A Multi-omics Data Resource for Frontotemporal Dementia Research

Peter Heutink, Kevin Menden, and Anupriya Dalmia

Introduction

Frontotemporal dementia (FTD) is a devastating early-onset dementia characterized by the deterioration of the frontal and temporal lobes, severe changes in social and personal behaviour and blunting of emotions [1]. Up to 40% of cases have a positive family history, and mutations in at least ten genes explain almost 50% of familial cases, and this has been the key to the remarkable progress in our understanding of the molecular basis of FTD. Among the familial cases, mutations in the microtubule-associated protein tau (*MAPT*), granulin (*GRN*) and *C9orf72* are responsible for the majority of cases [2]. Neuropathologically, mutations in *MAPT* are associated with neurofibrillary tangles consisting of hyperphosphorylated tau protein, and mutations in *GRN* and *C9orf72* lead to accumulation of the transactive response DNA-binding protein 43 kDa (TDP-43). Although all three genes are associated with a clinical FTD phenotype, their cellular functions are quite diverse, and how these different genes lead to a similar clinical phenotype is still an unanswered question. Currently, there is no cure for FTD, and for the development of successful therapies, it is essential to understand the role of all genetic and environmental risk factors in the disease process, and to investigate which factors are important in the progression of the disease in all patients and which are specific for subgroups of patients.

It is therefore of utmost importance to identify the regulatory mechanisms that lead to neurodegeneration as a consequence of the already identified mutations and novel genes that are being identified by whole-genome sequencing (WGS) and whole-exome sequencing (WES) studies and genome-wide association studies (GWAS).

Publicly available data resources such as Genotype-Tissue Expression (GTEx) (https://gtexportal.org/home/), Encyclopedia of DNA Elements (ENCODE) [3, 4] and the Functional Annotation of the Mammalian Genome (FANTOM) [5] provide excellent tools to investigate the molecular processes in which identified genes and candidate genes for FTD are involved and can help to determine the processes that regulate the expression of these genes, but an

P. Heutink (✉)
Department of Genome Biology of Neurodegenerative Diseases, German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany
e-mail: peter.heutink@dzne.de

K. Menden
Department of Genome Biology of Neurodegenerative Diseases, Deutsches Zentrum für Neurodegenerative Erkrankungen, Tübingen, Baden-Württemberg, Germany

A. Dalmia
Department of Genome Biology, Deutsches Zentrum für Neurodegenerative Erkrankungen, Tübingen, Baden-Württemberg, Germany
e-mail: anupriya.dalmia@dzne.de

B. Ghetti et al. (eds.), *Frontotemporal Dementias*, Advances in Experimental Medicine and Biology 1281, https://doi.org/10.1007/978-3-030-51140-1_16

important limitation is that all these resources have been generated from human tissues and cellular models of unaffected controls. To understand the role of identified genes in the disease situation, there is a need to generate a publicly available resource from affected cells and tissues obtained from patients and animal models. As part of the European Union (EU) Joint Programme – Neurodegenerative Diseases Research (JPND), we formed the Risk and modifying factors in FTD (RiMod-FTD) consortium with the aim to investigate common and distinctly affected processes in different groups of FTD patients, using a combination of genomic and cell biological approaches on tissues of selected patient groups and corresponding animal and cellular model systems. Our integrative approach allows an unbiased selection of the most suitable targets that can improve our understanding of disease progression and, in addition, will help identify the key genes in the disease process that are the most suitable targets to modify the disease phenotype, and thus provide better choices for therapy development. Here, we describe the current state of our resource and provide examples of how the data can be mined to understand the molecular processes associated with identified genes for FTD and help to prioritize candidate genes identified through WGS/WES and GWAS studies.

The Risk and Modifying Factors in Frontotemporal Dementia Resource

In order to generate a comprehensive multi-omics data resource, we collected frozen post-mortem brain tissue from seven regions (frontal, temporal and occipital lobes, hippocampus, cerebellum, putamen, caudate) of patients carrying mutations in the three most commonly mutated genes in FTD—*MAPT, GRN* and *C9orf72*—and controls without neurological disease for multi-omics characterization. Extensive quality control measures ensured we only included samples that provided us with high-quality ribonucleic acid (RNA), epigenetic and protein data. Because human post-mortem brain represents the disease end stage, we have also collected tissue at different time points of the development of pathology from the frontal lobes of established mouse models for the same three genes. In addition, we have used human immune pluripotent stem (iPS) lines carrying the same mutations, differentiated them into neurons and performed similar analyses. In this way, we have created a resource that can be used to mine molecular data at the end stage of disease but also during life and early differentiation. The inclusion of iPS lines provides us with the additional possibility to investigate and validate identified pathways by targeted perturbation studies with, for example, RNAi and CRISPR-Cas9 (Table 1).

To thoroughly characterize the molecular mechanisms in post-mortem human brain tissue, mouse models and induced pluripotent stem cell (iPSC)-derived neurons, we generated various omics-datasets. RNA-sequencing (RNA-seq), the most widely used omics-technology [6], allows to measure the gene expression of the entire transcriptome, and it thus represents a central dataset in the resource. Additionally, we generated Cap Analysis of Gene Expression sequencing (CAGE-seq) [7] data, which captures the 5′-end of transcripts and can thus be used to profile the transcription start site (TSS) of genes. The CAGE-seq data thus represents a complementary dataset to the RNA-seq data, as it can not only be used to measure gene expression but also to identify different TSS or promoter usage as well as enhancers [8]. The transcriptome is heavily influenced by the epigenome, for instance, by CpG methylation [9]. To assess potential epigenomic changes in FTD, and to help explain observed transcriptomic aberrations, we profiled over 800,000 CpG sites for methylation. Since for all protein-coding genes, the end-product of gene expression is a protein, we used proteomics technology to quantify the expression of thousands of proteins as an important complementary readout to the transcriptome. As both gene expression and translation are regulated, in part, by micro RNAs (miRNAs), we performed small RNA-sequencing (smRNA-seq) to identify important regulator miRNAs and potentially explain

Table 1 List of datasets that have already been generated and processed for RiMod-FTD

Post-mortem human brain tissue		
Data type	Brain region	Samples (control, MAPT, GRN, C9orf72, sporadic)
RNA-seq	Frontal	47 (16, 11, 7, 13, 0)
CAGE-seq	Frontal, temporal, caudate, hippocampus, occipital, cerebellum, and putamen	248 (66, 61, 42, 53, 24)
smRNA-seq	Frontal and temporal	87 (27, 25, 14, 21, 0)
Proteomics	Frontal and temporal	69 (16, 24, 12, 17, 0)
Methylation	Frontal	48 (14, 13, 7, 14, 0)
ChIP-seq H3K4me3	Frontal	16 (4, 4, 4, 4, 0)
ChIP-seq H3K4me3	Sorted neurons (frontal)	25 (8, 8, 3, 6, 0)
Mouse models		
Data type	**Model Mouse line**	**Samples**
CAGE-seq	MAPT-P301L rTg(TauP301L)4510	32 (control: 16, transgenic: 16)
CAGE-seq	GRN knockout $Grn^{tm1.1Pvd}$	33 (control: 17, knockout: 16)
CAGE-seq	C9orf72 knockdown C57BL/6j-Tg(C9orf72_i3)112Lutzy/J	29 (WT: 12, scramble: 9, knockdown: 8)
Proteomics	MAPT-P301L rTg(TauP301L)4510	33 (control: 16, transgenic: 17)
Proteomics	GRN knockout $Grn^{tm1.1Pvd}$	33 (control: 17, knockout: 16)
Proteomics	C9orf72 knockdown C57BL/6j-Tg(C9orf72_i3)112Lutzy/J	31 (WT: 12, scramble: 9, knockdown: 10)
iPSC-derived cells		
Data type	**Cell type**	**Samples (control, MAPT, GRN, C9orf72)**
smRNA-seq	Neurons	21 (9, 7, 4, 6)

changes observed in the transcriptome or proteome. Finally, Chromatin Immuno-Precipitation sequencing (ChIP-seq) was performed for the H3K4me3 protein to identify active promoters. All the above-mentioned genomics data types that have been generated for the RiMod-FTD resource focus on different parts of the cellular transcriptional machinery. By combining these different datasets, it is possible to generate better hypotheses about the disease-causing regulatory mechanisms or to validate existing hypotheses using multiple data modalities. A graphical overview of the datasets already generated and planned for future releases is depicted in Fig. 1.

Analysing Multi-omics Datasets

Generating a multi-omics data resource is, of course, only the first step on the path to gain new knowledge about the condition of interest. The next step is to rigorously analyse the data and/or integrate it with genetic data to generate new hypotheses about disease mechanisms. For large and complex datasets such as those found in a multi-omics data resource, there exists a plethora of bioinformatics methods that can be applied to gather new information. For conventional techniques like RNA-seq, there are several accessible and established tools. For others, the researchers might have to write new algorithms themselves. In recent years, specialized algorithms have been developed that allow the integration of multiple experiments from different technologies [10]. Combining the different datasets with the possibilities of modern bioinformatics can then lead to new insights. Moreover, having a central disease-specific data resource available is beneficial in more ways than just to create new insights based on the resource datasets alone. It depicts a valuable asset that FTD-researchers can use to better interpret their own experiments or test their hypotheses. For instance, a clinician or biologist may state a hypothesis about the involvement of a new gene in FTD pathology based on results from an experiment. Before investing more resources in further investigating the role of this gene, the researcher would like to see some more

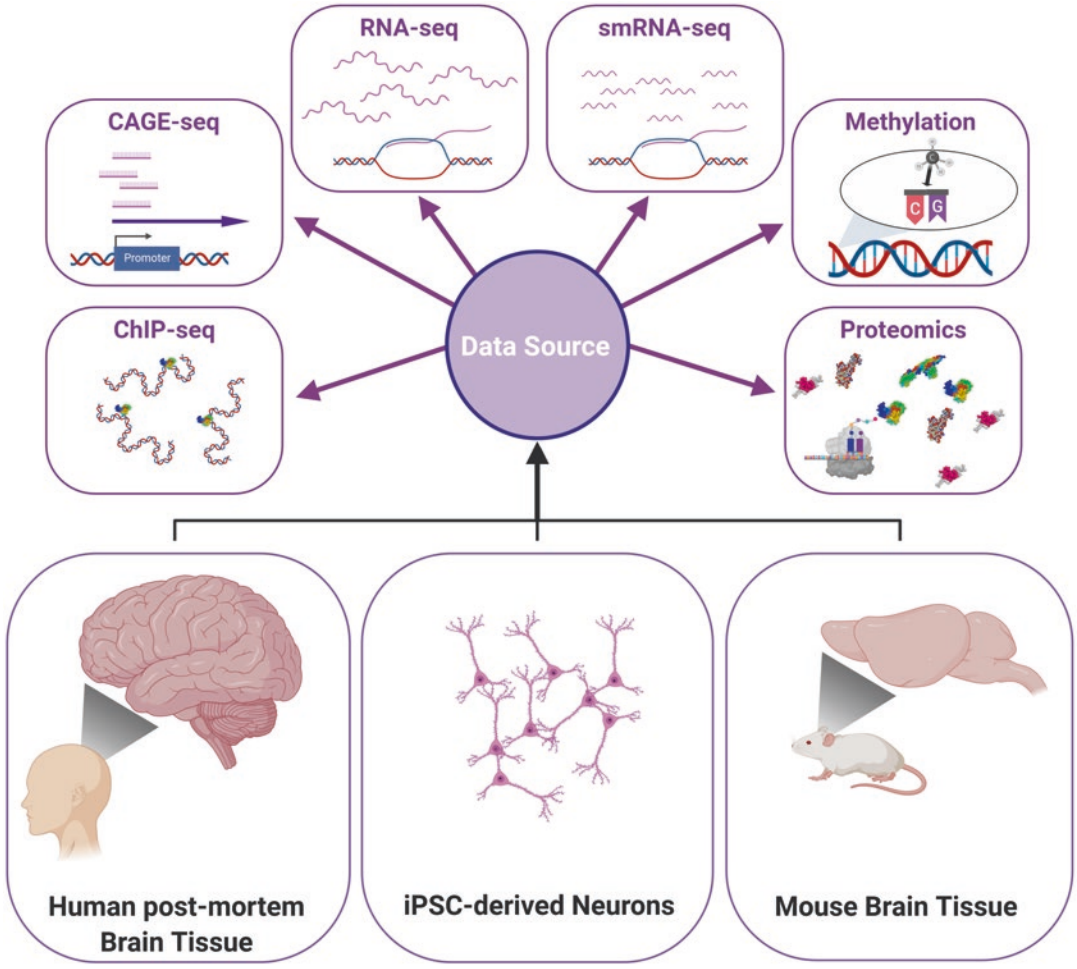

Fig. 1 The RiMod-FTD data resource consists of datasets generated from post-mortem human brain tissue, iPSC-derived neurons and brain tissue from mouse models covering FTD caused by MAPT, GRN and C9orf72. The multi-omics technologies used to generate the data cover ChIP-seq, CAGE-seq, RNA-seq, smRNA-seq, epigenetic arrays and proteomics

evidence. In such a case, RiMod-FTD allows to quickly check the transcriptional state of this gene in several FTD subtypes or whether the quantities of the protein product are changed in the disease. Additionally, the researcher could examine whether the gene is differentially methylated and, finally, check whether aberrant regulation of the gene can be observed in multiple model systems. With more datasets added to the resource in the future, the possibilities for validating experimental results will further increase. Being able to validate scientific findings from own experiments in public data is obviously of great value and helps to identify the best research paths to pursue and thus to accelerate the scientific progress. In the following, we cover the different technologies used to generate the datasets found in the resource, how these data can be analysed and, where suitable, we present some examples related to FTD.

Pre-processing

Before any dataset generated in the wet lab can be mined for interesting results, it first has to be processed and brought into a format suitable for analysis. While great efforts have been undertaken

to simplify this part of the analysis, it remains a very crucial and important step in bioinformatics. The process of converting the raw data that come, for instance, from a sequencing machine, into interpretable and biologically meaningful data points usually requires several steps, each of which is executed with a specialized algorithm. This sequence of steps is commonly called a processing or analysis pipeline. Writing such a pipeline for any omics-data type requires extensive technical knowledge about the data-generating process as well as a good understanding of bioinformatics algorithms capable of handling the respective data. All datasets in RiMod-FTD have been processed and analysed carefully and are available in raw data as well as processed data format. This makes the data more accessible for scientists without extensive domain knowledge, while preserving the raw data for any scientist who wants to process the data with a different pipeline.

Analysing the Transcriptome with Ribonucleic Acid Sequencing

The transcriptome is probably the most commonly studied 'ome' and plays a central role in many studies. Rightfully so, as regulation of gene expression underlies most cellular processes, it is aberrant in many diseases and depicts the closest readout for effects from genetic and epigenetic variation. While multiple technologies exist that can measure gene expression, RNA-seq is the most common one nowadays. Because of this, and because of the importance of the transcriptome, excellent tools exist that help to analyse RNA-seq data. Usually analysis of transcriptomic starts with identifying differentially expressed genes (DEGs) between different groups of samples. Several software packages for this purpose, called differential expression (DE) analysis, exist, such as DESeq [11] or edgeR [12], which allow to apply carefully developed statistical models to calculate fold-changes and *p*-values for every gene. Although DE analysis is a very standard approach and the above-mentioned software packages are easy to use, care must be taken by the user to specify the design matrix correctly and to account for confounding variables such as age, gender or experiment batches. The results of DE analysis constitute the basics of many downstream methods and help the experimenter to identify pathways that are most affected by a condition. Along with raw RNA-seq data, the RiMod-FTD resource contains pre-calculated fold-changes and *p*-values for the most important comparisons of the contained transcriptomic datasets. This makes it easy to quickly check the status of a specific gene in multiple FTD subgroups or model systems, without the need to first process and analyse the data.

The entire set of DEGs defined by DE analysis can be used in combination with public databases of pathways and gene sets that have been curated by experts to test for enrichment of DEGs in some of these pathways. Results from such analyses can be of great value, as they, if done correctly, immediately highlight the cellular processes different between conditions. In a recent study, Dickson et al. [13] performed RNA-sequencing on human brain samples of patients with *C9orf72* repeat expansion, patients without this mutation and control subjects. Using pathway analysis in combination with weighted gene co-expression network analysis (WGCNA), they found that vesicular transport pathways are especially affected by *C9orf72* repeat expansions. Using only transcriptomic data, the authors could highlight several affected pathways in *C9orf72* mutation carriers and identified biomarker candidate genes by applying LASSO regression. Importantly, RiMod-FTD contains datasets from patients not only with *C9orf72* but also with *GRN* and *MAPT* mutations, and it thus allows to test for commonalities between the disease subgroups in terms of affected pathways or WGCNA modules. For example, analysing the RNA-seq data from the RiMod-FTD resource, we have found that oxidative phosphorylation is impaired in both FTD-*GRN* and FTD-*MAPT*. However, membrane-trafficking-associated pathways appear to be strongly down-regulated in FTD-*MAPT*, while FTD-*GRN* shows a stronger enrichment for immune system–related pathways. Moreover, as lists of affected pathways are available in the resource, a scientist with an interest in a specific pathway can quickly investigate

whether this pathway is affected in some FTD subtype or model system.

Complex tissue, like post-mortem brain tissue, consists of several transcriptionally different cell types. When interpreting RNA-seq experiments on such tissues, it is important to keep in mind that systematic differences in cell-type compositions between sample groups can lead to false-positive DEGs in the analysis. To account for this problem, several cell deconvolution methods have been developed that allow to estimate the cellular composition of each sample from RNA-seq data. Not only does this help to control for false positives, but it can also uncover unknown cellular composition changes in a disease. Examples for cell deconvolution algorithms are MuSiC [14] and Scaden [15]. The latter has been developed for the analysis of data from the RiMod-FTD project and showed best performance on post-mortem brain tissue when compared to other algorithms.

Co-expression Analysis

If an expression dataset is sufficiently large, gene co-expression analysis can be used to obtain dataset-specific expression modules that are relevant to the disease. WGCNA, which was mentioned earlier, is the most popular algorithm for this task [16]. Briefly, WGCNA calculates co-expression values of genes across a dataset, which can then be used to cluster genes into co-expression modules. The underlying assumption is that genes with similar expression patterns tend to have similar functions or are involved in overlapping regulatory mechanisms. A module eigengene, which is the first principal component of the expression matrix, can be used to associate traits with modules—which allows to identify disease-associated modules. Other, module-internal metrics calculated by WGCNA help to identify module hub-genes that might be of special importance. In the study mentioned earlier by Dickson et al., WGCNA was used to identify co-expression modules that are associated with the *C9orf72* repeat expansion. Through module analysis, they identified a module that contained the gene *C9orf72* and was enriched for metabolic pathways, indicating that *C9orf72* might have a similar function or affect these pathways. Another study from Swarup and colleagues [17] performed WGCNA on RNA-seq data from brain tissue of mouse models for *MAPT* and *GRN* mutations. The authors identified two modules that are significantly correlated with tau hyperphosphorylation, a marker of disease progression in FTD and Alzheimer's disease (AD) [18]. By further analysing these modules, they were able to highlight multiple genes with potentially important roles in the pathways represented by the modules. These studies show how valuable information can be extracted from transcriptomic data alone using pathway- and module-based approaches. A great advantage of RiMod-FTD is the availability of transcriptomics datasets from several tissues and model systems. This allows us to evaluate the robustness of co-expression modules—which are often to some extent dataset-specific—longitudinally and across different model systems. Furthermore, modules or pathways that a researcher has identified in their own dataset can be tested for reproducibility in the various FTD-related datasets of RiMod-FTD. We believe that lacking reproducibility of results generated with genomics technologies is a major hurdle to the scientific progress, and public resources with easily accessible datasets like RiMod-FTD are one way of addressing this problem.

Alternative Splicing of Transcripts

While it is common to perform most analyses with RNA-seq data on the gene level, it is possible to infer transcript-level information from this data as well. However, estimating transcript abundances from RNA-seq data is substantially more challenging, as the sequence of isoforms overlaps to a large part, and, consequently, most reads could be assigned to multiple transcripts. Furthermore, the downstream analysis options are currently not as rich for transcripts as for genes, since many tools (e.g. pathway databases) operate mainly on the gene level. Nevertheless, various tools for the quantification of transcripts and the detection of alternative splicing have been developed. For instance, Leafcutter and MAJIQ are two modern examples of algorithms

that can identify alternative splicing events from RNA-seq data [19, 20]. Both tools circumvent the problem of transcript quantification by focusing on exon splice junctions, and thus the exclusion of introns, instead of the inclusion of exons [19]. Although differential splicing analysis is still not routinely done with RNA-seq data, it has long been known that aberrant splicing can have devastating effects and lead to disease. For instance, the authors of MAJIQ reported differential splicing of the *CAM2K* gene in Alzheimer's disease (AD) [20]. The gene *MAPT* is another prominent example. Mutations in *MAPT* lead to a ratio change of tau isoforms, the protein product of the gene. The isoforms have different chemical properties, and the disrupted balance between them can cause disease [21]. Mutations in the genes for TDP-43 and FUS have been associated with alternative splicing in amyotrophic lateral sclerosis (ALS) [22, 23], and a mutation in the gene *PINK1* was shown to activate a cryptic splice-site in Parkinson's disease [24]. Many other mutations can cause alterations in splicing and cause disease, showing that the interrogation of differential splicing represents an important aspect of RNA-seq data analysis. The RNA-seq datasets in the RiMod-FTD resource have been analysed for alternative splicing and can be easily queried for evidence of alternative splicing of a gene of interest in a specific FTD subgroup. Transcriptomic regulation via alternative splicing is a complex mechanism that certainly has not been fully interrogated, and we hope that the diverse RNA-seq data available in RiMod-FTD can help to elucidate the role of gene isoforms in FTD.

Detecting Regulatory Mechanisms

Once deregulated cellular pathways in a disease have been identified using methods such as DE analysis, pathway enrichment or WGCNA, it is often of great interest to identify the regulatory mechanisms that drive these changes. Indeed, this depicts the major goal of many studies. Understanding the regulatory mechanisms that underlie a disease greatly helps to identify druggable targets that can be further interrogated and potentially help to develop treatments. However, the regulation of the transcriptome involves numerous players that work with and against each other, and no single assay can capture all of them. Therefore, a multi-omics approach is essential. The great advantage of RiMod-FTD is that it contains multi-omics datasets from matching samples, which measure different aspects of transcriptomic regulation. This makes it possible to identify potential regulatory mechanisms or confirm or deny hypotheses about transcriptomic regulation. In the following, we cover different modes of regulation, assays available in RiMod-FTD that can be used to understand them and bioinformatics algorithms that help to extract the desired information.

Regulation by Transcription Factors

The most well-known players in the regulation of gene expression are transcription factors (TFs), which bind to promoters and can increase or repress the expression of one or several genes. Multiple bioinformatics tools have been developed to identify candidate TFs responsible for observed expression patterns. They differ in the data that they require as input and the information they use to generate TF rankings. One method to identify active TFs is to look for enrichment of transcription factor binding sites (TFBS) in the promotor region of a set of genes compared to a background. CAGEd-oPOSSUM [25] uses user-provided CAGE-seq data to generate promoter-proximal regions, which are then scanned for TFBS enrichment. Promoters, which are often in the vicinity of the TSS, are thus frequently enriched in the region around CAGE-peaks. A different approach is taken by ChEA3, which only needs a list of genes as input [26]. The algorithm then integrates information gathered from various sources to rank TFs according to consistent evidence across information sources. As this approach only relies on a list of, for example, up-regulated genes, which can be readily inferred from RNA-seq data, it is widely applicable. Because RiMod-FTD contains both CAGE-seq and RNA-seq data, both above-discussed methods can be applied, in complementary fashion, to the data. Chromatin Immunoprecipitation sequencing (ChIP-seq) is

another technology that can be used to study regulation by TFs [27]. With ChIP-seq, the experimenter can identify DNA elements to which a protein of interest binds. As TFs bind to DNA, a ChIP-seq experiment for a particular TF will identify promoters and enhancers that are bound by the TF of interest, which can be used to identify genes regulated by these promoters. The analysis of ChIP-seq data requires specialized algorithms that discriminate between real binding sites and background signal. A very popular tool for this purpose is MACS2 [28]. Although RiMod-FTD currently does not contain ChIP-seq data for specific transcription factors, it contains H3K4me3 ChIP-seq data. H3K4me3 is associated with active promoters and can thus be used to identify active genes and TFs that potentially drive the expression (similar to CAGE-seq). In addition to RNA-seq, CAGE-seq and ChIP-seq, RiMod-FTD also contains proteomic data that can be assessed for TF quantities, which give a more direct readout than using mRNA levels as proxy. However, on a more cautious note, we want to mention that TFs are usually of low abundance in the cell and are thus not always caught by proteomics experiments [29]. It is thus important to use all available datasets for inferring relevant TFs.

Regulation by Micro-RNAs

Micro-RNAs (miRNAs) are another type of important transcriptional regulator that mainly works by binding to the 3′-end of messenger RNAs (mRNAs) to decrease the mRNA stability or to repress the rate of translation [30]. Hence, they affect both the abundance of mRNA and the rate of protein production. Because miRNAs are very short (21–25 nucleotides), specialized protocols must be used for miRNA expression profiling, which is why their activity cannot reliably be inferred from a typical RNA-seq experiment, which measures mRNA or total RNA expression. RiMod-FTD contains smRNA-seq and RNA-seq data from matched samples. This is of great value, as it allows to identify potential miRNA-target pairings with greater confidence. First, candidate targets for each miRNA are predicted, a task for which several computational tools have been developed. These algorithms incorporate knowledge about miRNA-biology, such as the seed sequence of miRNAs—which must be complementary to a region in the target gene—or evolutionary information. However, as the seed regions used for binding to targets are very small, computationally predicted targets contain high numbers of false positives [31]. Paired information of gene and miRNA expression can be used to perform correlation analysis of miRNA-target pairs [32]. The assumption here is that a negative correlation should be observed when the miRNA regulates a target candidate. If no negative correlation is observed, then either the target prediction is wrong or the regulation by the miRNA is overshadowed by other regulatory effects.

As an example for this approach, we want to highlight a study by Swarup and colleagues, where the authors used protein coding gene and miRNA expression data to identify the miRNA—miR-203—as a potential regulator for a disease-associated co-expression module in mouse models of FTD [17]. After highlighting this miRNA as a potential regulator, the authors went further and overexpressed this miRNA in mouse neuronal cell cultures, where they could observe down-regulation of the predicted targets along with increased apoptosis, thus validating their findings from the transcriptomic data. Replication of such candidate miRNAs in other datasets is important. The RiMod-FTD resource contains several datasets of matched gene- and miRNA-expression, which can be used to infer potentially important regulator miRNAs or to validate findings from other studies, such as those from Swarup et al.

Regulation by Deoxyribonucleic Acid Methylation

The methylation of DNA residues can have strong regulatory effects on gene expression. Cytosine residues can be methylated at their fifth carbon molecule, usually in the context of CpG dinucleotides [9]. CpG methylation at the promoter of genes causes transcriptional repression of that gene. Aberrant methylation can therefore directly affect the transcriptome, and many human diseases have now been associated with

methylation [33]. Many technologies for measuring DNA methylation exist, of which methylation array chips are a popular method that nowadays cover over 850,000 different CpG sites across the genome. Specialized software packages have been developed to analyse this data. Like DE analysis, differentially methylated CpG sites between two conditions can be inferred. RiMod-FTD contains methylation data of the newest technology, covering over 850,000 different CpG sites. These data serve as an additional resource for identifying underlying regulatory mechanisms and can help to elucidate disease-related changes in the epigenome. As an example for the relevance of DNA methylation in FTD, repeat expansions in the *C9orf72* gene—a common cause of FTD and ALS—are associated with hypermethylation of the repeat itself and *C9orf72*-flanking CpG island [34]. Gijselinck and colleagues reported that the repeat size correlates with the degree of hypermethylation, with longer repeats leading to more methylation of the flanking CpG island [35]. Repeat size and methylation state are also correlated with age at onset, and the authors suggested that the increased methylation might be a factor explaining the differences in age at onset of the disease.

Proteomics

Being the end-product of gene expression, splicing and translation, proteins constitute the major functional molecules in the cell. Although higher gene expression generally leads to higher quantities of the protein product, the correlation of these two quantities varies significantly [36]. Measuring mRNA concentration is hence not enough to infer protein concentrations [37]. It is obvious that the interrogation of the proteome is a fundamentally important step on the path to understanding cellular pathways and diseases that complement transcriptomic and epigenomic profiling. While the mature RNA-seq technology can be readily used to measure the expression of the entire transcriptome, quantification of the proteome depicts a more difficult challenge. The current technology works by digesting proteins into smaller peptides, which are subsequently measured by lipid chromatography (LC) and mass spectrography (MS). Bioinformatic algorithms are then employed, in combination with databases, to translate the quantified peptides into protein-level information [38]. Like gene expression, differences of protein quantities between conditions can then be assessed. In addition to the transcriptomic and regulatory assays, RiMod-FTD contains several proteomics datasets from diverse resources, such as multiple brain tissues, patients with different causal mutations or different mouse models. While these datasets cannot cover the entire transcriptome, they represent valuable complementary measurements that help to examine how transcriptional aberrances translate into the proteome. As proteomics experiments are less often conducted than RNA-seq experiments, we believe that the proteomics datasets of RiMod-FTD will be of especially high value for scientists working in the field.

Advantages of Multi-Model Approaches

As shown earlier, the use of multiple omics technologies to profile a biological system and to understand a disease is of great value. It allows us to study several, albeit not all, parts of the highly interconnected regulatory machine that is the cell and is therefore indispensable for widening the systems-level understanding. However, most diseases, especially neurodegenerative diseases such as FTD, arise through complex mechanisms that lead from disease onset to the final disease stages. Understanding these temporal pathway activity patterns and interactions is essential for a complete understanding of a disease, and most probably necessary to eventually develop remedies. To study neurodegeneration, brain tissue is often used—which is only available post-mortem (with some exceptions) and therefore represents the very end stage of the disease. Especially for diseases that develop over many years, only examining the end stage will not allow us to fully understand how the disease develops. It is therefore crucial to use a multi-model approach to study a complex disease like FTD. For instance, mouse models of neurodegeneration allow to

profile the disease development over different temporal stages [39]. Of course, other ramifications exist for these models, as findings in mice rarely entirely translate to humans, and a mouse disease model never completely recapitulates the actual disease [40]. Nevertheless, they depict a valuable complementary model to human post-mortem brain tissue. To increase the value of using mouse models, modern machine learning–based approaches have been developed that help to translate the findings from mice to humans [41].

A further level of complexity arises when considering the complex multicellular nature of both human and mouse brain tissue. While many cell types are typically affected in neurodegenerative diseases, the dysregulated pathways likely differ from type to type. This has been increasingly recognized in recent years. As an example, microglia have been identified as being a major factor in the development of AD [42]. In addition to tissue-level models, studying specific cell types is therefore necessary to understand the causal mechanisms behind the development of neurodegenerative diseases. In the past decade, several methods have been developed that made it possible to differentiate patient-derived induced pluripotent stem cells (iPSCs) into all the major cell types found in the brain [43]. This makes it possible to study the effects of the patient-specific genetic background on specific cell types, for instance, neurons. IPSC-derived neurons thus represent a valuable approach to study cell type–specific effects under controlled conditions that cannot be examined in complex tissues. Zhang and colleagues differentiated iPSCs derived from a patient with a mutation in the FTD-causing *CHMP2B* gene into cortical neurons, which allowed them to study neuronal-specific effects of this mutation [44]. The authors identified abnormalities in endosomes and mitochondria as the most significant alterations caused by this mutation, providing insights into the causal mechanisms of *CHMP2B* mutations in neurons. The authors of a different study used iPSC-derived neurons from a patient with *MAPT* mutation and identified transcriptional changes of GABA receptor genes, which they verified in other data from mouse modes and human brain tissue [45]. These results show how iPSC-derived neurons can be used to study neuron-specific disease mechanisms that are directly caused by a genetic alteration.

The consideration of the above-mentioned advantages and disadvantages of different model systems and tissues led to the decision to make RiMod-FTD a disease-specific data resource that contains datasets from multiple model systems. Having these multi-model datasets facilitates the discovery of mechanisms that translate from model to model, or tissue to model and enables to derive much more robust hypotheses.

Genetics Analysis

Even though almost 40% of patients with FTD have a positive family history, there exists a large gap of missing heritability to explain close to half of these cases, with the rest carrying mutations in known FTD genes such as *MAPT, GRN* and *C9orf72* [2]. With a massive influx of advancement in genetic methodologies in the past two decades, the scope to identify and study disease-causing mutations has amplified and goes beyond linkage analysis and candidate gene studies. The human genome has 100 million single-nucleotide polymorphisms (SNPs) identified to date, which can quickly and cost-effectively be genotyped using arrays. Genome-wide association studies (GWAS) are a classic example of using genotyped data to compare SNPs between healthy and diseased individuals. Strides in next-generation sequencing have also helped identify novel genetic factors and rare damaging variants implicated in FTD.

Genome-Wide Association Studies

A GWAS is based on the concept of linkage disequilibrium, which allows for a subset of SNPs to be used as proxies to genotype the entire genome. It relies on the 'common variants' theory to identify risk factors with modest effect and, in turn, risk loci in the genome that may be used to identify genes that can be clumped together to confirm pathways and processes relevant to that disease [46]. In the largest FTD-GWAS cohort, to

date, alterations in the immune system, lysosomal and autophagic pathways were identified as associated to FTD risk [47]. Since GWASs rely on finding SNPs with moderate effects, it is important to have large cohorts to be able to achieve enough statistical power to see a true biological effect. This study included a two-stage GWAS (discovery phase and replication phase) for clinical FTD, utilizing samples from 44 international research groups. The most widely used tool for GWAS is PLINK [48, 49].

As a follow-up, they performed expression and methylation quantitative loci analysis to study their effect on the associated SNPs. These types of analyses are frequently clubbed together to help discriminate causation from association as it is an important point of note that while proxy SNPs are associated with traits, they are seldom causative. The RiMod-FTD resource of multi-omic data from different brain regions of FTD patients can be useful in mining the hits found in such large-scale GWAS studies and understand the biology lying underneath the association.

For example, a recent GWAS study, shows that the rs72824905-G allele in the gene PLCG2 is associated with decreased risk in FTD as well as increased changes of longevity [50]. Following up on this finding using the RiMod-FTD RNA-seq data, we found that PLCG2 is up-regulated in patients carrying a *GRN* mutation. Loss of *GRN* function has been associated with elevated microglial neuroinflammation [51]; this finding may lend evidence to the protective effect of PLCG2 in brain immune function.

To verify this link between genes involved in brain immune function analysis and FTD and the mechanism by which they act, integrative analysis involving the results from the different omics data under the RiMod-FTD resource can help utilize the plethora of information that all of these different techniques shed a light on.

Next-Generation Sequencing

Identification of rare variants that play a role in disease progression cannot be accomplished with GWA studies that rely on the 'common variants theory'. Association of rare variants with patient status can be assessed using burden tests using the SNP-set (Sequence) Kernel Association Test (SKAT) [52]. Such tests collapse variants into genetic scores and are extremely powerful at detecting high-impact variants that are causal in the same direction. Other tests that have been used are variance tests and combined variance tests that combine burden and variance tests. These tests rely on estimating the variance of genetic effects to uncover the missing heritability. PLINK can be used to perform all of these different types of tests to elucidate the effects of rare variants in FTD, which are often of higher impact than common variants.

In the FTLD-TDP whole-genome sequencing consortium [53], WGS data from 517 unrelated patients and 838 controls were used as a discovery cohort to perform a gene-level analysis of rare variants. The authors used gene-burden analyses to prioritize 61 genes in which LOF variants were observed in at least three patients. TBK1 showed the most LOF mutation carriers, along with genes involved in the TBK1-immunity pathway. TBK1 LOF mutations are also third most frequent in the Belgian FTD cohort from the BELNEU Consortium [54], after *C9orf72* and *GRN*. While this association has been confirmed by multiple studies, the mechanisms are yet to be confirmed. Using RNA-seq and CAGE-seq data from the RiMod-FTD resource, pathway and gene-set enrichment analysis can be performed to explain the mechanism in which TBK1 mutations implicate patient status for FTD. Interestingly, TBK1, unlike PLCG2 was down-regulated in patients carrying a *GRN* mutation in the RiMod-FTD RNA-seq data. These findings offer an opportunity at a deeper understanding at the mechanism behind these correlations and the potential to uncover therapeutic targets.

Public Resource

The primary goal of RiMod-FTD is to generate a versatile data resource that can help to accelerate and support the field of FTD research. To this end, all datasets generated during the project, accompanied by useful analysis results, are made available at the European Genome-phenome

Archive (EGA) [55]. Additional to making the data available in the central and well-known database EGA, it is our plan to develop a graphical user interface that facilitates to visually inspect the data directly in the browser, without any need to download it or analyse it. This will make RiMod-FTD further accessible, especially for scientists or clinicians who only want to check the expression of a single gene or pathway.

Concluding Remarks and Outlook

An ongoing effort of RiMod-FTD is to increase the number of diverse and useful datasets over time. In addition to completing the set of currently used multi-omics experiments for all tissues and model systems available, other experiments are planned as well. We aim to extend human post-mortem brain samples and mouse models to additional mutations, sporadic cases and spectrum disorders such as progressive supranuclear palsy (PSP) and amyotrophic lateral sclerosis (ALS). We also aim to extend over brain regions to be able to compare strongly affected regions with relatively preserved regions. The development of single-cell approaches and spatial transcriptomics has enabled us to examine changes at single-cell resolution, which is necessary to disentangle the cell-type-specific transcriptomic changes. Adding single-cell experiments to RiMod-FTD will therefore increase the value of the resource. Complementary to single-cell RNA-sequencing (scRNA-seq) approaches, we aim to differentiate patient-derived iPSCs into different relevant cell types, such as microglia and co-cultures. This will be done for additional mutations as well.

With these planned efforts and the already existing data, we hope to further untangle the cellular mechanisms behind the complex disease FTD and believe that the RiMod-FTD resource constitutes a significant contribution to the field of FTD research that will help to accelerate the scientific progress towards better disease understanding, diagnosis and eventually treatment.

Acknowledgements This study was supported, in part, by RiMod-FTD an EU Joint Programme – Neurodegenerative Disease Research (JPND) to PH, KM; BMBF Integrative Data Semantics for Neurodegenerative research (IDSN) to PH and the DZNE and NOMIS Foundation to PH, KM.

References

1. Rohrer JD, Guerreiro R, Vandrovcova J et al (2009) The heritability and genetics of frontotemporal lobar degeneration. Neurology 73(18):1451–1456
2. Bang J, Spina S, Miller BL (2015) Frontotemporal dementia. Lancet 386(10004):1672–1682
3. Davis CA, Hitz BC, Sloan CA et al (2018) The Encyclopedia of DNA elements (ENCODE): Data portal update. Nucleic Acids Res 46(D1):D794–D801
4. Dunham I, Kundaje A, Aldred SF et al (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74
5. Lizio M, Harshbarger J, Shimoji H et al (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol 16(1):22
6. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet
7. Shiraki T, Kondo S, Katayama S et al (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA 100(26):15776–15781
8. Andersson R, Gebhard C, Miguel-Escalada I et al (2014) An atlas of active enhancers across human cell types and tissues. Nature 507(7493):455–461
9. Greenberg MVC, Bourc'his D (2019) The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol. Nature Publishing Group 20:590–607
10. Argelaguet R, Velten B, Arnol D et al (2018) Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. Mol Sys Biol 14(6):e8124
11. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550
12. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England), 26(1):139–140
13. Dickson DW, Baker MC, Jackson JL et al (2019) Extensive transcriptomic study emphasizes importance of vesicular transport in C9orf72 expansion carriers. Acta Neuropathol Commun 7(1):150
14. Wang X, Park J, Susztak K et al (2019) Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun 10:1):1–1):9

15. Menden K, Marouf M, Dalmia A et al (2019) Deep-learning-based cell composition analysis from tissue expression profiles. bioRxiv 659227
16. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:1
17. Swarup V, Hinz FI, Rexach JE et al (2019) Identification of evolutionarily conserved gene networks mediating neurodegenerative dementia. Nat Med 25(1):152–164
18. Rademakers R, Cruts M, Van Broeckhoven C (2004) The role of tau (MAPT) in frontotemporal dementia and related tauopathies. Hum Mutat 24:277–295
19. Li YI, Knowles DA, Humphrey J et al (2018) Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet 50(1):151–158
20. Vaquero-Garcia J, Barrera A, Gazzara MR et al (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. elife 5(February):e11752
21. Buée L, Bussière T, Buée-Scherrer V et al (2000) Tau protein isoforms, phosphorylation and role in neurodegenerative disorders. Brain Res Rev. Elsevier B.V 33:95–130
22. Arnold ES, Ling SC, Huelga SC et al (2013) ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. Proc Natl Acad Sci USA 110(8):E736–E745
23. Sun S, Ling SC, Qiu J et al (2015) ALS-causative mutations in FUS/TLS confer gain and loss of function by altered association with SMN and U1-snRNP. Nat Commun 6:6171
24. Samaranch L, Lorenzo-Betancor O, Arbelo JM et al (2010) PINK1-linked parkinsonism is associated with Lewy body pathology. Brain 133(4):1128–1142
25. Arenillas DJ, Forrest ARR, Kawaji H et al (2016) CAGEd-oPOSSUM: motif enrichment analysis from CAGE-derived TSSs. Bioinformatics 32(18):2858–2860
26. Keenan AB, Torre D, Lachmann A et al (2019) ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res 47(W1):W212–W224
27. Johnson DS, Mortazavi A, Myers RM et al (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science (80-) 316(5830): 1497–1502
28. Gaspar JM 2018 Improved peak-calling with MACS2. bioRxiv 496521
29. Ding C, Chan DW, Liu W et al (2013) Proteome-wide profiling of activated transcription factors with a concatenated tandem array of transcription factor response elements. Proc Natl Acad Sci USA 110(17):6771–6776
30. Hausser J, Zavolan M (2014) Identification and consequences of miRNA–target interactions — beyond repression of gene expression. Nat Rev Genet 15(9):599–612
31. Sethupathy P, Megraw M, Hatzigeorgiou AG (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. Nat Methods 3(11):881–886
32. Borgmästars E, de Weerd HA, Lubovac-Pilav Z et al (2019) miRFA: an automated pipeline for microRNA functional analysis with correlation support from TCGA and TCPA expression data in pancreatic cancer. BMC Bioinformatics 20(1):393
33. Jin Z, Liu Y (2018) DNA methylation in human diseases. Genes Dis. Chongqing yi ke da xue, di 2 lin chuang xue yuan Bing du xing gan yan yan jiu suo 5:1–8
34. Xi Z, Zhang M, Bruni AC et al (2015) The C9orf72 repeat expansion itself is methylated in ALS and FTLD patients. Acta Neuropathol 129(5):715–727
35. Gijselinck I, Van Mossevelde S, Van Der Zee J et al (2016) The C9orf72 repeat size correlates with onset age of disease, DNA methylation and transcriptional downregulation of the promoter. Mol Psychiatry 21(8):1112–1124
36. Schwanhüusser B, Busse D, Li N et al (2011) Global quantification of mammalian gene expression control. Nature 473(7347):337–342
37. Liu Y, Beyer A, Aebersold R (2016) On the dependency of cellular protein levels on mRNA abundance. Cell. Cell Press 165:535–550
38. Altelaar AFM, Munoz J, Heck AJR (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet 14:35–48
39. Trancikova A, Ramonet D, Moore DJ (2011) Genetic mouse models of neurodegenerative diseases. In: Progress in molecular biology and translational science. Elsevier B.V, Amsterdam, pp 419–482
40. Seok J, Shaw Warren H, Alex GC et al (2013) Genomic responses in mouse models poorly mimic human inflammatory diseases. Proc Natl Acad Sci USA 110(9):3507–3512
41. Normand R, Du W, Briller M et al (2018) Found in translation: a machine learning model for mouse-to-human inference. Nat Methods 15(12):1067–1073
42. McQuade A, Blurton-Jones M (2019) Microglia in Alzheimer's disease: exploring how genetics and phenotype influence risk. J Mol Biol. Academic Press 431:1805–1817
43. Penney J, Ralveniu, WT, Tsai, L (2020) Modeling Alzheimer's disease with iPSC-derived brain cells. Mol Psychiatry 25:148–167
44. Zhang Y, Schmid B, Nikolaisen NK et al (2017) Patient iPSC-derived neurons for disease modeling of frontotemporal dementia with mutation in CHMP2B. Stem Cell Rep 8(3):648–658
45. Jiang S, Wen N, Li Z et al. (2018) Integrative system biology analyses of CRISPR-edited iPSC-derived neurons and human brains reveal deficiencies of presynaptic signaling in FTLD and PSP. Transl Psychiatry 8:265
46. Ferrari R, Grassi M, Salvi E et al (2015) A genome-wide screening and SNPs-to-genes approach to identify novel genetic risk factors associated

with frontotemporal dementia. Neurobiol Aging 36(10):2904.e13–2904.e26
47. Ferrari R, Hernandez DG, Nalls MA et al (2014) Frontotemporal dementia and its subtypes: a genome-wide association study. Lancet Neurol 13(7):686–699
48. Shaun Purcell. PLINK. 2017
49. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575
50. van der Lee SJ, Conway OJ, Jansen I et al (2019) A nonsynonymous mutation in PLCG2 reduces the risk of Alzheimer's disease, dementia with Lewy bodies and frontotemporal dementia, and increases the likelihood of longevity. Acta Neuropathol 138(2): 237–250
51. Martens LH, Zhang J, Barmada SJ et al (2012) Progranulin deficiency promotes neuroinflammation and neuron loss following toxin-induced injury. J Clin Invest 122(11):3955–3959
52. Wu MC, Lee S, Cai T et al (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82–93
53. Pottier C, Ren Y, Perkerson RB et al (2019) Genome-wide analyses as part of the international FTLD-TDP whole-genome sequencing consortium reveals novel disease risk factors and increases support for immune dysfunction in FTLD. Acta Neuropathol 137(6):879–899
54. Gijselinck I, Van Mossevelde S, Van Der Zee J et al (2015) Loss of TBK1 is a frequent cause of frontotemporal dementia in a Belgian cohort. Neurology 85(24):2116–2125
55. Lappalainen I, Almeida-King J, Kumanduri V et al (2015) The European genome-phenome archive of human data consented for biomedical research. Nat Genet. Nature Publishing Group 47:692–695