quantification estimates of Ribo-seq levels, were also inspected (Figure 18). Distributions of the two aforementioned types of unannotated ORFs appeared similar, with comparable medians, while exhibiting higher levels when compared to ORFs on non-coding RNAs.

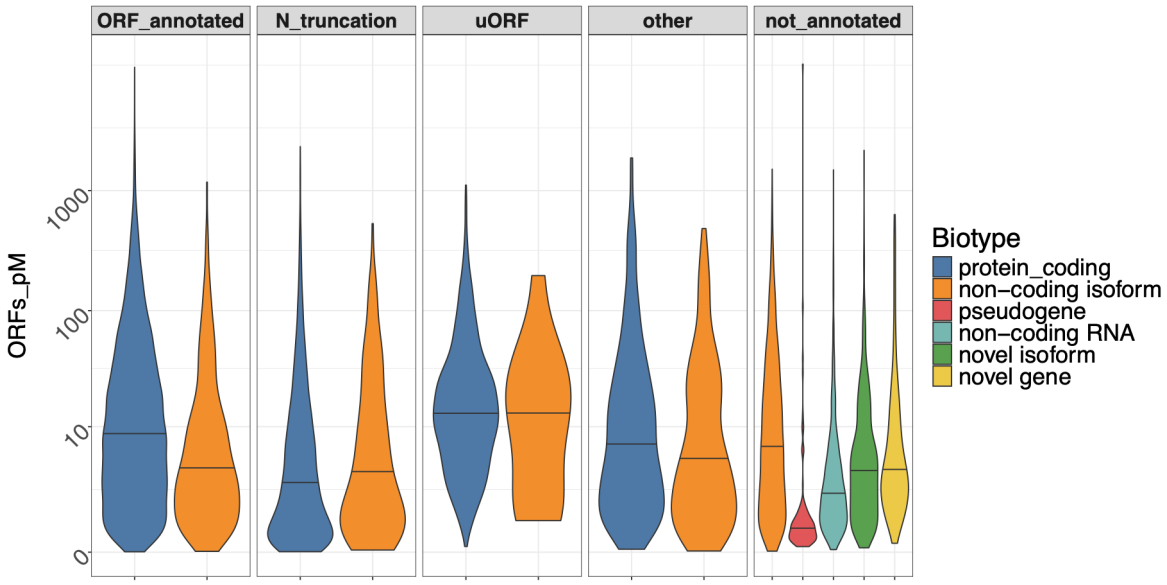


Figure 18 - Distributions of quantification estimates of the ORFquant-detected translated ORFs.

Altogether, isoform-level *de novo* ORF finding with *ORFquant* provided a comprehensive overview of the translatome, allowing us to focus on different statistics of annotated and novel ORFs. After obtaining this detailed characterization of the translated transcriptome, we were interested in analysing proteome composition and proteomic changes following *DUX4* activation (Results – Section 2.3.1).

2.3 Detecting *de novo* identified targets in the cellular proteome and in FSHD patients

This section concerns the analysis of TMT proteomic data of the FSHD model. Peptide-level identification and quantification results were obtained by running proteomic searches of the Protein module of R2T2P (Results – Section 1.4). Our analyses led to the discovery of novel proteins with upregulated peptides and with evidence for RNA expression in patient-derived myotubes and patient biopsies.

2.3.1 Novel peptides are upregulated following *DUX4* activation

We focused on characterizing the downstream consequences of *DUX4* activation at the proteome level. To this end, we examined the R2T2P results of the proteomic search that was performed using the protein sequence database including both annotated proteins (corresponding to CDSs of GENCODE v47 annotation) and *ORFquant* proteins (corresponding to *ORFquant*-detected ORFs).

This combined database allowed us to detect not only annotated but also unannotated translation products, albeit with an increased FDR cutoff resulting from a large database (Discussion). Peptides were classified as novel if they mapped only to *ORFquant* proteins.

Numbers of the detected annotated and novel peptides, excluding peptides mapping to multiple genes, are shown in Table 4 (Materials and Methods – Peptide-level analyses).

Peptide type	Number of peptides (genes)
Annotated	124941 (9432)
Novel	217 (144)

Table 4 - Numbers of detected annotated and novel peptides.

Peptides were detected by the proteomic search with the combined database (GENCODE v47 and ORFquant proteins), and peptides mapping to multiple genes were excluded. Numbers of corresponding genes are in parentheses.

We then analysed peptide-level quantification results of the proteomic search. For the quantification, the reference channel was used by running *TMTIntegrator* with reference approach (Materials and Methods - Proteomic searches). Before performing peptide-level differential expression analyses, peptide intensities were log_2-transformed.

Median-centering normalization was not applied to the log_2-transformed values, as it was deemed unnecessary based on careful inspection of the distributions (Figure 19).

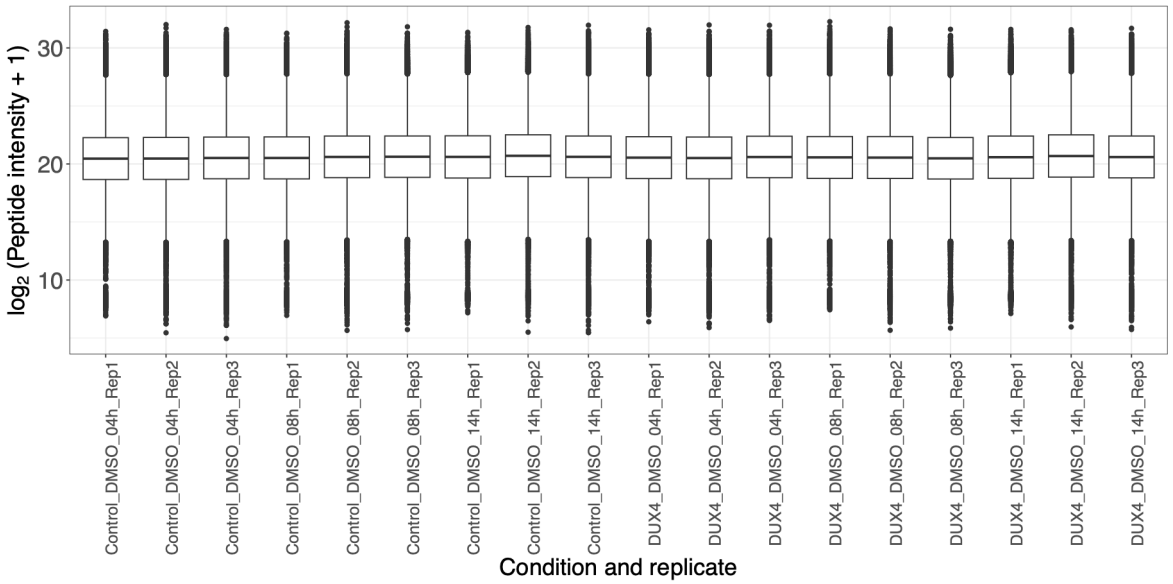


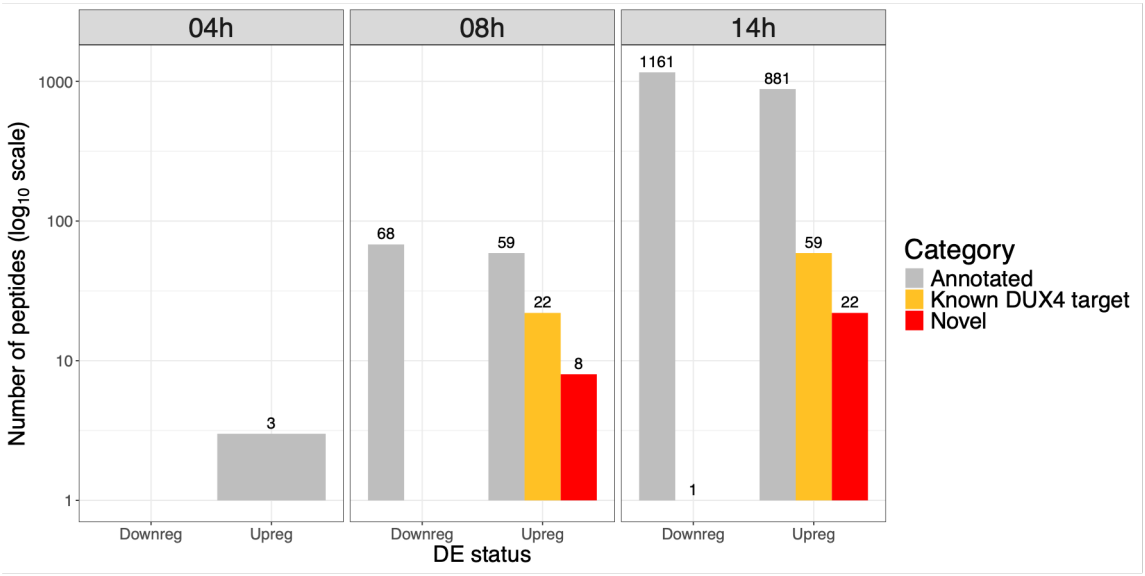
Figure 19 - Distributions of log_2-transformed peptide intensities.

A pseudocount of 1 was added to intensities before log_2-transforming.

Using the R package *limma* [98], we then performed differential expression analyses by comparing peptide intensities of each time point with those of the corresponding control condition (Materials and Methods – Peptide-level analyses).

As expected, we noticed an overall increase in the number of differentially expressed peptides over time, including a rise in the number of upregulated novel peptides (Figure 20a). We also obtained the numbers of corresponding genes with downregulated and upregulated peptides (Figure 20b). The number of genes with differentially expressed novel peptides was similar albeit lower to that of known DUX4 targets with differentially expressed peptides.

a



b

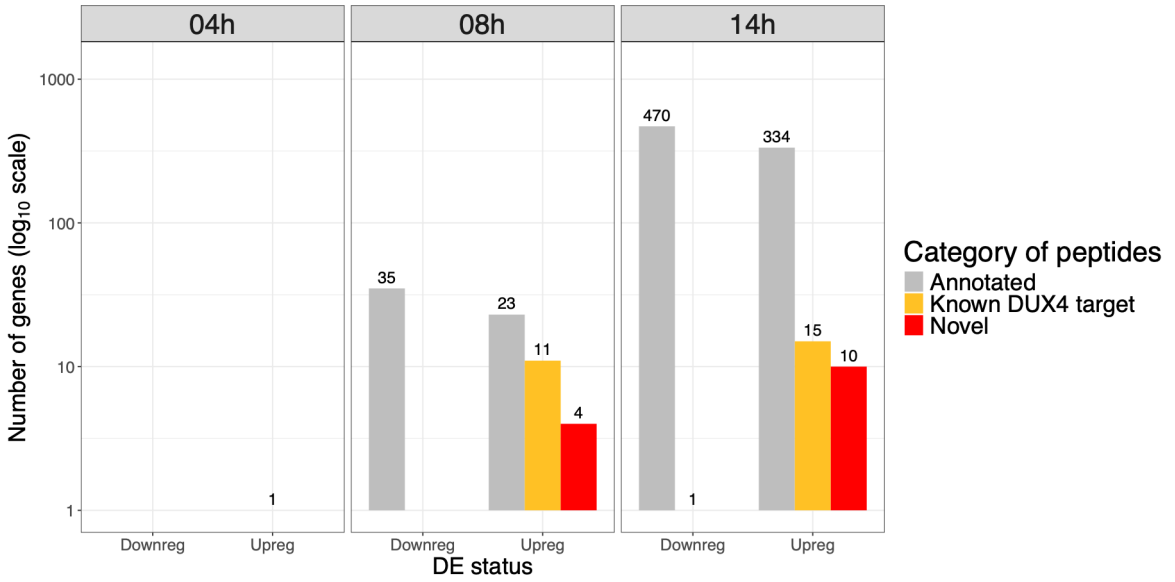


Figure 20 - Peptide-level differential expression analyses.

a. Numbers of downregulated and upregulated peptides of different timepoint-specific comparisons; b. Numbers of genes with downregulated and upregulated peptides of different timepoint-specific comparisons

We then focused on the comparison *DUX4* 14h-DMSO 14h by visualizing log2 fold changes and average expression levels of differentially expressed peptides (Figure 21). Interestingly, we found novel peptides with log2 fold changes that were comparable to those of peptides deriving from annotated transcripts of known DUX4 target genes.

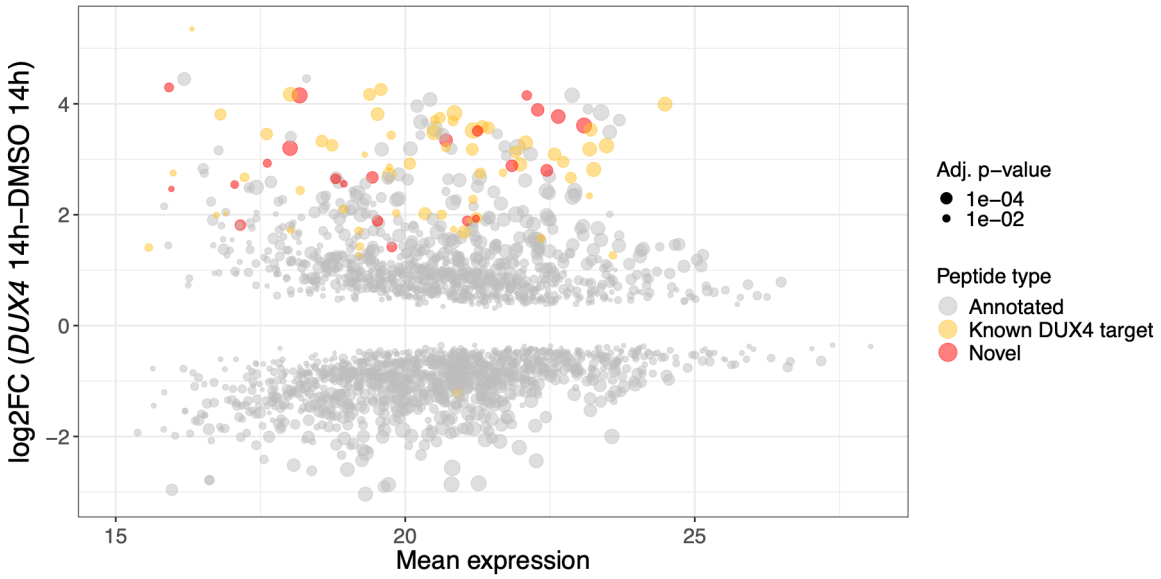


Figure 21 - MA plot of the differentially expressed peptides of the comparison DUX4 14h-DMSO 14h.

Mean expression is on the x axis, while log2 fold changes are on the y axis. Only significant peptides are shown.

2.3.2 Novel proteins with upregulated peptides

Our proteomic analyses led to the identification of previously uncharacterized novel proteins with peptides that were significantly upregulated at 14h after *DUX4* activation (Figure 22). We focused our attention on two novel proteins harbouring different upregulated peptides.

One of the two proteins was produced from a *StringTie3*-assembled transcript and had 3 upregulated peptides (Figure 22a). The novel transcript is antisense to the intron of an annotated gene, and it is in proximity to another annotated gene on the same strand.

The other was compatible with a transcript of a transcribed unprocessed pseudogene (with GENCODE transcript ID ENST00000550420.2) and had 3 upregulated peptides (Figure 22b).

Interestingly, DUX4 ChIP-seq peaks from [43] overlapped the transcripts producing the two novel proteins.

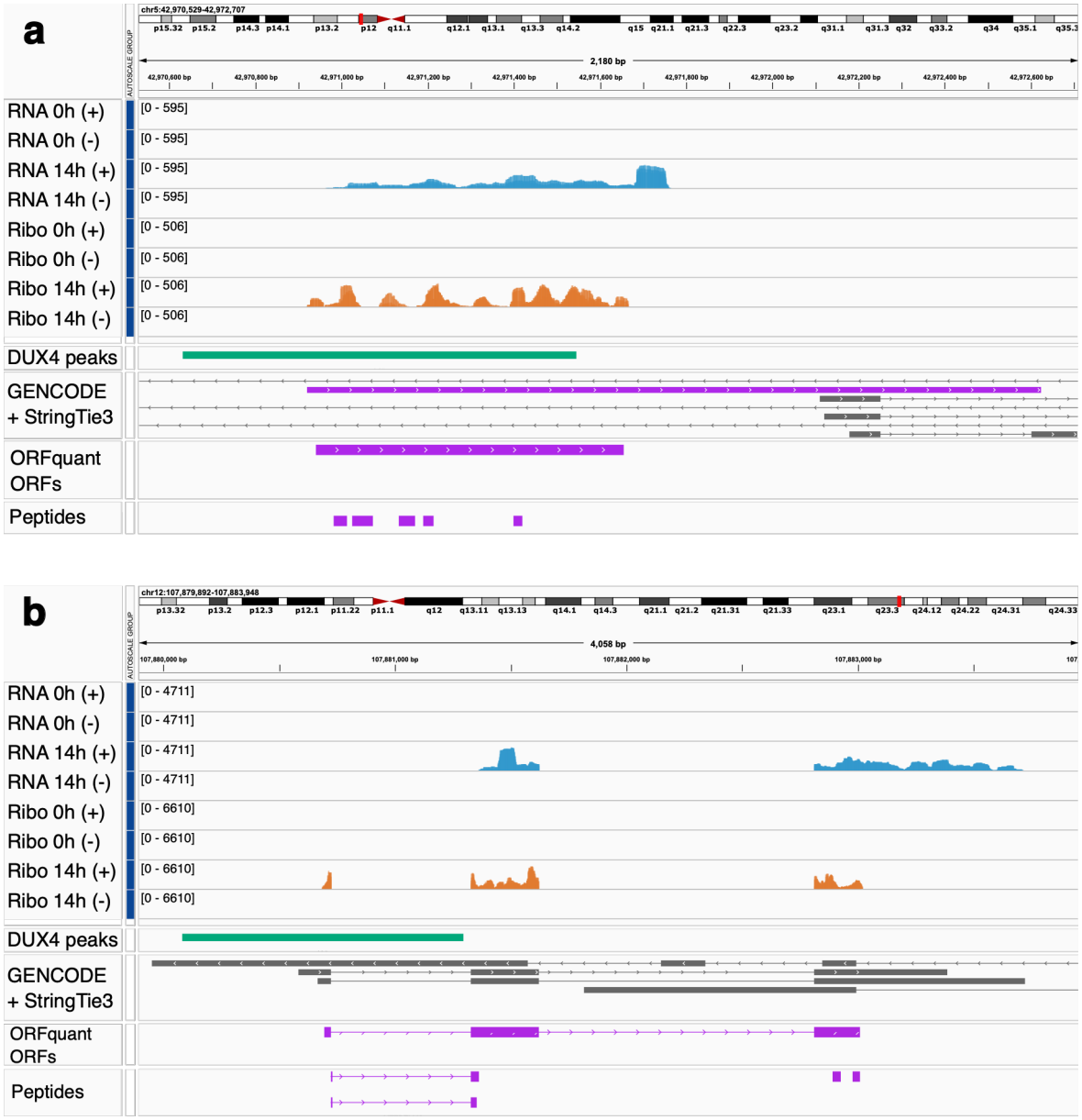


Figure 22 - IGV screenshots of the genomic loci of two novel proteins with upregulated peptides

a. novel protein which is produced from a novel transcript; b. novel protein which is produced from a transcript which is annotated as non-coding.

RNA-seq and Ribo-seq data of the 3 replicates at DUX4 0h and DUX4 14h are shown on top. Tracks from individual replicates are overlayed. RNA-seq data are shown in blue, whereas Ribo-seq data in orange. GENCODE v47 transcripts are represented in grey, while novel StringTie3-assembled transcripts, novel ORFquant-detected ORFs, and novel peptides in violet. DUX4 ChIP-seq peaks from Geng et al., Developmental Cell, 2012 are shown in green.

2.3.3 Novel genes and proteins with evidence for RNA expression in patient-derived myotubes and patient biopsies

We were interested in assessing whether there was evidence for RNA expression of these two loci in data from patient-derived myotubes and patient biopsies. For this reason, we analysed RNA-seq data of FSHD and control myotubes [42] as well as RNA-seq data of biopsies from FSHD patients with increasing levels of disease severity, i.e. from G1 to G4 [100]. Both RNA-seq datasets were not strand-specific. Details on mapping statistics from the two datasets are shown in Table 5.

	Raw reads	Mapped reads	Uniquely mapped reads	Dataset	Group
C20_tubes	118,164,630	111,574,826	100,040,370	Yao et al.	Control myotubes
C21_tubes	108,132,662	102,457,630	92,118,320	Yao et al.	Control myotubes
C22_tubes	131,783,191	124,920,743	110,766,686	Yao et al.	Control myotubes
F4_tubes	232,766,074	220,798,606	194,536,711	Yao et al.	FSHD myotubes
F6_tubes	117,966,914	112,424,028	99,808,594	Yao et al.	FSHD myotubes
F12_tubes	121,137,696	115,070,326	102,892,266	Yao et al.	FSHD myotubes
F14_tubes	197,576,453	190,490,264	170,977,268	Yao et al.	FSHD myotubes
F20_tubes	119,678,048	113,987,940	101,764,736	Yao et al.	FSHD myotubes
01_0041	81,630,949	78,995,849	74,750,507	Wang et al.	G1 controls
01_0042	90,898,443	88,345,805	83,475,109	Wang et al.	G1 controls
01_0043	95,658,547	91,978,718	86,552,395	Wang et al.	G1 controls
01_0044	92,757,859	90,160,513	86,049,235	Wang et al.	G1 controls
01_0045	89,425,838	86,389,483	81,311,802	Wang et al.	G1 controls
01_0046	33,231,218	31,750,325	29,925,733	Wang et al.	G1 controls
01_0047	32,026,478	30,220,359	28,375,616	Wang et al.	G1 controls
01_0048	88,613,889	86,145,692	81,938,667	Wang et al.	G1 controls
01_0049	33,032,370	31,865,425	29,916,863	Wang et al.	G1 controls
01_0023	83,470,060	80,953,614	75,937,967	Wang et al.	G1 patients
01_0032	32,832,010	31,475,225	30,022,809	Wang et al.	G1 patients
01_0036	49,373,856	47,049,796	44,450,368	Wang et al.	G1 patients
01_0037	40,234,986	30,594,157	28,653,188	Wang et al.	G1 patients
01_0038	36,001,195	34,624,274	32,873,005	Wang et al.	G1 patients
32_0011	82,146,364	78,856,782	74,543,498	Wang et al.	G1 patients
32_0012	87,145,916	84,545,250	79,945,367	Wang et al.	G1 patients
32_0014	80,109,321	77,101,024	70,960,085	Wang et al.	G1 patients
32_0015	35,577,118	34,173,069	32,453,613	Wang et al.	G1 patients
32_0016	29,293,124	23,994,103	22,516,799	Wang et al.	G1 patients
01_0024	71,129,829	67,492,232	63,490,507	Wang et al.	G2 patients
01_0030	90,848,898	87,866,750	83,353,398	Wang et al.	G2 patients
01_0033	93,599,426	90,852,638	86,383,215	Wang et al.	G2 patients
32_0007	44,754,074	42,819,734	40,652,357	Wang et al.	G2 patients
32_0013	86,484,161	84,047,549	78,328,735	Wang et al.	G2 patients
01_0021	37,204,008	34,661,581	32,953,958	Wang et al.	G3 patients
01_0022_1	86,320,159	84,041,439	79,959,617	Wang et al.	G3 patients
01_0026	78,497,241	76,048,777	72,026,578	Wang et al.	G3 patients
01_0027	86,431,862	83,861,137	79,198,789	Wang et al.	G3 patients
01_0034	83,877,544	80,842,545	76,697,327	Wang et al.	G3 patients
01_0035	82,710,632	80,039,273	75,612,335	Wang et al.	G3 patients
32_0002	56,370,447	53,347,189	50,361,681	Wang et al.	G3 patients
32_0008	31,400,924	19,395,839	17,861,196	Wang et al.	G3 patients
32_0009	42,323,996	36,383,044	33,084,720	Wang et al.	G3 patients
32_0010	89,025,146	86,130,797	81,528,505	Wang et al.	G3 patients
32_0017	29,199,857	28,347,792	26,851,055	Wang et al.	G3 patients
32_0018	35,801,601	33,826,154	31,919,130	Wang et al.	G3 patients
32_0019	34,321,744	32,868,859	31,463,401	Wang et al.	G3 patients
01_0025	84,202,038	81,260,010	77,372,737	Wang et al.	G4 patients
01_0029	77,755,117	74,589,128	70,750,946	Wang et al.	G4 patients
32_0003	52,850,126	50,490,495	48,097,095	Wang et al.	G4 patients
32_0004	60,687,594	57,906,259	55,080,179	Wang et al.	G4 patients
32_0005	52,241,342	48,113,956	44,694,156	Wang et al.	G4 patients
32_0006	52,355,740	50,106,498	47,057,591	Wang et al.	G4 patients

Table 5 - Mapping statistics of published RNA-seq datasets for FSHD samples

Datasets and groups are in the last two columns.

We used the transcriptome annotation that was generated by our FSHD model data analysis, and we performed gene-level differential expression analyses with the two RNA-seq datasets: RNA-seq data from patient-derived myotubes and FSHD patient biopsies were compared to data from control myotubes and G1 controls, respectively. Interestingly, several novel genes were upregulated in both datasets' comparisons (Figure 23).

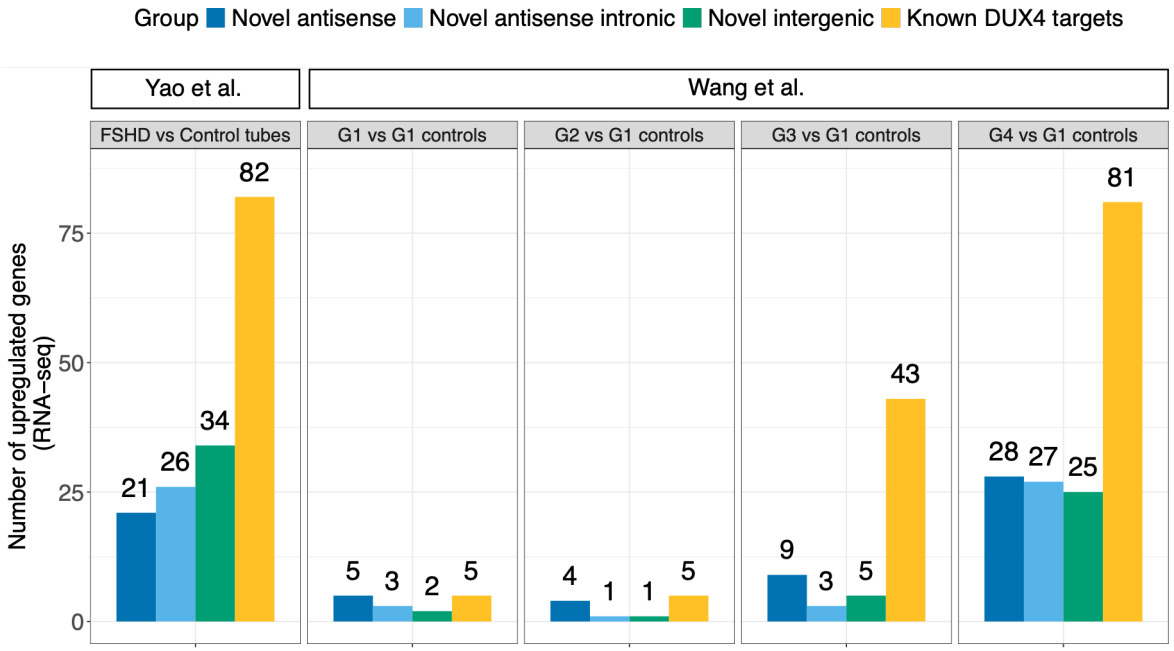


Figure 23 - Gene-level RNA-seq differential expression analyses (numbers of upregulated genes).

Numbers of RNA-seq upregulated genes in different comparisons: FSHD vs control myotubes (from Yao et al., Human Molecular Genetics, 2014); FSHD patient biopsies vs G1 controls (from Wang et al., Human Molecular Genetics, 2019).

We also investigated if there was evidence for RNA expression in the loci of the two novel proteins described in the previous section.

Interestingly, we found evidence for expression specifically in the patient-derived myotubes and in patient biopsies, whereas such expression was absent or considerably lower in control myotubes and in G1 patients (Figure 24).

While these findings are interesting, future studies will be required to elucidate the functional roles of these two novel proteins in FSHD pathogenesis and progression (Discussion).

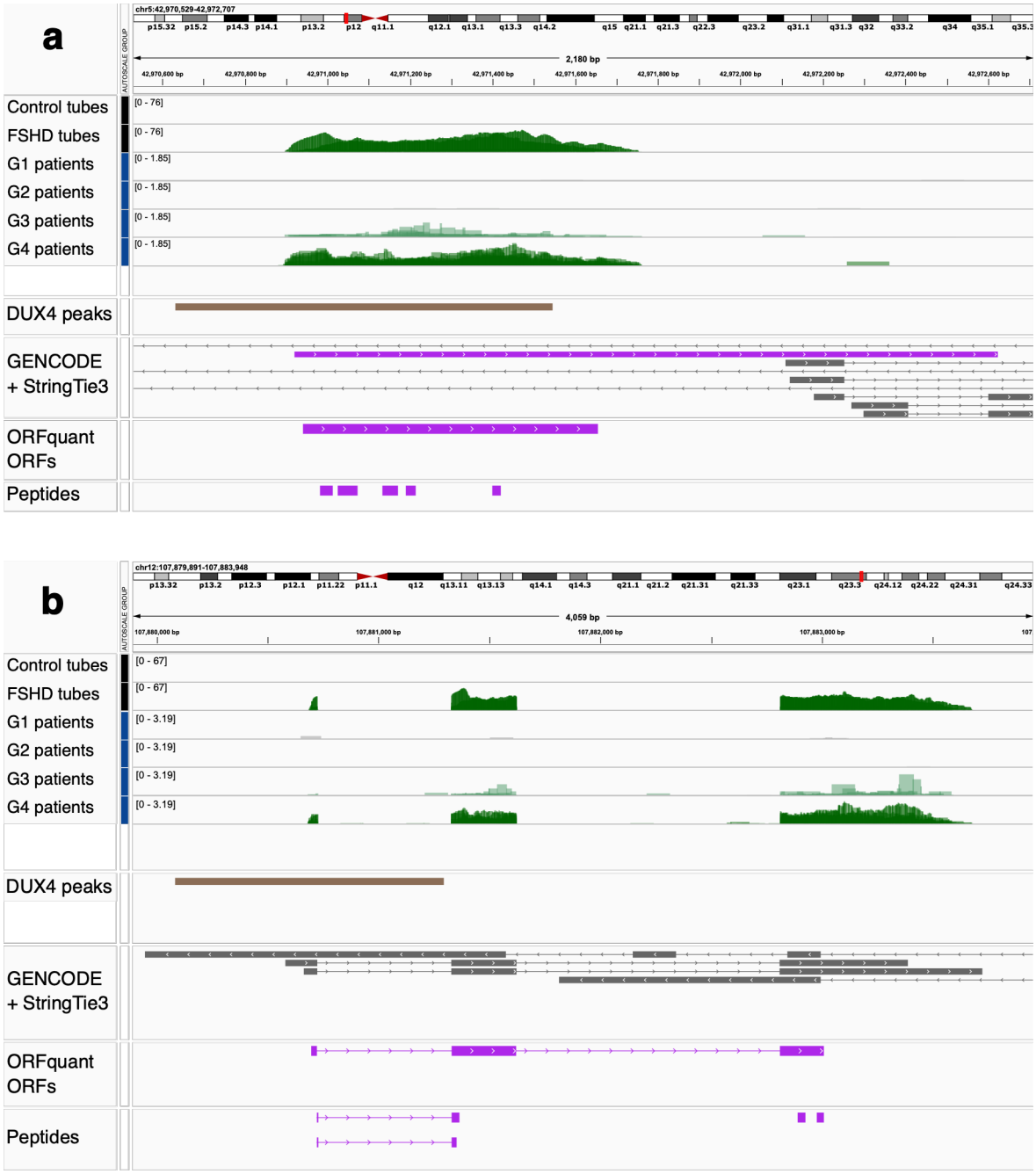


Figure 24 - IGV screenshots of the genomic loci of the two novel proteins (with RNA-seq data of myotubes and patient biopsies)

a. novel protein which is produced from a novel transcript; b. novel protein which is produced from a transcript which is annotated as non-coding.

RNA-seq data of control myotubes and patient-derived myotubes (Yao et al., Human Molecular Genetics, 2014, black side bar), as well as RNA-seq data of biopsies from FSHD patients (Wang et al., Human Molecular Genetics, 2019, dark blue side bar), are shown on top. Tracks from individual replicates are overlayed. GENCODE v47 transcripts are represented in grey, whereas novel StringTie3-assembled transcripts, novel ORFquant-detected ORFs, and novel peptides are in violet. DUX4 ChIP-seq peaks from Geng et al., Developmental Cell, 2012 are shown in brown.

Moreover, we visualized RNA-seq data in the genomic region of the first exon of gene *DDX10* (Results – Section 2.2.2). Interestingly, there was an increase in RNA-

seq levels in patient-derived myotubes and in G4 patients (Figure 25). As previously mentioned, both RNA-seq datasets were not strand-specific, but the increase in RNA-seq levels was particularly evident in genomic regions from which the antisense transcripts derived, suggesting their upregulation in patient-derived myotubes and patient biopsies.

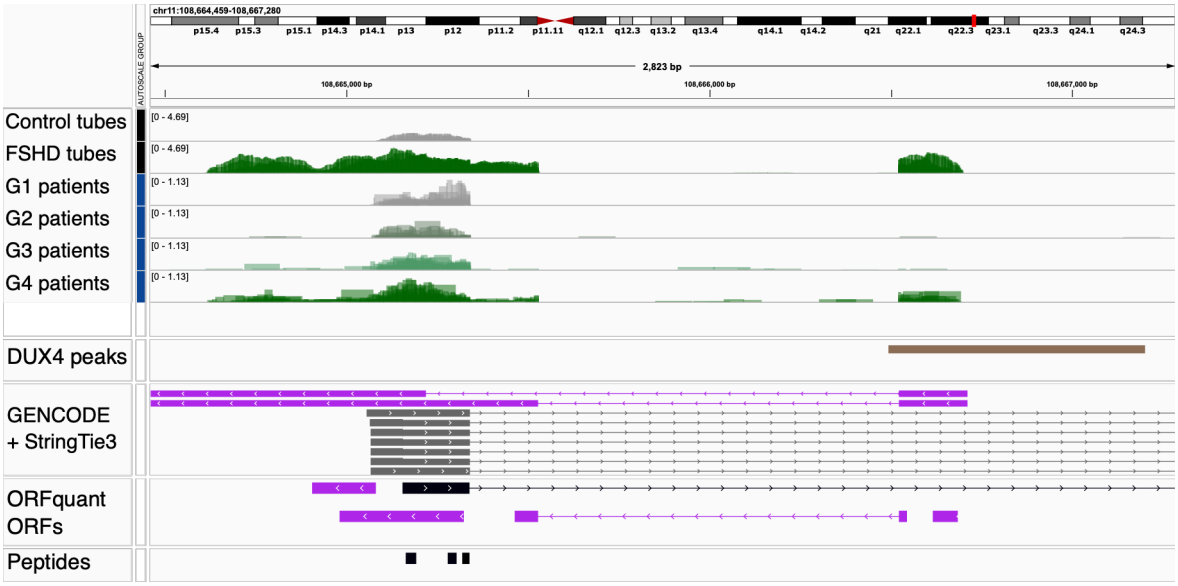


Figure 25 - IGV screenshot of DDX10 (with RNA-seq data of myotubes and patient biopsies).

RNA-seq data of control myotubes and patient-derived myotubes (Yao et al., Human Molecular Genetics, 2014, black side bar), as well as RNA-seq data of biopsies from FSHD patients (Wang et al., Human Molecular Genetics, 2019, dark blue side bar), are shown on top. GENCODE v47 transcripts are in grey, whereas novel StringTie3-assembled transcripts and novel ORFquant-detected ORFs are in violet. An annotated ORF and three peptides mapping to its corresponding protein sequence are in black. DUX4 ChIP-seq peaks from Geng et al., Developmental Cell, 2012 are shown in brown.

Altogether, by using R2T2P we discovered novel proteins which have evidence for RNA expression in FSHD patients. The significance of these findings is addressed and further explored in the Discussion (Discussion – Functional relevance of novel species).

2.4 A further expanded transcriptome annotation modestly improves identification and quantification

By using the R2T2P pipeline, we analysed data of the FSHD model by adding a long-read-derived transcriptome annotation [41] (Table 6) in the final merging step of the RNA analysis module. We were interested in including the additional, independent long-read-derived annotation to further validate transcripts of our short-read-derived annotation. However, more importantly, we also aimed to assess the extent to which its inclusion extended the sets of detected events from RNA to Protein.

Total number of genes	12,592
-- Protein-coding	9,717
-- Pseudogene	217
-- lncRNA	1,766
-- ncRNA	107
-- Novel	785
Total number of transcripts	84,687
-- Novel transcripts from annotated genes	50,089
-- Novel transcripts from novel genes	1,570

Table 6 - Statistics on genes and transcripts of the long-read-derived annotation

We obtained an annotation including novel genes and transcripts derived from short-read (SR) and long-read (LR) sequencing data. Moreover, unified gene models (hereafter referred to as “SR+LR”) were defined in cases in which long-read-derived transcripts were assigned to *StringTie3* genes when merging the two annotations (Materials and Methods – Combining GTF files).

Numbers of the identified novel transcripts, grouped according to the gene categories described above, are summarized in Table 7.

Transcript type	SR gene	LR gene	SR+LR gene
Novel antisense	288 (272)	205 (167)	489 (136)
Novel antisense intronic	459 (452)	210 (164)	346 (96)
Novel intergenic	478 (463)	268 (194)	431 (127)

Table 7 - Numbers of novel transcripts in the transcriptome annotation.

Novel transcripts are divided based on their biotypes and on their gene categories, i.e., short-read-derived (SR), long-read-derived (LR) or unified gene models (SR+LR). Numbers of genes are in parentheses.

By using the transcriptome annotation, we performed RNA-seq and Ribo-seq differential expression analyses to identify genes exhibiting significant changes at 14h after *DUX4* activation (Figure 26), akin to our previous analysis in Figure 12.

Upregulated novel genes were categorized according to their biotype, allowing for a more detailed understanding of the functional classes represented among the genes showing increased expression. In addition to this classification, the genes were further subdivided based on the transcriptome annotation files in which they were present, i.e, in the short-read-derived annotation only, in the long-read-derived annotation only, or in both annotations. This enabled a comparison of the contributions of each sequencing technology.

Importantly, while a subset of the upregulated genes was consistently detected by both short-read and long-read sequencing approaches, a substantial number of upregulated genes were exclusively identified through the short-read-based assembly.

This discrepancy could be related to the higher throughput and greater sequencing depth which is achievable with short-read technologies (Introduction – Section 1.2.1). However, it is also possible that some genes which were detected only with short-read technologies constitute assembly artefacts or pre-mRNA intermediates (Discussion).

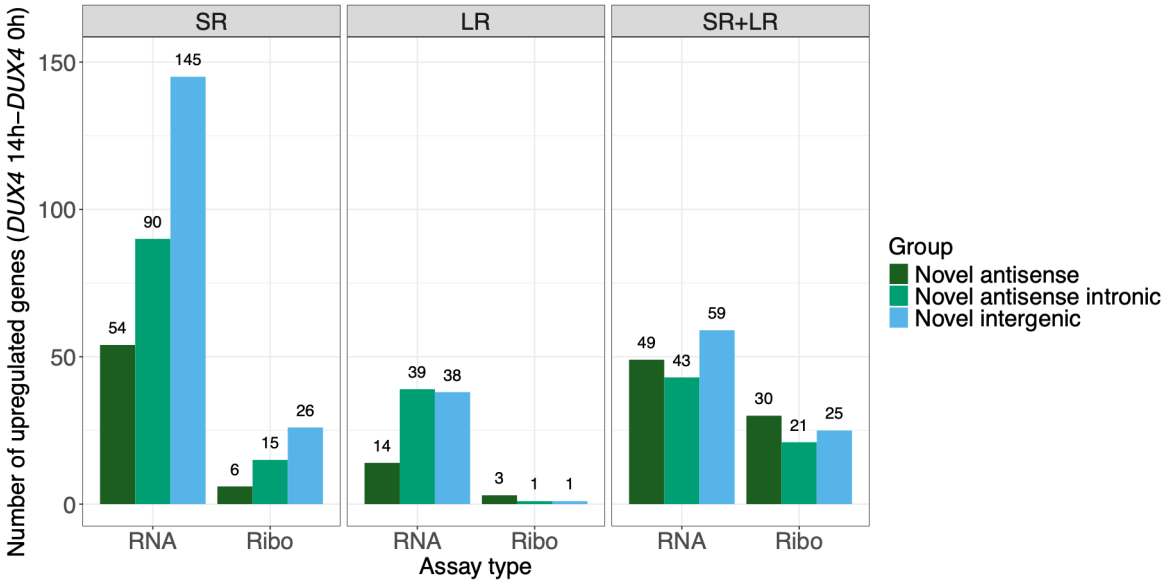


Figure 26 - Gene-level RNA-seq and Ribo-seq differential expression analyses (numbers of upregulated genes).

Numbers of upregulated novel genes which were detected when comparing expression levels of DUX4 14h and DUX4 0h. Novel genes are divided based on the transcriptome annotations including them (SR: short-read-derived annotation; LR: long-read-derived annotation; SR+LR: short-read-derived and long-read-derived annotations). Novel genes are also divided based on their biotypes.

To assess the relationship between statistical significance and the extent of expression changes, we visualized the adjusted p-values and the log2 fold changes of differentially expressed genes (Figure 27). This revealed that novel genes identified by both sequencing technologies tended to exhibit higher log2 fold changes compared to those detected by only one method.

Moreover, this highlighted the presence of some novel intergenic genes which were detected only with short-read sequencing data, and which had particularly low adjusted p-values in the RNA-seq differential expression analysis; however, most of these RNAs were not upregulated in the Ribo-seq data.

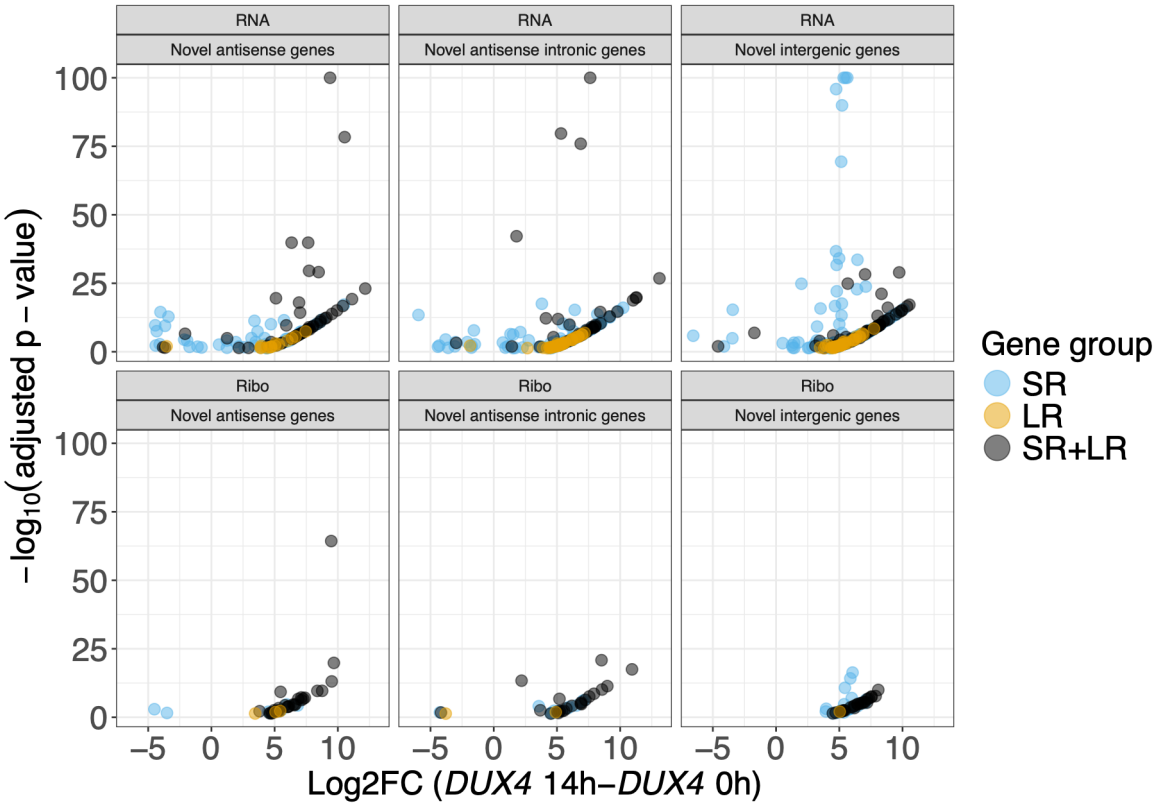


Figure 27 - Adjusted p-values and log2 fold changes of differentially expressed novel genes in the comparison DUX4 14h-DUX4 0h.

Novel genes are divided based on the transcriptome annotations containing them (SR: short-read-derived annotation; LR: long-read-derived annotation; SR+LR: short-read-derived and long-read-derived annotations). Novel genes are also divided based on their biotypes.

We then investigated the consequences of *DUX4* activation on the proteome. We analysed results of the proteomic search using the combined database with GENCODE v47 and ORFquant protein sequences. As in Results – Section 2.3.1, peptides were classified as novel if they mapped only to ORFquant proteins.

Numbers of detected annotated and novel peptides, excluding peptides mapping to multiple genes, are in Table 8 (Materials and Methods – Peptide-level analyses).

Peptide type	Number of peptides (genes)
Annotated	124294 (9449)
Novel (Annotated tx)	141 (95)
Novel (SR tx)	33 (26)
Novel (LR tx)	44 (33)
Novel (SR and LR tx)	18 (6)

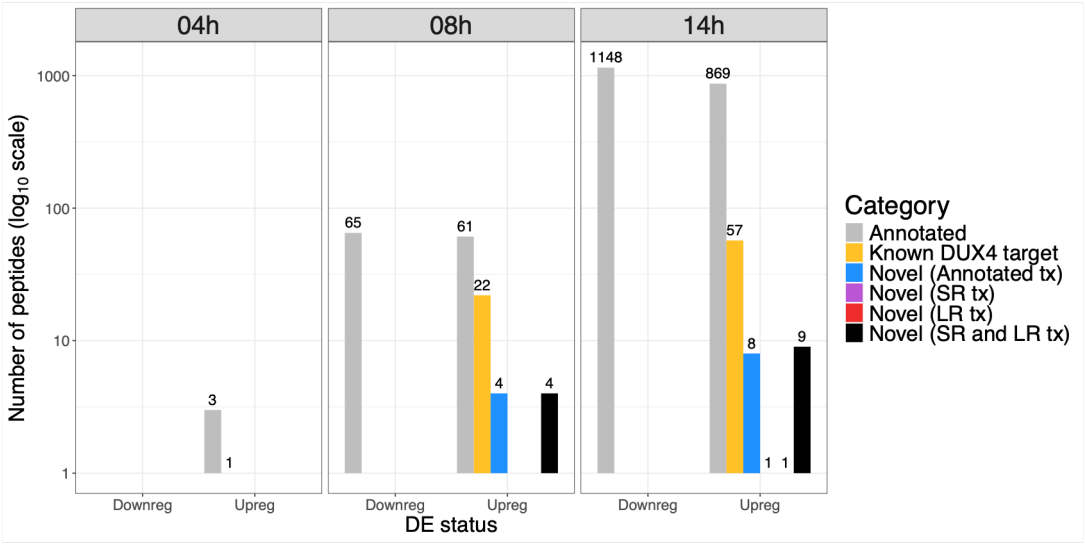
Table 8 - Numbers of detected annotated and novel peptides

Peptides were detected by the proteomic search with the combined database (GENCODE v47 and ORFquant proteins), and peptides mapping to multiple genes were excluded. Novel peptides are classified based on the types of transcripts producing them. Number of genes are in parentheses.

While different peptides were uniquely detected in different annotations, we performed differential expression analyses, following the same procedure shown in Results - Section 2.3.1 (Materials and Methods – Peptide-level analyses).

We detected novel peptides which were significantly upregulated in the last two time-point comparisons, and which were produced from different types of transcripts (Figure 28a). Numbers of genes with downregulated and upregulated peptides were also obtained (Figure 28b). Most of the upregulated novel peptides were detected by both annotations, leaving only one identification unique to one of the two annotations (Discussion).

a



b

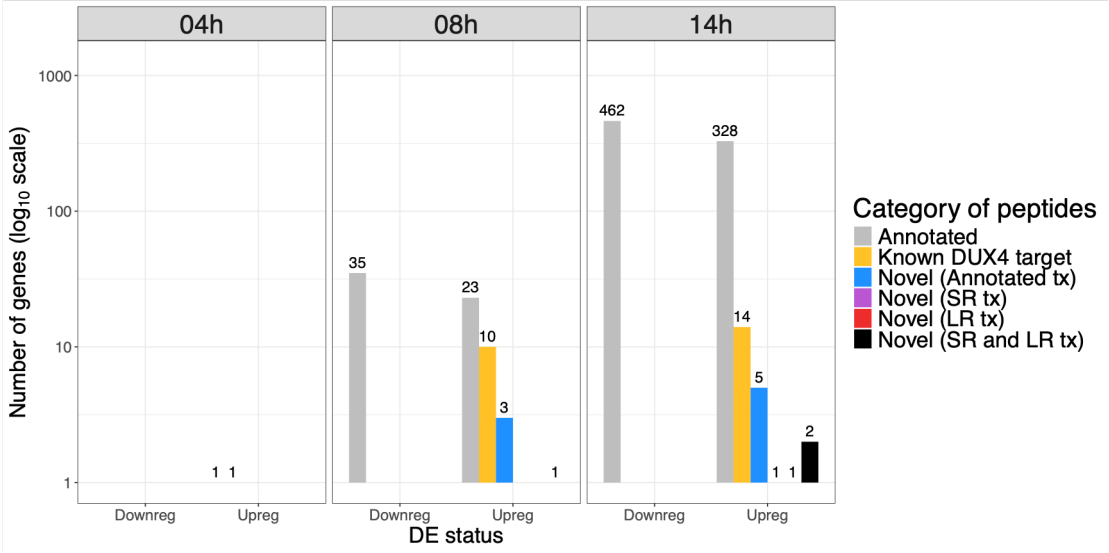


Figure 28 - Peptide-level differential expression analyses using long and short read-derived annotations.

a. Numbers of downregulated and upregulated peptides of different timepoint-specific comparisons; b. Numbers of genes with downregulated and upregulated peptides of different timepoint-specific comparisons

We then focused on the comparison *DUX4* 14h-DMSO 14h, visualizing log2 fold changes and mean expression of the differentially expressed peptides (Figure 29).

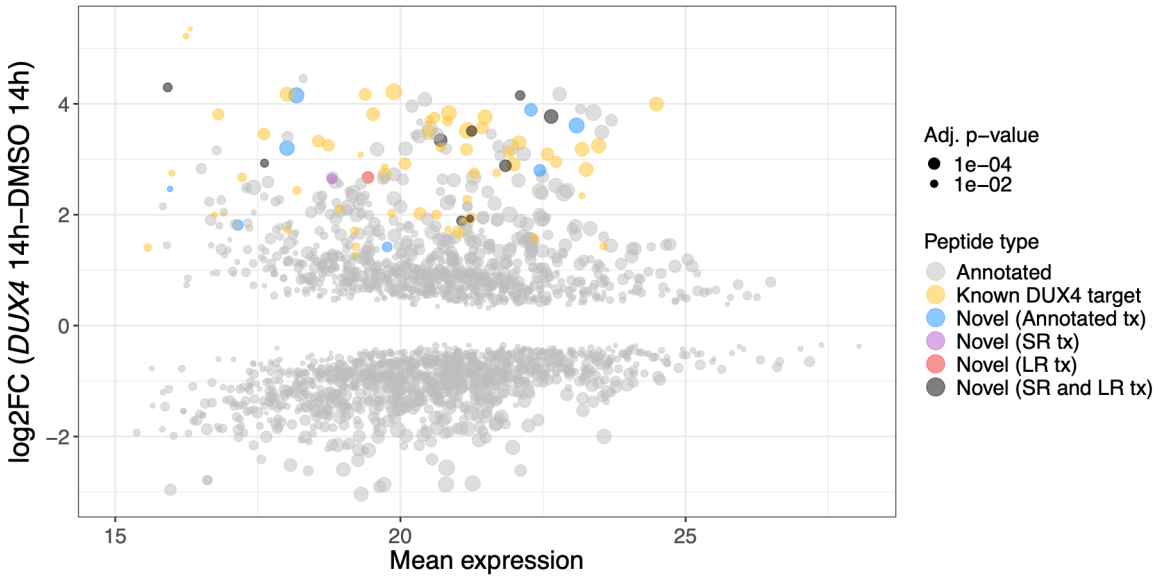


Figure 29 - MA plot of the differentially expressed peptides of the comparison DUX4 14h – DMSO 14h.

Peptides are classified based on the types of transcripts producing them.

In summary, including long-read-derived transcripts led to a modest increase in the number of detected changes, not only at the transcriptome but also at the translation and proteome level.

3. Extended benchmarking strategies for *de novo* detection

All the benchmarking methods which are included in this section are part of an ancillary R package to our *de novo* pipeline R2T2P, the package *R2T2P.Benchmark*. It comprises a workflow for the assessment of *de novo* transcriptome assembly and ORF finding results as well as methods for the comparison between proteomic search results and between transcriptome annotations.

3.1 Assessing results of *de novo* transcriptome assembly and isoform-aware *de novo* ORF finding

3.1.1 A computational strategy for the assessment of *de novo* results

We were interested in evaluating *de novo* transcriptome assembly and *de novo* ORF finding results, and in identifying the factors affecting the results. For this reason, we devised and implemented a workflow for the evaluation of the performances of *StringTie2* (to be reassessed with newer *StringTie* versions) and *ORFquant* at the level of transcript isoforms and at the level of genes.

By applying our workflow, we evaluated *de novo* transcriptome assembly and *de novo* ORF finding results. For this analysis, we used data of the K562 cell line, RNA-seq data from ENCODE as well as Ribo-seq data [16], [101] (Table 9; data of 2 replicates per assay type; Materials and Methods – Data description and availability). The GENCODE v46 annotation was used as reference annotation.

3.1.2 Evaluation of *de novo* results at the isoform level

For the assessment of results at the level of transcript isoforms, we remove coding transcripts from the reference annotation, then we check if *StringTie2* can detect them, or at least detect section of their CDSs (Figure 30a); to assess ORF detection, we remove transcript coding sequences (CDSs), then we check if *ORFquant* can detect them (Figure 30b). At the level of transcripts, *StringTie2* had modest performance (Figure 30a), while *ORFquant* had good performance (Figure 30b), reflecting differences in algorithm performances and data quality (Discussion). As expected, there was an inverse relationship between rank and detected percentage of removed transcripts/transcript CDSs, highlighting that lowly expressed novel transcripts and ORFs are more difficult to find.

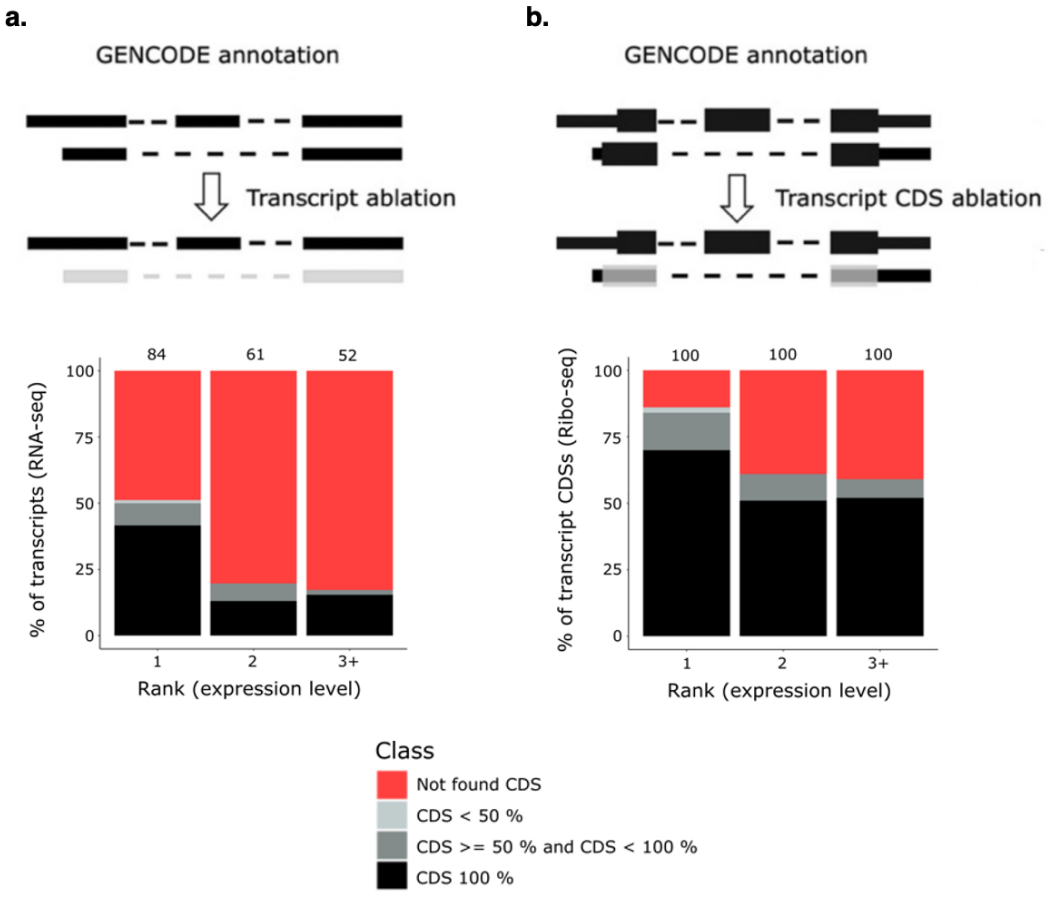


Figure 30 - Evaluation of de novo transcriptome assembly and de novo ORF finding results (transcript-level).

Workflow for the assessment of StringTie2 performance and corresponding results (a); Workflow for the assessment of ORFquant performance and corresponding results (b).

3.1.3 Evaluation of *de novo* results for small genes

To evaluate *de novo* transcriptome assembly results at the level of genes, we remove coding genes with small CDSs (genes with transcript CDSs which are longer than 100 nucleotides and with total gene CDS length which is smaller than or equal to 450 nucleotides) from the reference annotation, then we check if *StringTie2* can detect them, or at least detect section of their CDSs (Figure 31a); to evaluate ORF detection, we remove coding sequences (CDSs) of the selected genes, and we check if *ORFquant* can detect them (Figure 31b). Results show that, when using the two tools, it is more difficult to completely detect coding sequences of multi-exonic genes than coding sequences of mono-exonic genes (Figure 31). More than 50% of CDSs of mono-exonic genes were completely detected by *StringTie2* (Figure 31a) and *ORFquant* (Figure 31b).

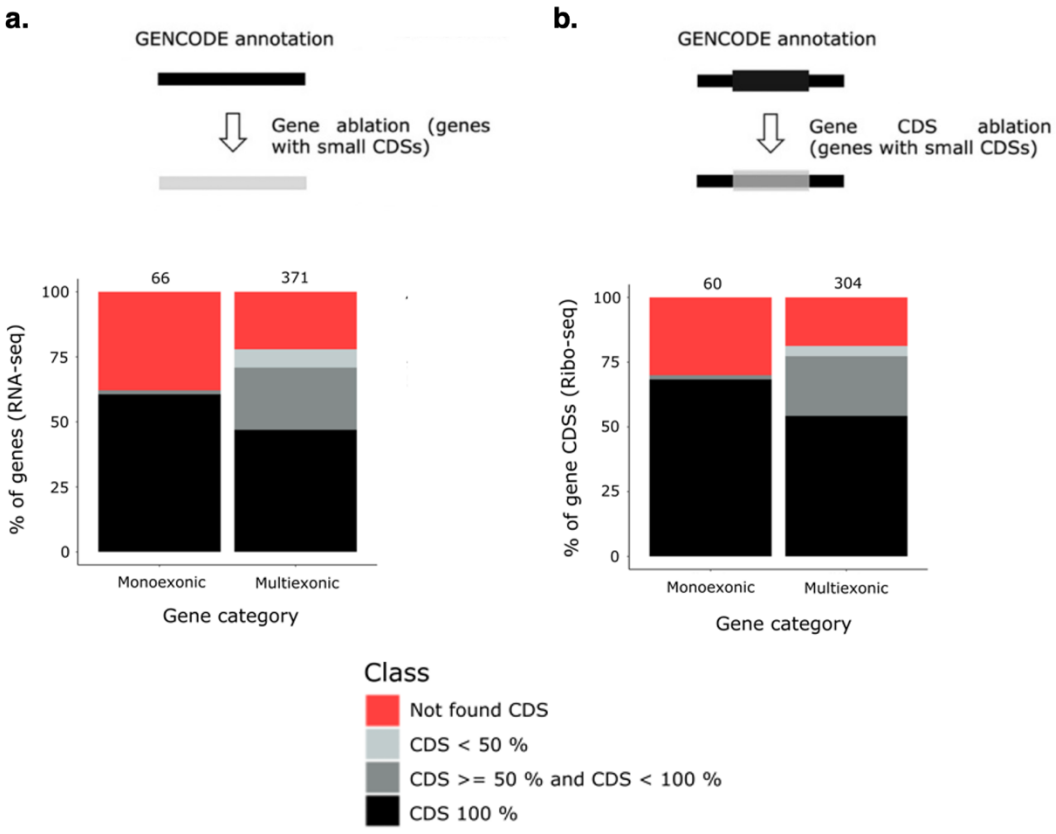


Figure 31 - Evaluation of de novo transcriptome assembly and de novo ORF finding results (gene-level).

Workflow for the assessment of StringTie2 performance and corresponding results (a); Workflow for the assessment of ORFquant performance and corresponding results (b).

Results of this section were obtained by running our assessment workflow with an example dataset (Table 9). In general, since evaluating the quality of *de novo* transcriptome assembly and *de novo* ORF finding results is challenging, our workflow will represent a useful benchmarking method (Discussion).

Replicate	Raw	Cleaned	Mapped	Uniquely mapped
RNAseq_rep1	119,053,315	119,053,315	104,422,961	96,437,399
RNAseq_rep2	113,588,758	113,588,758	101,824,335	94,581,737
Riboseq_rep1	84,777,431	40,509,627	39,289,585	27,401,812
Riboseq_rep2	75,398,258	34,931,298	34,018,524	23,164,611

Table 9 - Mapping statistics of the published K562 data

Mapping statistics are relative to alignments which were performed by providing GENCODE v46 annotation to the aligner STAR. Numbers of read pairs are shown for RNA-seq data.

3.2 Comparing results of custom protein database searches

We developed a method for the comparison of results of proteomic searches at the level of Peptide Spectrum Matches (PSMs). Our method was then applied with data and results of our *de novo* analysis of the FSHD model (Results – Section 2). Two proteomic searches were run by providing the TMT dataset and by setting the FDR threshold to 0.1 for PSMs (Materials and Methods – PSM-level comparison between results of proteomic searches). Two different databases, the GENCODE database and the ORFquant database, were used. The GENCODE database contained GENCODE v47 protein sequences, whereas the ORFquant database contained protein sequences corresponding to *ORFquant*-detected ORFs. We then focused on a specific group of spectra, i.e., 10 639 spectra that were present in results of both searches but were assigned to different peptides by the two searches (Fig. 32a). By performing paired Wilcoxon tests, we compared gene expression levels and numbers of PSMs of their corresponding protein groups.

The numbers of PSMs of proteins assigned in the ORFquant search results were significantly higher (Fig. 32b). RNA-seq expression values were also higher in the ORFquant search results (Fig. 32c), and the same for Ribo-seq TPM values (Fig. 32d).

In conclusion, for all features calculated, ORFquant search – based assignments had better evidence of expression from RNA to protein (Fig. 32b-d), showing a bias towards more expressed genes, or a reduced number of false identifications (Discussion).

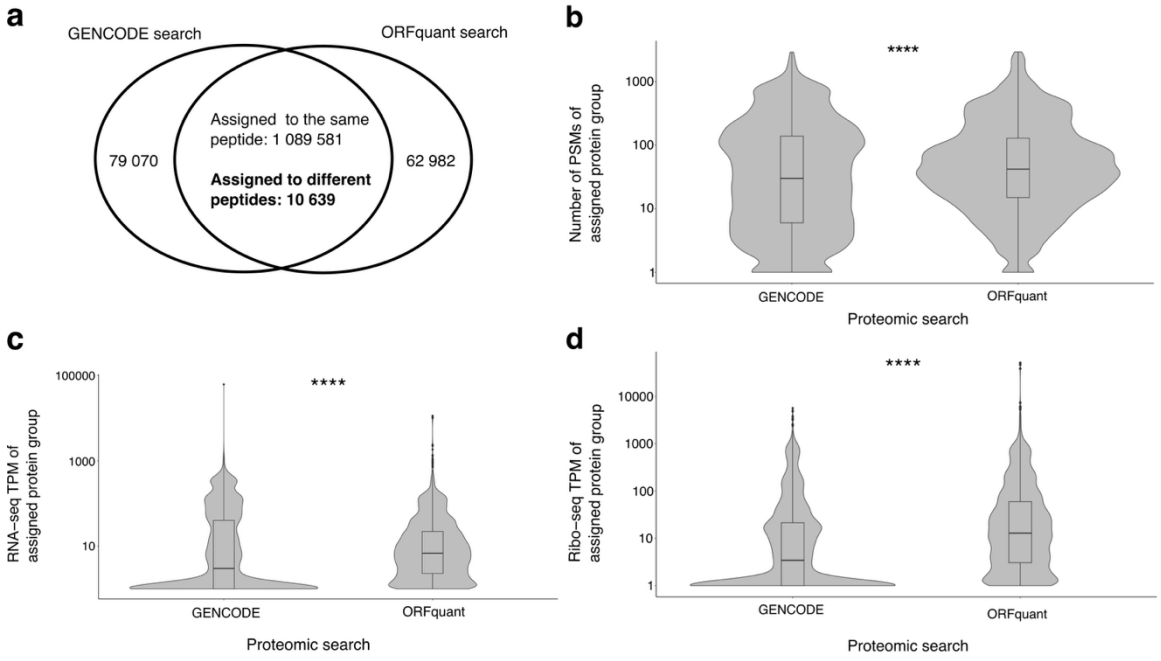


Figure 32 - PSM-level comparison between results of custom protein database searches.

a, Venn diagram showing the numbers of spectra that are present in results of GENCODE search and ORFquant search b-d, Numbers of PSMs, RNA-seq TPMs, Ribo-seq TPMs of protein groups which were assigned to spectra. In b-d, stars represent p-values of paired Wilcoxon tests.

3.3 Comparing long-read-derived transcriptome annotations

As part of the FANTOM consortium [83], we analyzed and compared four transcriptome annotations. To obtain these annotations our collaborators used Cap-trap full-length cDNA sequencing (or CFC-seq), coupling CAP-trapping with long read sequencing, and applied it on two differentiation series, from induced pluripotent stem cells (iPSCs) to Neurons and THP1-derived macrophage activation [102] (Table 10). They provided alignment results not only to the already published tools *bambu*, *IsoQuant* and *TALON*, but also to a new assembler *SALA (Start-site Aware Long-read Assembler)*, which has been recently developed by the consortium [102]. The *SALA* assembler creates 5' end clusters, checks junction consistency, and filters for internal priming [102].

	Raw reads	Complete reads	Aligned reads
iPSC (n=2)	104,715,204	84,037,434	83,567,511
NSC (n=2)	84,410,992	64,790,525	64,395,952
Neuron (n=2)	71,737,831	51,763,571	51,339,392
THP-1 (n=4)	27,315,803	10,720,858	10,634,038
dTHP-1 (n=4)	19,256,619	7,557,804	7,461,294
THP-1 without PAT (n=4)	31,406,353	10,572,355	10,526,128
dTHP-1 without PAT (n=4)	26,441,404	8,900,999	8,864,159
Total	365,284,206	238,343,546	236,788,474

Table 10 - Mapping statistics of the CFC-seq data

Numbers of replicates are in parentheses

To evaluate the four CFC-seq-derived transcriptome assemblies, we obtained, for each annotation, the number of transcripts which were annotation-specific or shared by multiple annotations (Figure 33; Materials and Methods – Comparison between long-read-derived transcriptome annotations). Transcripts which were found by 3 or 4 tools were considered as true positives, or high-confidence transcripts; on the other hand, transcripts which were detected by 1 tool only were considered low-confidence transcripts. Transcripts were also classified as annotated (in GENCODE v39) or novel (not in GENCODE v39).

SALA and *TALON* found similar numbers of high-confidence transcripts. However, *SALA* annotation contained fewer low-confidence transcripts than *TALON* annotation. Moreover, many of the novel transcripts which were found by all tools were low-confidence transcripts, highlighting lack of agreement between tools when the goal is the detection of novel transcripts.

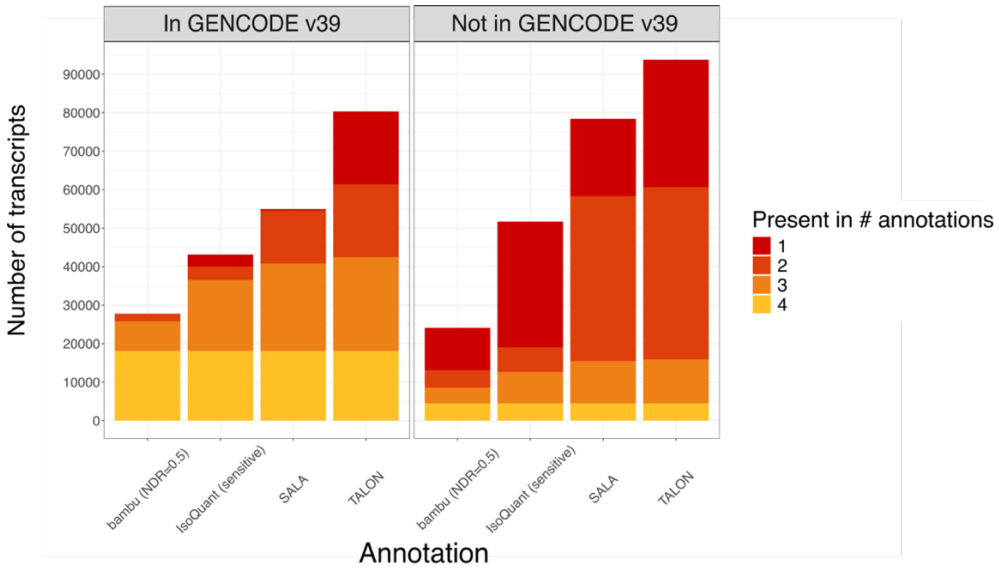


Figure 33 - Comparison between multiple long-read-derived transcriptome annotations.

Number of transcripts is on the y axis. Transcripts were classified as annotation-specific or present in multiple (2, 3, or 4) annotations. Transcripts were also classified as annotated (in GENCODE v39) or novel (not in GENCODE v39).

Discussion

Our *de novo* discovery pipeline from RNA to Protein

In this work we presented R2T2P, our computational pipeline combining *de novo* transcriptome assembly and isoform-level *de novo* ORF finding with proteomic data analysis. It allows to obtain a detailed and comprehensive characterization of transcriptomes, translated transcriptomes, and proteomes. The pipeline improves on traditional data analysis approaches which rely on reference annotations and reference protein databases: it uses data-driven methods to obtain sample-specific or condition-specific transcriptomes and translational landscapes; it builds custom protein databases containing protein sequences corresponding to the detected translated ORFs. By jointly using annotated and identified transcripts, ORFs, and proteins, the pipeline provides a more accurate overview of gene expression products and translation events. Importantly, R2T2P, which was originally developed in bash, has been entirely converted to the workflow manager Nextflow, facilitating not only code maintenance and extension but also monitoring of computational resources and execution times. Using Nextflow will also facilitate sharing of the pipeline, rendering it available to the scientific community as an important platform for integrative, data-driven analyses from RNA to Protein.

Isoform-level resolution and the protein inference problem

In proteomic data analysis, the identification of the specific proteins and protein isoforms from which the detected peptides derive represents an open challenge (Introduction – Section 1.2.3). As different studies have shown, the integration of transcriptomic and translation-related data can provide supporting evidence when characterizing proteomes (Introduction – Section 1.2.4). By working with transcriptome annotations and by using the R package *ORFquant* for *de novo* ORF detection, our pipeline R2T2P achieves isoform-level resolution, and this allows the

assignment of each ORF and peptide to specific transcript isoforms. Such high resolution is a fundamental characteristic of our pipeline, helping to solve the protein inference problem. An extension of the pipeline will include differential expression analyses at the level of protein isoforms (ongoing).

From detection to regulation

The potential of our pipeline R2T2P extends beyond detection of RNA isoforms, ORFs, peptides, and proteins, as it also performs quantification and differential expression analyses with all the used data types, providing information not only on expression and translational levels but also on regulation upon experimental conditions. In this way, it allows to fully appreciate the contribution of molecular heterogeneity and complexity to our biological questions. As our results showed*, de novo* transcriptome assembly can lead to detection of hundreds of unannotated genes and transcripts, including regulated events, even when using moderate to low sequencing depth in RNA-seq or Ribo-seq data (Results – Section 2.2). Analysis of *de novo* ORF finding results provided evidence for translation of unannotated ORFs and for RNA-seq and Ribo-seq upregulation of NMD target transcripts (Results – Sections 2.2.3 and Section 2.1). Moreover, expansion of the proteomic search space by adding *ORFquant* proteins allowed the identification of unannotated peptides, with some of them being regulated following *DUX4* activation in the FSHD model (Results – Section 2.3.1). These findings highlight the potential of our *de novo* pipeline to reveal otherwise undetectable molecular entities and regulated changes, thus shedding light on disease-specific alterations. For its potential to unveil the "dark proteome", R2T2P will represent a powerful approach for future applications and studies on cancer and other diseases.

Increase in transcriptomic complexity

As already mentioned, our pipeline comprises a step of *de novo* transcriptome assembly, which can lead to identification of several unannotated genes and

transcripts. This increases the granularity of the transcriptomic analysis, not only with respect to traditional RNA-seq bulk analyses relying on reference annotations but also with respect to several single-cell and spatial techniques which mostly ignore isoforms and unannotated species in their analysis workflows. However, this data-driven approach also presents some limitations. First, assembly artefacts deriving from lowly expressed genomic regions or fragmented transcripts might be included in the transcriptome annotation, leading to an overestimation of novelty. Importantly, Ribo-seq allows us to focus on cytoplasmic RNAs, effectively filtering out nuclear-retained species. Moreover, the increase in transcriptomic complexity makes assignment of novel transcripts to specific annotated or novel genes particularly challenging.

Adding long-read-derived transcripts and genes further expanded our annotation, potentially increasing the number of assembly artefacts and making the assignment of transcripts to genes even more complex and arbitrary (Materials and Methods – Combining GTF files). When adding the long-read-derived annotation, we found a higher number of detected changes, at the transcriptome and proteome level. However, the increase was only modest, with most of the changes being detected only with the short-read-derived annotation or with both annotations. This might suggest that the short-read-derived annotation captures most of the regulated events, but it might also be related to the fact that the long-read-derived annotation and our annotation were obtained by analysing data of different cells (iMB-clone-5 cells for the long-read-derived annotation; MB135-iDUX4 cells for the short-read-derived annotation). In general, in the absence of a ground truth for transcriptomes, translatomes, and proteomes, it is difficult to establish whether the identifications and detected changes represent true biological events rather than artefacts. To assess results of our pipeline, we worked on developing computational benchmarking strategies, which are addressed in the following section.

Benchmarking methods from RNA to Protein

We developed and implemented a workflow to test the quality of *de novo* transcriptome assembly and *de novo* ORF finding results. By removing some features from the annotation and by checking if the transcriptome assembler *StringTie2* and the R package *ORFquant* detected them, we derived information on their performances on the data at hand when identifying unannotated transcripts and ORFs. In general, since transcriptome assembly and ORF finding results depend not only on data quality but also on the used data analysis strategies, our evaluation workflow will be a useful and powerful tool to guide analysis choices.

With respect to proteogenomic results, we highlighted the importance of the protein database choice for a few spectra. The observed differences, in particular the higher RNA-seq, Ribo-seq and proteomic values of the ORFquant search results, might reflect the fact that when using the ORFquant database there is a lower number of false peptide identifications. Indeed, by assuming that proteins which are produced from highly expressed genes are more likely to be present, high RNA-seq and Ribo-seq expression levels might be considered as a proxy to support the presence of proteins. However, the higher values of the ORFquant search results might also indicate that the ORFquant search is biased towards more expressed genes and proteins. Future improvements of this comparison method might shed light on the accuracy of peptide identifications in multiple proteomic searches and provide better guidelines on custom database building and usage.

As shown in [103], a systematic comparison of long-read-derived transcriptome annotations can provide not only guidelines for the usage of currently available tools but also suggestions for future method development. With many datasets and algorithms for long-read transcriptomics, we highlighted the importance of different assembly strategies, finding a general lack of agreement between assemblers when detecting unannotated transcripts. However, we identified a set of novel transcripts which were found by all assembly tools. Their potential as novel protein-coding loci or functional RNAs remains to be elucidated.

Robustness and functional relevance of the detected novel species

When showing results of proteomic data analyses, we focused on peptides which were identified and quantified by using the combined database, which contained not only annotated proteins but also sequences corresponding to ORFquant-detected ORFs. As already highlighted in Results – Section 1.3, this approach is particularly robust, as it allows the simultaneous detection of annotated and previously unknown translation products, albeit at the cost of reduced sensitivity due to a large protein database. The control of False Discovery Rate (FDR) is important to obtain reliable peptide identifications (Introduction – 1.2.3). When working with databases comprising two protein sets, a possible strategy for FDR control consists in using group-specific FDR thresholds. For our specific application, this approach would have required to specify two different protein evidence levels for the two sets of protein sequences. However, this would have represented a technical and methodological challenge, as exemplified by the presence of novel protein isoforms sharing peptides with known ones, thus complicating the assignment of protein evidence levels. For this reason, we did not perform group-specific FDR estimation, and we instead relied on *FragPipe* default settings for FDR filtering. A potential improvement of our DDA TMT data analysis could consist in activating *MSBooster* (Results – Section 1.4): by rescoring PSMs, this tool might improve peptide identification results. Targeted approaches using synthetic peptides and custom antibodies might be useful for the experimental validation of the detected peptides and proteins, respectively.

By using our pipeline, we discovered novel genes and proteins with evidence for RNA-seq expression in FSHD patients. These findings are particularly intriguing, as these novel genes and proteins could represent biomarkers of the disease, facilitating disease diagnosis. They could also be useful for the development of therapeutic strategies. However, future studies will be needed to elucidate their biological roles by collecting data on their subcellular localization, their molecular interactions, and their ability to form molecular complexes, alongside loss of function or overexpression experiments. Even if they might exert specific molecular functions in the cell, novel translation products might be important for the immune response,

as shown by recent large-scale immunopeptidomics analyses [104]. Our analysis of FSHD patient-derived data could be improved not only by confirming the findings using larger cohorts but also by integrating patient-level proteomic data.

Limitations of R2T2P

Despite its potential, our pipeline has some limitations, which represent opportunities for future extensions and improvements. First, the pipeline does not take into account the presence of genetic mutations and DNA alterations, which can alter the sequences of transcripts, translated regions, peptides, and proteins. Moreover, the pipeline does not include analysis on chimeric transcripts, circRNA or trans-splicing events. In the future, the pipeline will include steps for *de novo* peptide sequencing [105], which would allow the detection of peptides that are not included in protein databases, thus reducing the dependence on database-based searches. Finally, improvement and refinement in terms of execution times and assigned computational resources of the different steps of the Nextflow implementation will be possible after running R2T2P with multiple datasets.

Materials and Methods

RNA-seq and Ribo-seq alignments

RNA-seq reads in FASTQ format were converted to FASTA. Ribo-seq reads were preprocessed before running the alignments which are described in this section: adapter sequences were removed from Ribo-seq reads using *cutadapt*, *fastx_collapser* was used to collapse identical Ribo-seq reads to FASTA format, and Unique Molecular Identifiers (UMIs) were removed using a Perl script. Contaminant sequences (rRNAs, snoRNAs and miRNAs, tRNAs) were then removed from Ribo-seq reads using *Bowtie2*. Finally, the remaining Ribo-seq reads were saved in FASTA files.

In the RNA module two different types of RNA-seq genome alignments are performed with *STAR 2.7.9a*.

In the first type of alignments (hereafter, two-pass alignments), RNA-seq reads are aligned by running STAR in two-pass mode (*--twopassMode Basic*) and by providing the reference annotation Gene Transfer Format (GTF) file to the *STAR* parameter *–sjdbGTFfile* to guide splice junction detection. Local alignment is used (*--alignEndsType Local*), and this enables soft-clipping at the read ends. Alignments are allowed up to 4 mismatches (*--outFilterMismatchNmax 4*), and up to 250 alignments for a read are considered (*--outFilterMultimapNmax 250*). Chimeric alignments are enabled with a minimum segment length of 15 (*--chimSegmentMin 15*), a minimum junction overhang of 10 (*--chimJunctionOverhangMin 10*), a minimum total score of the chimeric segments of 20 (*--chimScoreMin 20*), and a minimum difference between the best chimeric score and the next one of 10 (*--chimScoreSeparation 10*). The Sequence Alignment Map (SAM) output file includes the attributes NH, HI AS, nM, NM, MD, and XS (*--outSAMattributes NH HI AS nM NM MD XS*), while unmapped and partially mapped (i.e., mapped only one mate of a paired end read) reads are saved in separate files (*--outReadsUnmapped Fastx*). The maximum numbers of mismatches for stitching of the splice junctions were set to 0 for non-canonical motifs, GT/AG and CT/AC motif, GC/AG and CT/GC motif,

and AT/AC and GT/AT motif (*--alignSJstitchMismatchNmax 0 0 0 0*). The number of output alignments for multimappers is set to 1 (*--outSAMmultNmax 1*), and alignments for each multimapper are ordered using the *Old_2.4* option. Output wiggle files contain values that are normalized to reads per million (*--outWigNorm RPM*), and the strand is derived from the intron motif (*--outSAMstrandField intronMotif*). *STAR* is run with multi-threading, and 16 threads are used (*--runThreadN 16*). After alignment, *samtools* is used to sort the output SAM files, then to index the resulting BAM (Binary Alignment Map) files.

In the second type of alignments (hereafter, no-two-pass alignments) *STAR* is provided with the reference annotation (using the parameter *--sjdbGTFfile*) as well as with unannotated first-pass junctions of all RNA-seq two-pass alignments (using the parameter *--sjdbFileChrStartEnd*). To reduce spurious junctions a stringent minimum overhang of 500 nt is used for unannotated junctions (*--alignSJoverhangMin 500*), and the insertion of up to 5,000,000 junctions is allowed *(--limitSjdbInsertNsj 5000000*). As in two-pass alignments, the local mode is used (*--alignEndsType Local*), up to 4 mismatches per alignment are permitted (*--outFilterMismatchNmax 4*), and unmapped and partially mapped reads are saved in separate files (*--outReadsUnmapped Fastx*). Regarding handling of multimapping reads, chimeric alignments, SAM attributes, output wiggle files, and strand information from the intron motif, parameter values are the same as in two-pass alignments. The output SAM files are converted into sorted and indexed BAM files using *samtools*.

In the Translation module RNA-seq and Ribo-seq reads are aligned to the genome with *STAR 2.7.9a.*

In these final alignments, *STAR* is provided with an output GTF file of module RNA by using the parameter *--sjdbGTFfile*. As in no-two-pass RNA-seq alignments, a minimum overhang of 500 nt is used for unannotated junctions (*--alignSJoverhangMin 500*). Regarding local alignment, multithreading, SAM attributes, output wiggle files as well as strand information from the intron motif, parameter values are the same as in two-pass and no-two-pass RNA-seq alignments. The same settings are also used for chimeric alignments, except from parameter *–chimJunctionOverhangMin,* which is left at its default value. Specific

parameters are tuned differently for RNA-seq and Ribo-seq data: the maximum number of mismatches per alignment is set to 4 for RNA-seq data and to 3 for Ribo-seq data (parameter *--outFilterMismatchNmax*); up to 250 alignments for a read are considered when aligning RNA-seq reads, but only up to 50 when using Ribo-seq reads (parameter *--outFilterMultimapNmax*). As in previous alignments, only one alignment is reported per multimapping read (*--outSAMmultNmax 1*).

All downstream processing of the alignment files was conducted with Bioconductor R packages *GenomicRanges*, *GenomicFeatures*, *GenomicAlignments*, and *rtracklayer* [106].

Metaplots of RNA-seq and Ribo-seq profiles around stop codons

Pairs of stop codons from transcripts belonging to the same gene were used for the analysis. Each pair consisted of a premature termination codon (PTC) and a canonical stop codon. A window of 100 nt around such stop codons was defined, and genes with PTC windows overlapping other coding regions were excluded from the analysis. RNA-seq and Ribo-seq profiles over the 100nt windows were computed and normalized: when analysing data of each time point and assay (RNA-seq or Ribo-seq), we normalized values by library depth, then divided the resulting values by the total signal per gene. Finally, we 0-1 normalized the values and we computed averages of profiles across genes.

Transcriptome assembly

When analyzing data of the FSHD model, the transcriptome assembler *StringTie3* was used to build transcripts. The assembler was run in conservative mode, and the minimum length for the predicted transcripts was set to 100. *StringTie3* was provided with RNA-seq no-two-pass alignments and with GENCODE v47 annotation information.

Comparison between long-read-derived transcriptome annotations

The contents of four long-read-derived transcriptome annotations of the FANTOM consortium [83] were analyzed and compared by using *GFFCompare* [85] and *R*. First, *GFFCompare* was run by providing all the long-read-derived annotations. Then, it was run by providing GENCODE v39 annotation and each of the long-read-derived annotations. The *GFFCompare* output files were then analyzed in *R* to obtain the numbers of transcripts that were present in a single annotation and in multiple annotations. In this analysis, transcripts were classified as annotated (in GENCODE v39) and novel (not in GENCODE v39).

Combining GTF files

Transcriptome annotations are combined using two custom R scripts, whose details are explained below.

The first R script takes as input the reference transcriptome annotation and a *GFFCompare* output GTF file, which is produced by comparing the *StringTie3*-derived GTF to the reference annotation. Using these inputs, the script merges the reference and the *StringTie3* annotation to generate an annotation which is suitable for downstream analyses, as those performed by the R package *RiboseQC*. After loading the existing reference annotation object and importing the *GFFCompare* output GTF file, *GFFCompare* class codes are used to classify novel transcripts based on their positions relative to reference transcripts (e.g., identical, contained, intronic, antisense, and intergenic). The script then assigns gene and transcript biotypes, propagating them from the reference annotation when appropriate. When merging the annotations, the script excludes *StringTie3*-assembled transcripts which are present in the reference annotation: assembled transcripts with exact intron-chain matches (i.e., *StringTie3* transcripts with *GFFCompare* class code “=”) and assembled transcripts which are contained within reference transcripts and intron compatible with them (i.e., *StringTie3* transcripts with *GFFCompare* class code “c”) are not included in the result of the merging between annotations. The

script generates a merged annotation, while also obtaining genomic features such as exons, CDSs, UTRs, introns, junctions, intergenic regions, and exonic bins. The script also obtains transcript- and gene-level CDS coordinates as well as start and stop codons. Finally, the script saves an .RData file containing the merged annotation with information of reference and novel *StringTie3*-assembled transcripts.

The second R script takes as input the *StringTie3*-generated transcript annotation and a *GFFCompare* output GTF, produced by comparing the user-provided annotation to the *StringTie3* annotation. It loads both annotations as TxDb objects. The script uses *GFFCompare* class codes to classify user-provided transcripts based on their positions relative to *StringTie3* transcripts (e.g., exact matches, contained, intronic, antisense, or novel). The script then merges the annotations. Gene and transcript biotypes are assigned, propagating them from the *StringTie3* annotation when appropriate. When merging the annotations, the script includes user-provided transcripts which are present in the *StringTie3* annotation: user-provided transcripts with exact intron-chain matches (i.e., user-provided transcripts with *GFFCompare* class code "=") and user-provided transcripts which are contained within *StringTie3* transcripts and intron compatible with them (i.e., user-provided transcripts with *GFFCompare* class code "c") are included in the result of the merging between annotations. The script uses the *StringTie3* annotation as the primary annotation and the user-provided annotation as the secondary annotation: in cases where user transcripts are assigned to *StringTie3* genes, *StringTie3* and user gene IDs (identifiers) of such cases are concatenated and used to define unified gene identifiers with the R package *igraph*. The final output of the R script is a GTF file containing exons of *StringTie3* and user transcripts, along with a text file with details on which original *StringTie3* and user gene IDs were combined when defining the unified gene identifiers. The output GTF file is then compared to the reference transcriptome annotation with *GFFCompare*, and the two annotations are merged by using the first R script.

Gene-level differential expression analyses

Gene-level differential expression analyses were performed using the R package *DESeq2*. Before visualizing results of the differential expression analyses, we removed genes with biotype “artifact” as well as overlapping genes with no differential feature uniquely mapping to them. Differentially expressed genes were defined as genes with non-zero log2 fold changes and with adjusted p-values below a threshold of 0.05. Upregulated genes were defined as differentially expressed genes with positive log2 fold changes.

For the analyses of line plots with RNA-seq and Ribo-seq log2 fold changes, 1000 genes were pseudo-randomly sampled (using the *sample* function with seed set to 42) from the set of expressed genes, which were defined as those with RNA-seq and Ribo-seq Transcripts Per Million (TPM) values which were equal to or greater than 1 in all three replicates of the FSHD model.

Visualizations of the genomic region comprising the first exon of the gene *DDX10* (in Results – Section 2.2.2 and in Results – Section 2.3.3) were obtained using the *Integrative Genomics Viewer (IGV) version 2.16.0* [107].

Assessment of DUX4 binding with chromatin immunoprecipitation sequencing data

Transcript promoters were extracted with the *promoters* function of the R package *DESeq2* by using its default parameters: promoters comprised 2000 nucleotides before and 200 nucleotides after Transcription Start Sites.

A gene was defined as supported by chromatin immunoprecipitation sequencing (ChIP-seq) data if at least one of its transcript promoters contained a peak, but, when analysing transcript promoters of novel genes, peaks overlapping promoters of annotated transcripts were excluded, ensuring no ChIP-seq signal from them.

For each gene group, the proportion of ChIP-seq-supported genes was calculated: the number of genes with ChIP-seq support was divided by the total number of

genes in the group. Then, pairwise comparisons of proportions were computed by using the *pairwise.prop.test* function.

ORF finding

The R package *RiboseQC* was provided with Ribo-seq alignment results of the Translation module and with the “Rannot” output file of the RNA module, and it was used to get Ribo-seq read counts. Then, the “for_ORFquant” file, which was produced by *RiboseQC*, as well as the “Rannot” file were provided to the R package *ORFquant* to identify translated regions on transcripts. When analysing data of the FSHD model, translated regions on GENCODE v47, *StringTie3*, and long-read-derived transcripts were detected by using only the signal from uniquely mapping reads and by using only canonical start codons. Plots with numbers, lengths, and quantification estimates of detected ORFs within annotated and *StringTie3*-assembled novel transcripts were generated with a custom R script.

Evaluation of *de novo* transcriptome assembly and *de novo* ORF finding results

When using data of the K562 cell line for the *StringTie2* performance, the assembler was run in conservative mode, and the minimum length for the predicted transcripts was set to 100. The assembler was provided with RNA-seq two-pass alignments and with two modified versions of the GENCODE v46 annotation, i.e., annotation information without specific genes and annotation information without specific transcripts. When using data of the K562 cell line for the *ORFquant* performance, translated regions on GENCODE v46 transcripts were detected, and *ORFquant* was provided with two modified versions of the GENCODE v46 annotation, i.e., annotation information without specific gene CDSs and annotation information without specific transcript CDSs. When running *ORFquant*, the minimum percentage of total gene translation for an ORF to be selected was set to 0.

For the transcript-level assessment analyses test transcripts and test transcript CDSs were selected. Removed protein-coding transcripts (for *StringTie2* evaluation) had more than 30 uniquely mapped RNA-seq reads on unique features, i.e., exons and junctions, whereas removed transcript CDSs of protein-coding transcripts (for *ORFquant* evaluation) had more than 10 uniquely mapped Ribo-seq reads on unique features. Both removed transcripts and removed transcript CDSs had a total length of unique features which was greater than or equal to 60 nucleotides. Removed transcripts and transcript CDSs were ranked by gene according to their counts of uniquely mapped reads on unique features (RNA-seq reads for *StringTie2* evaluation, Ribo-seq reads for *ORFquant* evaluation).

For the gene-level assessment analyses a set of test genes was selected. We selected genes which had a total CDS length which was smaller than or equal to 450 nucleotides. This set of genes was filtered by excluding not only genes whose CDSs overlapped exonic parts which were shared by multiple genes, but also genes whose CDSs overlapped CDSs of genes with longer CDSs. Transcript CDSs of the selected genes had to be longer than 100 nucleotides, and they had to start at least 10 nucleotides downstream the transcript 5' end. Moreover, selected genes were required to have at least 10 uniquely mapped reads on exons (RNA-seq reads for *StringTie2* evaluation, Ribo-seq reads for *ORFquant* evaluation). Selected genes were classified as mono-exonic or multi-exonic. Finally, selected genes were removed from the annotation for *StringTie2* evaluation, and transcript CDSs of the selected genes were removed for *ORFquant* evaluation.

Proteomic searches

Protein sequences corresponding to CDSs of GENCODE v47 annotation were extracted from the GENCODE v47 Rannot file in a custom R script using functions from packages *GenomicFeatures*, *Biostrings*, and *GenomicRanges*. Protein sequences were then saved to a file in FASTA format. The same custom R script also saved a FASTA file containing protein sequences corresponding to *ORFquant*-detected ORFs and a FASTA file containing both sets of proteins. *Philosopher 5.1.2* was used to add decoys and contaminants to the three protein databases.

Proteomic searches were run by providing the TMT data of the FSHD model as well as the three protein databases to *FragPipe v23.1* (Results – Section 2.3.1). *MSFragger v4.3* and the pre-released *IonQuant v1.11.13* were used for the searches. The pre-released version of *IonQuant* was provided upon request via e-mail by the developers of *FragPipe* on October 20th 2025 at 2:45 PM (Italy time). The following search settings were used: the basic TMT 16-plex workflow was used, with quantification and identification from MS2; Met oxidation, protein N-term Acetyl, n-term TMT were specified as variable modifications; the default value for *Philosopher* filtering was used; *TMT-Integrator* was used with reference approach, performing data summarization at all levels, with no normalization or log2 transformation applied. *MSFragger v4.3* was run using *stricttrypsin* as search enzyme, a minimum peptide length of 7, a maximum length of 50, and allowing up to 2 missed cleavages.

Peptide-level analyses

Detected peptides mapping to multiple genes were removed before obtaining numbers of detected peptides; then, peptide intensities were $\log_2$-transformed and differential expression analyses were performed with the R package *limma*. Median-centering normalization was not applied as it was considered unnecessary based on inspection of distributions of $\log_2$-transformed peptide intensities. An adjusted p-value threshold of 0.05 was used when defining upregulated and downregulated peptides, which had positive and negative log2 fold changes, respectively.

Visualizations of the genomic loci of the two identified novel proteins (in Results – Section 2.3.2 and in Results – Section 2.3.3) were obtained using the *Integrative Genomics Viewer (IGV) version 2.16.0*.

PSM-level comparison between results of proteomic searches

Other two proteomic searches were performed by providing *FragPipe v23.1* with the previously described databases of annotated proteins and *ORFquant* proteins, with both databases including target sequences, decoys, and contaminants. The same search settings as in previous searches were used, except for the following parameters: neither sequential nor picked was specified for *Philosopher* filtering, the *ProteinProphet* file was not used, the FDR threshold for Peptide Spectrum Matches (PSMs) was set to 0.1, and the minimum probability in *Percolator* was set to 0. These parameter values were changed to use the same FDR threshold of 0.1 for PSMs in both searches, thus obtaining comparable PSM-level results. Results of the two proteomic searches were compared in *R* by focusing on spectra that were present in results of both searches but were assigned to different peptides by the two searches. In results of each proteomic search each spectrum was associated with a group of genes and with a group of proteins. RNA-seq TPMs and Ribo-seq TPMs of genes at 14h were obtained with *RiboseQC*. Numbers of PSMs of proteins were also obtained. In this comparison analysis, the highest values of each spectrum, i.e., the highest RNA-seq TPM, the highest Ribo-seq TPM, and the highest number of PSMs, were considered.

Data description and availability

RNA-seq and Ribo-seq data of human skeletal muscle cells are published, and they are available at GEO (GEO Accession: GSE178761). This dataset consists of RNA-seq and Ribo-seq reads of 3 replicates at four time points after *DUX4* activation (0, 4, 8, and 14 h). DUX4 ChIP-seq data were taken from [43]. The list of DUX4 targets which was used in our analyses was taken from [42]. The RNA-seq dataset of patient-derived and control myotubes was taken from [42], whereas the RNA-seq dataset of patient-derived biopsies and G1 controls was taken from [100].

The long-read-derived GTF annotation file which was used for the analysis of data of the FSHD model (Results – Section 2.4) was provided by authors of [41]. Before

using it in our analyses, we filtered the GTF file to keep only features on canonical chromosomes.

The proteomic dataset which was used in this study is currently not published. It was provided by our collaborators of the Jagannathan Lab (University of Colorado). It comprises TMT proteomic data of 3 human skeletal muscle replicates at four time points (0, 4, 8, and 14h) after *DUX4* activation or treatment with dimethyl sulfoxide (DMSO). The dataset also included conditions with proteasome inhibition, but they were not used in this study. The TMT 16plex was used to obtain the data, but the last 3 channels of the kit were not used.

Paired-end RNA-seq data of two K562 replicates are published, and they are available at GEO (GEO Accessions: GSM2308418 and GSM2308419; ENCODE accession number: ENCSR000CPH; ENCODE production laboratory: Thomas Gingeras, CSHL). Ribo-seq data of two K562 replicates are also available at GEO (Accessions: GSM3692340 and GSM3692341).

The four long-read-derived annotations of the FANTOM consortium [83] were obtained by providing *SALA (Start-site Aware Long-read Assembler)* [102], *TALON* [108], *IsoQuant* [55], and *bambu* [109] with CFC-seq (Cap-trap full-length cDNA sequencing) [102] alignment results of two differentiation series (Neuron series and THP1 series). These annotation files are currently not published, and they were provided by our collaborators of the FANTOM consortium.

Software implementation

The computational pipeline *R2T2P* (Results – Section 1) was implemented in bash and R scripts. A principal bash script sequentially launches other bash scripts to perform the different steps of the pipeline. The main script is provided with paths to RNA-seq, Ribo-seq, and proteomic data files and with additional file metadata, including data type (“RNA” or “Ribo” for transcriptomic data; “DDA” or “DIA” for proteomic data) and conditions to compare in the differential expression analyses. The script is also provided with the GTF and Rannot files of a reference annotation, a *FragPipe* workflow file specifying proteomic search settings and, optionally, an additional user-provided GTF file. The entire pipeline was converted to *Nextflow* [99]

to improve usability, reproducibility, portability, and resource allocation (Results – Section 1.5). Using the workflow manager *Nextflow* also facilitates code maintenance due to the modular structure of the pipeline.

Benchmarking methods were implemented in R as two different sets of functions, which were included in the R package *R2T2P.Benchmark* (Results – Section 3). The first set of functions allows evaluation of *de novo* transcriptome assembly and isoform-aware *de novo* ORF finding results (Results – Section 3.1). The second set of functions is useful for the comparison of proteomic search results and transcriptome annotations (Results – Sections 3.2 and 3.3).

References

[1] M. Pertea, 'The Human Transcriptome: An Unfinished Story', *Genes*, vol. 3, no. 3, pp. 344–360, Jun. 2012, doi: 10.3390/genes3030344.

[2] D. A. Glinos *et al.*, 'Transcriptome variation in human tissues revealed by long-read sequencing', *Nature*, vol. 608, no. 7922, pp. 353–359, Aug. 2022, doi: 10.1038/s41586-022-05035-y.

[3] B. B. Cummings *et al.*, 'Improving genetic diagnosis in Mendelian disease with transcriptome sequencing', *Sci Transl Med*, vol. 9, no. 386, p. eaal5209, Apr. 2017, doi: 10.1126/scitranslmed.aal5209.

[4] J. Tapial *et al.*, 'An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms', *Genome Res*, vol. 27, no. 10, pp. 1759–1768, Oct. 2017, doi: 10.1101/gr.220962.117.

[5] A. Reyes and W. Huber, 'Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues', *Nucleic Acids Research*, vol. 46, no. 2, pp. 582–592, Jan. 2018, doi: 10.1093/nar/gkx1165.

[6] L. E. Marasco and A. R. Kornblihtt, 'The physiology of alternative splicing', *Nat Rev Mol Cell Biol*, vol. 24, no. 4, pp. 242–254, Apr. 2023, doi: 10.1038/s41580-022-00545-z.

[7] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nat Genet*, vol. 40, no. 12, pp. 1413–1415, Dec. 2008, doi: 10.1038/ng.259.

[8] E. T. Wang *et al.*, 'Alternative isoform regulation in human tissue transcriptomes', *Nature*, vol. 456, no. 7221, pp. 470–476, Nov. 2008, doi: 10.1038/nature07509.

[9] T. Sterne-Weiler *et al.*, 'Frac-seq reveals isoform-specific recruitment to polyribosomes', *Genome Res*, vol. 23, no. 10, pp. 1615–1623, Oct. 2013, doi: 10.1101/gr.148585.112.

[10] E. de Klerk and P. A. C. ’t Hoen, ‘Alternative mRNA transcription, processing, and translation: insights from RNA sequencing’, *Trends Genet*, vol. 31, no. 3, pp. 128–139, Mar. 2015, doi: 10.1016/j.tig.2015.01.001.

[11] V. Pelechano and L. M. Steinmetz, ‘Gene regulation by antisense transcription’, *Nat Rev Genet*, vol. 14, no. 12, pp. 880–893, Dec. 2013, doi: 10.1038/nrg3594.

[12] Y. Malka *et al.*, ‘Post-transcriptional 3´-UTR cleavage of mRNA transcripts generates thousands of stable uncapped autonomous RNA fragments’, *Nat Commun*, vol. 8, no. 1, p. 2029, Dec. 2017, doi: 10.1038/s41467-017-02099-7.

[13] E. Eisenberg and E. Y. Levanon, ‘A-to-I RNA editing — immune protector and transcriptome diversifier’, *Nat Rev Genet*, vol. 19, no. 8, pp. 473–490, Aug. 2018, doi: 10.1038/s41576-018-0006-1.

[14] F. Valdivia-Francia and A. Sendoel, ‘No country for old methods: New tools for studying microproteins’, *iScience*, vol. 27, no. 2, p. 108972, Feb. 2024, doi: 10.1016/j.isci.2024.108972.

[15] M. W. Orr, Y. Mao, G. Storz, and S.-B. Qian, ‘Alternative ORFs and small ORFs: shedding light on the dark proteome’, *Nucleic Acids Research*, vol. 48, no. 3, pp. 1029–1042, Feb. 2020, doi: 10.1093/nar/gkz734.

[16] L. Calviello, A. Hirsekorn, and U. Ohler, ‘Quantification of translation uncovers the functions of the alternative transcriptome’, *Nat Struct Mol Biol*, vol. 27, no. 8, pp. 717–725, Aug. 2020, doi: 10.1038/s41594-020-0450-4.

[17] T. G. Johnstone, A. A. Bazzini, and A. J. Giraldez, ‘Upstream ORFs are prevalent translational repressors in vertebrates’, *The EMBO Journal*, vol. 35, no. 7, pp. 706–723, Apr. 2016, doi: 10.15252/embj.201592759.

[18] L. Calviello and U. Ohler, ‘Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome’, *Trends Genet*, vol. 33, no. 10, pp. 728–744, Oct. 2017, doi: 10.1016/j.tig.2017.08.003.

[19] S. Lykke-Andersen and T. H. Jensen, ‘Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes’, *Nat Rev Mol Cell Biol*, vol. 16, no. 11, pp. 665–677, Nov. 2015, doi: 10.1038/nrm4063.

[20] A. Kosti and G. J. Bassell, ‘Where to start? Activity-dependent alternative translation initiation generates multifunctional proteoforms in the brain’,

Molecular Cell, vol. 84, no. 20, pp. 3863–3865, Oct. 2024, doi: 10.1016/j.molcel.2024.09.029.

[21] G. Ren *et al.*, 'Ribosomal frameshifting at normal codon repeats recodes functional chimeric proteins in human', *Nucleic Acids Research*, vol. 52, no. 5, pp. 2463–2479, Mar. 2024, doi: 10.1093/nar/gkae035.

[22] K. Mangkalaphiban *et al.*, 'Extended stop codon context predicts nonsense codon readthrough efficiency in human cells', *Nat Commun*, vol. 15, no. 1, p. 2486, Mar. 2024, doi: 10.1038/s41467-024-46703-z.

[23] S. Tsour *et al.*, 'Alternate RNA decoding results in stable and abundant proteins in mammals', Aug. 26, 2024, *bioRxiv*. doi: 10.1101/2024.08.26.609665.

[24] V. Tretyachenko, T. Leiman, O. Asraf, O. Dahan, D. Dahary, and Y. Pilpel, 'Encoded and non-genetic alternative protein variants expand human functional proteome', Feb. 17, 2025, *bioRxiv*. doi: 10.1101/2025.02.17.638604.

[25] Y. Yang *et al.*, 'ASpdb: an integrative knowledgebase of human protein isoforms from experimental and AI-predicted structures', *Nucleic Acids Research*, vol. 53, no. D1, pp. D331–D339, Jan. 2025, doi: 10.1093/nar/gkae1018.

[26] The Consortium for Top Down Proteomics, L. M. Smith, and N. L. Kelleher, 'Proteoform: a single term describing protein complexity', *Nat Methods*, vol. 10, no. 3, pp. 186–187, Mar. 2013, doi: 10.1038/nmeth.2369.

[27] X. Yang *et al.*, 'Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing', *Cell*, vol. 164, no. 4, pp. 805–817, Feb. 2016, doi: 10.1016/j.cell.2016.01.029.

[28] T. Klein, U. Eckhard, A. Dufour, N. Solis, and C. M. Overall, 'Proteolytic Cleavage—Mechanisms, Function, and "Omic" Approaches for a Near-Ubiquitous Posttranslational Modification', *Chem. Rev.*, vol. 118, no. 3, pp. 1137–1168, Feb. 2018, doi: 10.1021/acs.chemrev.7b00120.

[29] J. A. Korchak, S. Stephen Yi, N. L. Kelleher, N. Sahni, and G. M. Sheynkman, 'Proteoform medicine: characterizing and targeting protein forms in human disease', *Nat Rev Genet*, Jan. 2026, doi: 10.1038/s41576-025-00915-1.

[30] B. J. Blencowe, ‘The Relationship between Alternative Splicing and Proteomic Complexity’, *Trends Biochem Sci*, vol. 42, no. 6, pp. 407–408, Jun. 2017, doi: 10.1016/j.tibs.2017.04.001.

[31] M. L. Tress, F. Abascal, and A. Valencia, ‘Most Alternative Isoforms Are Not Functionally Important’, *Trends Biochem Sci*, vol. 42, no. 6, pp. 408–410, Jun. 2017, doi: 10.1016/j.tibs.2017.04.002.

[32] C. Buccitelli and M. Selbach, ‘mRNAs, proteins and the emerging principles of gene expression control’, *Nat Rev Genet*, vol. 21, no. 10, pp. 630–644, Oct. 2020, doi: 10.1038/s41576-020-0258-4.

[33] D. C. Lynch *et al.*, ‘Disrupted auto-regulation of the spliceosomal gene SNRPB causes cerebro-costo-mandibular syndrome’, *Nat Commun*, vol. 5, p. 4483, Jul. 2014, doi: 10.1038/ncomms5483.

[34] E. Di Nisio *et al.*, ‘A truncated and catalytically inactive isoform of KDM5B histone demethylase accumulates in breast cancer cells and regulates H3K4 tri-methylation and gene expression’, *Cancer Gene Ther*, vol. 30, no. 6, pp. 822–832, Jun. 2023, doi: 10.1038/s41417-022-00584-w.

[35] G. R. LaForce *et al.*, ‘Suppression of premature transcription termination leads to reduced mRNA isoform diversity and neurodegeneration’, *Neuron*, vol. 110, no. 8, pp. 1340-1357.e7, Apr. 2022, doi: 10.1016/j.neuron.2022.01.018.

[36] M. Hutton *et al.*, ‘Association of missense and 5’-splice-site mutations in tau with the inherited dementia FTDP-17’, *Nature*, vol. 393, no. 6686, pp. 702–705, Jun. 1998, doi: 10.1038/31508.

[37] C. L. Lin *et al.*, ‘Aberrant RNA processing in a neurodegenerative disease: the cause for absent EAAT2, a glutamate transporter, in amyotrophic lateral sclerosis’, *Neuron*, vol. 20, no. 3, pp. 589–602, Mar. 1998, doi: 10.1016/s0896-6273(00)80997-6.

[38] H. Rhinn *et al.*, ‘Alternative α-synuclein transcript usage as a convergent mechanism in Parkinson’s disease pathology’, *Nat Commun*, vol. 3, no. 1, p. 1084, Sep. 2012, doi: 10.1038/ncomms2032.

[39] T. Gonatopoulos-Pournatzis *et al.*, ‘Autism-Misregulated eIF4G Microexons Control Synaptic Translation and Higher Order Cognitive Functions’, *Mol Cell*, vol. 77, no. 6, pp. 1176-1192.e16, Mar. 2020, doi: 10.1016/j.molcel.2020.01.006.

[40] M. S. Tihaya *et al.*, ‘Facioscapulohumeral muscular dystrophy: the road to targeted therapies’, *Nat Rev Neurol*, vol. 19, no. 2, pp. 91–108, Feb. 2023, doi: 10.1038/s41582-022-00762-2.

[41] D. Zheng *et al.*, ‘DUX4 activates common and context-specific intergenic transcripts and isoforms’, *Sci. Adv.*, vol. 11, no. 19, p. eadt5356, May 2025, doi: 10.1126/sciadv.adt5356.

[42] Z. Yao *et al.*, ‘DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle’, *Human Molecular Genetics*, vol. 23, no. 20, pp. 5342–5352, Oct. 2014, doi: 10.1093/hmg/ddu251.

[43] L. N. Geng *et al.*, ‘DUX4 Activates Germline Genes, Retroelements, and Immune Mediators: Implications for Facioscapulohumeral Dystrophy’, *Developmental Cell*, vol. 22, no. 1, pp. 38–51, Jan. 2012, doi: 10.1016/j.devcel.2011.11.013.

[44] A. Cortesi *et al.*, ‘4q-D4Z4 chromatin architecture regulates the transcription of muscle atrophic genes in facioscapulohumeral muscular dystrophy’, *Genome Res.*, vol. 29, no. 6, pp. 883–895, Jun. 2019, doi: 10.1101/gr.233288.117.

[45] A. E. Campbell *et al.*, ‘Compromised nonsense-mediated RNA decay results in truncated RNA-binding protein production upon DUX4 expression’, *Cell Rep*, vol. 42, no. 6, p. 112642, Jun. 2023, doi: 10.1016/j.celrep.2023.112642.

[46] S. Jagannathan, Y. Ogata, P. R. Gafken, S. J. Tapscott, and R. K. Bradley, ‘Quantitative proteomics reveals key roles for post-transcriptional gene regulation in the molecular pathology of facioscapulohumeral muscular dystrophy’, *eLife*, vol. 8, p. e41740, Jan. 2019, doi: 10.7554/eLife.41740.

[47] R. Stark, M. Grzelak, and J. Hadfield, ‘RNA sequencing: the teenage years’, *Nat Rev Genet*, vol. 20, no. 11, pp. 631–656, Nov. 2019, doi: 10.1038/s41576-019-0150-2.

[48] A. Dobin *et al.*, ‘STAR: ultrafast universal RNA-seq aligner’, *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.

[49] J. M. Heather and B. Chain, ‘The sequence of sequencers: The history of sequencing DNA’, *Genomics*, vol. 107, no. 1, pp. 1–8, Jan. 2016, doi: 10.1016/j.ygeno.2015.11.003.

[50] F. R. Ringeling *et al.*, ‘Partitioning RNAs by length improves transcriptome reconstruction from short-read RNA-seq data’, *Nat Biotechnol*, vol. 40, no. 5, pp. 741–750, May 2022, doi: 10.1038/s41587-021-01136-7.

[51] Y. Wang, Y. Zhao, A. Bollas, Y. Wang, and K. F. Au, ‘Nanopore sequencing technology, bioinformatics and applications’, *Nat Biotechnol*, vol. 39, no. 11, pp. 1348–1365, Nov. 2021, doi: 10.1038/s41587-021-01108-x.

[52] A. D. Tang *et al.*, ‘Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns’, *Nat Commun*, vol. 11, no. 1, p. 1438, Mar. 2020, doi: 10.1038/s41467-020-15171-6.

[53] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg, ‘StringTie enables improved reconstruction of a transcriptome from RNA-seq reads’, *Nat Biotechnol*, vol. 33, no. 3, pp. 290–295, Mar. 2015, doi: 10.1038/nbt.3122.

[54] I. Shinder, G. Pertea, R. Hu, Z. Rudnick, and M. Pertea, ‘StringTie3 Improves Total RNA-seq Assembly by Resolving Nascent and Mature Transcripts’, May 26, 2025, *bioRxiv*. doi: 10.1101/2025.05.21.655404.

[55] A. D. Prjibelski *et al.*, ‘Accurate isoform discovery with IsoQuant using long reads’, *Nat Biotechnol*, vol. 41, no. 7, pp. 915–918, Jul. 2023, doi: 10.1038/s41587-022-01565-y.

[56] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, ‘Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling’, *Science*, vol. 324, no. 5924, pp. 218–223, Apr. 2009, doi: 10.1126/science.1168978.

[57] L. Calviello *et al.*, ‘Detecting actively translated open reading frames in ribosome profiling data’, *Nat Methods*, vol. 13, no. 2, pp. 165–170, Feb. 2016, doi: 10.1038/nmeth.3688.

[58] A. A. Bazzini *et al.*, ‘Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation’, *EMBO J*, vol. 33, no. 9, pp. 981–993, May 2014, doi: 10.1002/embj.201488411.

[59] N. T. Ingolia *et al.*, ‘Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes’, *Cell Rep*, vol. 8, no. 5, pp. 1365–1379, Sep. 2014, doi: 10.1016/j.celrep.2014.07.045.

[60] A. M. Michel, K. R. Choudhury, A. E. Firth, N. T. Ingolia, J. F. Atkins, and P. V. Baranov, ‘Observation of dually decoded regions of the human genome using ribosome profiling data’, *Genome Res*, vol. 22, no. 11, pp. 2219–2229, Nov. 2012, doi: 10.1101/gr.133249.111.

[61] Z. Ji, R. Song, A. Regev, and K. Struhl, ‘Many lncRNAs, 5’UTRs, and pseudogenes are translated and some are likely to express functional proteins’, *Elife*, vol. 4, p. e08890, Dec. 2015, doi: 10.7554/eLife.08890.

[62] A. Varabyou, B. Erdogdu, S. L. Salzberg, and M. Pertea, ‘Investigating open reading frames in known and novel transcripts using ORFanage’, *Nat Comput Sci*, vol. 3, no. 8, pp. 700–708, Jul. 2023, doi: 10.1038/s43588-023-00496-1.

[63] Y. Jiang *et al.*, ‘Comprehensive Overview of Bottom-Up Proteomics Using Mass Spectrometry’, *ACS Meas. Sci. Au*, vol. 4, no. 4, pp. 338–417, Aug. 2024, doi: 10.1021/acsmeasuresciau.3c00068.

[64] A. I. Nesvizhskii, ‘Proteogenomics: concepts, applications and computational strategies’, *Nat Methods*, vol. 11, no. 11, pp. 1114–1125, Nov. 2014, doi: 10.1038/nmeth.3144.

[65] X. Zou *et al.*, ‘In-depth analysis of data characteristics and comparative evaluation of dda and dia accuracy in label-free quantitative proteomics of biological samples’, *Clin Proteom*, vol. 23, no. 1, p. 2, Dec. 2026, doi: 10.1186/s12014-025-09572-2.

[66] R. Lou and W. Shui, ‘Acquisition and Analysis of DIA-Based Proteomic Data: A Comprehensive Survey in 2023’, *Molecular & Cellular Proteomics*, vol. 23, no. 2, p. 100712, Feb. 2024, doi: 10.1016/j.mcpro.2024.100712.

[67] V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley, and M. Ralser, ‘DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput’, *Nat Methods*, vol. 17, no. 1, pp. 41–44, Jan. 2020, doi: 10.1038/s41592-019-0638-x.

[68] M. Kalhor, J. Lapin, M. Picciani, and M. Wilhelm, ‘Rescoring Peptide Spectrum Matches: Boosting Proteomics Performance by Integrating Peptide Property Predictors Into Peptide Identification’, *Molecular & Cellular Proteomics*, vol. 23, no. 7, p. 100798, Jul. 2024, doi: 10.1016/j.mcpro.2024.100798.

[69] R. M. Miller *et al.*, ‘Enhanced protein isoform characterization through long-read proteogenomics’, *Genome Biol*, vol. 23, no. 1, p. 69, Mar. 2022, doi: 10.1186/s13059-022-02624-y.

[70] M. The, P. Samaras, B. Kuster, and M. Wilhelm, ‘Reanalysis of ProteomicsDB Using an Accurate, Sensitive, and Scalable False Discovery Rate Estimation Approach for Protein Groups’, *Mol Cell Proteomics*, vol. 21, no. 12, p. 100437, Dec. 2022, doi: 10.1016/j.mcpro.2022.100437.

[71] B. Bogdanow, H. Zauber, and M. Selbach, ‘Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides’, *Mol Cell Proteomics*, vol. 15, no. 8, pp. 2791–2801, Aug. 2016, doi: 10.1074/mcp.M115.055103.

[72] L. Fancello and T. Burger, ‘An analysis of proteogenomics and how and when transcriptome-informed reduction of protein databases can enhance eukaryotic proteomics’, *Genome Biol*, vol. 23, no. 1, p. 132, Jun. 2022, doi: 10.1186/s13059-022-02701-2.

[73] Y. Zhu *et al.*, ‘DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis’, *Mol Cell Proteomics*, vol. 19, no. 6, pp. 1047–1057, Jun. 2020, doi: 10.1074/mcp.TIR119.001646.

[74] L. Zhang and J. E. Elias, ‘Relative Protein Quantification Using Tandem Mass Tag Mass Spectrometry’, *Methods Mol Biol*, vol. 1550, pp. 185–198, 2017, doi: 10.1007/978-1-4939-6747-6_14.

[75] J. Li *et al.*, ‘TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples’, *Nat Methods*, vol. 17, no. 4, pp. 399–404, Apr. 2020, doi: 10.1038/s41592-020-0781-4.

[76] J. Li *et al.*, ‘TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing’, *J Proteome Res*, vol. 20, no. 5, pp. 2964–2972, May 2021, doi: 10.1021/acs.jproteome.1c00168.

[77] Y. Li *et al.*, ‘Proteogenomic data and resources for pan-cancer analysis’, *Cancer Cell*, vol. 41, no. 8, pp. 1397–1406, Aug. 2023, doi: 10.1016/j.ccell.2023.06.009.

[78] S. Verbruggen *et al.*, ‘PROTEOFORMER 2.0: Further Developments in the Ribosome Profiling-assisted Proteogenomic Hunt for New Proteoforms’,

Molecular & Cellular Proteomics, vol. 18, no. 8, pp. S126–S140, Aug. 2019, doi: 10.1074/mcp.RA118.001218.

[79] F. Huber *et al.*, 'A comprehensive proteogenomic pipeline for neoantigen discovery to advance personalized cancer immunotherapy', *Nat Biotechnol*, vol. 43, no. 8, pp. 1360–1372, Aug. 2025, doi: 10.1038/s41587-024-02420-y.

[80] T. Ouspenskaia *et al.*, 'Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer', *Nat Biotechnol*, vol. 40, no. 2, pp. 209–217, Feb. 2022, doi: 10.1038/s41587-021-01021-3.

[81] E. Vieira De Souza *et al.*, 'Rp3: Ribosome profiling-assisted proteogenomics improves coverage and confidence during microprotein discovery', *Nat Commun*, vol. 15, no. 1, p. 6839, Aug. 2024, doi: 10.1038/s41467-024-50301-4.

[82] C.-Y. Kim *et al.*, 'FusionPro, a Versatile Proteogenomic Tool for Identification of Novel Fusion Transcripts and Their Potential Translation Products in Cancer Cells*', *Molecular & Cellular Proteomics*, vol. 18, no. 8, pp. 1651–1668, Aug. 2019, doi: 10.1074/mcp.RA119.001456.

[83] I. Abugessaisa *et al.*, 'FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs', *Nucleic Acids Research*, vol. 49, no. D1, pp. D892–D898, Jan. 2021, doi: 10.1093/nar/gkaa1054.

[84] J. M. Mudge *et al.*, 'GENCODE 2025: reference gene annotation for human and mouse', *Nucleic Acids Research*, vol. 53, no. D1, pp. D966–D975, Jan. 2025, doi: 10.1093/nar/gkae1078.

[85] G. Pertea and M. Pertea, 'GFF Utilities: GffRead and GffCompare', *F1000Res*, vol. 9, p. 304, Sep. 2020, doi: 10.12688/f1000research.23297.2.

[86] L. Calviello, D. Sydow, D. Harnett, and U. Ohler, 'Ribo-seQC: comprehensive analysis of cytoplasmic and organellar ribosome profiling data', Apr. 08, 2019, *bioRxiv*. doi: 10.1101/601468.

[87] M. I. Love, W. Huber, and S. Anders, 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biol*, vol. 15, no. 12, p. 550, Dec. 2014, doi: 10.1186/s13059-014-0550-8.

[88] S. Anders, A. Reyes, and W. Huber, 'Detecting differential usage of exons from RNA-seq data', *Genome Res.*, vol. 22, no. 10, pp. 2008–2017, Oct. 2012, doi: 10.1101/gr.133744.111.

[89] A. Reyes, S. Anders, R. J. Weatheritt, T. J. Gibson, L. M. Steinmetz, and W. Huber, ‘Drift and conservation of differential exon usage across tissues in primate species’, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 110, no. 38, pp. 15377–15382, Sep. 2013, doi: 10.1073/pnas.1307202110.

[90] F. Da Veiga Leprevost *et al.*, ‘Philosopher: a versatile toolkit for shotgun proteomics data analysis’, *Nat Methods*, vol. 17, no. 9, pp. 869–870, Sep. 2020, doi: 10.1038/s41592-020-0912-y.

[91] F. Yu *et al.*, ‘Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform’, *Nat Commun*, vol. 14, no. 1, p. 4154, Jul. 2023, doi: 10.1038/s41467-023-39869-5.

[92] A. T. Kong, F. V. Leprevost, D. M. Avtonomov, D. Mellacheruvu, and A. I. Nesvizhskii, ‘MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics’, *Nat Methods*, vol. 14, no. 5, pp. 513–520, May 2017, doi: 10.1038/nmeth.4256.

[93] K. L. Yang *et al.*, ‘MSBooster: improving peptide identification rates using deep learning-based features’, *Nat Commun*, vol. 14, no. 1, p. 4539, Jul. 2023, doi: 10.1038/s41467-023-40129-9.

[94] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss, ‘Semi-supervised learning for peptide identification from shotgun proteomics datasets’, *Nat Methods*, vol. 4, no. 11, pp. 923–925, Nov. 2007, doi: 10.1038/nmeth1113.

[95] F. Yu, S. E. Haynes, and A. I. Nesvizhskii, ‘IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs’, *Molecular & Cellular Proteomics*, vol. 20, p. 100077, 2021, doi: 10.1016/j.mcpro.2021.100077.

[96] H.-Y. Chang *et al.*, ‘Analysis of isobaric quantitative proteomic data using TMT-Integrator and FragPipe computational platform’, *Nat Commun*, Mar. 2026, doi: 10.1038/s41467-026-70118-7.

[97] C.-C. Tsou *et al.*, ‘DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics’, *Nat Methods*, vol. 12, no. 3, pp. 258–264, Mar. 2015, doi: 10.1038/nmeth.3255.

[98] M. E. Ritchie *et al.*, ‘limma powers differential expression analyses for RNA-sequencing and microarray studies’, *Nucleic Acids Res*, vol. 43, no. 7, p. e47, Apr. 2015, doi: 10.1093/nar/gkv007.

[99] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, 'Nextflow enables reproducible computational workflows', *Nat Biotechnol*, vol. 35, no. 4, pp. 316–319, Apr. 2017, doi: 10.1038/nbt.3820.

[100] L. H. Wang *et al.*, 'MRI-informed muscle biopsies correlate MRI with pathology and DUX4 target gene expression in FSHD', *Human Molecular Genetics*, vol. 28, no. 3, pp. 476–486, Feb. 2019, doi: 10.1093/hmg/ddy364.

[101] ENCODE Project Consortium, 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012, doi: 10.1038/nature11247.

[102] C. W. Yip *et al.*, 'CFC-seq: identification of full-length capped RNAs unveil enhancer-derived transcription', Oct. 31, 2024, *bioRxiv*. doi: 10.1101/2024.10.31.620483.

[103] F. J. Pardo-Palacios *et al.*, 'Systematic assessment of long-read RNA-seq methods for transcript identification and quantification', *Nat Methods*, vol. 21, no. 7, pp. 1349–1363, Jul. 2024, doi: 10.1038/s41592-024-02298-3.

[104] E. W. Deutsch *et al.*, 'High-quality peptide evidence for annotating non-canonical open reading frames as human proteins', Sep. 09, 2024, *bioRxiv*. doi: 10.1101/2024.09.09.612016.

[105] K. Liu, Y. Ye, S. Li, and H. Tang, 'Accurate de novo peptide sequencing using fully convolutional neural networks', *Nat Commun*, vol. 14, no. 1, p. 7974, Dec. 2023, doi: 10.1038/s41467-023-43010-x.

[106] W. Huber *et al.*, 'Orchestrating high-throughput genomic analysis with Bioconductor', *Nat Methods*, vol. 12, no. 2, pp. 115–121, Feb. 2015, doi: 10.1038/nmeth.3252.

[107] J. T. Robinson *et al.*, 'Integrative genomics viewer', *Nat Biotechnol*, vol. 29, no. 1, pp. 24–26, Jan. 2011, doi: 10.1038/nbt.1754.

[108] D. Wyman *et al.*, 'A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification', Jun. 18, 2019, *bioRxiv*. doi: 10.1101/672931.

[109] Y. Chen *et al.*, 'Context-aware transcript quantification from long-read RNA-seq data with Bambu', *Nat Methods*, vol. 20, no. 8, pp. 1187–1195, Aug. 2023, doi: 10.1038/s41592-023-01908-w.