The use of systems biology to study the role of TDP-43 pathology in disease

Claire Green

Registration Number

150122122

Supervisors

Professor Winston Hide and Dr Johnathan Cooper-Knock

Department

Neuroscience

Table of Contents

List of Figures

List of Abbreviations

AD	Alzheimer's Disease
ALS	Amyotrophic Lateral Sclerosis
AxD	Alexander's Disease
C9orf72	Chromosome 9 open reading frame 72
CFTR	Cystic fibrosis transmembrane conductance regulator
CHMP2B	Charged multivesicular body protein 2B
CTE	Chronic traumatic encephalopathy
fALS	Familial amyotrophic lateral sclerosis
FTLD	Frontotemporal lobar degeneration
FUS	Fused in sarcoma
GFAP	Glial fibrillary acidic protein
GoF	Gain of function
GWAS	Genome wide association study
HIV-1	Human immunodeficiency virus type 1
*h*NFL	Human neurofilament light
hnRNP	Heterogeneous nuclear ribonuclearprotein
IBMPFD	Inclusion body myopathy associated with Paget disease of bone and frontotemporal dementia
iCLIP	Individual nucleotide-resolution ultraviolet cross-linking and immunoprecipitation
LBD	Lewey body dementia
LoF	Loss of function
MAPT	Microtubule-associated protein tau
NFT	Neurofibrillary tangles
PA	Pathway analysis
PGRN	Progranulin
RRM	RNA recognition motif
sALS	Sporadic amyotrophic lateral sclerosis
TARDBP	TAR DNA binding protein
TDP-43	TAR DNA binding protein 43
VCP	Valosin-containing protein

Introduction

1. TDP-43

1.1 TDP-43 Structure and Function

Transactive response DNA-binding Protein (TDP-43) is a highly conserved, ubiquitously expressed protein, named for its molecular weight of approximately 43-kDa. *TARDBP*, the gene which encodes TDP-43, is located on Chromosome 1 and can be alternatively spliced into 11 variants, however only TDP-43 proteins have been detected through western blotting (Wang et al. 2004). TDP-43 itself has three protein isoforms; two 43-kDa proteins (of which one is lacking 6 nucleotides), and one 28 kDa isoform missing exon 3 and a large section of exon 6. The roles of the two non-full length isoforms are currently unknown (Strong et al. 2007). Due to its structure, TDP-43 is considered a member of the heterogeneous nuclear ribonucleoprotein (hnRNP) family. hnRNPs are known to be multifunctional, with distinct roles in mRNA biogenesis and gene expression regulation (Chaudhury et al. 2010). As an hnRNP, TDP-43 is characterised by its ability to bind single-stranded RNA and DNA using binding domains known as RNA recognition motifs (RRM). TDP-43 possesses two such binding domains, RRM1 and RRM2. RRM1 is particularly affinitive to UG/TG dinucleotide repeats of 6 or more, and has been shown to be necessary and sufficient for nucleotide binding (Buratti & Baralle 2001). The RRMs of TDP-43 are flanked by an amino-terminal (N-terminal) domain and a glycine-rich carboxy-terminal (C-terminal) domain. The N-terminal domain has been shown to be required for formation of TDP-43 homodimers and for regulation of TDP-43's splicing activity (Zhang et al. 2013). Alternatively, the C-terminal domain of TDP-43 has been implicated in the mediation of protein-protein interactions (Buratti et al. 2005). TDP-43 has the ability to use its RRM1 and C-terminal binding domains to auto-regulate its own expression via a negative feedback loop; an ability often held by proteins whose overexpression is toxic (Ayala et al. 2011).

TDP-43 has a variety of known roles within the cell (see Figure 1). This widespread functionality is reflected by the discovery that over 6000 protein-coding genes have been shown to contain binding site sequences for TDP-43 (Polymenidou et al. 2011). Its role as a transcriptional regulator was the first to be documented. In a study investigating proteins that bind to the TAR DNA region of human immunodeficiency virus type 1 (HIV-1), expression of TDP-43 was found to cause the repression of HIV-1 transcription by interfering with assembly of transcription complexes that respond to transcription activators. (Ou et al. 1995). Since then, TDP-43 has been shown to repress expression of the mouse gene *SP-10*, by preventing enhancer-promoter interactions (Abhyankar et al. 2007). TDP-43 has also been implicated in the regulation of mRNA splicing. Whilst searching for modulators which inhibit splicing of exon 9 of the cystic fibrosis transmembrane conductance regulator (CFTR) gene, Buratti and colleagues succeeded in isolating the glycine-rich C-terminal domain of TDP-43 (Buratti et al. 2005). Through use of individual nucleotide-resolution ultraviolet cross-linking and immunoprecipitation (iCLIP), TDP-43 was later shown to have splicing interactions with mRNAs known to have important functions in the brain, including *MEF2D, CFTR, hNFL* and *FUS* (Tollervey et al. 2011). In addition, mRNAs associated with TDP-43's own regulation were shown to contain splice sites, giving further explanation for TDP-43's ability to self-regulate.

Due to its importance in numerous nuclear processes, it is unsurprising that the majority of TDP-43 protein is localised to the nucleus (with exception of the nucleolus). However, low levels of TDP-43 have also been detected in the cytoplasm. This cytoplasmic expression has been partially explained by TDP-43's relationship with the mRNA for human low molecular weight neurofilament (*hNF*L). Neurofilament proteins make up the internal cytoskeleton of neurons, subsequently *hNFL* mRNA is required throughout the cell. TDP-43 has been suggested to stabilise this mRNA whilst it is shuttled along the axon, giving rise to TDP-43's cytoplasmic localisation (Strong et al. 2007; Ayala et al. 2008). At a similar time, hnRNPs became implicated in the processing of microRNAs (miRNA) Initial investigation of TDP-43 uncovered interactions with the miRNAs *let-7b*, *miR-663*, *miR-574-5p* and *miR-558*, however further investigation implicated a more significant role within Drosha and Dicer complexes for miRNA biogenesis. The role of the Drosha complex is to crop miRNA precursors, called pri-miRNAs, into another intermediary molecule called a pre-miRNA. The suspected role of TDP-43 is the recruitment of pri-mRNAs into the Drosha complex. Once created, the pre-miRNAs are then exported out of the nucleus into the cytoplasm where the Dicer complex initiates their conversion into fully mature miRNAs. TDP-43 has also been shown to bind to pre-miRNAs during this maturation, suggesting a role within the Dicer complex process (Buratti et al. 2010; Kawahara & Mieda-Sato 2012). As Dicer processing of pre-miRNAs occurs in the cytoplasm, this further explains TDP-43's presence outside of the nucleus.

The downstream effects of TDP-43 activity have been illustrated in several cellular processes. TDP-43 has been shown to regulate protein quality control during cellular stress by disinhibition of the transcription factor FOXO, allowing clearance of misfolded proteins (Zhang et al. 2014). As is suggested by its high concentration in brain tissue, TDP-43 also has neuron-specific roles, including regulation of axon growth and neuronal plasticity (Tripathi et al. 2014; Wang et al. 2008).

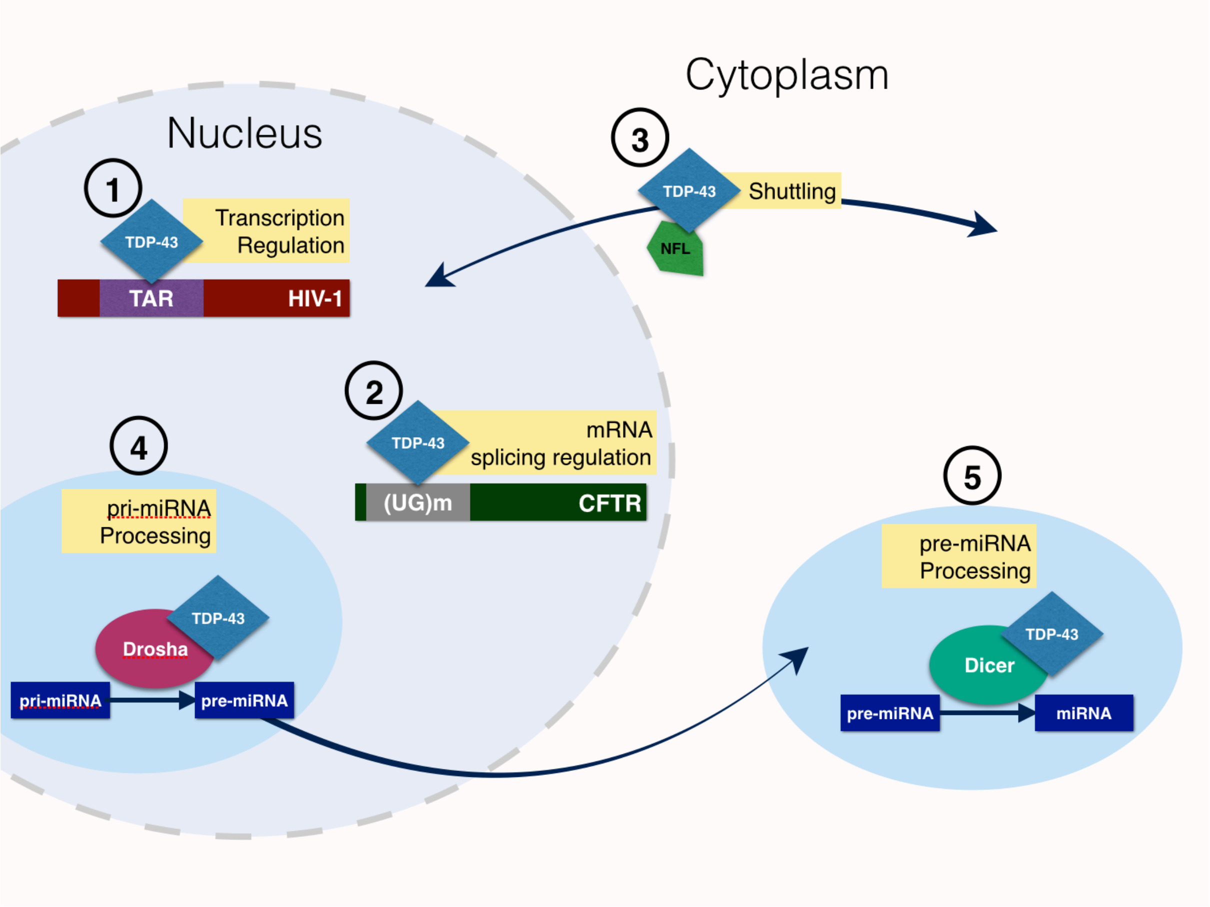

Figure 1: A diagram depicting the roles of TDP-43 within a healthy cell. **1** TDP-43 is known to regulate transcription, including HIV-1 DNA. HIV1 contains a TAR region to which TDP-43 binds and represses transcription. **2** TDP-43 engages in splicing regulation, including suppression of CFTR exon 9 splicing by binding to the 3' repeated UG sequence. **3** TDP-43 stabilises hNFL protein whilst it is shuttled from the nucleus. **4** within the nucleus, TDP-43 aids the Drosha complex in the conversion of pri-miRNAs into pri-mRNAs. **5** Within the cytoplasm, TDP-43 aids the Dicer complex in the conversion of pre-miRNAs into miRNAs.

1.2 TDP-43-associated diseases

Despite all that has been uncovered about the activity of TDP-43, it is likely that there are many more cellular processes to which it contributes. Research into TDP-43 function has increased rapidly over the last decade due to the discovery that aggregated TDP-43 protein is present in the neuronal cells of multiple neurodegenerative diseases. TDP-43 aggregations were first identified during investigation of the contents of ubiquitinated inclusions in post-mortem brain cells of frontotemporal lobar degeneration (FTLD) patients (Neumann et al. 2006). Two unknown proteins of 24kDa and 26kDa were discovered within these aggregations, both of which turned out to be truncated portions of TDP-43's C-terminus. In addition, anti-TDP-43 antibodies identified a 45kDa protein, which was later shown to be hyperphosphorylated full-length protein. Since this discovery, TDP-43 pathology has been identified in the post-mortem tissue of multiple diseases, all linked by the common phenotype of neurodegeneration. Here, we provide a summary of 6 diseases associated with TDP-43 pathology. For a full list, see Table 1.

1.2.1 Amyotrophic Lateral Sclerosis

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease characterised by the atrophy of both upper and lower motor neurons. Symptoms of ALS include muscle weakness, paralysis, and wastage. Those suffering from ALS have a life expectancy of 2-5 years post-diagnosis, and as of yet there is only one treatment option, Riluzole, which achieves a modest 2 month increase in survival (Miller et al. 2009).

ALS can be subdivided into two groups; sporadic (sALS) or familial (fALS). Those with sALS have no family history of the disease, and constitute 95% of ALS cases (Byrne et al. 2011). The remaining 5% are patients with fALS, who tend to acquire disease through inheritance of a mutated autosomal dominant gene. These genes include *C9orf72*, *SOD1*, *FUS*, *CHMP2B* and *TARDBP*. It is believed that approximately 97% of all ALS cases present with TDP-43 proteinopathy, with the exception of two groups; those with mutations in the *superoxide dismutase 1* (*SOD1*) gene, or in the *fused in sarcoma* (*FUS*) gene. For this subset of ALS patients, their pathology is classified by the SOD1 and FUS proteins themselves (Mackenzie et al. 2007; Vance et al. 2009).

1.2.2 Frontotemporal Lobar Degeneration

FTLD is characterised by atrophy of the frontal and temporal lobes, resulting in symptoms such as unsocial behaviour, aphasia, and dementia. There are three subtypes of FTLD, of which TDP-43 pathology is observed in one – ubiquitin-positive, tau-negative FTLD (FTLD-U). Like ALS, FTLD-U can also be familial or sporadic, and approximately 50% of all FTLD patients exhibit TDP-43 pathology. Mutated genes associated with TDP-43 pathology in FTLD-U include *C9orf72*, *PGRN,* *TARDBP* and *VCP*. Interestingly, *CHMP2B* mutations do not cause TDP-43 pathology in FTLD, but do in ALS (Cairns et al. 2007).

1.2.3 Multisystem Proteinopathy

Multisystem proteinopathy (MSP) is one of a group of inclusion body myopathies, and is characterised by progressive muscle weakness of both proximal and distal muscles, Paget's disease of bone, and frontotemporal dementia. Six mutations in *VCP*, or *valosin-containing protein*, have been associated with the disease, and account for nearly 50% of affected families (Watts et al. 2004). 45% of MSP patients, however, are not associated with a disease-causing mutation (Le Ber et al. 2014). TDP-43 pathology is present as both intranuclear and cytoplasmic inclusions, and unlike other neurodegenerative diseases, is present in muscle tissue.

1.2.4 Alexander's Disease

Alexander's disease (AxD) is a rare neurodegenerative disease that predominantly affects the white matter of the brain. Onset can occur at any point in life, with the infantile form being most prevalent. Symptoms include delayed motor and intellectual development, macrocephaly, and seizures. In 2001, Brenner and colleagues discovered 95% of AxD patients had a mutation in the gene encoding *glial fibrillary acidic protein* (*GFAP*), a protein previously identified in astrocytic inclusions known as Rosenthal fibers (Brenner et al. 2001). Recently, investigation of other proteins aggregated in AxD revealed the presence of TDP-43 protein (Walker et al. 2014). Interestingly, the TDP-43 pathology was not present in the neurons, but was in fact present within the astrocytes, suggesting that TDP-43 aggregation may have effects spanning further than just neurons.

1.2.5 Alzheimer's Disease and Lewy Body Dementia

Alzheimer's disease (AD) is the most common cause of dementia, and its neuropathology is defined by global cerebral atrophy as well as degeneration of some midbrain areas. Other than memory loss, symptoms can include personality changes, aggression and delusions. Only a very small proportion of AD cases can be attributed to genetics, despite over 20 genetic loci being associated with the disease (Lambert et al. 2013). Histologically, cells from patients with AD show two distinct proteinopathies; amyloid plaques and neurofibrillary tangles (NFT). In approximately one third of AD cases, TDP-43 pathology has been shown to localize within the NFTs, predominantly in the amygdala, hippocampus and dentate gyrus (Higashi et al. 2007).

Lewy body dementia (LBD) is sometimes considered a hybrid of AD and Parkinson's Disease (PD), due to the presence of AD psychological symptomology and Lewy bodies. Lewy bodies are aggregations of the proteins alpha-synuclein and ubiquitin, and in 45% of cases, TDP-43 (Higashi et al. 2007). As in AD, inclusions in LBD were localized to midbrain areas but not the cortex, which contrasts with pathology identified in other TDP-43 proteinopathies, such as FTLD and ALS.

1.2.6 Chronic Traumatic Encephalopathy

Chromic traumatic encephalopathy (CTE) is neurodegeneration caused by repeated trauma to the brain. It has been documented in dozens of contact sports, as well as in ex-military personnel (McKee et al. 2013). CTE can only be diagnosed post-humously, but retroactive associations have suggested that symptoms include mood changes, cognitive dysfunction, and ataxias. Neuropathy includes aggregation of various proteins, including tau NFTs, beta-amyloid plaques, and importantly, TDP-43. In CTE, TDP-43 pathology is evident in both cortical and subcortical structures, as well as in lower motor neurons of the spinal tract (McKee et al. 2010). The fact that TDP-43 pathology can be caused by external traumatic events has led to stronger implications that a stress response dysfunction is likely to play a large role in the formation of TDP-43 inclusions.

1.3 TDP-43 dysfunction

Since TDP-43 pathology is observed in so many devastating and often untreatable neurological diseases, understanding the processes that lead to its dysfunction has become imperative to the search for effective drug therapies. In recent years, the molecular processes leading this dysfunction have begun to emerge (Figure 2). During episodes of cellular stress (such as disease), TDP-43 is ubiquitinated and phosphorylated at two serine sites, 409 and 410. It is then cleaved by the enzyme caspase, producing the C-terminal fragments (CTFs) as observed by Neumann and colleagues (Zhang et al. 2009). The caspase family are a group of endoproteases that have been shown to have significant roles in apoptosis, and have been associated with several diseases such as Alzheimer's disease, inflammatory disease, and a variety of different cancers (McIlwain et al. 2013). Once cleaved, the TDP-43 fragments, as well as full-length TDP-43, are exported from the nucleus into the cytoplasm, a process recently shown to be regulated by the activity of the protein kinase AMPK (Liu et al. 2015). This process results in a significant mislocalisation of the predominantly nuclear TDP-43 protein (Barmada et al. 2010).

Once in the cytoplasm, full-length TDP-43 is recruited into stress granules (SGs), whereas TDP-43 CTFs appear to aggregate separately, likely due to a lack of the RRM1 domain (Colombrita et al. 2009; Yang et al. 2010). SGs are protective ribonucleoprotein compartments formed to silence and protect partially translated housekeeping mRNAs during episodes of cellular stress. Other than mRNA, they often also contain proteins involved in mRNA stabilisation, a possible explanation for the presence of TDP-43. SGs are usually reversible, returning to normal once the stress has subsided. However if this stress does not subside, SGs can form permanent aggregations, such as the inclusions seen with TDP-43. The difference between reversible and irreversible SGs has recently been attributed to the differing phase-states of RNA-binding proteins, such as TDP-43 (Murakami et al. 2015). Normally, such proteins undergo a reversible phase transition into liquid or hydrogel states during SG formation. It is hypothesised that in the presence of mutations, the proteins instead transition into an irreversible, fibrillar state, leading to permanent protein aggregation.

Though aggregation and mislocalisation of TDP-43 has been strongly associated with multiple neurodegenerative diseases, its active contribution to neuronal death is still unclear. There are two main sources of information supporting a toxic role of TDP-43 in disease; firstly, mutations within *TARDBP* itself are known to be associated with ALS, and produce ALS-like symptoms in transgenic mice (Kabashi et al. 2008; Wegorzewska et al. 2009). Secondly, mutant TDP-43 protein has been shown to correlate with increased cell death, and has prion-like properties (Barmada et al. 2010; Zhang et al. 2009; Nonaka et al. 2013).

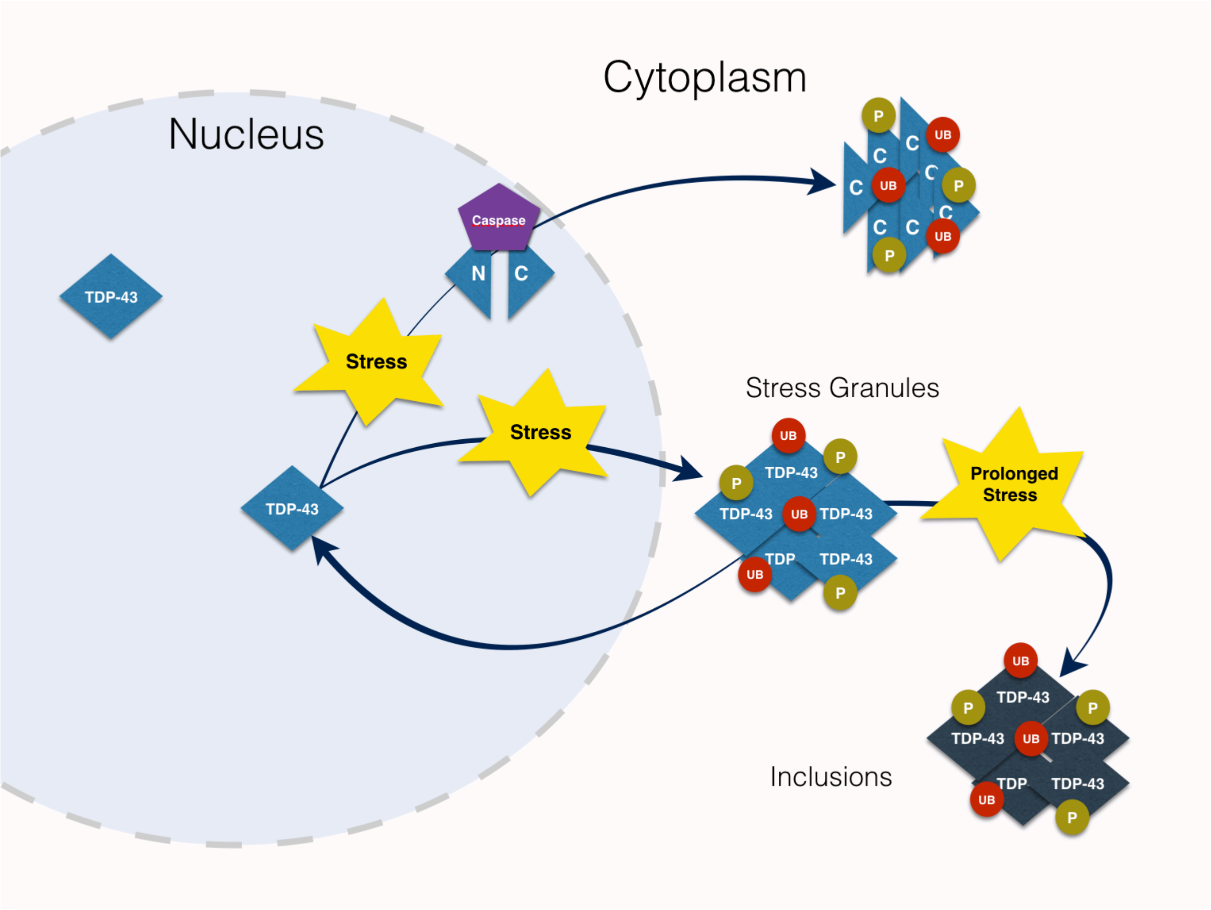

Figure 2: A diagram showing the production of TDP-43 inclusions during disease. The presence of environmental stressors or mutations causes TDP-43 to be phosphorylated and cleaved by caspase, creating N- and C-terminal fragments. The C-terminal fragments are exported into the cytoplasm and congregate into stress granules. Formation of stress granules is a reversible process, however if the stress continues, these granules form into permanent inclusions. C = C-terminal, N = N-terminal, UB = ubiquitin, P = phosphoryl

2. Approaching the mechanism of TDP-43 pathology

Since aberrant TDP-43 is suspected to contribute towards a large population of neurodegenerative diseases, it has become an extremely attractive target for therapeutic intervention. To provide such a treatment, however, requires more extensive knowledge of the mechanisms leading to such toxicity, and currently these mechanisms are unknown. The difficulty faced in addressing this problem arises from the fact that TDP-43 aggregation and mislocalisation occurs in largely variable environments. What we are considering here is a phenotype associated with multiple diseases, observed in multiple different cell types and tissues, and as the result of many different mutations. Traditional laboratory methods such as cell culture and animal models are likely to find investigation of TDP-43 extremely challenging, as the time and resources required to encapsulate the multiple contexts of TDP-43 pathology are vast. Consequently, what is required is a technique that allows the identification of a consensus molecular signature that is disrupted in all cases presenting with TDP-43 pathology

The aim of this literature review is to propose how a computational systems biology approach could be effectively recruited to identify a molecular signature associated with TDP-43 pathology, and how such an approach can provide a higher level of understanding when compared to traditional gene expression methodologies alone.

2.1 'Omic data

In the last three decades, the collection of data concerning the molecular structure and function of organisms has increased at an exponential rate. In 1990 it took a team of twenty Universities 13 years and 2.7 billion dollars to sequence the human genome. In 2016, 1000 dollars will allow an entire genome to be sequenced, and in a matter of days. Such vast improvements in technologies relating to genes and gene expression have created huge amounts of high dimensional data, and the current challenge is now to interpret it. In the field of computational biology, statistical analysis is recruited to extract meaning from such data, usually in relation to a particular phenotype. This data can come from a variety of sources, each relating to different populations of molecules, or 'omes' (Figure 3).

The genome refers to the raw sequence of DNA, and consists of approximately 3 billion base pairs. The coding regions of DNA are referred to as the exome. Sequencing of the genome can provide information on the structure of the protein each gene codes for, for example whether it contains a nuclear localisation signal or an RNA binding site, or predictions of its secondary or tertiary structure. It can also provide information on non-coding regions, such as transcription and translation regulatory regions, and sequences coding miRNAs. In disease research, genome wide association studies (GWAS) use genetic sequencing to identify high frequency single nucleotide polymorphisms (SNPs) that relate to a disease phenotype. Closely linked to the genome is the epigenome. The epigenome refers to molecules that can bind to DNA or histone proteins and modify how the genome functions. This can involve events such as methylation of promoters to silence a gene, or acetylation of histones to enhance them. Epigenomic modifications have been shown to be heritable, and are currently implicated in multiple research fields including cancer and monoallelic diseases such as Prader-Willi and Angelman's syndrome (REF).

Expression of a particular gene is usually denoted by the transcriptome, and later the translatome. The transcriptome refers to the group of mRNA molecules that have been recently transcribed from the genome. RNA can be sequenced in the same way as the genome, allowing identification of events such as post-transcriptional modifications, and splice variations. By measuring RNA expression, this gives insight into which genes are being expressed, and whether this expression changes in relation to a particular phenotype. RNA expression techniques have been employed in tens of thousands of disease investigations, and have produced field-changing results. Microarray technologies are even beginning to be employed in clinical diagnostics, although tentatively. Following the transcriptome is the translatome. The translatome is largely similar to the transcriptome, however it refers to mRNA molecules that are in the process of translation i.e. they are associated with ribosomes. Detecting molecules within the translatome has proved to be more challenging than the transcriptome, due to the complex, and therefore costly, nature of extracting mRNA from the ribosomal complexes. However, analysis of the translatome has proved it to be a more accurate representation of protein expression as it correlates more significantly than the transcriptome (REF). Subsequently, continued development in translatome detection and analysis techniques will likely result in an increase of its use in research.

At a higher 'omic level is the proteome. The proteome refers to the all proteins that are expressed within the sample. Other than a more accurate depiction of protein expression, analysis of the proteome can also indicate post-translational modifications such as phosphorylation or ubiquitination. Proteomic data can be produced from mass spectrometry or protein microarrays, however it is notoriously difficult to analyse. This is due to the fact that the proteome is constantly changing, resulting in hugely variable interpretations. As a consequence, proteomic studies are much less common than studies of the genome or transcriptome. The 'omes described so far have referred to the different phases between gene transcription and protein formation, however there are many other sources of 'omic data, many representing more abstract processes. The metabolome for instance is used to describe the small-molecule chemicals present within a sample. This can include metabolites created endogenously within the cell such as sugars and vitamins, or exogenous molecules such as toxins and drugs. Conversely, the interactome contains all known interactions between any molecules within a cell. Interactomes often encapsulate biological pathways, and can be represented by complex networks.

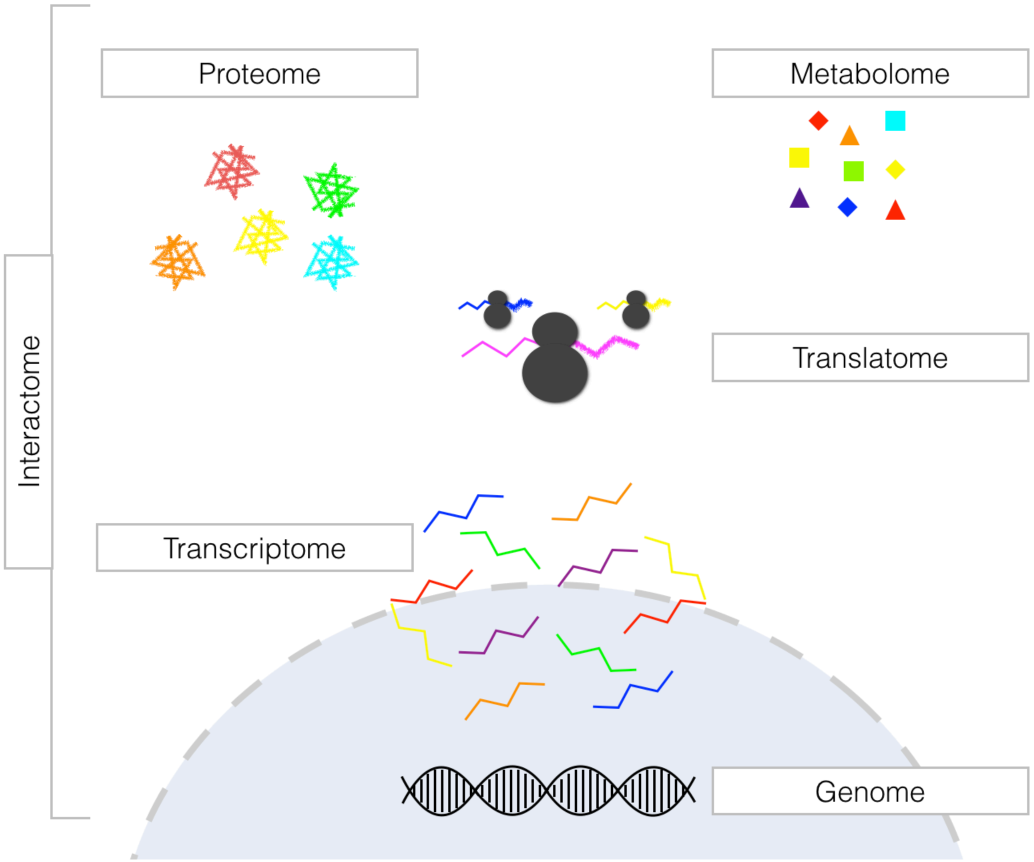
Proteome
Metabolome
Translatome
Interactome
Transcriptome
Genome

Figure 3

2.2 A systems approach to TDP-43

When using bioinformatics techniques to analyse biological data, many studies are restricted to only one of the data sources covered in the section above - usually the genome or transcriptome. While this has proved successful on many accounts, it has also been criticised. For example, the rate of successfully translating genetic hits to effective drug treatments is remarkably low. In a study investigating the power of GWAS for detecting drug targets in disease, only 20 out of over 850 drug targets corresponded with disease-associated loci (Cao & Moult 2014). Similar criticisms have been made of studies using RNA expression analysis to produce lists of disease-associated differentially expressed genes (DEGs). Submaranian and colleagues critiqued this method, claiming that even if such a list is generated - which often is not possible - the constituent genes rarely have any biological relevance to one another (Subramanian et al. 2005). Subsequently, the successful translation of such information into viable treatments is not often achieved.

Since many studies are utilising this mono-‘omic approach to complex diseases, it is understandable why successful clinical trials are so rare. By inferring mechanisms of disease from genomic or transcriptomic data alone, a problem of biological extrapolation arises; in between a genetic sequence and the presentation of a phenotype, there are a huge number of intermediate steps that are unimaginably complex and constantly changing. The nature of such a system, therefore, requires an analytical approach that is able to accommodate this complexity. By combining data from multiple ‘omic levels, one can more accurately understand the interactions and mechanisms contributing to disease. This field of research is referred to as systems biology. A ‘system’ can refer to a group of molecules, cells, tissues, or even organisms. In the context of this review, systems biology refers to the statistical analysis of the relationships and interactions between different levels of ‘omic data, from genotype to phenotype.

A multi-‘omic approach for investigation of a molecular signature for TDP-43 pathology is appropriate for multiple reasons. Firstly, since TDP-43 pathology is associated with many different mutations, and samples may come from a variety cell types or tissues, one-dimensional analysis of only genomic or transcriptomic data is unlikely to generate meaningful results. Instead, a systems approach would allow analysis of the molecular mechanisms surrounding TDP-43 pathology at multiple ‘omic levels, encapsulating the complex network of genes, proteins, metabolites and environmental factors. Systems approaches are also data driven; i.e. the elements of interested are highlighted by their statistical rather than biological significance. This avoids the bias of prior biological knowledge, producing results that are more relevant to disease mechanisms, and are subsequently more likely to produce translational results. Another benefit of a systems approach is higher replicability. Studies using single ‘omic analysis have been criticised for the significant lack of replicability and high false positive reporting. This has been attributed to underpowered experimental designs, inaccurate multiple hypothesis corrections, and inadequate control of extraneous variables (Rietveld et al. 2014; Hewitt 2012).

For investigating the mechanisms behind TDP-43 pathology, an experimental design based on a systems approach would incorporate data from multiple sources. At a genomic level, available data includes SNPs or loci identified from GWAS, linkage analysis or quantitative trait loci (QTL) mapping. Out of the many genes associated with TDP-43 pathology, a vast number of disease-associated SNPs have already been identified. For example, the TARDBP gene alone contains 33 known SNPs. A problem often faced at this point is gene prioritisation. To tackle this, SNPs can be combined with expression data from oligonucleotide microarrays or RNA seq. By identifying SNP-containing genes that are also abnormally expressed in all cases presenting with TDP-43 inclusions, this would help to prioritise genes that may have an active role in disease. The translatome is less commonly incorporated into systems research, mainly because the collection of data is still convoluted and expensive. However, since the correlation between mRNA and protein expression is remarkably low at approximately 40%, the addition of translatomic expression data could provide a much more realistic picture of downstream processes. This is also true of the proteome, as protein expression data derived from protein assays and mass spectrometry can give the most accurate depiction of which processes are dysregulated in association with TDP-43 pathology.

Incorporating of all this data is extremely challenging, but is most accurately accomplished by the generation of networks. Networks consist of nodes joined by edges (Figure 4). A node can be any kind of molecule, including genes, RNA, proteins, or even other networks (as is seen in meta-networks). Edges represent relationships between nodes, and can include co-expression, co-regulation, and direct interactions. When networks have directionality, this is often described as a pathway. Well known pathways include the Wnt pathway, or metabolic pathways such as glycolysis. Information on pathways is stored on databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), Wikipathways, or Reactome. The expression of pathways can be calculated in the same way as genes, using bioinformatics tools such as Pathprint (REF). Identification of commonly dysregulated pathways across all cases exhibiting a TDP-43 phenotype could provide an important start point in determining the effect of such pathology. Subsequent incorporation of genetic data on SNPs and gene expression could be used to validate targets within such a pathway, and prioritise them as drug targets. To support this proposition, here we summarise example case studies in which systems approaches have already been employed to produce statistically robust and biologically meaningful results for diseases, like those associated with TDP-43 pathology, that have proven to be extraordinarily complex.

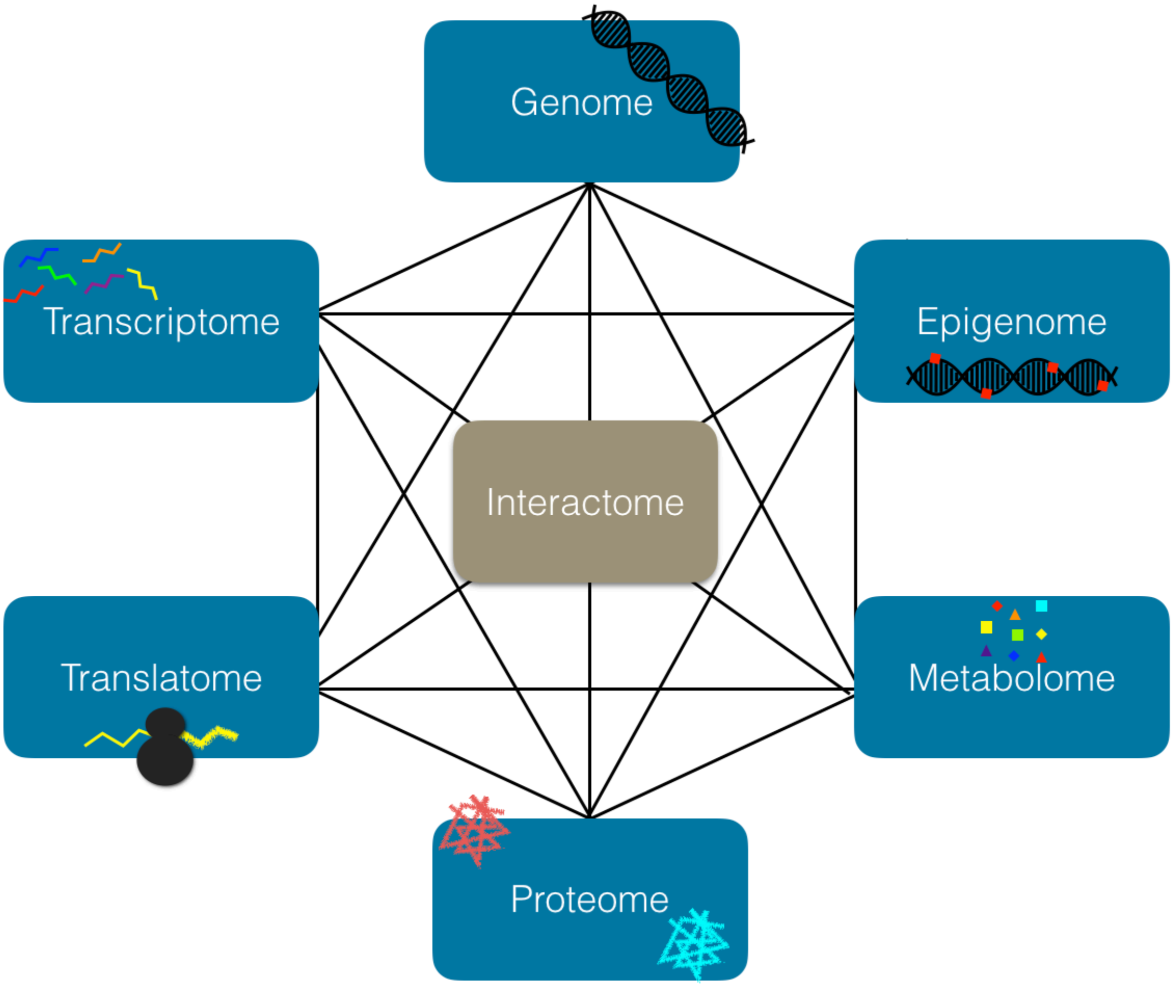
Genome
Transcriptome
Epigenome
Interactome
Translatome
Metabolome
Proteome

Figure 4

2.3 Systems approaches in disease research

2.3.1 Obesity

Diseases relating to metabolic processes and autoimmunity are notoriously complex. Previous to a study conducted by Chen and colleagues, one half of chromosome 1 had been identified as a suspected contributor to phenotypic variables such as weight, fat mass and cholesterol. However, at the time, only two genes in that region, *Apoa2* and *Tnfsf4* had been identified. As a result, Chen *et al* constructed co-expression networks for both liver and adipose tissue from a transgenic mouse known to have metabolic traits closely linked to chromosome 1 (Chen et al. 2008). Within this network, they performed enrichment analysis for Gene Ontology (GO) terms, finding enrichment of terms relating to insulin signalling and inflammation. The network was then divided into distinct sub-networks of closely clustered expression traits, and enrichment for 6 metabolic traits was calculated. In one particular network, enrichment was extremely significant for all 6 traits. This sub-network was then compared to an atlas of gene expression across 60 different tissue types. Interestingly, the sub-network was most related to bone marrow macrophages and the spleen, indicating enrichment of macrophage-related processes in the three diseases. Cross sectioning of this network with previously established data on obesity identified three candidate genes, two of which (*lpl* and *Lactb*) had no previous association with obesity. An additional method of ranking the genes in the network based on causal relationships with obesity identified the gene *Ppm1l* that, through generation of a knockout mouse model, was shown to mediate traits of metabolic syndrome. Subsequently, use of network analysis was able to identify two previously unknown genes associated with obesity, as well as link a gene already associated with drug compounds to metabolic syndrome.

2.3.2 Coronary Heart Disease

According to the 2012 report from the World Health Organisation, coronary heart disease (CHD) is now the leading cause of death worldwide (Finegold et al. 2013). Like many diseases, GWAS has identified a number of genetic loci associated with CHD, however the implication of these loci in disease development is unclear. In response, Huan and colleagues developed a systems approach using blood samples collected from 188 CHS patients and an equivalent number of age, sex and lipid treatment-matched controls. RNA expression was calculated and differential expression analysis identified 35 DEGs. Weighted Co-expression Network Analysis (WGCNA) methods were used to construct gene co-expression networks for cases and controls. When compared, two modules differed between the CHD and control groups. When compared to the genes identified by differential expression analysis there was no overlap, suggested by the authors to result from different methods capturing different biological signals. Next, the two modules were analysed for enriched GO terms. Overall, results suggested a loss of enrichment for B-cell activation in CHD cases compared to controls, with apparent enrichment for ion transport instead.

To identify causative factors, SNP set enrichment analysis (SSEA) was used to identify which SNPs known to be associated with CHD were enriched in the two modules. This analysis indicated significant enrichment for multiple cholesterol-related traits in one of the modules. To identify potential causal genes, precompiled networks representing tissue-specific Bayesian networks and non-directional protein-protein interaction networks were combined with the CHD module. This combination highlighted 59 genes, of which 37 contained CHD-related eSNPs. These 59 genes were passed through 4 prioritisation criteria, resulting in a sub-network of 20 regulatory genes. Within this tightly linked sub-network there was significant enrichment for immune-related GO terms, as well as 5 genes previously unassociated with CHD. Overall, their findings highlight the role of immune responses in CHD, and suggest novel targets for therapeutic intervention.

2.3.3 Late-Onset Alzheimer's Disease

Like many neurodegenerative diseases, the mechanisms behind AD have proven to be extremely challenging to comprehend. Late-onset AD (LOAD) is the most common form of AD, and currently there is no cure or treatment. Only a handful of genes had been associated with LOAD, subsequently Zhang *et al* undertook an impressive study that utilised over 370 LOAD and 170 healthy samples (Zhang et al. 2013). Tissue samples from the pre-frontal cortex, visual cortex and cerebellum were extracted, and both mRNA expression for nearly 40,000 genes and sequencing for over 800,000 known SNPs were conducted. After calculating differential gene expression, the top DEGs from all three tissues were use to construct an integrative co-expression network. To compare the network reorganisation between patients and controls, a metric called Modular Differential Connectivity was developed to calculate the ratio of average connectivity within a group of genes forming a cluster, or module. Approximately 60% of modules were abnormally connected in LOAD as compared to controls.

When looking for enrichment of GO terms, functional categories including immune response, synaptic transmission and nerve myelination were significantly associated with the dysregulated modules in LOAD. Using causal probabilistic Bayesian networks combined with *cis* expression SNPs (eSNPs), they discovered that the immunity and microglia modules were significantly enriched for functional categories and eSNPs, as well as possessing the highest association with LOAD neuropathological features. Next, prioritisation by Bayesian inference produced a lists of genes suspected to be causal regulators of the microglia node. By cross-referencing the microglia modules in each of the three tissues, 8 genes were found to be common to all. Of these 8, the gene *TYROBP* was found to score highest in regulation effect and differential expression. The pathological effect of *TYROBP* was validated *in vitro*, and was linked to amyloid-β turnover and neuronal damage. *TYROBP* had no previous associations with AD, and subsequently represents a new target for treatment of dementia-related neurodegeneration.

3. Discussion

In this review, we aimed to present the current dilemma of TDP-43 dysfunction, and propose a previously unconsidered methodology to approach it. What is currently clear is that in health, TDP-43 is a multi-functional protein with wide-reaching roles throughout the cell. During the development of certain neurodegenerative diseases however, TDP-43 is mislocalised from the nucleus and aggregated in proteinaceous inclusions. There is still a great deal of debate over the downstream effects of this pathology, and how it might be prevented or reversed. Subsequently, we proposed that a systems approach to TDP-43 pathology could uncover previously unknown information on these effects. By observing the dysregulation that occurs at multiple ‘omic levels across multiple diseases, one could piece together the common changes in cellular function that coincide with formation of TDP-43 inclusions. As we have seen, similar approaches to obesity, heart disease and Alzheimer’s disease have been used to discover novel genes associated with disease, and therefore new potential drug targets. Since many of the diseases associated with TDP-43 pathology are severely lacking in treatment options, this approach could produce vital information that leads to effective therapeutic interventions.