PhD degree in Molecular Medicine (curriculum in Molecular Oncology)
European School of Molecular Medicine (SEMM),
University of Milan and University of Naples "Federico II"
Settore disciplinare: MED/04

Propagation of dysregulation across gene expression layers in 7q11.23 CNV-associated developmental disorders

Giuseppe Alessandro D'Agostino

European Institute of Oncology, Milan

Matricola n. R10305

Supervisor: Prof. Giuseppe Testa

European Institute of Oncology, Milan

University of Milan, Milan

Anno accademico 2015-2016

Io volevo addiventare,
addiventare uno scienziato
Invece maggio accapputtato
E mo sto ittat mmiezz a na via
(G. Marziano)

Everything can go better,
but not everything can go worse.
(A.L.T.)

a P, che c'è.

List of abbreviations

7dup 7q11.23 microduplication Syndrome

ADHD Attention Deficit-Hyperactivity Disorder

AMPA α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid

ANOVA ANalysis Of Variance

AS Angelman Syndrome

ASD Autistic Spectrum Disorders

ATP Adenosine-Tri-Phosphate

CDS Coding sequence

CHX Cycloheximide

CNV Copy Number Variation

CrPV Cricket Paralysis Virus

DEG Differentially Expressed Gene

DEP Differentiallye Expressed Protein

EBV Epstein-Barr Virus

ED Exponentially Degraded

ESC Embryonic Stem Cell

FC Fold Change

FDR False Discovery Rate

fMRI Functional MRI

FPKM Fragments Per Kilobase of exon per Million reads mapped

GEF Guanine nucleotide Exchange Factors

GLM Generalized Linear Model

GO Gene Ontology

GTP Guanine-Tri-Phosphate

ID Intellectual disability

iPSC Induced Pluripotent Stem Cell

IQ Intelligence quotient

IRES Internal Ribosomal Entry Site

IVF In Vitro Fertilization

LCL Lymphoblastoid Cell Lines

LFQ Label-Free Quantification

LOESS Local Regression

MEF Mouse Embryonic Fibroblast

MRI Magnetic Resonance Imaging

mRNA messenger RNA

mTOR Mammalian Target Of Rapamycin

NAHR Non-Allelic Homologous Recombination

NED Non-Exponentially Degraded

NMD Nonsense-Mediated Decay

ORF Open Reading Frame

PABP Poly-A Binding Protein

PcG Polycomb Group

PCR Polymerase Chain Reaction

PIC Pre-Initiation Complex

pSILAC pulsed-Stable Isotope Labeling of Aminoacids in Culture

RBP RNA-binding Protein

RIA Relative Isotope Abundance

RISC RNA-Induced Silencing Complex

RPF Ribosome-protected fragment

RRM RNA Recognition Motif

rRNA Ribosomal RNA

RT Reverse Transcription

SCNT Somatic Cell Nuclear Transfer

shRNA short hairpin RNA

SMA Standardized Major Axis

SNP Single Nucleotide Polymorphism

SVAS Supravalvular Aortic Stenosis

SWATH-MS Sequential Window Acquisition of all Theoretical Mass Spectra

TE Translation Efficiency

TF Transcription factor

tRNA transfer RNA

TrxG Trithorax Group

UTR Untranslated Region

WBS Williams-Beuren Syndrome

WBSCR Williams-Beuren Syndrome Chromosome Region

Index of figures and tables

Figures in the Introduction chapter

Figures in the Materials and Methods chapter

Figures in the Results chapter

Tables in the Introduction chapter

Tables in the Materials and Methods chapter

Tables in the Results chapter

Abstract

Williams-Beuren Syndrome (WBS) and 7q11.23 microduplication syndrome associated to autistic spectrum disorders (7dup) , two multi-systemic developmental disorders, arise from symmetrical copy number variations of the same region on chromosome 7q, comprised of 26-28 genes. In WBS patients this region is deleted, whereas it is duplicated in 7dup individuals. These syndromes display a striking combination of shared and opposite clinical manifestations at the level of neuro-cognitive, craniofacial and cardiovascular features, thus pointing to a remarkable dosage-sensitive effect of a small group of genes on the development and maintenance of complex traits such as sociality, language and facial morphology.
We harnessed the power of somatic cell reprogramming to derive a large cohort of induced pluripotent stem cells (iPSCs) from samples with WBS, 7dup and from healthy controls, and we demonstrated that, already at the pluripotent state, the transcriptome is dysregulated in pathways that map onto clinical aspects such as neuronal, cardiovascular and craniofacial development. Moreover, these pathways were selectively dysregulated in differentiated lineages, thus demonstrating an unforeseen anticipatory power of the pluripotent state. Indeed, genes in the region that are involved in the regulation of transcription and translation are highly expressed in iPSCs, and their expression mirrors the CNV dosage across samples.
Building on these results, my PhD project is aimed at expanding the view on the dysregulation in pluripotency, by measuring three layers of gene expression: transcriptome, translatome and proteome. We mapped the propagation of differences across layers by integrating two innovative technologies, ribosome profiling and SWATH-MS, and we probed the extent to which a translation initiation factor included in the CNV, EIF4H, was responsible for the regulation of translation. We found that each layer of gene expression has its own differentially expressed genes,

and the degrees of propagation can change between layers. Moreover, differentially expressed genes can cluster by different ways of propagation when they are compared to the levels of EIF4H. We then showed that protein degradation, measured by pulse-Stable Isotope Labeling in Culture (pSILAC) coupled to SWATH-MS, does not explain changes in protein abundance across samples in this system, and we provided some examples of gene expression modeling in which degradation rates do not increase the precision of the model. Finally, we set up a large cohort of scalable, homogeneous inducible neurons that are amenable to high-throughput experiments and promise to greatly enhance our understanding of the molecular and cellular basis of Williams-Beuren Syndrome and 7q11.23 microduplication syndrome

Introduction

Williams-Beuren Syndrome (WBS – OMIM 194050) and 7q11.23-microduplication syndrome (7dup, also known as Somerville-van der Aa Syndrome – OMIM 609757) arise from a symmetrical copy number variation (CNV) of a 1.5-1.8 Mbp region, containing 26-28 genes, and named Williams Beuren Syndrome Chromosome Region (WBSCR). WBS and 7dup are caused by the hemizygous deletion and duplication of the WBSCR respectively. The prevalence estimates for WBS are of 1 every 7,500 live births, whereas only a few cases (around 100) of 7dup have been reported.

Features of WBS and 7dup

Both Williams-Beuren Syndrome and 7q11.23 microduplication syndrome are multisystemic developmental disorders, with an autosomal dominant inheritance, in which many systems are affected. These CNVs give rise to altered neurocognitive, craniofacial, cardiovascular, metabolic, skeletal and genito-urinary features, in a combination that presents both shared and opposite abnormalities between the two syndromes (fig. 1).

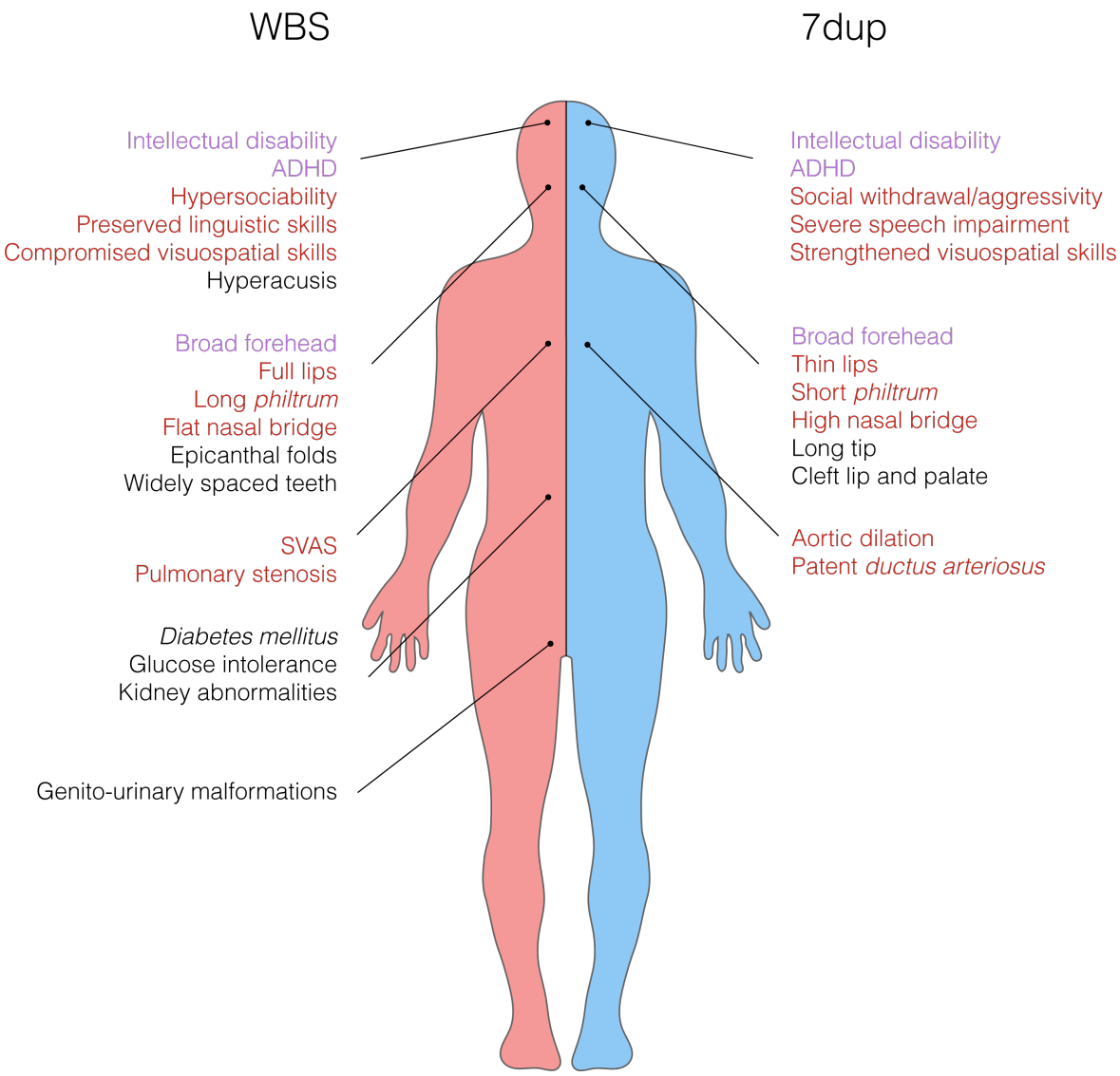

Figure 1: A schematic representation of clinical features in both syndromes. Purple text: shared features; red text: opposite features. SVAS: supravalvular aortic stenosis; ADHD: attention deficit-hyperactivity disorder.

The combination of shared and opposite features in many different systems and, especially, in high-order neurological functions such as language, sociality and visuospatial construction, points to the WBSCR region as a small cluster of genes that is able to control a diverse set of developmental features in a remarkably dosage-sensitive fashion. I will briefly summarize the key clinical aspects of both syndromes to give the reader a flavour of both the peculiar aspects of these pathologies and the variability across patients.

Craniofacial features

WBS patients have been historically identified for a characteristic facial morphology that, in the early 60s, was defined as *elfin*[1]. Epicanthal folds, a broad forehead, swollen lips, long *philtrum* (the area between nose and upper lip), widely spaced teeth, collapse of the nasal bridge, and squint all contribute to a *facies* that had been initially described by Williams (Williams et al., 1961) and Beuren(Beuren et al., 1962). Some cases of microcephaly have also been reported(Knudtzon et al., 2008; Pérez Jurado et al., 1996). Initial attempts at linking craniofacial dysmorphism with hypercalcemia have been disproved already in the early days of the definition of the syndrome; we now know that many genes in the WBSCR are likely to have a large impact on the craniofacial morphology (see Genes involved in the the CNV). Conversely, 7dup patients have a combination of facial traits that, although having different expressivity, is in many cases the opposite of the WBS *gestalt*: thin lips, a short *philtrum*, high nasal bridge and long nasal tip, retrognathia and macrocephaly. Although not always present, the broad forehead seems to be the shared craniofacial feature between the two syndromes; a few reports of cleft lip and palate have been published, mainly for 7dup patients (Torniero et al., 2008). Clinicians only attempt at correcting dental dysmorphisms by means of dentistry interventions.

Cardiovascular features

The predominant condition of WBS, affecting 70% of patients (Pober, 2010a), is supravalvolar aortic stenosis (SVAS), due to the haploinsufficiency of the elastin (ELN) gene. As the name suggests, it consists in the narrowing of the aorta just above

[1] Although there have been some calls to stop describing patients with this term (*Whether or not these children have elfin facies is difficult to establish, for while examples of the syndrome are common, this author has never seen an elf. The term should be dropped.* Burn, 1986) it has survived and even adapted to different cultural contexts: in Germany WBS is also called "koboldgesicht syndrom", in Spain "síndrome del duende", and so on.

the aortic valve, and is often accompanied by stenosis of the pulmonary artery. It is the major cause of death for WBS patients and is generally treated with surgical interventions ranging from aortoplasty to autografts. Conversely, 7dup patients display aortic dilation in 46% of the cases (Mervis et al., 2015) and, in ~9% of the cases, patent *ductus arteriosus*, i.e. failure, shortly after birth, to close a vessel connecting the aorta and the pulmonary artery.

Neuro-cognitive features

The unique neuro-cognitive profile of WBS patients had gained them yet another nickname: *cocktail party personality*. The openness and lack of social inhibition towards strangers is indeed one of the most striking behavioural features of these patients (Pober, 2010). Although they are affected by mild to severe intellectual disability (ID) - their intelligence quotient (IQ) is often compared to that of Down syndrome patients - their linguistic abilities are spared, as they are able to express themselves with a rich vocabulary and complex syntactic constructs. Many WBS patients have attention deficit – hyperactivity disorder (ADHD), are anxious and develop obsessions, especially concerning objects. Another remarkable feature of WBS patients is their severely compromised ability to construct space. They are not able to mentally rotate objects, or to discern relative distances and positions of elements, even simple ones; this feature seems to be spared in IQ-matched Down syndrome patients (fig. 2)(Bellugi et al., 1999), pointing again to a remarkably selective action of these genes when it comes to specific areas of cognition[2]. WBS

[2] Jerry Fodor's idea of *modularity of mind* (Fodor, 1983), i.e. the existence of discrete parcels of cognition with functional independence, information encapsulation and a fixed neural architecture has been proposed to be validated by cases such as William's Syndrome. However, the variability and nuances in cognition defects, their inter-dependence, their developmental origin combined with the inescapable fact that affected skills, however relatively strengthened, are still affected, led to a confutation of this theory. For a review see (Meyer-Lindenberg et al., 2006)

patients perform poorly on tasks in which planning and counting are required (Paterson et al., 2006).

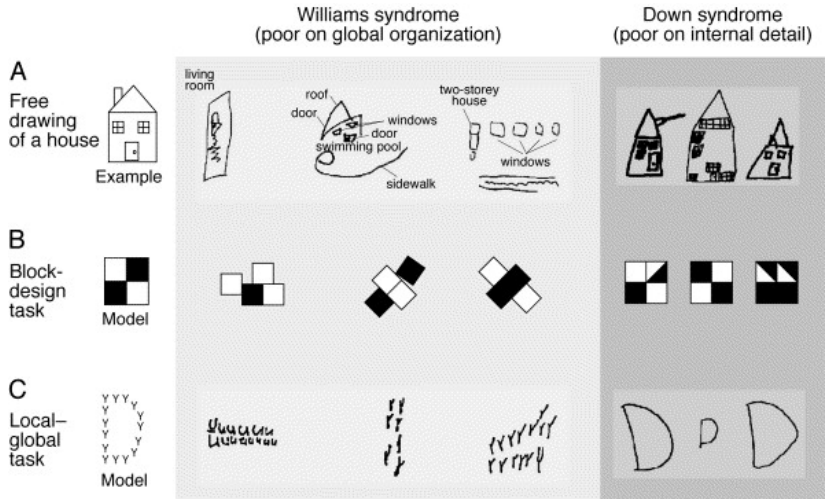

Figure 2: A comparison of drawing tasks performed by WBS patients and Down syndrome patients. From Bellugi et al., 1999.

Other more variable features involve hyperacusis – many WBS patients have "perfect pitch", the ability to discern a musical note only based on its absolute frequency – and sensorineural hearing loss (Marler et al., 2005). Conversely, while 7dup patients still retain intellectual disability and ADHD, their language abilities are severely compromised, and can often reach the harshness of full blown autism (Mervis et al., 2015); moreover, 7dup patients are socially withdrawn or aggressive, and engage in repetitive behaviour. Taken together, these features represent hallmarks of autistic spectrum disorders (ASD), for which the 7q11.23 duplication has been reported as a strong risk factor (Sanders et al., 2011). Indeed, 20% of 7dup patients can be considered as frankly autistic based on gold standard ASD-scoring methods (Mervis et al., 2015)[3].

[3] A previous interpretation of the behavioural features of WBS patients led the medical community to classify it as "the polar opposite of autism" (Jones et al., 2000). While this characterization was dropped over time, the boundaries of this definition have been made sharper by the association of 7dup with ASD, for which the symmetry of the genetic aberration warrants a comparison between WBS and ASD. However, the existence of WBS patients with ASD (Tordjman et al., 2012) blurs again the boundaries, and perhaps urges us to rethink the diagnostic criteria, scales and scoring methods we use to define ASD. A genotype-first approach, in which the classification of different subsets of ASD depends on the mutation underlying it (Stessman et al., 2014), may provide a valuable alternative.

Neurological abnormalities probed by magnetic resonance imaging (MRI) have shown that WBS patients have a 10-15% reduction in encephalon volume, with a proportional reduction in cerebrospinal fluid (Capitão et al., 2011). Interestingly, the cortex of WBS patients appears to have local reductions in thickness and an increase in the number and length of *gyri* compared to healthy cortices (Torniero et al., 2007). Functional MRI studies have shown how the activation of specific areas of the visual cortex is compromised in patients: when presented with specific visuospatial tasks, the ventral stream circuit (the "what") is activated as in healthy people, whereas the dorsal stream circuit (the "where") is not (Meyer-Lindenberg et al., 2004). Similarly, the amygdala, deputed to emotion processing, is not activated when patients are presented with aggressive faces, while it is strongly responsive towards non-social, menacing stimuli (Meyer-Lindenberg et al., 2005). On the contrary, MRI on 7dup patients shows an overall increase in cortical thickness and in the total brain volume, simplified *gyri* complexity, ventriculomegaly and other brain malformations (Torniero et al., 2007); functional MRI, albeit on a single case to date, has revealed a normal functioning of the visual cortex, contrasted by an inactivation of the limbic system and the emotion-processing areas (Prontera et al., 2014). These findings underscore again, this time from a neuroanatomical and even functional perspective, the symmetry between clinical manifestations of these syndromes.

Other features

Abnormal calcium metabolism: between 5 and 50% of WBS patients have a severe form of hypercalcemia and hypercalciuria in their infancy, which then progresses to a milder manifestation as they age. A link with vitamin D metabolism has been proposed, but never fully demonstrated. No calcium abnormalities have been reported in 7dup patients to our knowledge.

Diabetes mellitus and glucose intolerance: impaired glucose intolerance affects around 75% of adults with WBS, many of which also have been diagnosed with *diabetes mellitus* (Pober et al., 2010). Some studies in animal models suggest a role for MLXIPL and STX1A (see Genes involved in the CNV). As in the case for calcium abnormalities, no reports have shown occurrence of neither glucose intolerance nor diabetes in 7dup patients.

Genetic rearrangements at chromosome 7q11.23

The presence and peculiar arrangement of *duplicons*, i.e. duplicate elements with a high sequence homology within the WBSCR, is responsible for the aberrant alignment of sister chromatids for crossing over during meiosis I. During gamete formation, non-allelic homologous recombination (NAHR) events taking place between centromeric and medial or telomeric duplicons result in an incorrect crossing-over that causes the hemizygous deletion – and simultaneous duplication – of the WBSCR (fig. 3). The inversion of this region has been reported as a polymorphism that, when occurring in a healthy individual, is more likely to cause a deletion in his or her children (Hobart et al., 2010; Morris et al., 2011).

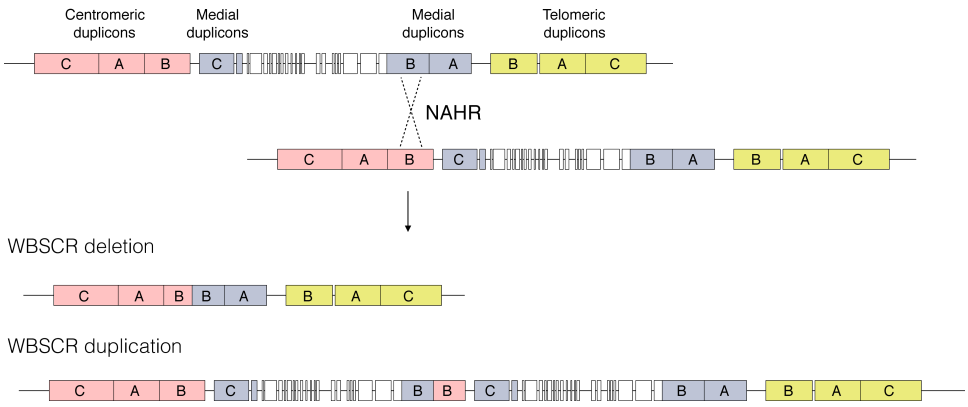

Figure 3: Schematic representation of NAHR occurring at 7q11.23 causing deletion and duplication of the WBSCR. Adapted from (Pober, 2010).

Turner and colleagues (Turner et al., 2008) have demonstrated how the frequency of NAHR does not, in principle, give rise to an equal number of gametes with the

deletion and gametes with the duplication. The deletion should rather be twice as frequent as the duplication. However, the prevalence of WBS patients is much higher than that of 7dup patients (1 in 7,500 against a few dozen cases reported) with a proportion that is bigger than what would be expected by taking in consideration only NAHR frequencies. This inequality can have different explanations. From a purely cellular perspective, haploid cells lacking the WBSCR may be more fit than their duplicated counterparts, thus making it more unlikely for a 7dup gamete to effectively give rise to a fertilized egg. There are however, to our knowledge, no studies in which the impact of 7q11.23 CNVs was investigated in gametes; moreover, there are reports of individuals bearing 7q11.23 duplication who were diagnosed only after receiving the same diagnosis for a child/grandchild (Mervis et al., 2015; Patil et al., 2015). These adults had only minor learning disabilities and/or sociality issues that were not previously categorized in any particular intellectual disability syndrome (Patil et al., 2015)[4]; other probands that tested positive for the duplication had no intellectual disability at all (Mervis et al., 2015). This points to a milder, or more nuanced effect of the duplication compared to the deletion, which makes its clinical manifestations less evident and therefore less amenable to a correct diagnosis. Finally, taking in consideration also a historical point of view, while WBS has been first defined in the 60s as a very characteristic set of clinical traits, the genetics of which were only later precisely identified by means of a traditional or "forward" genetics approach, 7dup was identified only recently in a small cohort of

[4] It is interesting to note, in the context of this reference, the description of the father transmitting the duplication: "On history, it was found that the father had a history of delayed speech. He never had any formal education and now he works as daily wage laborer.". The socioeconomical context in which such diagnoses of intellectual disability are made is likely to play an important role: how did the lack of formal education impact on the cognitive abilities of this individual, whose genetic aberration went undetected for years? Or, from another point of view, how did the socioeconomic environment impact on the assessment of his disability? This recognition points to the socially constructed nature of (intellectual) disability, and its interdependence with cultural, biomedical and socio-economic factors. For an analysis of the social construction of disability, see Susan Wendell's "The Rejected Body" (Wendell, 1996), especially chapter 2 and the introduction to Lennard J. Davis' "The Disability Studies Reader" (Davis, 1997).

samples (Kirchhoff et al., 2007; Kriek et al., 2006; Somerville et al., 2005; Thomas et al., 2006) and later (Sanders et al., 2011) in a genome-wide association study on a cohort of ASD patients. A single event of *de novo* triplication of the region has been reported in a patient displaying even more severe dysmorphisms and speech impairment (Beunders et al., 2010) , strengthening the link between the amplification of this region and a neurocognitive/morphological syndromic outcome. Taken together, all these reports point to the duplication as pathological variant that, although being less diagnosed, showing less penetrance and more variable expressivity when compared to the deletion, is responsible for a clinical phenotype that has both shared and opposite features of Williams-Beuren Syndrome. The size of the genetic rearrangement is usually between 1.55 Mbps (95% of cases) and 1.8 Mbps (5% of cases), and the position of its breakpoints are highly variable within the duplicons(Bayés et al., 2003). A few interesting cases of atypical deletions have been reported (Chailangkarn et al., 2016; Ferrero et al., 2010; Fusco et al., 2014), which spare specific regions of the WBSCR and allow more precise correlations between WBS genes and the clinical phenotype: depending on the portion of the region that is missing, these patients display a subset or a milder manifestation of pathological traits, such as intellectual disability with no dysmorphism, or exclusively cardiovascular problems.

Genes involved in the CNV

As briefly mentioned before, duplicons contain tandem repeats of duplicated sequences with very high homology (~99.7%). These sequences are mostly pseudogenes or duplications of sequences that are enclosed between the centromeric and medial duplicons, i.e. the *functional*[5] region of the WBSCR, with the sole exception

[5] I employ here a liberal definition of function, that postulates the existence of a transcript and/or protein that is expected to yield an effect on other molecules and on cellular functions. Yet,

of the medial B duplicon, which contains the functional GTF2I, NCF1 and GTF2IRD2 sequences, and the medial C duplicon, which contains the functional FKBP6, TRIM60. All the sequences contained in the functional region are duplicated in at least one duplicon, with the exception of ELN (Pober, 2010a). The functional WBSCR is in turn made of 28 protein-coding genes 1 micro-RNA and 1 antisense lncRNA (fig. 4).

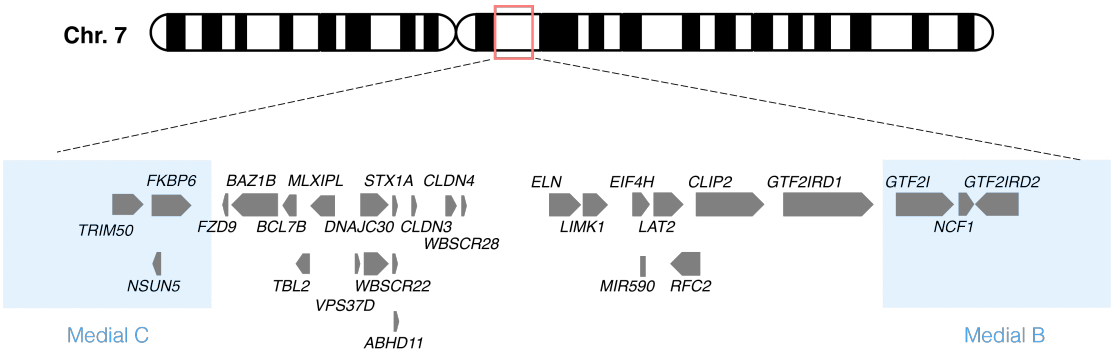

Figure 4: Schematic representation of the genes comprised in the WBSCR. Blue boxes: duplicons. Grey arrows: genes.

The WBSCR harbours a diverse set of genes, whose function and involvement in WBS is summarized for the reader's comfort in table 1.

pseudogenes in the duplicons are transcribed and in some cases translated, even though their contribution to a molecular or cellular phenotype, if any, has not yet been investigated. For an updated read on the notions of function regarding genomic loci see Germain, Ratti and Boem, *Junk or functional DNA? ENCODE and the function controversy*, Biology and Philosophy (Germain et al., 2014).

Gene	Functional product	Involvement in WBS	References
TRIM50	E3 ubiquitin ligase involved in the alimination of misfolded proteins via the formation of aggresomes	Unknown	(Fusco et al., 2012)
FKBP6	Protein co-chaperone involved in the PIWI-mediated transposone silencing in spermatogenesis and in cellular proteostasis	Unkown	(Taipale et al., 2014; Xiol et al., 2012)
NSUN5	28S rRNA methylase, influences the diet-associated lifespan of *C.elegans*	Unknown	(Schosserer et al., 2015)
FZD9	Wnt receptor, negative regulator of the beta-catenin pathway	Involved in apoptosis of neural precursors and in the determination of dendritic spine number, calcium oscillation and connectivity in upper layer cortical neurons of WBS patients.	(Chailangkarn et al., 2016)
BAZ1B	Transcription factor, tyrosine phosphatase for histone H2A.X, chromatin remodeler in WICH and SWI/SNF complexes. Involved in neural crest formation and migration	Deletion in mice causes craniofacial dysmorphisms comparable to those of WBS patients. Homozygous KO mice have major heart defects, which are milder but still present in heterozygosity. Accounts for ~40% of the transcriptional dysregulation in WBS neurons.	(Ashe et al., 2008; Lalli et al., 2016; Soldi and Bonaldi, 2013)
BCL7B	Tumor suppressor, negative regulator of Wnt signaling, member of SWI/SNF complexes	Unknown	(Kadoch et al., 2013; Middeljans et al., 2012; Uehara et al., 2015)
TBL2	Binds 60S ribosomal subunit and enhances translation of ATF4 during ER stress	Unknown	(Tsukumo et al., 2015, 2016)
MLXIPL	b-HLH transcription factor, binds carbohydrate-responsibe elements and regulates triglyceride synthesis	Possibly involved in the metabolic aspects of WBS including diabetes and glucose intolerance.	(Palacios-Verdú et al., 2015)
VPS37D	Member of the ESCRT-I complex, involved in removal of poly-ubiquitinated proteins	Unknown	(Schuh and Audhya, 2014)
DNAJC30	Chaperone, contains a DNA-J domain	Unknown	
WBSCR22	Methylates 18S rRNA in the nucleus, necessary for ribosome biogenesis	Unknown	(Õunap et al., 2015; Zorbas et al., 2015)
STX1A	Involved in the docking of synaptic vesicles in the presynaptic membrane	Possibly involved in white matter development in the encephalon. Expression levels are statistically associated to intelligence in WBS patients. Also involved in the regulation of insulin secretion in KO mice models.	(Gao et al., 2010; Hoeft et al., 2014; Lam et al., 2005)
ABHD11	Poorly characterized. Reports of serine hydrolase activity.	Unknown	(Navia-Paldanius et al., 2016)
ABHD11-as	Antisense lncRNA in the ABHD11 locus	Unknonw; neuroprotective in mice models of Huntington's disease	(Francelle et al., 2015)
CLDN3	Claudin involved in tight junctions	Unknown	
CLDN4	Claudin involved in tight junctions and regulation of Ca2+ paracellular concentration	Unknown	
WBSCR28	Poorly characterized. Repressed by androgen-receptor mediated pathways.	Unknown	(Prescott et al., 2007)
ELN	Elastin, involved in the formation of connective tissue with contractile features	Responsible for SVAS and cutis laxa	(Li et al., 1997; Micale et al., 2010; Tassabehji et al., 1997)
LIMK1	LIM-domain kinase involved in the formation of dendrites and synaptic transmission	Involved in the visuospatial construction deficit.	(Frangiskakis et al., 1996; Gray et al., 2006)
EIF4H	Translation initiation factor involved in the scanning of the 5' untranslated region of mRNAs	KO mice models display a reduction in number and complexity of CNS neurons, and learning deficits	(Capossela et al., 2012)
mir590	Included in an intron of EIF4H. Inhibits EMT by upregulating E-cadherin. Involved in cardiac regeneration	Unknown	(Eulalio et al., 2012; Liu et al., 2015)

LAT2	Aminoacid transporter involved in immune system activation (B and T cells)	Unknown	(Orr and McVicar, 2011)
RFC2	DNA replication factor C subunit, necessary for the elongation of primed regions by DNA polymerase	Unknown	(Gupte et al., 2005)
CLIP2	Links microtubules and organelles in dendrites	Undecided. Atypical deletions implicate it in neurocognitive phenotypes, but individuals with hemizygous deletion are healthy.	(van Hagen et al., 2007; Vandeweyer et al., 2012)
GTF2IRD1	Transcription factor involved in chromatin regulation	Involved in visuospatial processing, in sensorineural hearing loss and in craniofacial dysmorphisms.	(Antonell et al., 2010a; Broadbent et al., 2014; Canales et al., 2015; Edelmann et al., 2007; Hirota et al., 2003; Tassabehji, 2005)
GTF2I	Transcription factor involved in chromatin regulation	Involved in intellectual disability, sociality, visuospatial processing, craniofacial dysmorphisms, calcium intake. Accounts for ~ 20% of transcriptional dysregulation in iPSCs	(Adamo et al., 2014; Antonell et al., 2010a; Edelmann et al., 2007; Hirota et al., 2003; Malenfant et al., 2011; Sakurai et al., 2011)
NCF1	Member of the NADPH-oxidase complex	When deleted in hemizygosity it is a protective factor against hypertension	(Del Campo et al., 2006)
GTF2IRD2	Transcription factor arising from a fusion of a portion of GTF2I and a retrotransposon	Deleted only in the 1.8 Mbp CNV, which shows more severe clinical manifestations	(Bayés et al., 2003; Tipney et al., 2004)

Table 1: Summary of genes in the WBSCR and their function

Groups of genes in this region converge towards CNS-specific functions (FZDN9, STX1A, LIMK1, CLIP2, GTF2I), cardiovascular development (ELN, NCF1), metabolic regulation (MLXIPL, ABHD11), craniofacial development (GTF2I, BAZ1B, GTF2IRD1), or towards more general processes such as transcription, translation and degradation (GTF2I, WBSCR22, NSUN5, EIF4H, TRIM50, VPS37D).

It is beyond the scope of this introduction to provide an in-depth review of the characterization of all these genes, but I will summarize some important aspects on EIF4H and GTF2I that have emerged from genetic studies, work on animal models and *in vitro* patient-derived cellular models, including work performed in this laboratory.

EIF4H

The eukaryotic Initiation Factor 4H is a small (25-27 KDa) protein involved in the regulation of the initiation step of translation, which has been regarded as the rate limiting step where most of the regulation is thought to occur (Sonenberg and Hinnebusch, 2009) [6]. It interacts with the EIF4F complex at the stage of unwinding of the secondary structures of the 5' untranslated region (5'UTR) of mRNAs, a necessary step to allow the 43S pre-initiation complex (PIC) to scan the UTR and reach the AUG start codon (Parsyan et al., 2011). The PIC is composed by the small ribosomal subunit 40S and a series of initiation factor complexes (fig. 5). EIF4H has a paralog, EIF4B, which regulates translation initiation in the same complex. The sequence of EIF4B is homologous to the whole length of EIF4H, with the addition of longer N- and C- terminal domains that can be phosphorylated as a result of signaling cascades (Dennis et al., 2012a).

[6] This tenet has been challenged lately by the idea that also the initial clearance and elongation steps of translation (i.e. the joining of the first tRNAMet and the subsequent traslocation along the open reading frame) are closely related to initiation rate-limiting steps (Chu et al., 2014). The optimization of codons, the availability of free 40S subunits, the recycling of monosomes all play an important part in the regulation of translation kinetics. I will however consider a more traditional model for simplicity, keeping in mind that it is unlikely to be the only explanation to changes in protein synthesis rates.

EIF4H has been shown to physically interact with the mRNA helicase EIF4A and to compete with EIF4B for its binding (Rozovsky et al., 2008), however, *in vitro* studies have shown that EIF4B stimulates translation 5 times more than EIF4H; moreover, their relative stoichiometry varies across different tissues (Richter et al., 1999), with EIF4B being less expressed in the heart and in the brain. The exact order of events for the EIF4A-EIF4H/EIF4B interaction has not yet been fully described, but a few models have been built on experimental data. EIF4H enhances the processivity of the unfolding by acting simultaneously on both EIF4A and the mRNA. It binds unwound, single-strand mRNA via its RNA Recognition Motif (RRM), thus stabilizing it and preventing the formation of secondary structures (Marintchev et al., 2009; Sun et al., 2012); it also stimulates ATP hydrolisis by EIF4A, which makes its helicase activity more processive and unidirectional (Marintchev et al., 2009; Spirin, 2009). Another model has been proposed in which EIF4A is seeded on the 5'UTR together with EIF4B and EIF4H, polymerizing on the mRNA and allowing for faster initiation on already processed 5'UTR (Lindqvist et al., 2008). Beside its biochemical characterization, *EIF4H* has been studied in cancer as an oncogene (Vaysse et al., 2015; Wu et al., 2011), building into the notion of translation initiation as a pathway that can be targeted by anticancer drugs (Bhat et al., 2015). These reports also show that EIF4H is able to influence, in cultured tumoral cell lines, translation at a global level. The expression of EIF4H across many tumoral lines seems to be remarkably stable (Macrae et al., 2013), to the extent that it is being proposed as a replacement for GAPDH or TBP as a normalizer in qRT-PCR measurements. Only one report has linked EIF4H to intellectual disability by generating and phenotyping *Eif4h* null mice (Capossela et al., 2012). These mice presented an overall reduction in size and weight, but more importantly showed a decrease in encephalon size, and a reduced number and complexity of cortical neurons. Neurological abnormalities were indeed reflected

by a series of behavioural phenotypes that are reminiscent of some WBS neurocognitive traits: inability to develop spatial memory, hyperreactivity to novelties, defects in fear-associated learning. Quite paradoxically, the authors report that translation in bulk brain extracts – as assessed by polysome profiling – does not change in *Eif4h* null mice compared to WT, leaving the tissue-specificity of the translational effect of EIF4H open to debate.

GTF2I

The General Transcription Factor II-I, together with GTF2IRD1 and GTF2IRD2, belongs to the TF-II-I family of transcription factors, characterized by the presence of multiple helix-loop-helix domains termed I-repeats. GTF2I and GTF2IRD1 are always deleted in WBS cases, while GTF2IRD2 is only deleted in large (~1.8 Mbp) deletions. GTF2I has been characterized as a basal transcription factor able to bind initiator (Inr) elements at core promoters to initiate transcription (Roy, 2001). It was later discovered that it can also integrate the response from extracellular signals upon phosphorylation and translocation in the nucleus, where it would bind E-boxes and exert either a positive or negative effect on transcription of specific genes (Hakre et al., 2006). Five isoforms of GTF2I exist (α, β, γ, Δ, ε) with different expression levels and patterns across tissues, and different subcellular localizations; GTF2I γ is more expressed in the brain, in which α and Δ isoforms are instead absent. Several lines of evidence converge towards GTF2I as one of the major determinants of many clinical outcomes in Williams-Beuren Syndrome and in 7q11.23 microduplication Syndrome. Patients with atypical deletions that always encompass GTF2I still show the neurocognitive profile (Antonell et al., 2010a; Dai et al., 2009; Edelmann et al., 2007), whereas a patient with an atypical deletion sparing GTF2I has a normal IQ (Ferrero et al., 2010). Moreover, two single nucleotide polymorphisms (SNPs) in GTF2I have been associated to autism (Malenfant et al., 2012) (and, in the healthy population, one

of these SNPs has been associated with social anxiety, reduced social communication skills, and amygdala activation by avversive stimuli (Crespi et al., 2014; Swartz et al., 2015).

Many animal models have proven how homozygous KO, and heterozygous KO or duplication of *Gtf2i* lead to behavioural, neuronal and craniofacial phenotypes reminiscent of the clinical traits of either syndrome (Mervis et al., 2012; Osborne, 2010). Further linking the role of *Gtf2i* with the neurocognitive phenotype, an interesting work has recently been carried out in mice models in which the *Gtf2i* deficiency is rescued by administering an adeno-associated viral (AAV) vector overexpressing *Gtf2i* cDNA directly in the mouse encephalon by intracisternal injection (Borralleras et al., 2015). The authors report changes in both the expression of endogenous *Gtf2i* and in the motor, social and behavioural features of mice, thus even pointing to a potentially therapeutic use.

Our lab has demonstrated how, in induced pluripotent stem cells (iPSCs) derived from patients carrying either the deletion or the duplication, GTF2I can be held accountable for up to 20% of the transcriptional dysregulation (Adamo et al., 2014). By co-immunoprecipitation we were able to determine that in iPSCs GTF2I assembles a repressive complex by physically interacting with the histone demethylase LSD1 and the histone deacetylase HDAC2. Among its targets there are genes involved in many clinical manifestations typical of WBS and 7dupASD, such as neuronal development, cardiovascular structure development, smooth muscle contraction and presynaptic membrane organization. This suggests that GTF2I, among other factors, is able to seed at the earliest stages of development a sizeable portion of transcriptional dysregulation that maps onto pathways relevant for the disease.

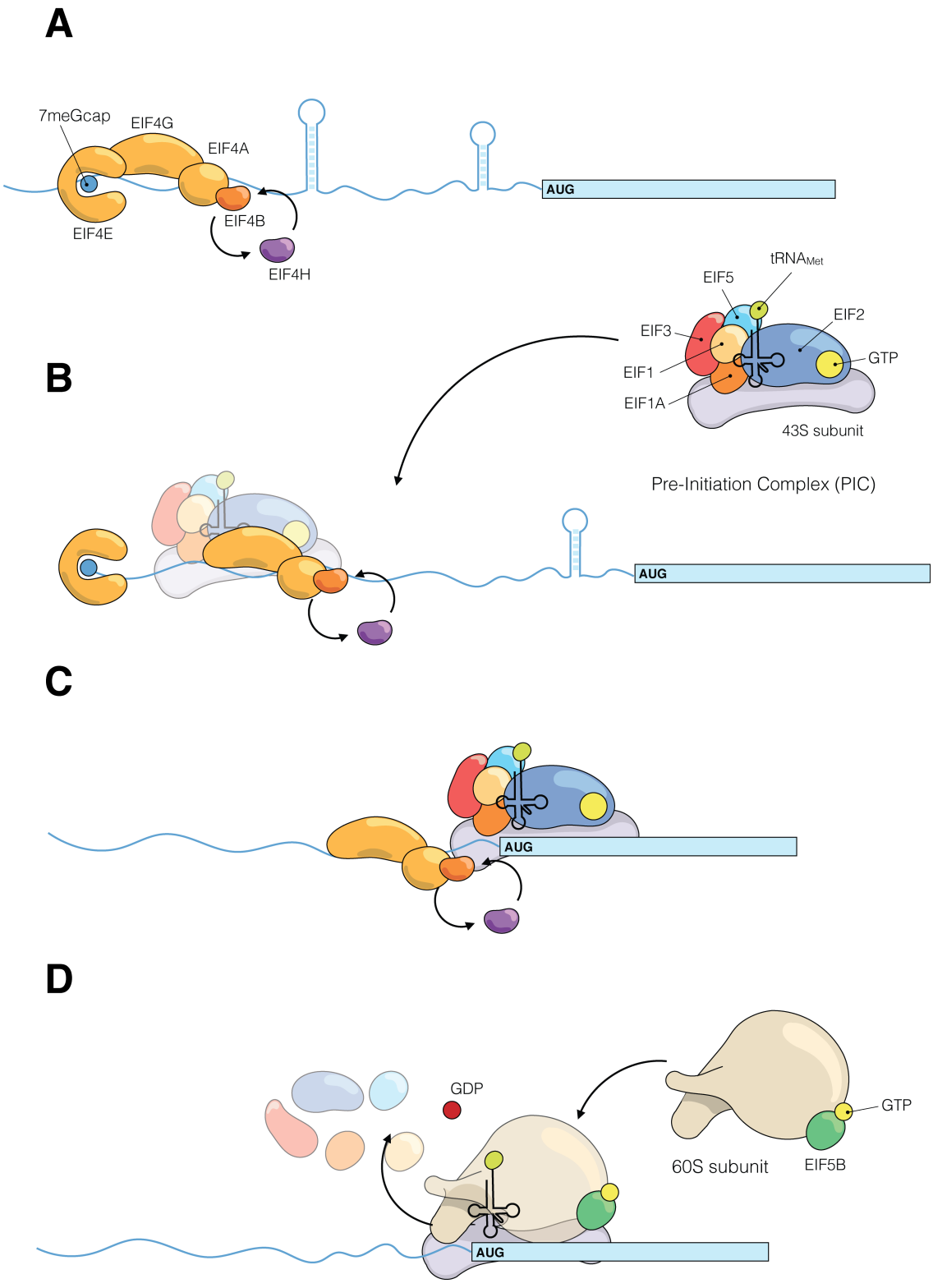

Figure 5: A simplified model of translation initiation (adapted from Parsyan et al., 2011).A: the cap-binding protein EIF4E recognizes the 5' 7meG cap and docks EIF4G and the RNA helicase EIF4A on the 5' untranslated region. EIF4A starts unwinding the 5'UTR in a 5'-> 3' direction in an ATP-dependent fashion, stimulated by EIF4H, EIF4B and EIF4G. B: the pre-initiation complex, composed by the small ribosomal subunit, the initiator Met-tRNA$_i^{Met}$ in the P (peptidyl) site, and a series of multi-subunit eukaryotic initiation factors lands on the 5'UTR, now stabilized in a form that can be scanned. C: The AUG in the open reading frame is recognized by the initiator Met-tRNA$_i^{Met}$. D: upon hydrolization of GTP, the PIC is disassembled. The large ribosomal subunit, helped by the GTP-ase activity of EIF5B, joins the small subunit for initiation clearance and the first translocation. Translation of the ORF can now begin.

Among WBSCR genes, EIF4H and GTF2I are the two most expressed ones at the pluripotent state in WBS and 7dup patients (Adamo et al., 2014), and their expression closely mirrors gene dosage. It is then reasonable to expect, by means of their involvement in general pathways of transcription and translation, that these genes have a broad effect on gene expression from the very first stages of development.

Regulation of gene expression

The central dogma of molecular biology states that there is a unidirectional transfer of information from DNA to RNA to protein (Crick, 1970). A properly functioning cell (and even moreso a properly functioning organism) has to carefully control three aspects of this information transfer: *what* information is needed, *when* it is needed, and *how much* of it is needed. The implementation and maintenance of this control has been studied for decades under the name of "gene expression regulation".

The notion of regulation implies that a set of instructions is being used to steer the combination of the three aspects towards a specific goal, which can be a cell state, a cell type, its localization, the replicative status to name a few; borrowing from computer science, these sets of instructions have been defined as "programs".

In fact, accompanying the unidirectional flow of information defined by the central dogma, there is a multidirectional flow of information that integrates, among many other processes, feedback loops, intracellular and extracellular signals, covalent modifications and subcellular localization in order to ensure that RNA, proteins and metabolites are produced and placed according to the needs of every program.

Although for historical and technological reasons the transcription of RNA had been placed on the central stage for a long time, it has now emerged clearly how gene

expression programs are tightly regulated at every possible step in the flow of information transfer. We can in fact represent gene expression as a discrete number of processes, each of them encompassed by a particular set of molecular species, series of actors and peculiar mechanisms that identify a "mode of regulation" (fig. 6). As the very first studies on gene expression already pointed out (Jacob and Monod, 1961), each process is able to cross-talk and influence the regulation of the others, converging towards a cellular phenotype that, depending on the program, can be thought of as static (homeostasis, quiescence, etc.) or dynamic (differentiation, proliferation, apoptosis, etc.).

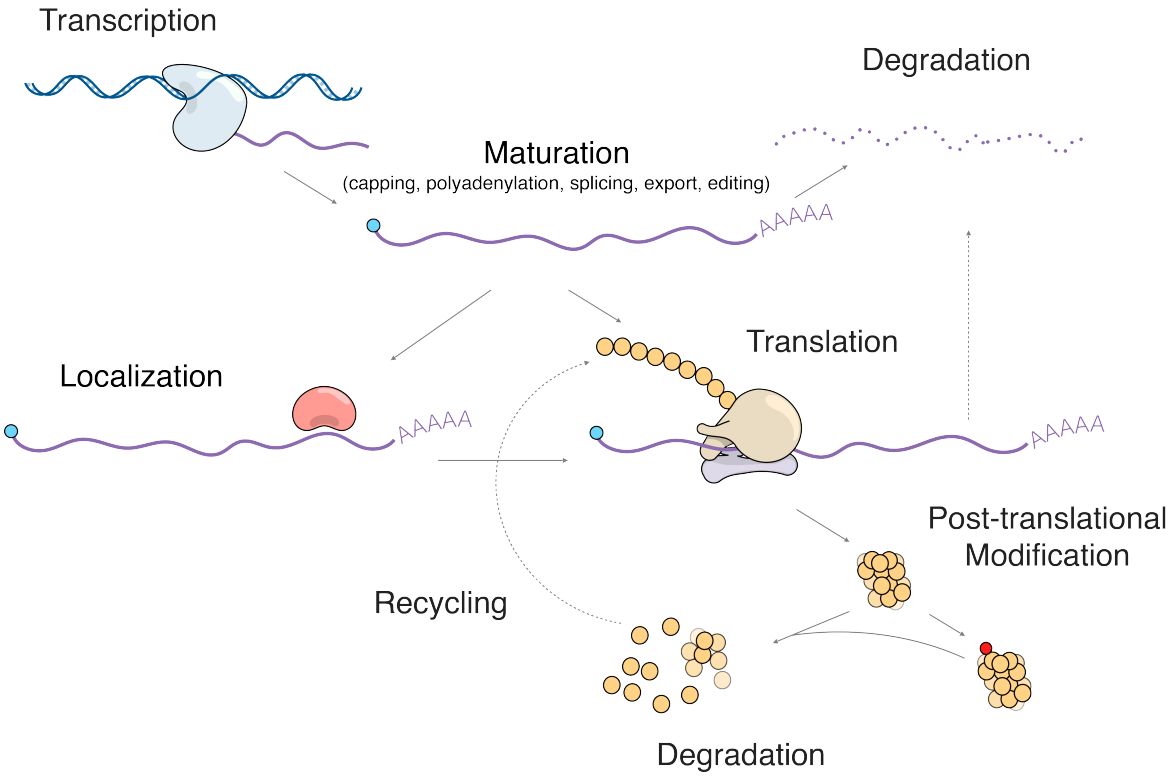

Figure 6: A simple model of gene expression.

I will touch briefly upon some of the main axes of regulation (transcription, translation and protein degradation), with a special focus on how they relate to neurodevelopmental and craniofacial disorders.

Transcriptional regulation

The regulation of transcription was initially studied as the interaction between transcription factors (TFs) and specific DNA sequences on the genome. TFs would bind these control elements and recruit the transcription machinery (composed by RNA polymerases and multi-subunit general transcription factors) on the transcription start site. Shortly after the initial description of transcriptional control by Jacob and Monod, studies predicted and then demonstrated a correlation between chromatin state (accessibility, presence of post-translational modifications, 3D organization) and the regulation of mRNA synthesis, leading to what is now the current consensus model (reviewed in Lee and Young, 2013). In this model, the chromatin state is deeply interconnected with the recruitment and function of transcription factors, with the initiation, pause or stalling of RNA polymerases and with transcriptional elongation. A "histone code" had been initially compiled according to which specific histone modifications (or "marks") correspond to different regulatory elements and to different activities of these elements (tab. 2). Chromatin remodelers, broadly classified as readers, writers and erasers, are able to decipher and modify this code by means of physical interactions with histone post-translational modifications. Depending on the mark, remodelers locally modify chromatin in order to compact or relax DNA packaging, or recruit factors for the establishment of long-range interactions between distal nucleosomes. Subsequent studies have demonstrated how the combinatorial nature of histone marks is not as simple as initially thought, revealing an intricate combination of modifications that confuted the existence of anything resembling a "code" (Wang et al., 2008) . However, functional genomics studies still make wide use of histone modifications as predictors of transcriptional activity, which become especially relevant in combination with high throughput RNA sequencing (RNA-seq) and ChIP-seq for chromatin remodelers with

a known activity. Histone modifications are highly dynamic and reversible, but they are heritable through cell divisions and, in some cases, through generations (Brykczynska et al., 2010; Hammoud et al., 2009)

Another reversible modification of DNA that is able to influence transcription regulation is the methylation of cytosines in CpG dinucleotides. Deposed by DNA methyltransferases (DNMTs), these covalent modifications cover roughly 70 to 80% of CpGs in vertebrates, with the exception of "CpG islands". In these loci the dinucleotide is over-represented and under-methylated, and they tend to cluster in the vicinity of promoters, suggesting an anti-correlation between methylation of CpGs near promoters and transcriptional activation. Hypermethylated promoters repress transcription by means of steric hindrance. The analysis of the genome-wide distribution of methylated CpGs reveals cell-type specific patterns (Deaton et al., 2011; Illingworth et al., 2008), which can be reshaped according to the expression program needed for another cell type. As already mentioned for histone marks, CpG methylation is heritable and is indeed responsible for the parental imprinting of genomic loci (Felsenfeld and Bell, 2000).

The set of reversible, heritable modifications in chromatin and DNA methylation are commonly referred to as "epigenetic marks"[7].

The history of the discovery of many of chromatin remodelers is deeply entwined with the history of developmental biology. Two key chromatin remodeling complexes, the Polycomb group (PcG) and the Trithorax group (TrxG) are named after *Drosophila melanogaster* developmental phenotypes caused by their loss of function (Ingham, 1983; Lewis, 1949), the homeotic transformation of body segments. Polycomb group proteins are assembled in repressive complexes that

[7] Although the term *epigenetics* is an ambiguous term that has gone through many rounds of re-definition (Meloni and Testa, 2014), mainly as a function of the technological advances that capture with increasing depth the identities and abundances of biomolecules, it is still tightly linked to transcriptional regulation through heritable, reversible chromatin and DNA modifications.

catalyze the ubiquitination of lysine 119 on histone H2A (performed by Polycomb Repressive Complex 1, PRC1) and the trimethylation of lysine 27 on histone H3 (performed by Polycomb Repressive Complex 2, PRC2). Conversely, Trithorax group proteins catalyze all the three methylations of lysine 4 on histone H3, which are marks that activate gene expression and are deposed at enhancers (H3K4me1), promoters (H3K4me3) and in broader regions of chromatin around enhancers (H3K4me2). These catalytic activities classify these two complexes as *chromatin writers*, which depose histone marks able to influence gene expression by means of steric hindrance, and, more importantly, by recruiting *chromatin readers*. Briefly, chromatin readers i) can induce architectural changes in chromatin (relaxation or compaction, looping, distal interactions), ii) can change nucleosome occupancy and expose or withdraw traits of DNA, iii) can mask nucleosomes to other readers or to RNA polymerase, iv) can propagate specific histone marks along loci by recruiting other writers or erasers, and v) can recruit other factors involved in transcription, replication or DNA damage repair. As briefly mentioned before, the existence of sharply distinct activating/repressing chromatin marks has been questioned by the experimental evidence on many combinations of marks that do not ascribe to the notion of an on/off-like switch of transcription. For instance, the presence of both H3K27me3 and H3K4me3, a repressive and an active mark, has been identified on the promoters of genes that are poised for activation in mouse embryonic stem cells (Bernstein et al., 2006). These "bivalent domains" allow for a more rapid activation or silencing of specific loci upon differentiation, thus shaping the transcriptome and ultimately cell identity.

The activity of RNA polymerase is controlled at the initiation of transcription and at its elongation. TFs bind DNA at specific loci and recruit cofactors, which in turn interact directly with the RNA polymerase or through the Mediator complex, a big

(1.3 MDa) multi-subunit complex that bridges the RNA polymerase and its associated GTFs with more distal transcription factors and cofactors. The RNA polymerase transcribes then a stretch of 25-50 nucleotides, after which it pauses (Rougvie and Lis, 1988). The release from the pause depends on the interaction with elongation factors that phosphorylate the RNA polymerase carboxy-terminal domain; if this clearance is not granted, the RNA polymerase will detach from DNA releasing a small fragment of RNA, otherwise it will go on transcribing until reaching a transcription termination sequence.

The regulation of transcription can be then summarized as the product of the concerted action of chromatin remodelers, which act through their modifications to influence the accessibility of DNA and its protein interactors, which in turn activate or repress the transcription of downstream sequences by positioning and modifying the RNA polymerase. A wealth of literature has demonstrated how the precise regulation of each of these steps is crucial for development. Naming a few cases: the removal of the PRC2 mark, catalyzed by the Jmjd3 demethylase, is essential for neuronal development (Burgold et al., 2008, 2012; see Testa, 2011 for a review); the deposition of activating marks by TrxG proteins is essential for gastrulation, spermatogenesis, and corticogenesis (Andreu-Vieyra et al., 2010; Glaser et al., 2009); bivalent domains in ES cells and in development have been extensively studied, with a special focus on neuronal development (Azuara et al., 2006; Bernstein et al., 2006); WICH complexes, containing BAZ1B, are involved in cardiovascular development (Han et al., 2011); histone demethylase KDM1A/LSD1, an interactor of GTF2I, regulates early stages of neuronal differentiation (Laurent et al., 2015). The number of evidences is unsurprisingly long, and steadily grows longer.

It is then interesting to note how many disorders affecting neuronal, cardiovascular and craniofacial development are caused by mutations in genes that are involved in

virtually all the steps of transcription regulation: readers and writers of DNA methylation, chromatin remodelers, specific and general transcription factors and RNA polymerase subunits (tab. 2).

As an increasing number of genome-wide association studies on trios and exomes of patients with autism spectrum disorder and intellectual disability are being performed and published, a staggering amount of genes is being associated to these conditions, prompting some researchers and clinicians to propose a genotype-based approach to classify ASD and ID, rather than the commonly used psychiatric categories.

Gene	Complex	Function	Associated syndrome	Clinical phenotypes	Inheritance	References
EED, EZH2	PRC2	Trimethylation of H3K27 (repressive)	Weaver Syndrome	ID, craniofacial dysmorphisms, overgrowth	Autosomal dominant	(Cohen et al., 2015; Gibson et al., 2012)
KANSL1	MLL1/NSL1	Acetylation of H4K5 and H4K16 (activating)	Koolen-de Vries syndrome	ID, craniofacial dysmorphism, epilepsy	Autosomal dominant	(Koolen et al., 2012)
KMT2D, KDM6A	COMPASS (KMT2D)	Mono-, di- and tri-methylation of H3K4 (KMT2D), demethylation of H3K27 (KDM6A) (activating)	Kabuki syndrome	ID, craniofacial dysmorphism, heart defects, genito-urinary abnormalities	Autosomal dominant	(Van Laarhoven et al., 2015; Miyake et al., 2013)
NSD1		Methylation of H3K36 and H4K20 (repressive or activating)	Sotos syndrome	ID, overgrowth	Autosomal dominant	(Höglund et al., 2003)
POLR1C, POLR1D	Subunits of RNA polymerases I and III	Core components of the transcriptional machinery	Treacher-Collins-Franceschetti Syndrome	Severe craniofacial dysmorphisms	Autosomal dominant (POLR1C), autosomal recessive (POLR1D)	(Bowman et al., 2012; Vincent et al., 2016)
ADNP	SWI/SNF	Chromatin remodeler and transcription factor	Helmsmoortel-van der Aa syndrome, ADNP-associated autism spectrum disorder	ID, ASD, craniofacial dysmorphism	Autosomal dominant	(Helsmoortel et al., 2014)
CHD7, CHD8		Chromatin remodeler with repressive function	Autism spectrum disorder, CHARGE syndrome	ID, ASD, craniofacial dysmorphism, heart defects	Autosomal dominant (CHARGE)	(Bernier et al., 2014; Lalani et al., 2006)
BAZ1B	WICH, B-WICH	Chromatin remodeler involved in transcriptional regulation, DNA damage response, DNA replication	Williams-Beuren syndrome, 7q11.23 duplication syndrome	ID, ASD, craniofacial dysmorphism, heart defects, genito-urinary anomalies, metabolic anomalies	Autosomal dominant	(Pober, 2010a)
GTF2I	In complex with HDAC2 and LSD1	General transcription factor	Williams-Beuren syndrome, 7q11.23 duplication syndrome	ID, ASD, craniofacial dysmorphism, heart defects, genito-urinary anomalies, metabolic anomalies	Autosomal dominant	(Pober, 2010a)
MED12	Mediator	Bridging of transcription factors/cofactors and RNA polymerase	Lujan-Fryns syndrome	ID, craniofacial dysmorphism, heart defects	X-linked dominant	(Schwartz et al., 2007)
ARID1A, ARID1B, SMARCA4, SMARCB1, SMARCE1	SWI/SNF	Chromatin remodeler complex with ATPase activity	Coffin-Siris syndrome	ID, craniofacial dysmorphism, heart defects, genito-urinary anomalies	Autosomal recessive	(Santen et al., 2012; Tsurusaki et al., 2012)
DNMT3B		Methylates CpG dinucleotides	Immunodeficiency, Centromere instability and Facial anomalies (ICF) syndrome	Craniofacial dysmorphism, immunodeficiency, predisposition to cancer	Autosomal recessive	(Xu et al., 1999)
HEMT1	E2F6	Mono- and di-methylation of H3K9 (repressive)	Kleefstra syndrome	ID, severe speech impairment, craniofacial dysmorphism	Autosomal dominant	(Kleefstra et al., 2012)
MECP2		Binding of methylated CpG	Rett syndrome	ID, ASD, microcephaly, epilepsy	X-linked dominant	(Amir et al., 1999; Wan et al., 1999)

Table 2: Transcription regulators involved in developmental disorders.

Translation regulation

The regulation of protein synthesis represents another crucial step of gene expression control. Cells regulate translation to rapidly adapt to extrinsic and intrinsic cues, to maintain homeostasis, and to finely tune gene expression programs in a spatio-temporally regulated fashion. Translation control was initially characterized in biochemistry and molecular biology experiments on cells and cell-free extracts, resulting in landmark discoveries on the function and structure of ribosomes (Ban et al., 2000), the identification of translation initiation, elongation and termination factors (Pestova and Kolupaeva, 2002), and the circular topology of actively translated mRNA (Wells et al., 1998). Subsequent technological advances have shed light on the complexity and the variety of protein and RNA factors that exert this regulation (fig. 7). As mentioned earlier, initiation is regarded as the chief regulatory locus of translation, being the rate-limiting step for protein synthesis kinetics.

The cap-binding protein EIF4E recognizes the 5' 7meG cap of mature mRNAs and recruits the EIF4F complex, composed by EIF4G and the mRNA DEAD-box helicase EIF4A. EIF4A is an ATP-dependent helicase tasked with unwinding the secondary structures of the 5' untranslated regions; its directionality and processivity are enhanced by EIF4H and EIF4B as already described in the first section of the introduction. EIF4G interacts with the poly-A binding protein PABP, bound at the 3' of the mRNA, in order to create a closed-loop conformation. This conformation enhances the binding of mRNA to the PIC, a multi-subunit complex composed by the 40S small ribosomal subunit, initiation factors 1, 1A, 2 and 3, the initiator Met-tRNA, and a GTP molecule bound to eIF2. The PIC scans the 5' untranslated region until the Met-tRNA matches the AUG start codon: upon this interaction, scanning is halted and

GTP is hydrolized to GDP, thus stimulating the disassembly of the PIC and leaving room for the large ribosomal subunit 60S to constitute the 80S initiating ribosome.

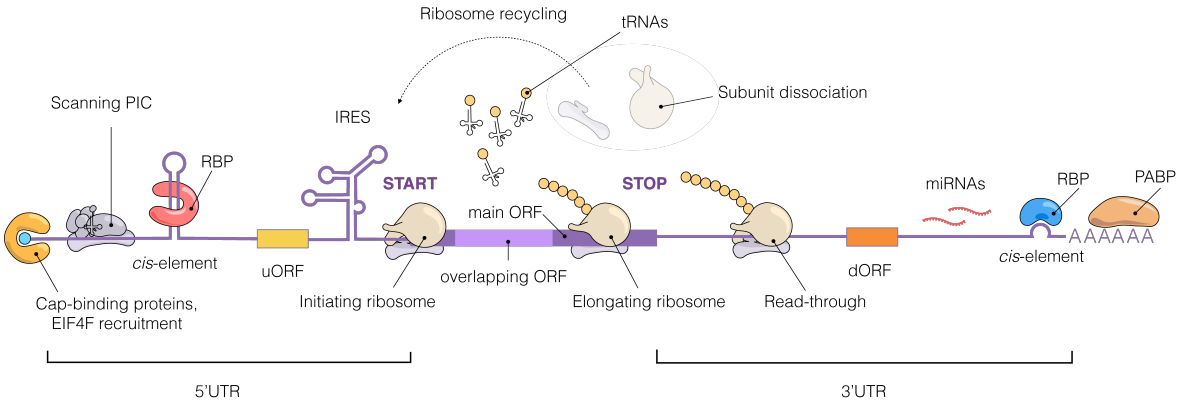

Figure 7: Schematic representation of the main cis- and trans-acting elements involved in translation regulation. UTR: untranslated region; PIC: pre-initiation complex; RBP: RNA-binding protein; uORF: upstream open reading frame; IRES: internal ribosomal entry site; dORF: downstream open reading frame; RBP: RNA-binding protein; PABP: poly-A-binding protein.

The initiation step can be regulated at many different roadblocks. EIF4E can be inhibited by the hypophosphorylated form of the EIF4E-binding protein 1 (EIF4EBP1). Stimuli such as insulin signaling and nutrient starvation activate the mTOR pathway which ultimately results in the phosphorylation of EIF4EBP1, its release from EIF4E and the assembly of the EIF4F complex. A recent report (Lee et al., 2016) identified EIF3D as a novel cap-binding protein that is required for the translation of specific mRNAs involved in cell proliferation and development and competes with EIF4E. EIF4B can be phosphorylated by the S6K kinase as a result of mTOR (mammalian Target Of Rapamycin) pathway activation (Dennis et al., 2012b), thus enhancing 5'UTR unwinding. mRNA helicases such as DDX3X or DHX29 can assist the unwinding by directly binding and relaxing secondary structures. 5'UTR secondary structures can also be recognized by RNA-binding proteins in order to repress translation. Upstream AUGs and upstream open reading frames (uORFs) can be recognized by the PIC and trigger initiation. Upstream AUGs and uORFs are considered to act mostly by negative regulation of translation in non-stressful conditions, by creating a decoy initiation that greatly reduces translation of the main

ORF, or recognizing a premature stop codon that initiates nonsense-mediated decay (NMD) of the transcript (reviewed in Barbosa et al., 2013). Conversely, in stressful conditions such as starvation, uORFs can enhance the translation of the main ORF by leaky scanning and bypass of the premature stop codon, thus reducing NMD and increasing protein synthesis. Although uORFs have been detected in 49% of the human transcriptome (Calvo et al., 2009), they are over-represented in genes involved in differentiation and proliferation; however, very few experimental evidences have been provided for the mechanisms of translation regulation via uORF. Complex secondary structures such as Internal Ribosomal Entry Sites (IRES) can be directly recognized by ribosomes, bypassing cap-dependent initiation. IRESs were initially identified in viral transcripts as a means to increase the translation without depending on the host eIFs. However, many studies have demonstrated how IRESs are also present in a large fraction of human transcripts, and are used in the contest of survival to stress (Schepens et al., 2005). The recognition of the AUG start codon is influenced by *cis*-elements, such as a favorable sequence context (the Kozak sequence or other cognate sequences), and by eIFs that are able to control for "poor AUG contexts" (eIF1, eIF2, eIF3; reviewed in Hinnebusch 2011) and increase the fidelity of the recognition of proper start sites. The clearance of initiation is heavily dependent on eIF2 and the hydrolysis of its associated GTP molecule. When the EIF2 subunit α is unphosphorylated, it can be bound by its guanine nucleotide exchange factor (GEF), eIF2B, that is able to replace GDP with GTP and thus allow for a new cycle of translation initiation. Phosphorylated EIF2α increases the affinity for EIF2B and sequesters it, effectively abolishing the GEF activity on EIF2 itself and therefore repressing translation. EIF2 phosphorylation is, as for many other mechanisms of translation regulation, the result of nutrient or oxidative tress.

The elongation phase starts with the initial translocation from the AUG (fig. 8). Initiating ribosomes host the initiator Met-tRNA in the P site. The second tRNA enters the A (aminoacyl) site, so that the peptidyl transferase activity of the 60S subunit can catalyze the formation of the first peptidic bond. The first Met is then relocated in the E (exit) site by translocation of the ribosomal subunits, and the initiator tRNA is released. During elongation, aa-tRNAs enter the A site and are stabilized by codon-anticodon pairing with the mRNA. The A-site aminoacid then forms a peptidic bond with the aminoacid positioned in the P site; ribosomal subunits translocate along the mRNA so that aminoacids follow an A-P-E site procession. This process is controlled by elongation factors EEF1 and EEF2. EEF1, in complex with GTP, delivers aa-tRNAs and is released upon GTP hydrolisis, caused by a conformational change in the ribosome. EEF2 acts as a translocase, coordinating the movements of tRNA, the two ribosomal subunit and the mRNA. EEF2 is fundamental for dictating the codon-wise pace of translation, which must be tightly regulated to avoid shifts in the reading of the ORF (Kaul et al., 2011). Recent data on ribosome profiling experiments revealed that the average speed of elongation is ~6 aminoacids per second (Ingolia et al., 2011). The elongation rate is also controlled by the codon optimality of conding sequnces. Recent work in yeast has shown non-optimal codons slow down translocation and increase mRNA degradation, unveiling a previously unrecognized layer of gene expression control at the level of elongation (Presnyak et al., 2015).

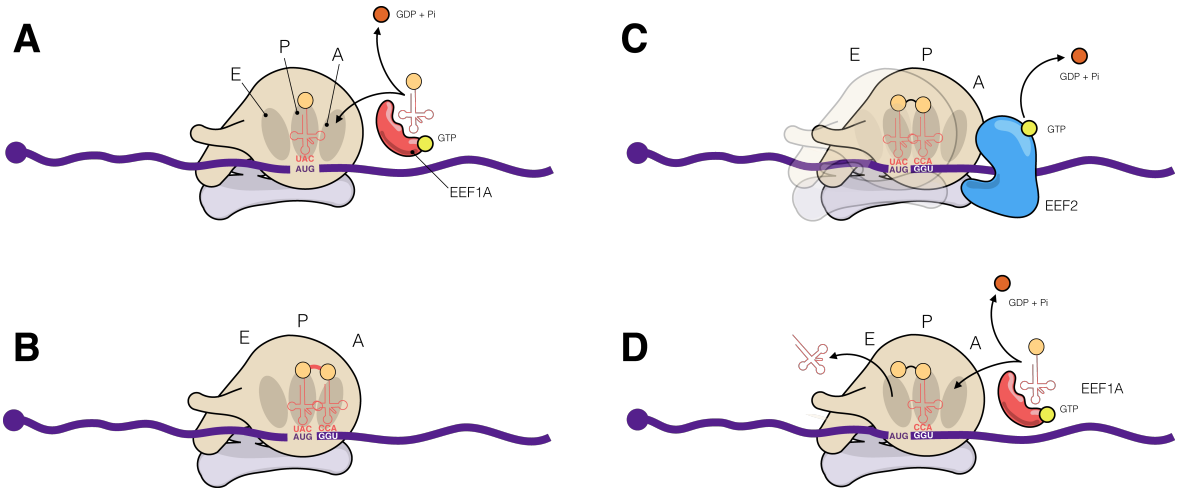

Figure 8: schematic representation of translation elongation. A: the first elongation step consists in the insertion at the A site of the first aa-tRNA after the initiator, mediated by the elongation factor EEF1A via a GTP-dependent mechanism. B: Once the aa-tRNA is accommodated in the A site, the peptidyltransferase activity of the 28S ribosomal subunit forms a peptidic bond between the two aminoacids. The two ribosomal subunit translocate along the mRNA and shift the positions of the tRNAs in the P and E sites, with the assistance of EEF2. D: immediately after translocation the E-site tRNA is discharged and another cycle of elongation begins.

The maintenance of the reading frame is important to ensure the fidelity of translation. However, alternative, overlapping ORFs can be encoded in the same main ORF and can be used by regulating the action of elongation factors, as in the case of many viral transcripts and, in animals, of ornythine decarboxylase (Shantz and Pegg, 1999). This enzyme catalyzes a key step of polyamine synthesis. When intracellular polyamine levels are high, they are able to induce a +1 ribosomal frameshifting that replaces the enzyme ORF with the reading frame for the antizyme, a polypeptide that th enzyme and targets it for ubiquitin-mediated degradation. This is another example of how translation regulation, in this case at the elongation step, allows to control for quick changes in the environment and to adapt the gene expression program accordingly. Upon reaching the stop codon, ribosomal subunits are relieved by the entry in the A site of class I release factors, which hydrolize the aa-tRNA bond of the last aminoacid in the P site and release the complete polypeptide. The two subunits are then dissociated and get recycled for subsequent rounds of translation.

Failure to recognize the stop codon results in read-through translation, by which proteins gain an extended C-terminus which alters their function causing abnormal aggregation or localization. Several safeguards are in place to avoid this outcome: if read-through translation encompasses the whole 3'UTR until the poly-A tail, mRNA and protein undergo degradation via nonstop-mediated decay (van Hoof, 2002). If, however, another stop codon is encountered before the poly-A tail, proteins with extended C-termini may still be produced. A recent study has shown how these extended termini enhance the instability of the protein, thus blunting the detrimental effects of read-through translation (Arribere et al., 2016).

A fraction of 3'UTR regions can also harbour downstream ORFs, which are more evolutionarily conserved than their neighboring untranslated sequences, but are translated at a much lower efficiency than their upstream counterparts (Ji et al., 2015). The function of dORFs is still unknown.

The 3' UTR has been studied extensively as the region hosting most of the post-transcriptional regulatory elements: miRNAs and RBPs bind specific sequences and secondary structures in order to exert spatio-temporal control of translation. MiRNAs bind target sequences by complementarity and regulate translation in mainly three ways: 1) they recruit the RISC (RNA-Induced Silencing Complex) that initiates endonucleolytic cleavage of the mRNA, which is rapidly degraded. This pathway, however, requires perfect pairing between miRNA and target mRNA, a very rare occurrence in mammals. 2) they deadenylate the polyA tract, making the target mRNA less stable and less efficiently translated. 3) they interact with components of the eIF4F complex, greatly reducing translation initiation (Meijer et al., 2013). The expression programs of miRNAs are as crucial as those of mRNAs, since they are heavily interweaved and represent an additional layer of control of the information flow, and are tightly regulated in an evolutionarily conserved fashion.

RBPs interact with the 3'UTR to influence the subcellular localization of mRNAs and also their translation status, acting as sensors of environmental hazards or even of defects in translation efficiency. Stalling initiation at various steps results in the formation of repressive ribonucleoprotein complexes termed "stress granules" (reviewed in Buchan and Parker, 2009), which can also include the mRNA decay machinery (in which case they are termed "P-bodies"). The architecture of the translation machinery is kept intact so that, upon disassembly of the granule, protein synthesis can rapidly continue. As in the case of chromatin remodelers, the first discoveries of this class of translation regulators were made in *Drosophila* mutation screens. In the earliest stages of *Drosophila* embryo development, the morphogens Bicoid and Pumilio bind the mRNAs of *caudal* and *hunchback* to ensure the initial fate specification of frontal and posterior segments by repressing their translation (Murata and Wharton, 1995; Rivera-Pomar et al., 1996). The *Drosophila* RBP UNR regulates X-linked inactivation in males by repressing the translation of male-specific lethal 2 (*msl-2*)(Abaza, 2006). In mammals, *PRM1* and *PRM2* mRNAs, encoding for sperm-specific protamines (the functional equivalent of histones), are kept in a repressed state by the TB-RBP (testis-brain RBP) protein, which binds to their 3'UTR until spermatogenesis reaches an advanced stage. TB-RBP is also expressed in the brain, where it is involved in translation repression and transcript localization (Han et al., 1995; Wu and Hecht, 2000). Perhaps the most interesting example for this study is FMRP, the protein econded by *FMR1*, whose mutation causes fragile-X syndrome. FMRP is localized in the postsynaptic space of neuronal dendrites, together with synapse-specific mRNAs. Upon activation of the metabotrope glutamate receptor (mGluR), FMRP is phosphorylated and binds to the 3'UTR and 5'UTR of mRNAs coding for the AMPA receptor (AMPAR). FMRP inhibits their translation in two ways: it establishes a physical, inhibitory interaction with eIF4E (bridged by

CYFIP1)(Napoli et al., 2008) and it recruits RISC (Didiot et al., 2009). mRNAs and proteins are then assembled in a repressive ribonucleoproteic complex, akin to ribosomal stress granules. This regulation finely tunes the presence of AMPA neurotransmitter receptors in synapses and heavily influences the connectivity of neurons. When other ligands, such as the brain-derived neurotrophic factor (BDNF), reach the synapse, a signalling cascade results in the dephosphorylation of FMRP, disassembly of the mRNP and restart of translation (Bear et al., 2004). Loss of function of *FMR1* results in the abrogation of this activity-dependent control of translation at the synapses and therefore in aberrant connectivity. A link with dysregulation of translation an developmental disorders, especially ASD, is warranted by a series of evidences, both in animal models and in human syndromes (tab. 3A and 3B). Apart from all the insights gained from fragile X syndrome, two recent studies have shown ASD-like features in mice models in which *Eif4e* expression was constitutively upregulated (Gkogkas et al., 2013; Santini et al., 2012), pointing to a "hyper-connectivity" phenotype in brain neurons and to an imbalance between excitatory and inhibitory neuronal activities. One of the studies has also shown that the behavioural phenotype could be rescued in adult mice with a drug targeting Eif4e, hinting at the possibility that translation could be a treatable axis of regulation for ASD patients. Upstream negative regulators of the mTOR pathway such as TSC1/2, NF1 and PTEN cause, when mutated, syndromes that include ASD among their clinical features (see table 3B)(Kumar et al., 1995; van Slegtenhorst et al., 1997), albeit with different penetrance.

Gene	Complex	Function	Associated syndrome	Clinical phenotype	Inheritance pattern	References
FMR1	CYFIP1-EIF4E-FMR1	Translation repressor	Fragile X syndrome	ASD, ID, craniofacial dysmorphisms,	X-linked dominant	(Knight et al., 1993)
EIF4H	Interacts with the EIF4F complex and mRNA	Translation initiation factor	Williams-Beuren syndrome, 7q11.23 duplication syndrome	ID, ASD, craniofacial dysmorphism, heart defects, genito-urinary anomalies, metabolic anomalies	Autosomal dominant	(Pober, 2010)
NSUN5		Methylates 28S rRNA possibly changing translation programs	Williams-Beuren syndrome, 7q11.23 duplication syndrome	ID, ASD, craniofacial dysmorphism, heart defects, genito-urinary anomalies, metabolic anomalies	Autosomal dominant	(Pober, 2010)
WBSCR22		Methylates 18S rRNA, necessary for ribosome biogenesis	Williams-Beuren syndrome, 7q11.23 duplication syndrome	ID, ASD, craniofacial dysmorphism, heart defects, genito-urinary anomalies, metabolic anomalies	Autosomal dominant	(Pober, 2010)
SBDS		Triggers release of 60S pre-ribosomes	Schwachman-Diamond syndrome	Skeletal abnormalities, heart defects, speech impairment, pancreas and bone marrow abnormalities	Autosomal dominant	(Boocock et al., 2002)
PUS1		Converts uridine in pseudouridine during tRNA maturation	Myopathy, lactic acidosis, and sideroblastic anemia 1 (MLASA1)	Moderate ID, myopathy, anemia, skeletal abnormalities	Autosomal recessive	(Bykhovskaya et al., 2004)
TCOF1	In complex with NOLC1, CUL3 and KBTBD8	Covalently modifies ribosomes to enhance translation of specific mRNAs	Treacher-Collins-Franceschetti syndrome	Severe craniofacial dysmorphisms	Autosomal dominant	(Bowman et al., 2012)

Table 3A: Translation regulators involved in developmental disorders

NF1		RAS GTPase activating protein, inhibits mTOR pathway	Neurofibromatosis – Noonan Syndrome	ASD, craniofacial dysmorphisms, heart defects, skin and connective tissue abnormalities	Autosomal dominant	(Bertola et al., 2005)
PTEN		Antagonist of PI3K, inhibits mTOR pathway	PTEN hamartoma tumor syndrome (PHTS)	ASD, macrocephaly, predisposition to tumor formation	Autosomal dominant	(Varga et al., 2009)
TSC1, TSC2		Inhibits mTORC1	Tuberous sclerosis complex	ASD, ID, predisposition to benign tumors	Autosomal dominant	(Kumar et al., 1995; van Slegtenhorst et al., 1997)

Table 3B: Upstream regulators of mTOR involved in developmental disorders.

The mTOR pathway is a central node that integrates external signals of differentiation, growth, inflammation, environmental stress and acts on translation mainly by positively regulating the eIF4F complex activity (reviewed in Hay and Sonenberg, 2004). This is an additional line of evidence that converges towards the lack of negative regulation of translation, or excess thereof, as one of the possible mechanisms underlying ASD.

Taken together, the phenotypes of both patients and animal models point to translation regulation as a process that is able to greatly influence development by modifying the response to external cues or by altering protein abundance homeostasis in a cell-intrinsinc manner, with particular regard to its spatial and temporal regulation. Moreover, since protein levels have undergone a greater selective pressure than their corresponding transcripts during evolution (Khan et al., 2013), translation regulators may also be extensively involved in buffering protein levels, serving as a safeguard for abnormal fluctuations in protein abundance that may have a transcriptional origin.

Regulation of protein degradation

Spatio-temporal regulation of protein abundances is not only carried out by increasing or decreasing their production rates, but also by actively dissociating them into their component aminoacids. Degradation is a fundamental process for maintaining cellular homeostasis, controlling the quality of protein folding, presenting antigens on the cell surface and enabling modifications to cell state and cell fate, such as cell division and differentiation. Initially thought to occur exclusively in specialyzed organelles called lysosomes, it is now understood as a multi-tiered process carried out mainly by the Ubiquitin-Proteasome system (UPS) (reviewed in (Glickman and Ciechanover, 2002)). Other processes such as chaperone-mediated degradation of misfolded proteins and autophagy of long-lived proteins and

organelles can be coupled to ubiquitination, but also exist as ubiquitin-independent pathways. Ubiquitin is a small (76 aminoacids) protein that is covalently attached to specific lysine residues of proteins by E3 ubiquitin-protein ligase enzymes, often as poly-ubiquitin chains. Polyubiquitinated proteins were initially thought to be targeted for degradation by the multi-subunit 26S proteasome, but further studies in the biochemistry of ubiquitin, which has now the dignity of a field *per se*, have revealed the existence of a complex "ubiquitin code" that depends on the length, complexity and geometry of polyubiquitin chains. Monoubiquitination of lysines is indeed a signal that regulates the recruitment of interactors of the tagged protein rather than its degradation; however, it accounts for a minority of ubiquitinations (possibly also because it is a rather small modification to measure by standard biochemical and even mass-spectrometry methods). In many cases the first ubiquitin is further tagged with other ubiquitin molecules on any of its 7 lysines (K6, K11, K27, K29, K33, K48 or K63), giving rise to a wide variety of combinations of polyubiquitin chains – termed after the residues they are linked to – with different functional outcomes (reviewed in Yau and Rape, 2016). Fast protein degradation is mostly achieved through K48- and K11-polyubiquitin tags (Chau et al., 1989; Meyer and Rape, 2014).

Ubiquitination is carried through a 3-component system. E1 ubiquitin activating enzymes consume an ATP molecule to create a highly reactive, high-energy thioesther bond between ubiquitin and a cysteine residue. Ubiquitin is then transferred to an E2 ubiquitin conjugating enzyme on another cysteine residue with a similar thioesther bond. E2-Ub then interacts with E3 ubiquitin-protein ligases that confer the actual substrate specificity, bridging E2-Ub with target proteins and allowing the final transfer of ubiquitin. This tiered system has an increasing complexity: 9 genes enconding E1 enzymes exist in humans, whereas the repertoire is slightly larger for

E2 enzymes (41 genes in humans) and maximally expanded for E3 ligases (2821 genes in humans), so as to cope with the diversity of substrates that must be recognized and targeted. Targeted proteins are recognized by the 26S proteasome, a large complex composed of a 20S cylinder in which the proteolytic activity takes place, enclosed by two 19S regulatory subunits. The proteasome dissociates ubiquitin from its targets, thus allowing it to be recycled by other E1 enzymes, and cleaves targeted proteins in small peptides that can be further dissociated in single aminoacids by non-specific peptidases.

Defects in protein degradation have been associated with several diseases, especially those arising with cytotoxic protein aggregation such as Parkinson's disease. In fact, mutations in *PARK2* (parkin), an multi-subunit E3 ligase, cause familial juvenile parkinsonism (Tanaka et al., 2004). Parkin fails to ubiquitinate target proteins that accumulate in dopaminergic neurons, thus causing cytotoxicity and neurodegeneration.

Very few cases of developmental disorders have been linked to mutations of genes in the UPS or other protein degradation pathways, the most notable of which is Angelman syndrome (AS). AS is a neurodevelopmental disorder caused by the deletion or mutation of the maternal, non-imprinted copy of UBE3A, an E3 enzyme. UBE3A is paternally imprinted and is not expressed in the brain, thus making the maternal copy the only one active. AS patients have a severe intellectual disability, speech impairment, hypotonia, craniofacial dysmorphisms and a conspicuously happy disposition (reviewed in Clayton-Smith and Laan, 2003). Two genes of the WBSCR, *TRIM50*, an E3 enzyme, and *VPS37D,* a member of the ESCRT complex involved in the removal of poly-ubiquitinated proteins, are involved in the UPS pathway (table 1), pointing to a possible role in these pathologies for the dysregulation of the degradation of specific targets during development.

Somatic cell reprogramming as a platform for disease modeling

The molecular and cellular basis of human genetic diseases have been historically modeled using two approaches, which can be roughly subdivided in *in vivo* and *in vitro*. On one hand, animal models were engineered with mutations in orthologs of disease-causing genes. These models allow to observe the effect of mutations in a complex, developing system, in which disease phenotypes can be scored at various scales, from systems genomics levels, to molecular and subcellular features, up to behavioural patterns. The evolutionary conservation of sequences and pathways allows to greatly enhance our ability to test hypotheses in different complex model systems, although mice models have been the most widely used so far. Moreover, the use of syngenic animals allows to isolate the specific contribution of a gene, or a group of genes, to a developmental phenotype without the confounding effect of the genetic background. Animal models have been and still are an invaluable source of information on disease and development.

However, many important differences must be taken into account when trying to port findings in mice to humans. Besides obvious macro-differences in reproductive, physical and cognitive abilities, which pose a great challenge for the modeling of speech impairment, sociality and learning, other less evident differences mar the attempts at painting the complete picture of human disease by exclusively using mice models. From a genetic point of view, up to 20% of essential human genes can be disrupted in mice without causing lethality (Liao and Zhang, 2008). Neuroanatomy and neurophysiology of mice differ from those of humans, as the human brain is gyrencephalic (i.e. presents convolutions of the cortex) whereas the mouse brain is lissencephalic (no convolutions), possibly as a secondarily acquired trait during

evolution rather than the maintenance of an ancestral one (Kelava et al., 2013). As stated above, the syngeny of inbred litters can be a great advantage when trying to isolate the contribution of genes to developmental features, but does not recapitulate the genetic reality of the vastly outbred human population, which is for instance reflected in the variable penetrance and expressivity of phenotypes associated to 7q11.23 CNVs. The other approach, instead, consists in using easily accessible patient-derived cells, such as fibroblasts or Epstein-Barr Virus (EBV)-immortalized lymphoblasts. By performing comparative genomic analyses on cells derived from patients and from healthy controls, one can infer disease-specific genetic signatures that arise from their genotype, in a human genomic context. The main disadvantage of this approach is that, due to their tissue of origin, tissues used in these studies (Antonell et al., 2010b; Henrichsen et al., 2011) are hardly affected in their phenotypic manifestations by developmental disorders, thus effectively decreasing their informative power regarding disease-specific alterations of developmental trajectories. The ground-breaking technology of somatic cell reprogramming managed to finally overcome these limitations by bridging the boundaries of human genotypes and their developmental phenotypes.

The seminal work of Sir John Gurdon with *Xenopus laevis* (Gurdon, 1962; Gurdon et al., 1975) demonstrated that the fate acquired by terminally differentiated cells was reversible. By fusing enucleated oocytes of a tadpole with the nucleus of gut cells of another – a procedure termed somatic cell nuclear transfer (SCNT) - he showed that the somatic nucleus could be "reprogrammed" to the totipotent state, insofar as being able to give rise to a new individual with exactly the same genome of the nucleus donor tadpole. It was the first proof that cells with a high degree of potency harboured some elements in the cytoplasm – RNA or proteins – which could act in a dominant fashion over the epigenome of the donor nucleus (fig. 9). This plasticity of

cell fate was showed to be conserved in mammals, albeit SCNT was not as easily attainable: initial reprogramming was obtained only by cell fusion of pluripotent cells with somatic ones (Kahan and Ephrussi, 1970), but it was eventually accomplished in the famous case of the SCNT cloning of Dolly the sheep (Wilmut et al., 1997), thus definitively opening the door to the new paradigm of therapeutic (and reproductive) cloning. The understanding of what *brings* cells back to pluripotency came through the understanding of what *keeps* them in the pluripotent state, and how to isolate them. This quest had started shortly after Gurdon's work, with the isolation of pluripotent cells from teratocarcinomas, benign tumors that can differentiate in the three germ layers (Kleinsmith and Pierce, 1964). It was however only with the pioneering work of Martin Evans and Matthew Kaufman that mouse embryonic stem cells (ESCs) were isolated from the inner cell mass of blastula-stage embryos and stabilized in culture (Evans and Kaufman, 1981), a feature that was reproduced in the human setting by James Thomson using the inner cell mass of blastulae coming from discarded in-vitro fertilization attempts (Thomson et al., 1998). Cell fusion experiments performed using mouse ESCs and differentiated cells showed, as John Gurdon did, that pluripotency is a dominant state over differentiation (Tada et al., 2001); these findings were also confirmed in humans (Cowan, 2005) (fig. 9). Another important field of stem cell biology had been then reactivated in the 90s, with landmark contributions from Austin Smith, Hans Schöler and colleagues, who identified and characterized the proteins that were responsible for the maintenance of pluripotency in both mice and humans: Oct4 (Nichols et al., 1998), Sox2 (Avilion et al., 2003), Nanog (Chambers et al., 2003), Klf4 (Li et al., 2005), c-Myc (Cartwright et al., 2005) to name a few. It is important to note that the vast majority of these genes code for transcription factors that instruct a tightly regulated pluripotency expression program.

The revolutionary intuition of Shinya Yamanaka, scantily described in the introduction to his landmark 2006 paper (Takahashi and Yamanaka, 2006), was that the ectopic over-expression of these pluripotency-maintenance proteins could revert the differentiated state of somatic cells and re-establish pluripotency. In this work, Takahashi and Yamanaka defined the minimal set of factors required to achieve this resetting: four transcription factors (Oct4, Sox2, Klf4 and c-Myc) which were aptly termed "Yamanaka factors". Cells in which these factors were able to reinstate pluripotency were named induced pluripotent stem cells (iPSCs). Not only iPSCs display morphology, replicative ability and expression profiles very similar to those of ESCs, but they have the same degree of developmental potential, as they can give rise to all three germ layers and, when implanted in a blastula, can contribute to a chimera up to a certain extent (Takahashi and Yamanaka, 2006). A year after, Yamanaka demonstrated that reprogramming could be achieved in human cells as well (Takahashi et al., 2007), thus revolutionizing the field of human disease modeling. In fact, since reprogramming protocols and pluripotency culture conditions have been perfected, it is now possible to harvest skin, blood or dental pulp biopsies from patients and derive several iPSC lines from them. These iPSCs have all the genetic information of their donors, with the added advantage of having the intrinsic potential to recapitulate the development of virtually all tissues in the body. In other words, they grant a privileged access to two axes: space - tissues that would be otherwise almost physically impossible to reach, such as neurons of the cortex or smooth muscle cells of the heart, and time – the witnessing *in vitro* of their developmental stages, making effectively visible what would be otherwise invisible (Nowotny and Testa).

After the pluripotent state was isolated, a part of the stem cell community set out to identify and standardize efficient and reproducible ways to derive differentiated

tissues. Among them, the derivation of neurons has been assiduously pursued, as a way to obtain cells that would serve as disease models, drug-screening platforms and cell replacement therapy for neurodegenerative and neurodevelopmental disorders (fig. 9).

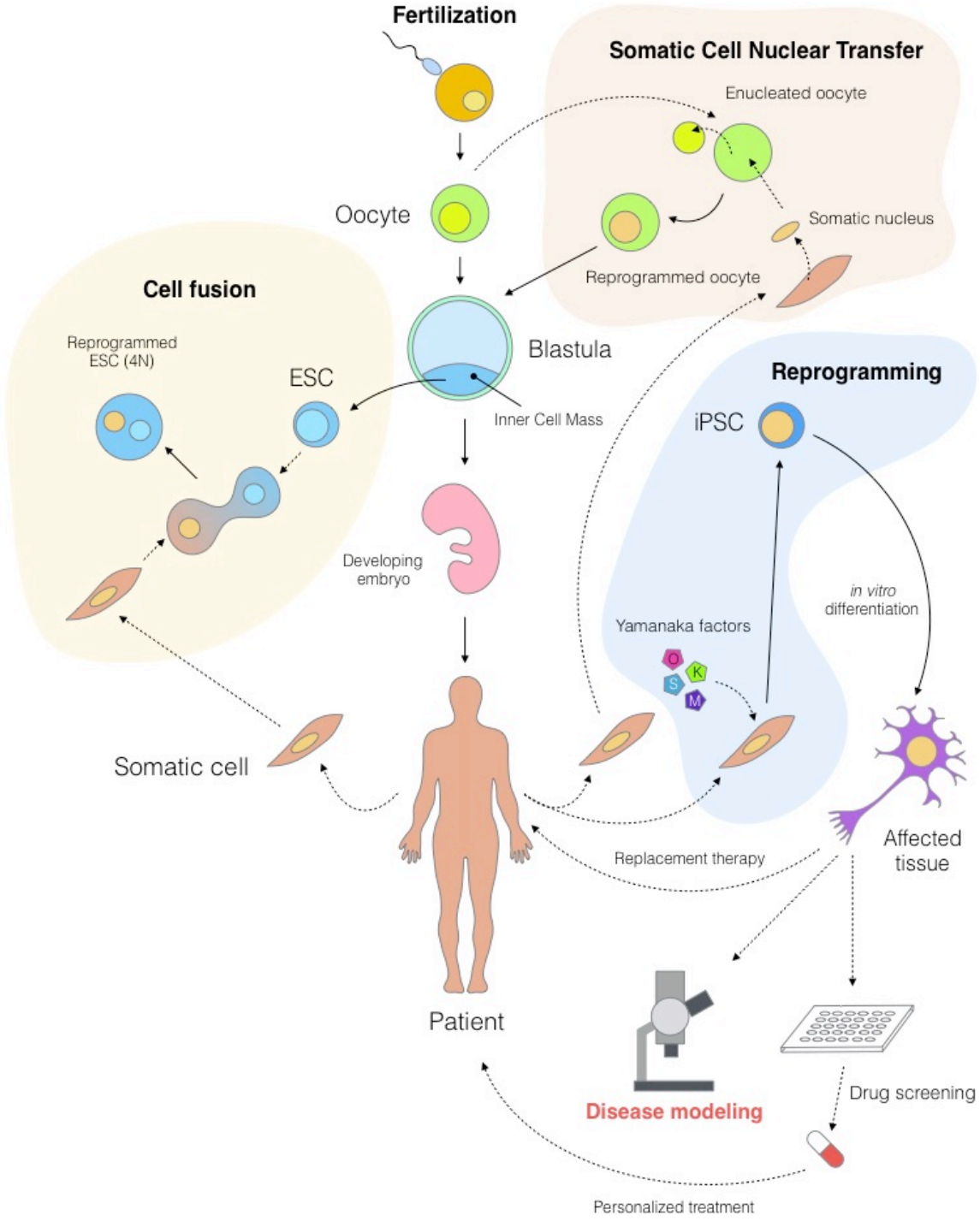

Figure 9: Pathways to pluripotency and their potential as a biomedical platform. Black solid lines indicate changes in potency, while dotted lines indicate experimental procedures.

Systems based on small-molecule inhibition of signaling pathways have been proposed to obtain monolayer cultures of neural progenitors (Chambers et al., 2012), cortical glutamatergic neurons (Shi et al., 2012a, 2012b), dopaminergic neurons (Kriks et al., 2011) and interneurons (Nicholas et al., 2013). Neurons obtained with these protocols can be probed with all the set of techniques that are normally employed in cellular neuroscience, such as morphological analysis by imaging, analysis of connectivity via multi-electrode arrays, study of electric currents and calcium influx, and many more. Moreover, they can be transplanted in mice brains to assess their ability to integrate and form physiologically relevant interactions in a more controlled model.

These protocols have offered great insights into the pathological phenotypes of diseases such as schizophrenia (Brennand et al., 2011), Rett syndrome(Marchetto et al., 2010), Timothy syndrome (Pasca et al., 2011), Alzheimer's disease (Kondo et al., 2013) to name a few. However, these systems display a high variability across lines, require long-term culturing and several manipulations and are, sometimes, difficult to reproduce. An alternative is represented by the ectopic overexpression of master regulators of lineage commitment, such as neurogenin-2 (Ngn2) (Zhang et al., 2013). By infecting iPSCs with a lentiviral vector which carries an inducible Ngn2 cDNA coupled with an antibiotic resistance, cells can be readily converted to a cortical neuronal fate in as little as 21 days, and only require very simple manipulation steps. Moreover, the antibiotic resistance allows to select for the activation of the transgene, thus increasing the homogeneity of the conditions. Ngn2 neurons express cortical markers, have a reproducible gene expression signature, form synapses and are able to integrate in the mouse brain, thus becoming a valuable tool for disease modeling and drug screening in the context of neurodevelopmental and neurodegenerative diseases.

As mentioned earlier, iPSCs give us an unprecedented temporal resolution in the study of developmental disorders. This means that the iPSC stage in itself can be already very informative, as it captures a transient but crucial early stage of development (Adamo et al., 2014; Quadrato et al., 2016).

Aim of the thesis

During the first part of my PhD I have contributed to the establishment of iPSC lines derived from WBS, 7dup patients and healthy controls, and to their transcriptional characterization. We have demonstrated that WBS and 7dup patient-derived iPSCs are already dysregulated, at the level of the transcriptome, in pathways that greatly overlap with those involved in the onset of the pathology (Adamo et al., 2014). Moreover, we showed that the dysregulation of these pathways is selectively amplified in differentiated lineages, further underscoring the anticipatory potential of iPSCs in capturing early effects at a systems level.

The core of the thesis builds upon these findings and on the realization that many WBSCR genes impinge on pathways that regulate gene expression at the level of transcription, translation and protein degradation.

The aim of my work is therefore to answer three fundamental questions:

- how is transcriptional dysregulation propagated to translation and protein abundance?
- what new information do we gain by integrating all layers of gene expression at the pluripotent state?
- what is the role of the translation initiation factor EIF4H in this regulation?

New technologies such as ribosome profiling and data-independent proteomics have emerged in the last few years and allow us to probe more in depth the extent to which transcriptional differences are propagated from the transcriptome to the proteome, yet systems-wide perspectives that integrate different layers of expression in human samples are still scarce .

With this work we want to fill the gap in the understanding of how 7q11.23 CNVs disrupt gene expression during pluripotency, and lay the practical and theoretical basis to explore this issue in more differentiated cell types.

Contributions

This project is a highly collaborative endeavour in which many different skillsets have been involved, and it is important to underline individual contributions at the experimental and analytical level, with each figure in which each individual has participated.

Contributions for the first part of the project (reprogramming, transcriptomic analysis, nanostring) are already stated in Adamo et al. 2014.

For the second part of the thesis, the experiments were carried out by Marija Mihailovic and me, and in particular:

- The establishment of feeder-free lines (fig. 2) was performed by Sina Atashpaz, Antonio Adamo, Matteo Zanella and me.
- The luciferase plasmids (referred to in figs. 6 and 7) were cloned by both Marija Mihailovic and me.
- The sineUP plasmids were cloned by me. The knock-down plasmids were cloned by Marija Mihailovic who also validated them by western blot (fig. 14)
- The viral particles for luciferase, sineUP and knock-down were prepared by me (figs. 6, 7, 14)
- The cell culture for the lable-free proteomics and the ribosome profiling experiments was carried out by both Marija Mihailovic and me (figs. 4, 8 through 13, 15 through 30)
- I carried out the cell culture for both pSILAC experiments (figs. 31 through 40).
- The Ribosome profiling protocol was set up by both Marija Mihailovic and me, and most of the samples were prepared by Marija (figs. 8, through 12, 15 through 30)
- For the label-free proteomics dataset, mass-spec processing, acquisition and primary data analysis, were performed by Yansheng Liu in the lab of Rudolf Aebersold at the ETH Zurich (figs. 4, 13, 19, 24, 25, 26, 28 through 40)

- Secondary data analysis of label-free and pulse-SILAC proteomics data was performed in our lab by Pierre-Luc Germain and myself (figs. 4, 13, 33 through 40)
- I performed all the secondary data analysis (figs. 8 through 13, 15 through 31, 33 through 40) with suggestions from Pierre-Luc Germain and Alessandro Vitriolo.
- I established the monoclonal NGN2 line establishment together with Matteo Zanella (figs. 41, 42)
- Immunofluorescence stainings were performed by Maddalena Lazzarin and Francesca Cavallo (fig. 42 C, D, E, F) and myself (42 A, B).
- All schemes and cartoons in the introduction (with the exception of figure 2), figure 1 in the materials and methods, and figures 6, 31, 41 in the results section) were designed and drawn by me.
- This study was conceived and supervised by Giuseppe Testa.

Materials and methods

Cell culture

iPSCs were initially cultured on a feeder layer of mouse embryonic fibroblasts inactivated with mitomycin C, as described in Takahashi et al., 2007 in an incubator set at 37°C, 3% O_2, 5% CO_2. For this condition, the medium is composed by DMEM/F12 (Gibco) in a 1:1 ratio, 20% Knockout Serum Replacemente (Gibco), 1% non-essential aminoacids, 1% Pennicyllin – Streptomycin (Pen-Strep), 1% L-Glutamine, 0.1% 2-mercaptoethanol, 10 ng/ml basic fibroblast growth factor (bFGF, Gibco). Colonies were manually passed on other feeder layers by physical fractionation with a 22G sterile needle. Upon stabilization, iPSCs were adapted to grow without the feeder layer in the commercial medium mTeSR-1 (StemCell Technologies) supplemented with 1% Pen-Strep and on hESC-qualified Matrigel (BD Biosciences) diluted 1:40 in DMEM/F12 and supplemented with 1% L-Glutamine and 1% Pen-Strep. Mechanical separation of differentiating patches of colonies was achieved using a sterile p200 pipette tip under a brightfield microscope. Cell lines growing in feeder-free conditions were further passaged by dissociation to single cells using accutase for 3 minutes at 37°C, and diluting cells 7 to 8 times for every passage. Single cells were resuspended in mTeSR-1 supplemented with 5 uM Y-27632 (Sigma) to avoid anoikis (Watanabe 2007).

RNAseq and Nanostring measurements

Cells were lysed in the plate using the RNeasy RLT+ buffer supplemented with 2-mercaptoethanol (Qiagen), centrifuged at 4°C and processed according to manufacturer instructions. Library preparation for RNA sequencing was performed using Poly-A and RiboZero kits (Illumina) according to manufacturer instructions) and libraries were sequenced on a HiSeq 2000 sequencer (Illumina) using 100 bp, paired-end reads.

Nanostring samples were prepared and analyzed according to manufacturer instructions.

Cloning

In order to insert the CrPV IRES in the 5' UTR of the Firefly luciferase, an additional restriction site was added upstream of the Firefly cDNA in the commercial plasmid pMirGLO (Promega). A PCR was performed using pMirGLO as a template adding the restriction site PacI, on the 5' end of the amplicon using oligo pair n. 1 in table 1. The PCR product was digested with EcoRI and ScaI and was cloned in the pMirGLO plasmid, now termed pMirGLO-host. The CrPV IRES was obtained as a dsDNA fragment by custom synthesis (GeneArt – Thermo Fisher) using a deposited sequence (id number 40 on IRESite) and it was amplified by PCR (oligo pair n. 2 in table 1). The PCR was digested using PacI and ApaI and it was cloned in the pMirGLO-host plasmid.

The luciferase construct was obtained by performing a PCR on the cDNA of Firefly luciferase in the pMirGLO-host, adding the restriction sites for BamHI (5') and NheI (3'), using oligo pair n. 3 in the table. The PCR fragment was then digested and subcloned in the pCDH-UbC-MCS-Ef1a-Hygromycin lentiviral backbone using BamHI and NheI, resulting in pUbC-Luc and pUbC-CrPV-Luc. Short hairpins for EIF4H RNA interference were cloned in the TRC pLKO.1 lentiviral backbone using AgeI and EcoRI sites and ssDNA oligo pairs 4, 5 and 6 from table 1 annealed in vitro.

No.	Clone name	Forward oligonucleotide	Reverse oligonucleotide
1	pMirGLO-host	AGAGAATTCTTAATTAAACCATGGAAGATGCCAAAAA	AGCGAGCTCGTTTAAACAACTAGA
2	pMirGLO-CrPV	AGATTAATTAACAACAACAAAAAGC	TCTGGGCCCTTCTTAAT
3	pUbC-Luc/pUbC-CrPV-Luc	ATATAGCTAGCAGCCCAAGCTTGGCAAT	TATATGGATCCAACTAGAATTACACGGCGATCTT
4	EIF4H sh1	CCGGGACTCCAGCTTAAACCTCGAACTCGAGTTCGAGGTTTAAGCTGGAGTCTTTTTG	AATTCAAAAAGACTCCAGCTTAAACCTCGAACTCGAGTTCGAGGTTTAAGCTGGAGTC
5	EIF4H sh2	CCGGGATCTCAGCATAAGGAGTGTACTCGAGTACACTCCTTATGCTGAGATCTTTTTG	AATTCAAAAAGATCTCAGCATAAGGAGTGTACTCGAGTACACTCCTTATGCTGAGATC
6	EIF4H shSCR	CCGGAACTTGCTATGAGAACAAATTCTCGAGAATTTGTTCTCATAGCAAGTTTTTTTG	AATTCAAAAAAACTTGCTATGAGAACAAATTCTCGAGAATTTGTTCTCATAGCAAGTT
7	pLKO-host	CCGGTCAACAACAACTGCAGCAACAACAAG	AATTCTTGTTGTTGCTGCAGTTGTTGTTGA
8	EIF4H binding domain	TCGAGGGCCCGATCGTCGTAGGTGTCGAAGTCCGCCATTTGCCGTCTCCGCTCCGAGAGGAACCAGGGTGAGCGAGGA	GGCCCGATCGTCGTAGGTGTCGAAGTCCGCCATTTGCCGTCTCCGCTCCGAGAGGAACCAGGGTGAGCGAGGA
9	EIF4H sineUP	TATAACCGGTGGCCCGATCGTC	TATACTGCAGAAGAGACTGGAGCTAAAGAG

Table 1: Oligonucleotides used for cloning.

To clone the sineUP constructs in the same pLKO.1 TRC backbone, it had to be slightly modified by adding an additional PstI restriction site between the AgeI and EcoRI sites. This was achieved by cloning two annealed ssDNA oligos (n. 7 in table 1) between AgeI and EcoRI, bearing the additional PstI site. The resulting clone was termed pLKO-host.

The EIF4H sineUP construct was obtained by first subcloning the EIF4H binding domain, obtained by two annealed ssDNA oligonucleotides (n. 8 in table 1), in the pCDNA 3.1(-) miniSINEUP plasmid (a gift of S. Gustinchich) using EcoRV and XhoI restriction sites. The full lenght of the sineUP (binding domain + inverted B2 domain) was then isolated by PCR using oligonucleotides (n. 8 in table 1) that changed the restriction sites from XhoI to AgeI and EcoRV to PstI. PCR fragments were digested and subcloned in the pLKO-host. All PCR reactions were carried out using Phusion DNA polymerase with HF buffer (New England Biosciences). All clones were checked by Sanger sequencing.

HeLa transfection of luciferase constructs

$5x10^5$ 60% confluent HeLa cells were transfected with 2 ug of the pMirGLO, pUbC-CrPV-Luc, pUbC-Luc and pUbC-host-Luc plasmids using Lipofectamine 2000 (Thermo Fisher) and assayed 24 hours after using a Glomax 96 plate reader (Promega) and the Dual-Glo Luciferase Assay kit (Promega).

Lentivirus preparation

All plasmids were extracted through the Nucleobond Xtra Maxi kit (Macherey-Nagel), according to manufacturer's instructions. Vectors were produced using a second generation system (envelope plasmid: pMD2-VSV-G; packaging plasmid: pCMV-Δ8.9).

All lentiviral vectors were generated through calcium phosphate transfection of human embryonic kidney 293T (HEK293T) cells. 5x106 of HEK293T cells were plated in a 10-cm dish in Iscove's Modified Dulbecco's Medium (IMDM) (Sigma Aldrich), 10% FBS, 1% Pen-Strep and 1% L-Glutamine and incubated at 37°C, 21% O2, 5%CO2. The day after, 1h before transfection, medium was replaced with 10 ml of fresh IMDM. In the meanwhile the following mix was prepared:

transfer vector: 10 μg

pVSV-G: 3.5 µg

pCMV-Δ8.9: 6 µg

ddH2O to a final volume of 500 µl.

Finally, 61 µl of 2.5 M CaCl2 were added to the mix and the tube was put on a rotating wheel for at least 20 minutes.

DNA precipitate was obtained by drop wise addition, on vortex at full speed, of 50 µl 2X HBS solution (281 mM NaCl, 100mM HEPES, 1.5 mM Na2HPO4 pH 7.12, 0.22 µM filtered) to the 500 µl mix previously prepared. This preparation was immediately added to HEK293T cells supernatant.

Cells were successively incubated at 37°C, 21% O2, 5% CO2 for 14 hours and afterwards medium was replaced with 10 ml of fresh IMDM medium. 30 hours after medium changing, the supernatant was filtered through a 0.22 µm pore nitrocellulose filter and ultracentrifuged at 20000 rpm in SW32Ti rotor (Optima L-60 preparative Ultracentrifuge; Beckman) for 2 hours at 20 °C. Pellets were resuspended in 80 ul of PBS and stored at -80°C in 10 ul aliquots.

Infection

Cells were split using accutase and resuspended in mTeSR-1 supplemented with 5 uM Y-27632. Approximately $4x10^5$ cells were infected using 2 ul of concentrated lentiviral particles while cells were still in suspension. Cells were infected overnight and fresh media was supplemented after ~12 h. Cells infected with pLKO-derived vectors were selected with puromycin (10 ug/ml) for 2 days, whereas cells infected with the luciferase reporter were selected with hygromycin (100 ug/ml) for up to 7 days.

Western Blot

Around $1x10^6$ cells were harvested by detaching with accutase, resuspending in ice-cold PBS and centrifuged for 5 minutes at 5000 rpm at 4°C on a refrigerated benchtop centrifuge.

Proteins were extracted using RIPA buffer (10 mM Tris-HCl pH 8, 1% Triton X-100, 0.1% SDS, 0.1% Sodium Deoxycholate, 140 mM NaCl, 1 mM EDTA). 4 volumes of RIPA buffer

were added to cell pellets and tubes were put at 4 °C on a rotating wheel for 30 minutes. Extracts were centrifuged for 30 minutes at 13000 rpm at 4 °C and supernatants were transferred in new tubes and store at -80 °C.

Before usage the protease inhibitor cocktail (PIC) and PMSF (1mM) were added.

Proteins were quantified at the spectrophotometer (λ 595nm), using the Bradford protein assay as follows: 200µl of protein Assay Dye reagent Concentrate (BioRad), 800µl of ddH2O, 1 µl of protein extract, using BSA (NEB) to derive a standard curve.

For blotting, 40 µg of protein extract were loaded on a 4-12% bisacrilamide-trisacrilamide using the Novex Sharp Pre-stained Protein Standard (ThermoFisher Scientific) as molecular weight marker. The electrophoretic run was executed in the NuPAGE SDS Running buffer (20X) (Thermo Fisher) at room temperature, initially at 80 V and, when the bands started to separate, at 100 V. Wet transfer was performed at 4 °C, 30V on a nitrocellulose 0.2 µm membrane (Sigma Aldrich) in 700 ml transfer buffer (200 ml methanol, 100 ml TB buffer 10X (Tris base 0.15M, glycine 1.9M), 400 ml ddH_2O). Transfer efficiency and quality was checked by Ponceau staining and membranes were washed with TBS-T (Tris 25 mM, NaCl 150 mM, KCl 2 mM, Tween-20 0.1%). The primary rabbit anti-EIF4H (Abcam, ab112966) and mouse anti-GAPDH antibodies were incubated in 5% milk in agitation overnight at 4°C , using 1:1000 (EIF4H) and 1:200 (GAPDH) ratios respectively. After 3 washes in TBS-T, the membrane was incubated with the secondary antibodies, diluted 1:10000 in 5% milk, 1h in agitation at room temperature. Afterwards the membrane was washed 3 times with TBS-T and then the ECL Prime Western Blotting detection reagent was used (Sigma-Aldrich) for detection with a BioRad Chemidoc imaging system.

Ribosome profiling

Ribosome profiling allows the quantification of translation at a genome-wide scale by sequencing of mRNA fragments occupied by ribosomes (Ingolia et al., 2009, 2011). This protocol allows the isolation of and preparation of a sequencing library containing small ribosome-protected fragments (RPF) of RNA. The protocol consists in the stalling of

ribosomes using cycloheximide, a drug that inhibits ribosome translocation, and the purification by size selection of monosomes (fig. 1). RNA is then enzimatically fragmented using RNase I, so that the ribosome protects a fragment of ~28-30 nt. The resulting fragments are depleted of ribosomal RNA using magnetic beads, and they are size-selected on a denaturing poly-acrilamide gel, after which their ends are repaired, they are ligated to sequencing adaptors and they are converted to cDNA by reverse transcription. The cDNA is further purified on a denaturing polyacrylamide gel and is then circularized. Circularized cDNA is amplified by PCR, in which sequencing indexes are inserted. The PCR is further size-selected on a native polyacrylamide gel and its quality is assessed using high sensitiviy electrophoresis systems such as the Bioanalyzer.

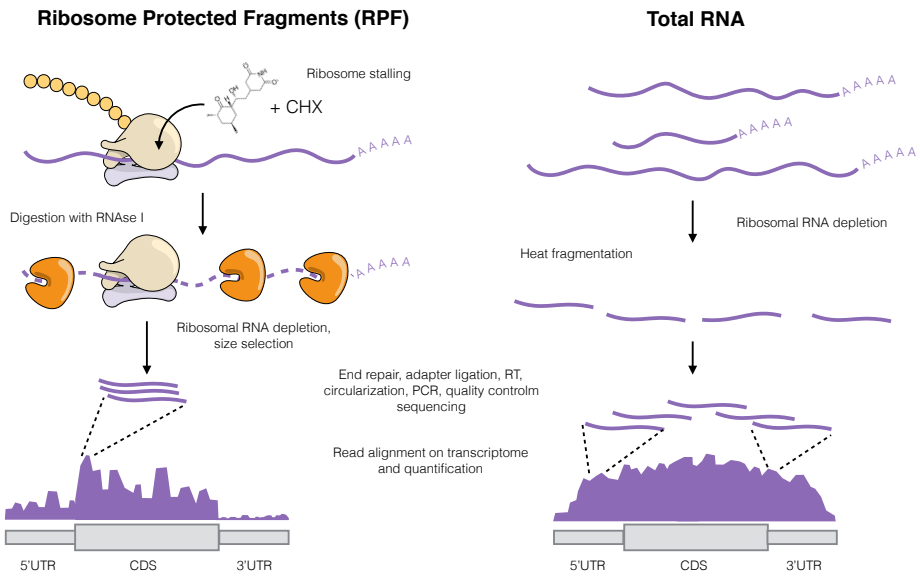

Figure 1: Schematic representation of the ribosome profiling protocol for the generation of ribosome-protected fragments and total RNA libraries. CHX: cycloheximide. CDS: coding sequence. RT: reverse transcription. PCR: polymerase chain reaction. UTR: untranslated region.

The ribosome profiling protocol allows to prepare in parallel and with very similar steps a total RNA library, which is used to measure the transcriptome. The only difference is the absence of RNase I digestion and the presence of a heat-fragmentation step before end repair. In this study, ~1.5-2x10^7 cells were treated with 10 ug/ml cycloheximide

(Sigma-Aldrich) while detaching with accutase at 37°C. Cells were then centrifuged at 4°C and cell pellets were frozen in dry ice at -80°C. Samples were then processed according to the manual instructions for the TruSeq Ribosome Profiling Kit (Illumina). RNA depletion was performed using the RiboZero Magnetic Gold kit (Illumina). All three PAGE separations were performed for RPF preparation. The first separation was performed to separate ribosome-protected fragments on a denaturing 15% acrilamide/bisacrilamide, TBE-Urea 8M gel. Gel bands were cut between 26 and 32 nucleotides approximately and retained for further processing. For both RPF and total RNA libraries, a denaturing 10% acrilamide/bisacrilamide TBE-Urea 8M gel was used to isolate bands between 80-100 nt (total RNA) and 70-80 nt (RPF). PCR products were further purified in a native 10% acrilamide/bisacrilamide TBE gel, where bands at around 150 nucleotides were cut for further processing. All gels used were precast Novex gels (Thermo Fisher).

Library quality was assessed with a Bioanalyzer High Sensitivity DNA chip (Agilent) and libraries of both total RNA and RPF RNA were sequenced using 50 bp, single end, 30 million reads in a HiSeq 2000 sequencer (Illumina).

Ribosome profiling and RNA-seq data analysis

Read preparation

As ribosome protected fragments (RPF) are smaller than 50bp, the majority of RPF reads contain adapter sequences. To eliminate these without creating biases between mRNA and RPF, we trimmed adapter sequences from all reads of both types of libraries using scythe 0.981 (and the set of adapter sequences specified by the protocol), discarding trimmed reads with 15 nucleotides or less. We then proceeded with two different kinds of analyses, aligning the reads either on the genome or the transcriptome (below).

Genome alignment

We used HISAT2 2.0.1 to perform SNP-aware spliced alignment of the reads on the hg38 genome using the RefSeq transcript annotation for known splice junctions. We used this data for sample quality control on the basis of codon periodicity. Codon periodicity was assessed by taking the 1000 most expressed transcripts and counting proportion of reads

falling on each base pair starting from the beginning of the coding sequence. Samples with very low number of aligned reads or which did not display a clear codon periodicity were considered failed and discarded.

Transcriptome alignment and quantification

To quantify genes and transcripts, we used RSEM 1.2.22 (with empirical read start position distribution estimate), which according to our benchmark (Germain et al., 2016) showed a very good accuracy, especially for absolute quantification, which is critical for our present purposes. Since the visual inspection of the reads on the genomic alignment showed a non-negligible proportion of the RPF reads falling in allegedly untranslated regions, we decided to perform quantifications on two types of indexes: 1) on the whole RefSeq transcripts, and 2) splitting coding sequence (CDS, adding 15bp on each side to account for the size of the ribosome), 3' UTR and 5'UTR of each transcript (which were considered all together for the purpose of calculating normalization factors).

Differential expression analysis

For total RNA and RPF data, we tested for differential expression analysis using edgeR v.3.12.1 (Robinson et al., 2010) on the TMM-normalized estimated fragment counts, which performed best in our benchmark. We included in the analysis features that had at least 10 reads in at least 3 samples. When doing pairwise comparisons of the genotypes, we use edgeR's classical dispersion model (based on a negative binomial model) and chose differentially expressed genes that passed a false discovery rate (FDR) threshold of < 0.1.

Slope calculation

For total RNA and RPF, slopes were calculated as log2(FC) values from edgeR's classical dispersion model using the log-transformed values of label-free protein intensity (see below) as the covariate.

GO enrichment

Gene ontology enrichments were performed using WebGestalt (Zhang et al., 2005), by using the hypergeometric test and choosing the top 10 significantly enriched categories.

Treemaps were drawn by taking only the enriched children categories that had non-overlapping parents.

Pulse-SILAC

Pilot experiment

Cells from a control line (3391-S) were growing in single-cell condition in mTeSR-1 and then adapted for 2 passages in a custom medium, composed as follows: DMEM, Knockout Serum Replacement 15% (Sigma), Pen-Strep 1%, Non-essential aminoacids 1%, Glutamine 1%, Probumin 0.5% (Millipore), 2-mercaptoethnaol 0.1 mM, L-Proline 500 mg/l (Sigma), FGF2 10 ng/ml (Peprotech). The medium was conditioned for 24 hours on a mouse embryonic fibroblast layer inactivated with mitomycin-C and filtered before use.

In the SILAC version of the medium a custom DMEM (Lonza) without arginine and lysine was complemented with 84 mg/l $^{13}C_6$ $^{15}N_4$ Arg10 (Sigma) and 146 mg/l $^{13}C_6$ $^{15}N_2$ Lys8 (Sigma). Cells were scraped and washed in cold PBS upon reaching 70% confluence approximately for protein harvest at 1.5, 4.5 and 13.5 hours after medium swap. Each plate was seeded and harvested in duplicate.

pSILAC experiment

Cells growing in mTeSR-1 (light condition) were swapped with a custom mTeSR-1 medium without arginine and lysine, complemented with 84 mg/l $^{13}C_6$ $^{15}N_4$ Arg10 (Sigma) and 146 mg/l $^{13}C_6$ $^{15}N_2$ Lys8 (Sigma). Cells were scraped and washed in cold PBS upon reaching 60% confluence approximately for protein harvest at 2, 4, 6, and 8 hours after medium swap. Each plate was seeded and harvested in triplicate.

Proteomic analysis

To better quantify subtle differences in protein abundance we harnessed ther power and resolution of a data-independent mass-spectrometry approach, Sequential Acquisition of all Theoretical MS2-spectra (SWATH-MS). In SWATH-MS, precursor ions are scanned at a high speed across small (25 Da) windows of mass to charge

(m/z) that cover a range of m/z values from 40 to 1200. All the fragmented ions coming from these small precursor windows are acquired, thus allowing a high resolution measurement of the peptides present in the precursor windows. The mass spectrometer cycles continuously through these windows, thus allowing a time-resolved measurement of chromatographic peaks (Gillet et al., 2012). SWATH-MS allows to reproducibly identify and quantify proteins at a low abundance with a high dynamic range.

Protein extraction and in-solution digestion

Around 10^7 cells were harvested by dissociating them with accutase, washing with ice-cold PBS and centrifuging at 4°C at 5000 rpm for 5 minutes.

All the cell pellets were suspended in 10M Urea lysis buffer and complete protease inhibitor cocktail (Roche), ultrasonically lysed by sonication at 4°C for 2 minutes using a VialTweeter device (Hielscher-Ultrasound Technology), and then centrifuged at 18,000 g for 1 hour to remove the insoluble material. The supernatant protein mixtures were reduced by 10mM Tris-(2-carboxyethyl)-phosphine (TCEP) for 1 hour at 37°C and 20 mM iodoacetamide (IAA) in the dark for 45 minutes at room temperature. All the samples were further diluted by 1:6 (v/v) with 100 mM NH_4HCO_3 and digested with sequencing-grade porcine trypsin (Promega) at a protease/protein ratio of 1:25 overnight at 37°C. The amount of the purified peptides was determined using Nanodrop ND-1000 (Thermo Scientific) and 1.5 µg peptides were injected in each LC-MS run.

SWATH and shotgun mass spectrometry

Peptide samples after digested were measured by SWATH mass spectrometry (SWATH-MS) or shotgun analysis as previously published (Collins et al., 2013; Gillet et al., 2012; Liu et al., 2013). Specifically a 120-min liquid chromatographic (LC) gradient was used for shotgun analysis and SWATH-MS measurement on the pSILAC samples, whereas a 60-min LC gradient was used for steady proteome expression

SWATH analysis. In the present SWATH-MS mode, the SCIEX 5600 plus TripleTOF instrument was specifically tuned to optimize the quadrupole settings for the selection of 64 variable wide precursor ion selection windows. SWATH MS2 spectra were collected from 50 to 2,000 m/z. The collision energy (CE) was optimized for each window according to the calculation for a charge 2+ ion centered upon the window with a spread of 15 eV. An accumulation time (dwell time) of 50 ms was used for all fragment-ion scans in high-sensitivity mode and for each SWATH-MS cycle a survey scan in high-resolution mode was also acquired for 250 ms, resulting in a duty cycle of ~3.45 s.

SWATH-MS data analysis: steady state expression data

The SWATH-MS identification was performed by OpenSWATH software (Röst et al., 2014) searching against a previously established SWATH assay library which contains mass spectrometric query parameters for 10,000 human proteins with unique Swiss-Prot identities (Rosenberger et al., 2014). OpenSWATH firstly identified the peak groups from all individual SWATH maps at a target FDR=1% and then aligned between SWATH maps with extension FDR=5% using a novel TRIC (TRansfer of Identification Confidence) algorithm that was specifically developed for targeted proteomic data analysis (Röst et al., 2016). Peptide intensities were first normalized using median normalization of log-transformed intensities, and technical replicates were aggregated using median values. For each protein that had at least 5 peptides, peptides poorly correlated to the others (<0.7 median correlation) were excluded. Protein intensity was then calculated as the median of the top 3 most intense peptides after filtering. To increase the protein-level confidence, only those peptide signals identified in at least ten out of 49 samples were accepted (requantification enabled).

SWATH-MS data analysis: pSILAC data

To analyze pSILAC SWATH data, we first generated a sample specific library containing the light and heavy peptide assays. All the shotgun runs of the "Light" samples (those from steady proteome as well as the first time point- samples in pSILAC experiment) were firstly searched against human Swissprot database using the iPortal pipeline (Kunszt et al., 2015). Profile-mode .wiff files from shotgun data acquisition were centroided and converted to mzML format using the AB Sciex Data Converter v.1.3. iPortal utilized iProphet schema (Shteynberg et al., 2011) to integrate the search results from X!Tandem, Omessa, Myrimatch, and Comet at peptide level FDR =1%. Especially, peptide tolerances at MS and MS/MS level were set to be 50 ppm and 0.1 Da respectively. Up to two missing trypsin cleavages were allowed. Oxidation at methionine was set as variable modification whereas carbamidomethylation at cysteine was set as fixed modification. The light version of the raw spectral library was generated from all valid peptide spectrum matches and then refined into the non redundant consensus libraries using SpectraST (Lam et al., 2007). Using the spectrast2tsv.py function in OpenSWATH (Röst et al., 2014) we then generated the light and heavy MS assays as the final library constructed from top 3-6 most intense fragments with Q3 range from 400 to 1200 m/z excluding those falling in the precursor SWATH window were used for targeted data analysis of SWATH maps. The light library contained 55,857 peptide sequences (light and heavy forms with the consideration of modifcations) of 3317 unique proteins. OpenSWATH analysis were run with the same options as above for protein expression data, however, as in the shotgun proteomic analysis, requantificated data points were discarded for protein turnover calculation (Lam et al., 2007).

Differential protein expression

As the error in protein intensities is log-normally distributed, we tested for differential protein expression using t-test (for categorical conditions) on log-

transformed intensities. For regression on measured quantitative variables, such as effective expression levels, we tested for a standardized major axis, SMA (using the smatr 3 R package) between the log-transformed expression and the independent variable. Contrarily to least-squares regression, which minimizes error on the Y axis and assumes that the x axis is determined, SMA simultaneously minimizes error on both axes, thereby accounting for potential errors in the measurement of the independent variable.

Determination of degradation rates

Relative isotope abundance (RIA) was calculated for each peptide and fitted to an exponential decay model using nonlinearleast-squares to estimate the rate of loss of the light isotope (Kdeg). To establish protein-level Kdeg, we tried three different methods: 1) using the median of the peptide-level kloss values, 2) refitting a model on the RIA of all peptides, and 3) doing a weighted mean of the peptide-level Kdeg values, using as weights the number of datapoints divided by the standard deviation of the coefficient estimation. (We also tried a combination of the RIA-based approach with a linear model on the ln(Heavy/Light+1) values). The weighted mean method was selected because it minimized the median deviation between replicates.

Generation of NGN2 monoclonal lines

Lentiviral particles containing the NGN2-EGFP and UbC-rtTA transgenes were prepared as described above using Tet-O-FUW-NGN2-EGFP-Puro and UbC-rtTA plasmids (a gift of Thomas Suedhof). $3x10^4$ cells were infected with 0.5 ul of concentrated viral particles and cells were expanded until reaching an amount of $\sim 6x10^6$. During this period, exhausted medium was kept and filtered. Filtered medium was mixed 1:1 with fresh medium to create a “conditioned” medium for post-sorting recovery. Cells were then incubated with DAPI to check for vitality, and only DAPI-negative cells were sorted as single cells using a FACSAria II sorter (BD Biosciences) in 96 well plates coated with 1:40 hESC-qualified matrigel and

containing the conditioned medium. Cells were kept in culture in this format, with daily media change starting from day 7 with conditioned medium, until a single colony was easily visible; based on the colony shape and morphology, which should suggest a single-cell origin, 3-5 colonies per line were chosen and expanded. The expansion was performed by stepwise increases of surface culture, from 96 well plates to 48, then to 24. While passaging cells from 48-well plates to 24-well plates, two wells were seeded per line. One of the wells was kept for induction with doxycycline (2 μg/ml) and scored for GFP positivity the day after. Positive clones were further expanded in a stepwise fashion until reaching the 6 cm format, after which they were frozen. Cells were then differentiated as in (Zhang et al., 2013).

Results

1. Generation and stabilization of feeder-free iPSC lines from a cohort of WBS, control and 7dup patients

iPSC lines were generated starting from patient-derived skin fibroblasts using a non-integrating reprogramming method based on daily transfections of mRNA coding for reprogramming factors (Adamo et al 2015, Warren et al. 2011). Successfully reprogrammed cells were stabilized in the pluripotent state by co-culturing them with a mouse embryonic fibroblast (MEF) feeder layer in a chemically defined medium. However, in order to perform high-throughput genomics experiments such as RNA-seq, Ribo-seq, proteomics without the confounding effect of murine RNA or proteins, iPSC lines must be adapted and stabilized in feeder-free conditions, which require a change of culture substrate and a different chemically defined culture medium.

Previously reprogrammed lines from 4 WBS patients, 1 atypical patient, 3 healthy controls, of which 1 relative of a WBS patient, and 3 7dup patients (fig. 1) were adapted to grow in feeder-free conditions by passaging them from the MEF layer to 1:40 matrigel-coated plates, thus creating pure iPSC lines (fig. 2A)

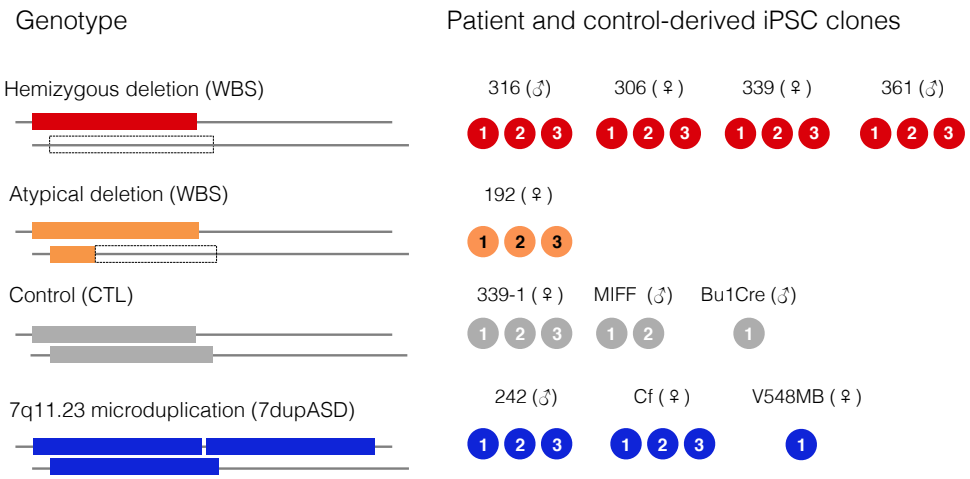

Figure 1: Schematic representation of the cohort of iPSC lines used in this study. Each clone is represented by a full circle, whereas each patient is identified by its alphanumeric code.

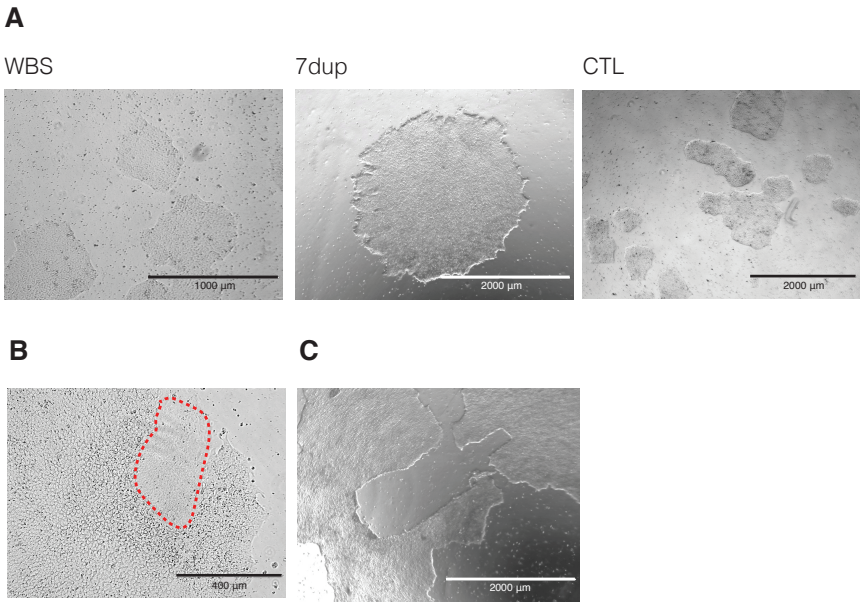

Figure 2: establishment and stabilization of feeder-free iPSC lines. A) representative images of feeder-free iPSC lines from a WBS, a 7dup and a CTL patient. B) feeder-free line show patches of differentiating cells (circled by a red dashed line). C) mechanical ablation of differentiated patches from the colonies.

Some lines, regardless of the phenotype, were prone to spontaneous differentiation more frequently in feeder-free condition than on feeders (fig. 2B). This effect was reduced by a combination of passaging and mechanical separation of differentiated regions (fig. 2C).

With few exceptions, after an average of 5-6 passages combined with separation, feeder-free iPSC lines greatly reduced occurrences of spontaneous differentiation.

2. Williams-Beuren Syndrome Chromosomal Region genes are expressed at the pluripotent stage and mirror gene dosage.

The expression of WBSCR genes in iPSC lines was measured at both the mRNA (fig. 3) and protein level (fig. 4) across a wide panel of lines.

All these genes are expressed at a detectable level and show an expression pattern that remarkably mirrors gene dosage (fig. 3A, 4A), especially for genes expressed at a high level (fig 3B, 4B). mRNA levels were measured with Nanostring, a hybridization-based technology that quantifies single mRNA molecules in a chip and quantifies them with an optical signal. mRNA are precisely quantified by counting the fields of view (FOV) in which the signal is detected. The expression pattern is still clearly detectable at the protein level, although slightly more variable and with a less clear-cut reflection of gene dosage. This could be due to technical variability arising from differences in the type of measurements performed in mass spectrometry-based technologies, in which, instead of sequence reads, different peptides from different proteoforms are measured (see Discussion).

The expression levels and patterns of these genes prompted the question whether a molecular phenotype, *i.e.* transcriptional dysregulation, could be captured already at the iPSC stage.

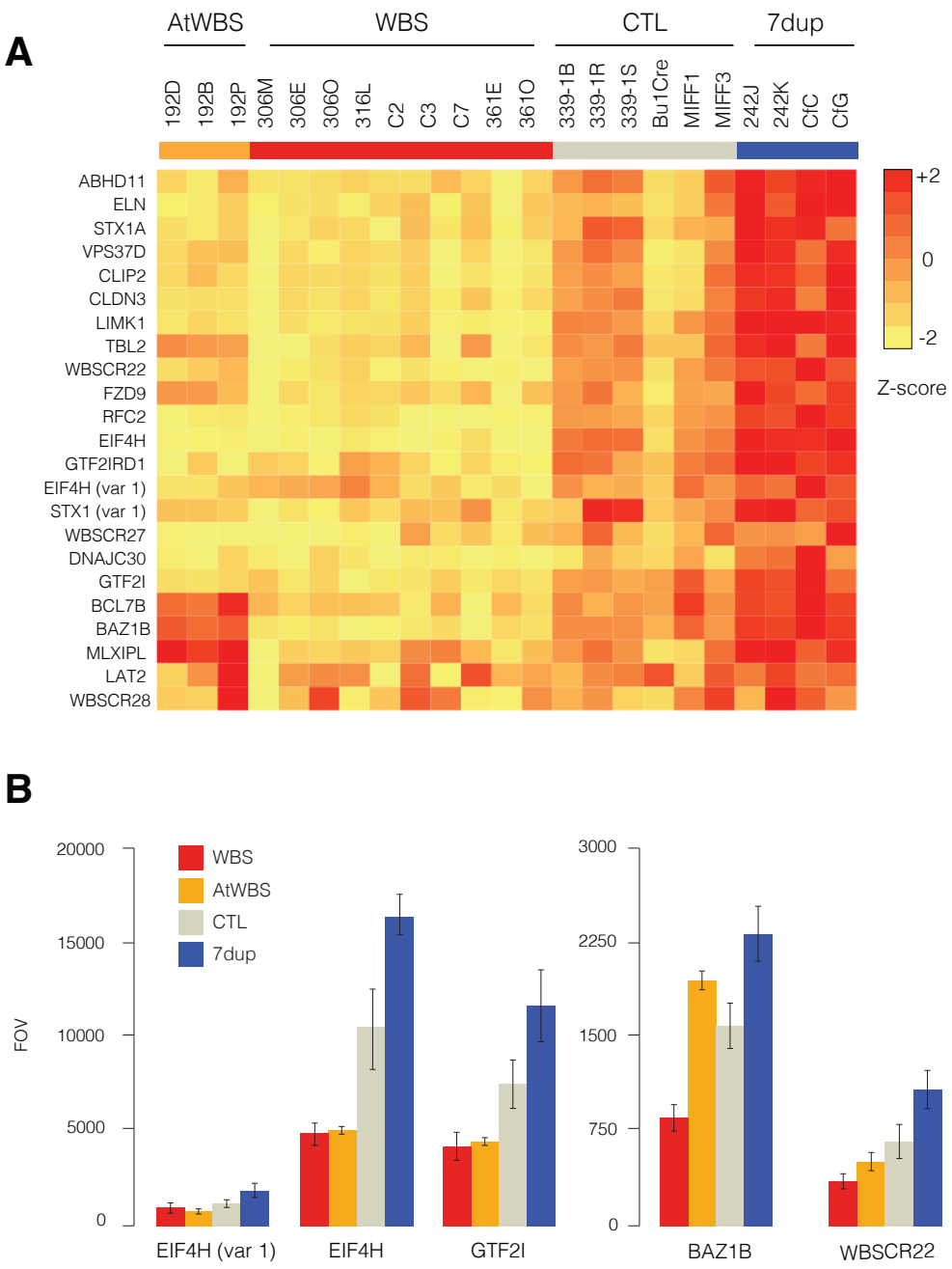

Figure 3: WBSCR genes mirror gene dosage at the mRNA level in iPSCs. A) heatmap showing Z-score of expression levels for all the WBSCR transcripts, measured by Nanostring. B) average mRNA expression levels of EIF4H, WBSCR22, BAZ1B and GTF2I in each genotype. Error bars represent standard deviation. FOV: Field Of View.

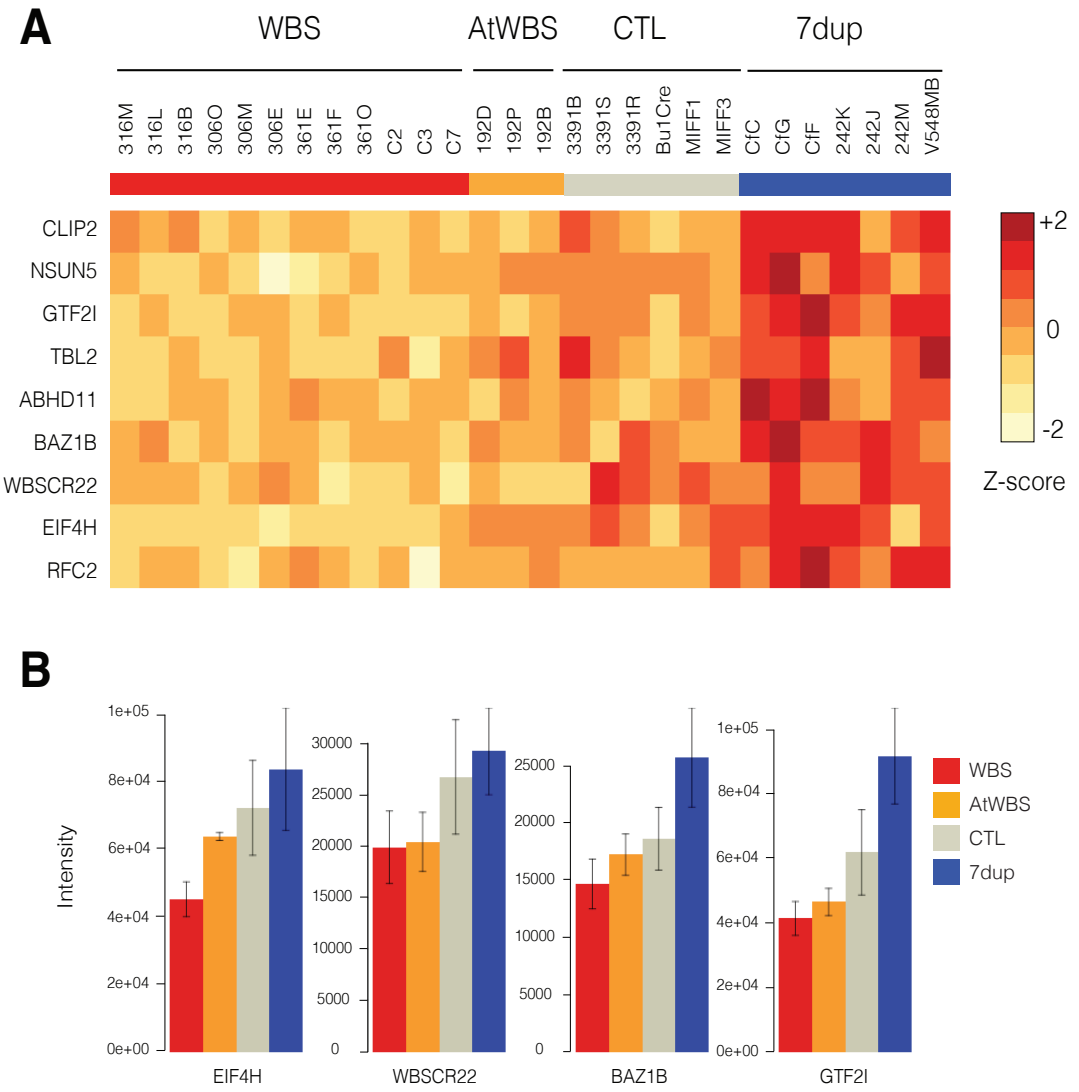

Figure 4: WBSCR genes mirror gene dosage at the protein level in iPSCs. A) heatmap showing Z-score of expression levels for the uniquely identified WBSCR proteins, measured by SWATH-MS label-free quantification. B) average protein expression levels of EIF4H, WBSCR22, BAZ1B and GTF2I in each genotype. Error bars represent standard deviation.

3. Transcriptional programs are already dysregulated in pluripotency and map onto disease-associated pathways

When performing differential gene expression analysis on transcriptomes obtained by RNA-seq of the 27 iPSC lines we found 757 differentially expressed genes (DEGs - see Annex 1 for the list).

Interestingly, we found that DEGs were enriched for Gene Ontology categories entailing biological processes and molecular functions closely related to the systems and the developmental pathways that are likely perturbed in WBS or 7dup (figure 5).

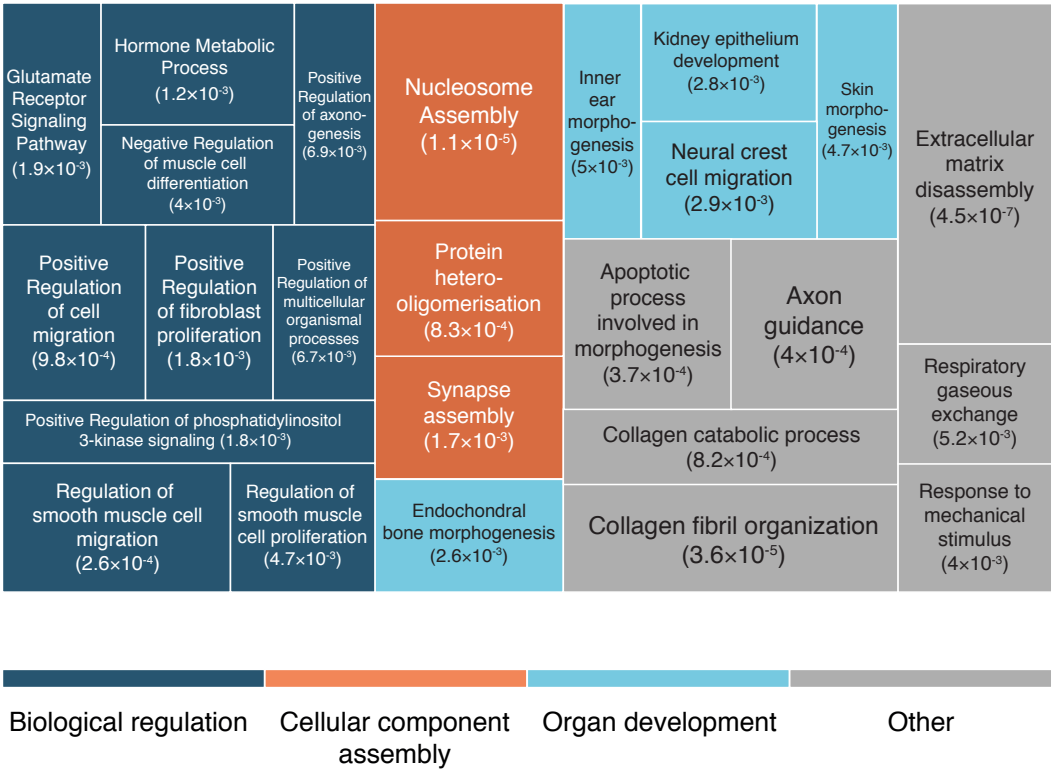

Figure 5: treemap representing Gene Ontology terms for which there is a statistically significant enrichment among RNA-seq DEGs. Parent categories with enriched children were removed. The size of each box is proportional to the statistical significance of the enrichment. Colours are assigned based on the top-most, non-overlapping parent category. Redrawn from Adamo et al. 2015

Owing to the functional involvement in translation of EIF4H and WBSCR22, and based on their dosage, we then set out to investigate the broad effect of the CNV in the translatomes and proteomes of these iPSC lines, by ribosome profiling and SWATH-MS label-free quantification respectively.

4. Generation of reporter iPSC lines to assess global changes in translation

Since normalization methods used in ribosome profiling data analysis work under the assumptions that there is no global effect in translation, we devised a reporter-based strategy that would allow us to measure, and possibly normalize by, any such global differences. The initial strategy envisaged the presence of a cricket paralysis virus (CrPV) IRES in the 5'UTR of the reporter (fig. 6A). CrPV IRES can initiate translation by directly recruiting an 80S monosome to the translation start site, effectively bypassing all initiation factors. We reasoned that reporters with the IRES would be able to show any global effect that was not depending on changes in EIF4H abundance. We subcloned in a lentiviral backbone the luciferase cDNA with and without the CrPV IRES, under the control of the UbC promoter. Since only the *Firefly* luciferase cDNA was cloned, the lentiviral backbone does not allow to normalize on another signal such as *Renilla* luciferase. To avoid this issue, we use the total RNA level of *Firefly* for normalization and use differences in translation efficiency as a measure of the global effect. To test their activity, we transfected both constructs in HeLa, together with the original dual-luciferase commercial constructs as a positive control (pMirGLO) and the modified dual-luciferase construct harbouring an additional pair of restriction sites to insert the IRES DNA (pMirGLO-host). Unfortunately, only the lentiviral construct containing the IRES did not expresses *Firefly* luciferase (fig. 6B), making our initial strategy unsuitable for our purposes. Nonetheless, we decided to use the lentiviral construct lacking the IRES, reasoning that a short 5'UTR would still have little sensibility to changes in EIF4H abundances. Moreover, the hygromycin antibiotic resistance, under the control of the EF1a promoter, can serve as an additional reporter in RNA-based measurements.

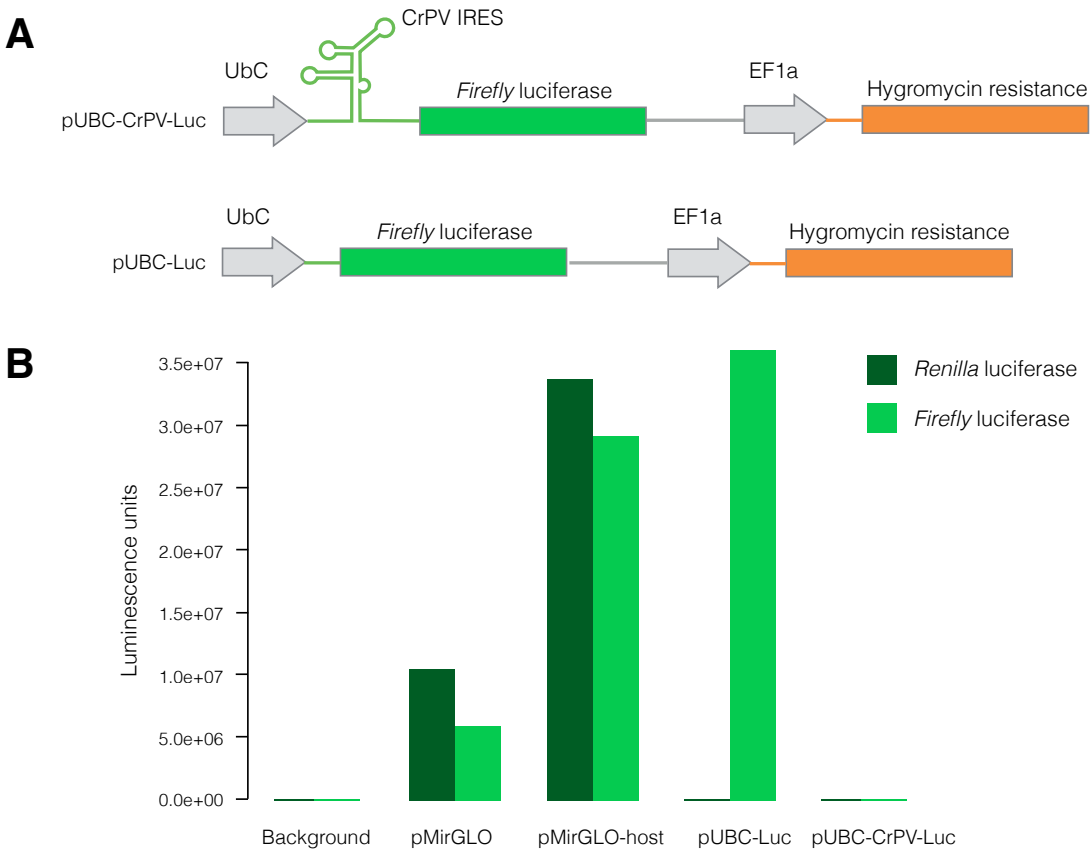

Figure 6: Reporter-based strategy to assess global differences in translation. A) two constructs were derived by subcloning the *Firefly* luciferase cDNA in a lentiviral backbone under the control of the UbC promoter. In pUBC-CrPV-Luc, a cricket paralysis virus (CrPV) IRES has been inserted, together with the obligatory alternative start codon CTG, needed to initiate translation with this IRES. In pUBC-Luc, only the luciferase cDNA has been inserted. B) Luciferase assay in HeLa cells transfected with pMirGLO, pMirGLO-host, pUbC-Luc and pUbC-CrPV-Luc. The levels of the IRES-containing reporter are indistiguishable from background. Light green: *Firefly* luciferase. Dark green: *Renilla* luciferase. Each bar represents one experiment.

All iPSC lines were infected with lentiviral particles containing the pUbC-Luc construct, and, upon selection, a subset of them was assayed for luciferase to make sure that the reporter was not being silenced (fig. 7).

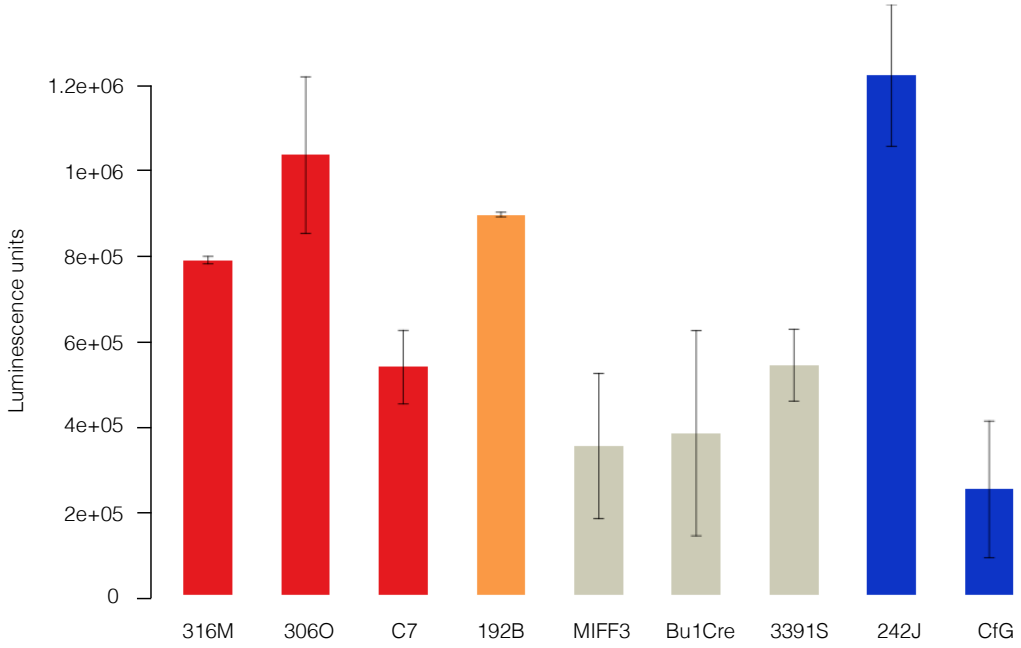

Figure 7: reporter expression in a sample of iPSC lines infected with pUbC-Luc lentiviral particles. All lines were antibiotic-selected. The reporter was expressed at a detectable level, showing successful integration and expression of the construct. Each measurement has been done twice. Error bars represent standard deviation.

5. Assessment of transcriptome and translatome dysregulation in iPSC lines by ribosome profiling

In order to gain a specific insight on the extent and the magnitude of translation dysregulation at the pluripotent stage, we performed ribosome profiling on a subset of 11 iPSC lines, each derived from a different patient (3 WBS, 1 atypical WBS, 3 CTL, 3 7dup), with the exception of a 7dup patient for whom we included 2 lines (because at the time we did not have available cells from a third 7dup patient). We used the ribosome profiling protocol to generate total RNA and ribosome-protected fragment (RPF) libraries simultaneously. Only uniquely mapping reads were retained. RPF reads were subdivided *in silico* by unequivocally assigning reads to 5'UTR, coding sequence (CDS) and 3'UTR of each transcript, in order to avoid confounding effects due to ribosome footprints in untranslated regions with a regulatory function, or arising as experimental artefacts. As expected, most of the reads map on the CDS,

followed by a fraction mapping on the 5'UTR and a remarkably smaller portion on the 3'UTR (fig. 8).

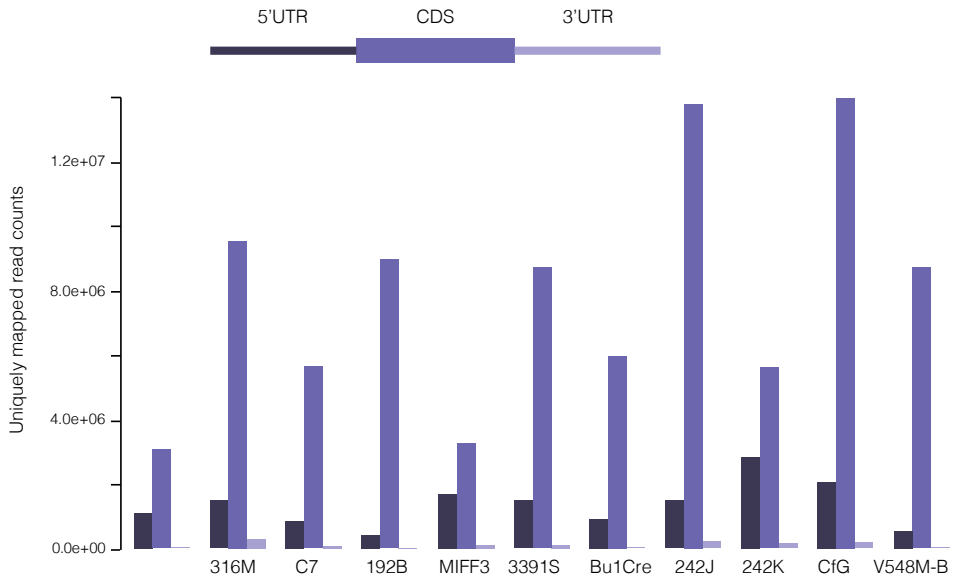

Figure 8: distributions of ribosome-protected fragment reads on different regions of the transcript. The majority of reads is expectedly mapped on the CDS portion of transcripts. Dark blue: 5'UTR; purple: CDS; light purple: 3'UTR.

To better appreciate the influence of translation regulation, a simple calculation that is routinely performed is translation efficiency (TE), *i.e.* the ratio between the abundance of RPF reads over their respective total RNA abundances. Since only reads aligned in the CDS are used to quantify translation, the same *in silico* division has been carried out on total RNA transcripts, so that TE can be calculated using exactly the same portions of transcripts. TE can serve as a readily interpretable value of the extent to which a gene is regulated on translation.

While the quantification of WBS genes expression at the level of RNA (fig. 9A) and RPF (fig. 9B) still mirrors gene dosage, there appear to be no major differences in TE across genotypes (fig. 9C). However, as will be discussed later on, the high variability of these measurements invites some caution in their interpretation.

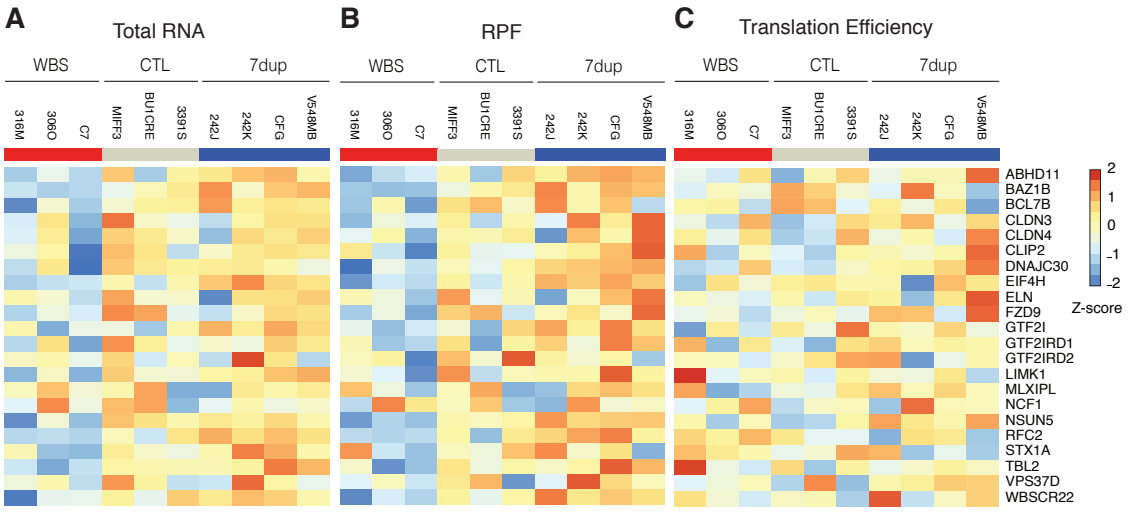

Figure 9: Heatmaps of Z-scores for WBSCR genes at the total RNA level (A), the RPF level (B) and Translation Efficiency (C).

We performed differential gene expression analysis by doing pairwise comparisons of groups of patients according to their condition: WBS vs control, 7dup vs control, 7dup vs WBS. When including the atypical WBS sample in the analysis, the number and identity of DEGs would greatly change based on whether it was grouped together with the WBS or the control; in any case it would be reduced. Since we do not have more than 1 atypical patient in our cohort, we do not reach a minimum amount of replicates that would allow for statistically significant comparisons with other samples. For this reason we decided to exclude the atypical WBS sample at this stage of the analysis. We then considered the union of differentially expressed genes passing a false discovery rate (FDR) threshold of 0.1 to represent the set of genes dysregulated by dosage imbalances in the CNV. We found 96 differentially expressed genes at the RNA level (fig. 10). About half of them show an expression pattern that goes in the same direction as that of WBSCR genes. GO enrichment analysis using the whole transcriptome as a background did not score any statistically significant enrichment. However, several differentially expressed genes are involved in pathways that map onto disease-associated phenotypes and functions (tab. 1). The high level of variability across samples with the same genotype invites caution, as it is

possible that technical isssues in library preparation may inflate variability. Nevertheless, it is also possible that differences between genetic backgrounds (inter-individual) account for a portion of the observed variability, in a way that is remarkably high at the mRNA level.

Gene	Function	Reference
SOX3	SRY-box transcription factor, regulates neural differentiation, when mutated causes X-linked intellectual disability with panhypopituitarism	(Laumonnier et al., 2002; Woods et al., 2005)
RPS16, RPLP1	40S ribosomal subunits	
MRPL12, MRPL26	Mitochondrial ribosomal subunits	(Serre et al., 2013)
TRDN	Ion channel involved in cardiovascular defects causing arrhythmogenic tachycardia	(Altmann et al., 2015)
PCDHA3, PCDHB3, PCDHB5	Protocadherins possibly involved in the formation of neural connections in the CNS	(Zipursky and Sanes, 2010)
HISTH1A	Histone H1, associated with schizophrenia in a GWAS on Ashkenazi jews	(Goes et al., 2015)
BANF1	Involved in nuclear envelope formation, when mutated causes Nestor-Guillermo progeria syndrome	(Puente et al., 2011)
ALCAM	Cell adhesion molecule involved in neurite extension, mesenchymal stem cell differentiation and cardiac morphogenesis	(Burns et al., 1991; Gessert et al., 2008)
HSPB1	Actin-binding protein involved in stress response, when mutated causes Charcot-Marie-Tooth neuropathy, axonal, type 2f	(Evgrafov et al., 2004)
RFX3, IER2	Regulate left-right polarity in embryogenesis, important for proper neural tube formation	(Hong and Dawid, 2009; Magnani et al., 2015)
GPR1	Found in GWAS for schizophrenia	(Bergen et al., 2012)
PDPK1	Master kinase in the mTOR pathway Essential for neuronal development	(Lawlor, 2002; Watatani et al., 2012)

Table 1: DEGs found in the total RNA dataset that show overlaps with clinical or molecular features of the two syndromes.

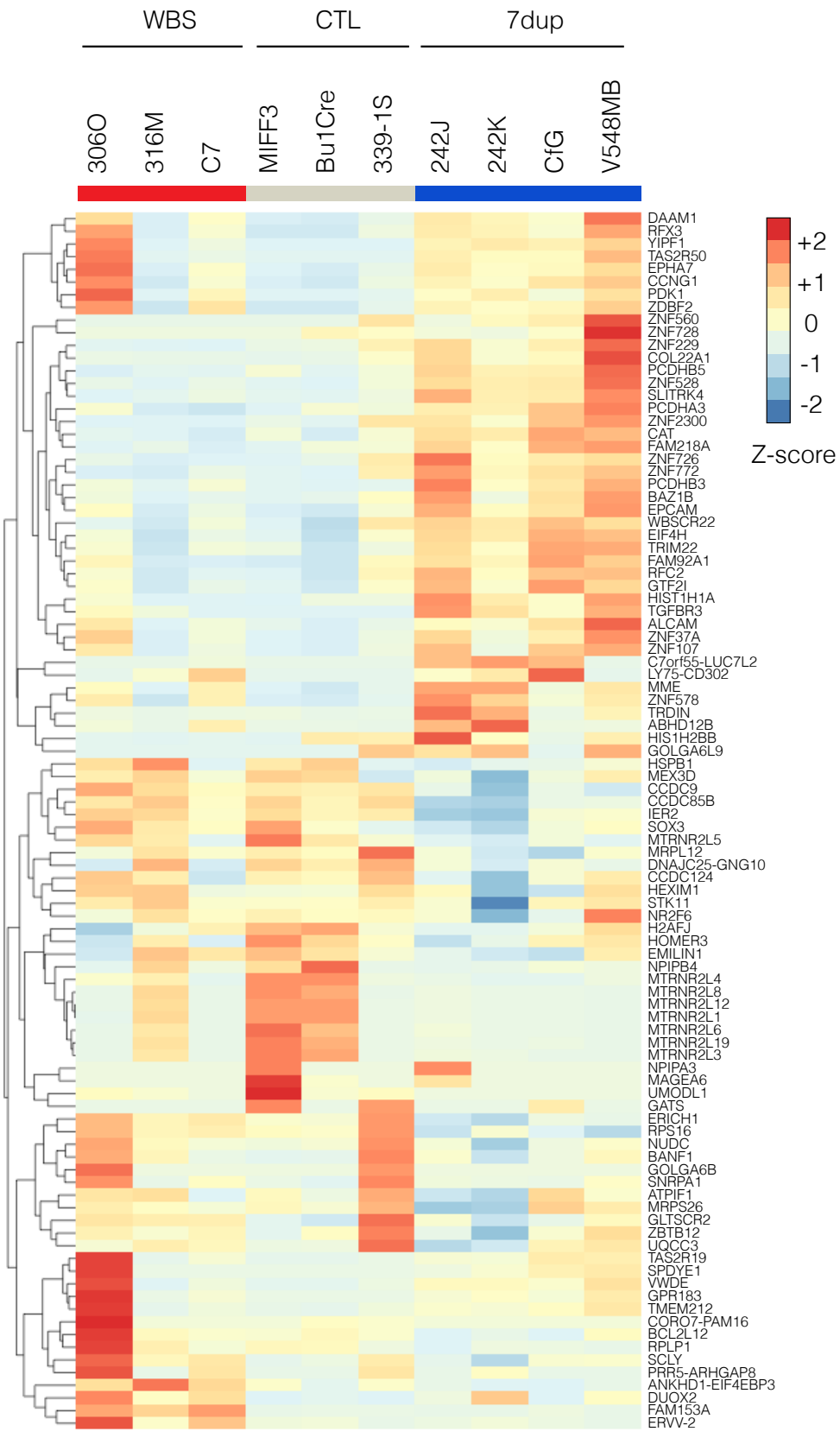

Figure 10: Heatmap of the Z-scores for differentially expressed genes (FDR < 0.1) at the Total RNA level according to edgeR's generalized linear model.

When performing differential gene expression analysis on the translatome in the RPF layer, the number of DEGs passing a FDR threshold of 0.1 is nearly halved (42 DEGs – fig. 11). Almost all the DEGs in this dataset have an expression pattern that is concordant with that of WBSCR genes. Interestingly, only 16 of them were also called as DE in the total RNA dataset, among which 6 are WBSCR genes. These findings hint at buffering, at the translation level, of perturbations originating in the transcriptome. The remaining 26 genes, among which 3 are WBSCR genes, are called as DE exclusively in the RPF dataset (fig. 12). Among these 26 genes we found some important regulators of neuronal development (tab. 2) associated to phenotypes reminiscent of the clinical characteristics of these syndromes. Importantly, neither Luciferase nor Hygromycin were detected as differentially expressed in any of the two datasets, thus pointing to a lack of global effects affecting the reporters that would be dependent on initiation. For this dataset, variability seems to be much less prominent, either pointing to buffering of inter-individual differences, or a less severe impact of technical artifacts. The fact that 242K and 242J, two samples coming from different iPSC clones of the same individual, have a similar pattern at the RNA level, would point to the first hypothesis.

Gene	Function	Reference
ANKRD1	Probable transcription factor that also interacts with sarcomeric proteins, involved in dilatative cardiomyopathies	(Moulik et al., 2009)
POU3F3	Homeobox transcription factor involved in neuronal development and in the development of inner ear epithelium	(Dheedene et al., 2014; Dominguez et al., 2013)
CPEB2	RNA-binding protein involved in dorso-ventral axis formation	(Hafer et al., 2011; Turimella et al., 2015)
CEBPB	Neuronal transcription regulator	(Sterneck and Johnson, 1998)
PCDHGB1, PCDHGB3, PCDHGB5 PCDHGA3	Protocadherins possibly involved in the formation of neural connections in the CNS	(Zipursky and Sanes, 2010)

Table 2: DEGs exclusively found in the RPF dataset that show overlaps with clinical or molecular features of both syndromes.

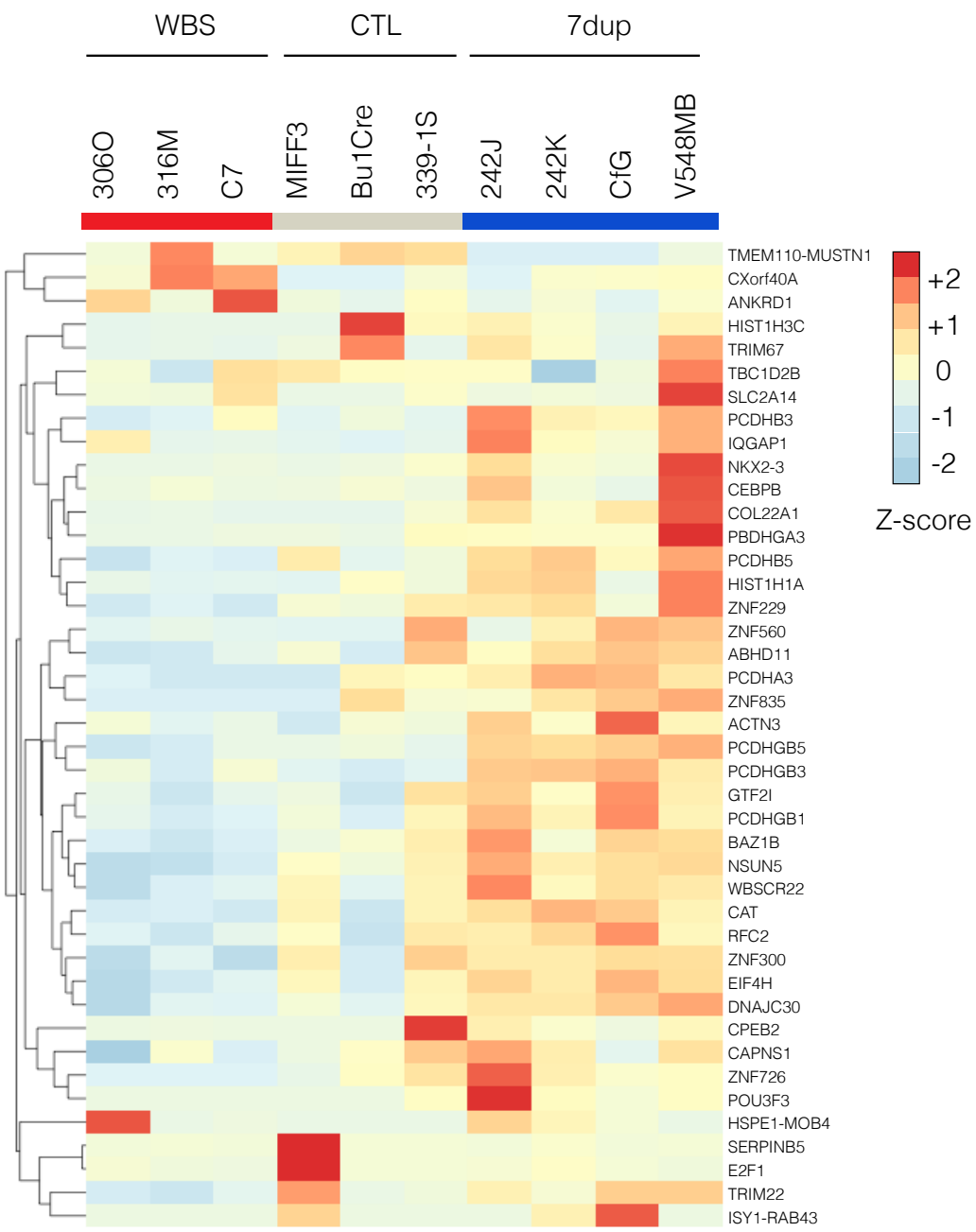

Figure 11: heatmap of the Z-scores for differentially expressed genes (FDR < 0.1) at the RPF level according to edgeR's generalized linear model.

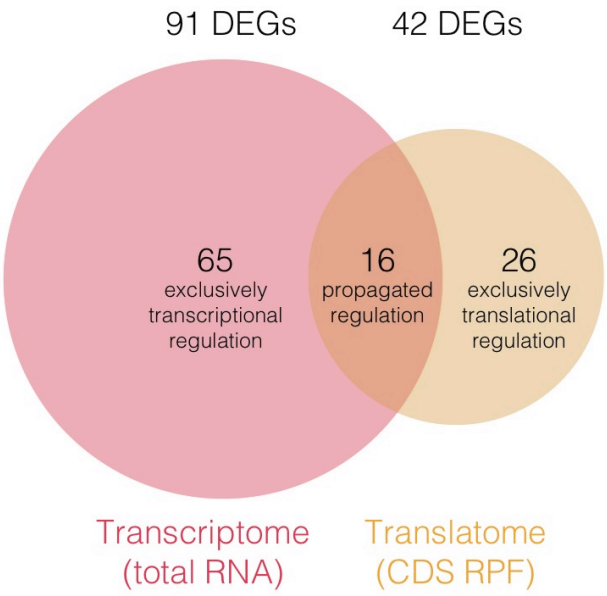

Figure 12: Venn diagrams representing the overlaps of DEGs across molecular layers.

6. Assessment of proteomic dysregulation by mass spectrometry

To gain further insight into the dysregulation caused by 7q11.23 CNVs in the proteome, we performed high-throughput mass spectrometry using a data-independent mass-spectrometry approach, SWATH-MS. We analyzed the proteome of 25 iPSC lines derived from 10 individuals (4 WBS patients, 3 controls, 3 7dup patients). Differential expression analysis was performed, as for transcriptome and translatome, by making pairwise comparisons between genotypes and considering the union of differentially expressed proteins (DEPs) passing the FDR threshold of 0.1. We found 41 DEPs (fig. 13), among which roughly 60% follows WBSCR dosage. Only 8 DEPs are also DEGs in the translatome. Besides 6 WBSCR proteins, only CAT was differentially expressed in both transcriptome and translatome, whereas EPCAM was differentially expressed in the transcriptome only, and ABHD11 in the translatome only. Lowering the FDR threshold for DEP identification to 0.2 did not increase the number of overlapping proteins across layers. These observations hint again at a layer-specific regulation, buffering the transition from actively translated

mRNA to final protein abundance, and introducing other perturbations exclusively in the proteome. It is possible, however, that the intrinsic differences between sequencing-based measurments and mass-spectreometry measurements, together with unavoidable differences in the statistical tests used to measure differential expression, may mask the overlap of some other differentially expressed genes and proteins. Further validation steps are needed to assess the technical and the biological differences, and how each of them impacts gene expression measurements.

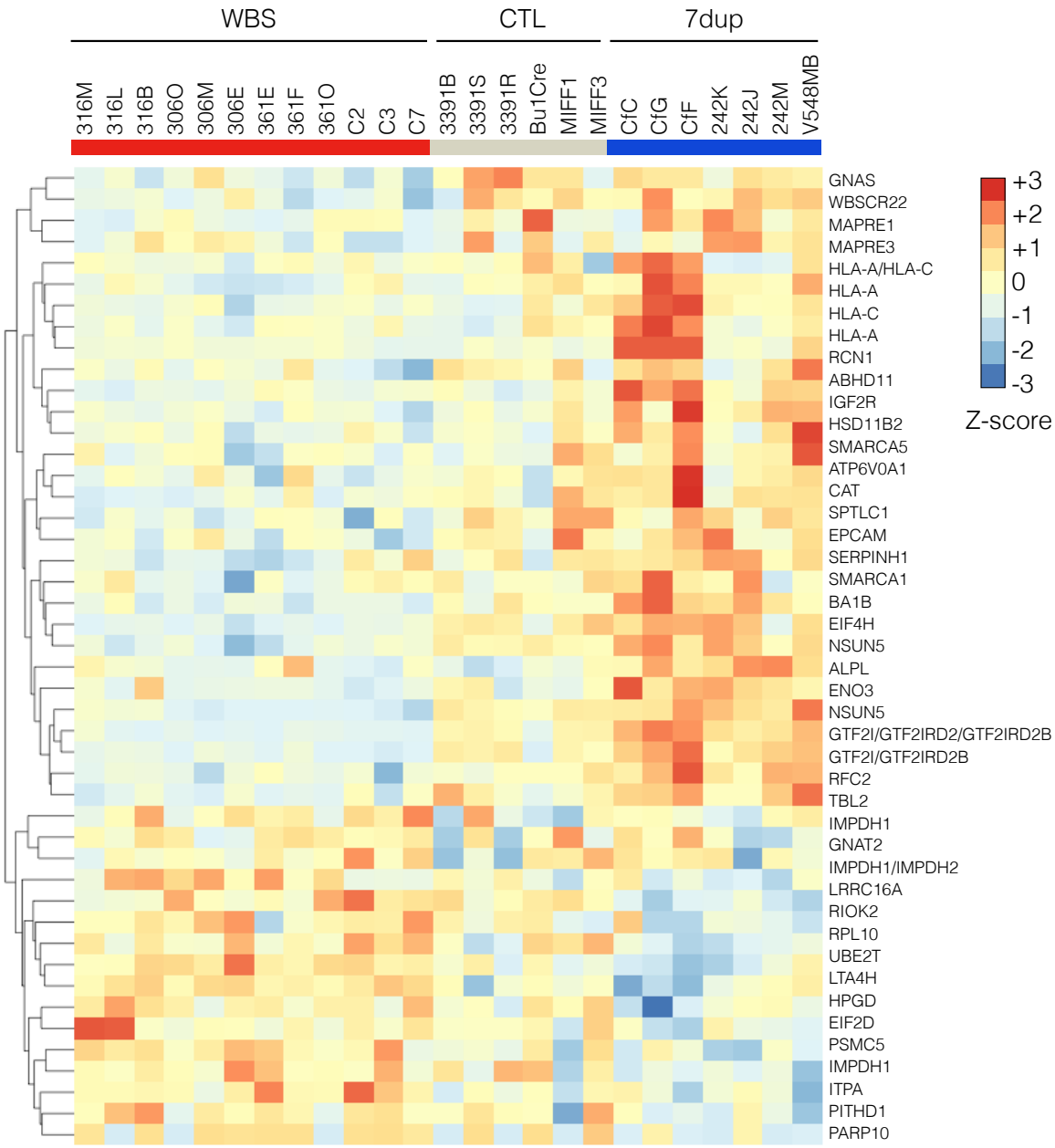

Figure 13: heatmap of the Z-scores for differentially expressed proteins (FDR < 0.1) according to pairwise categorical t-tests.

As already observed in both total RNA and RPF datasets, several differentially expressed proteins map onto functions and pathways that overlap with the clinical and molecular characteristics of both syndromes (tab. 3).

Protein	Function	Reference
HLA-A/HLA-C	Members of the Major Histocompatibility Complex, class I, reported as risk factor for type I diabetes	(Noble and Valdes, 2011)
SMARCA1/ SMARCA5	Member of SWI/SNF chromatin remodeling complex. SMARCA1 promotes neurite outgrowth.	(Barak et al., 2004)
HPGD	Involved in prostaglandin metabolism, when mutated causes cranioosteoarthropathy and patent *ductus arteriosum*	(Seifert et al., 2009)
ITPA	Involved in nucleotide metabolism, when mutated causes early infantile encephalopathy, with brain abnormalities, seizures and cardiac defects	(Kevelam et al., 2015)
SERPINH1	Serine peptidase inhibitor, when mutated causes type-X *osteogenesis imperfecta* with bone fragility, craniofacial dysmorphisms and hearing loss	(Christiansen et al., 2010)
SPTLC1	Enzyme of the sphingolipid biosynthesis pathway, when mutated causes peripheral neuropathy	(Auer-Grumbach et al., 2013)
ALPL	Alkaline phosphatase specific for liver, bone and kidney, when mutated causes infantile hypophosphatasia, a severe skeletal disorder due to abnormal bone mineralization	(Weiss et al., 1988)
ATP6V0A1	V-ATPase subunit involved in neural crest migration	(Tuttle et al., 2014)
IMPDH1	Involved in nucleotide metabolism, when mutated causes Leber Congenital Amaurosis with retinal dystrophy	(Mortimer and Hedstrom, 2005)
PSMC5	Regulatory subunit of the 26S proteasome	(Makino et al., 1997)
UBE2T	E2-ubiquitin conjugating enzyme	(Longerich et al., 2009)
RPL10	Ribosomal subunit, when mutated is a risk factor for autistic spectrum disorders, especially X-linked autism	(Brooks et al., 2014; Klauck et al., 2006)
EIF2D	Translation initiation factor involved in Met-tRNA delivery	(Dmitriev et al., 2010)

Table 3: DEPs found in the protein dataset and their overlaps with clinical or molecular features of both syndromes.

7. Generation of knock-down and sineUP lines for EIF4H

Given the changes in differential expression at each level, it becomes interesting to gain more mechanistic insight into these differences by artificially manipulating the dosage in iPSCs of a key contributor to translational regulation that has already been

linked to several WBS-related phenotypes, EIF4H. The interfered lines, especially when analyzed at the level of translatome, are particularly useful to deconvolute the transcriptional and translational effect of EIF4H on differential gene expression. To lower the abundance of EIF4H we cloned 2 short hairpins for RNA interference (shRNA) in a lentiviral backbone. Conversely, to increase its abundance, we cloned sineUP constructs in the same lentiviral backbone. SineUPs (Zucchelli et al., 2015) are small, single-stranded RNA molecules that recognize target mRNAs at their 5' by sequence complementarity and enhance their translation by mechanisms still under investigation. While shRNAs target mRNAs for degradation, thus lowering their abundance, sineUPs increase their translation, leaving their mRNA abundance unchanged. For each construct a scrambled construct was included (shSCR and sineUP-SCR) as a control. A single control iPSC line (MIFF3) was infected with sineUP and shRNA lentiviral particles and selected with puromycin. Unfortunately, the sineUP construct did not cause the expected up-regulation (fig. 14B), whereas both shRNAs were successfully down-regulating their target (fig. 14A). The failure of the sineUP modulation may be due to an absence of the necessary molecular mechanisms for its function specifically in iPSCs, or to a suboptimal design of the sineUP construct that would require additional tweaking. For the time being, we decided to proceed with shRNAs only.

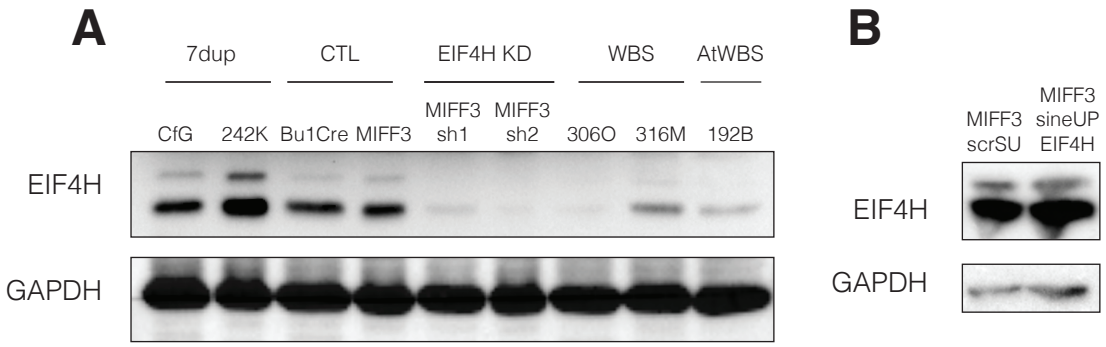

Figure 14: Validation of EIF4H knock-downs and sineUP. A) Western blot on EIF4H and GAPDH to evaluate the effect of both short hairpins. B) Western blot on EIF4H and GAPDH to evaluate the effect of the sineUP and its scramble counterpart.

Since both EIF4H shRNAs worked well, we included both lines (sh1 and sh2) for further downstream analysis by ribosome profiling and SWATH-MS.

Owing to small sample size, we decided not to perform differential gene expression analysis on the two hairpins compared to the scramble, but rather use these datasets as a validation tool for differential expression. By assessing the effect of EIF4H on the RNA level of DEGs found in the transcriptome (fig. 15) we can appreciate a substantial down-regulation of more than half of the transcripts, with 9 genes having a log2(FC) over scramble of less than -1, corresponding to a negative 2-fold change, using both hairpins. Two genes (MTRNRL1 and MTRNRL 8) are instead up-regulated by more than 1 log2(FC) upon EIF4H knockdown using both hairpins.

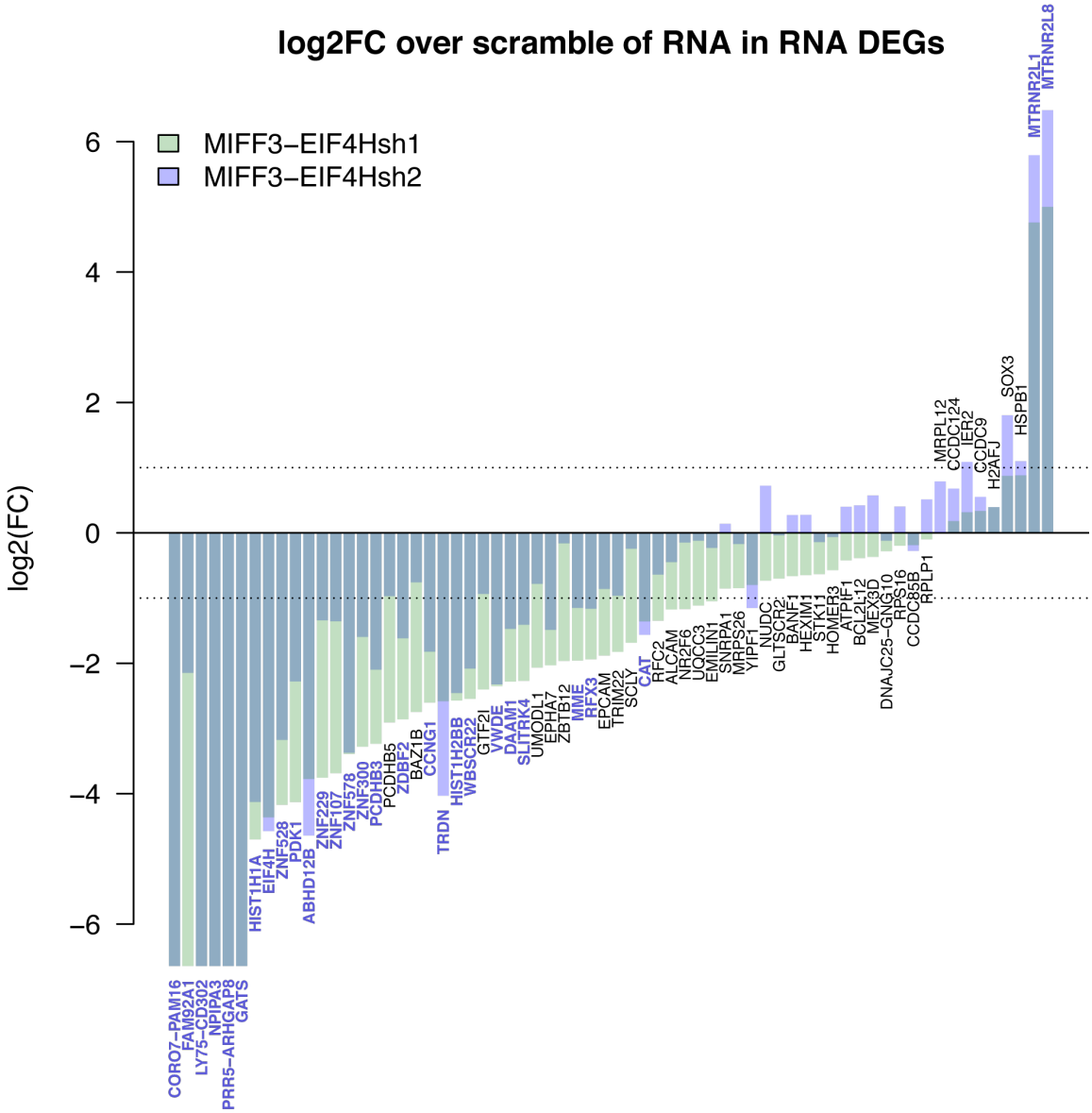

Figure 15: Barplot of log2(FC) at the RNA level of DEGs found in the RNA dataset. In blue, genes that are down- or up-regulated more than 2-fold by both EIF4H short hairpins compared to the scramble hairpin. The black dotted line represents +/- 1 log(FC), corresponding to a 2-fold change.

The down-regulation is also visible, albeit at a lesser extent, when computing the log2(FC) of these genes in the translatome (fig. 16), and it is counterbalanced by a remarkable part of up-regulated DEGs upon EIF4H knock-down.

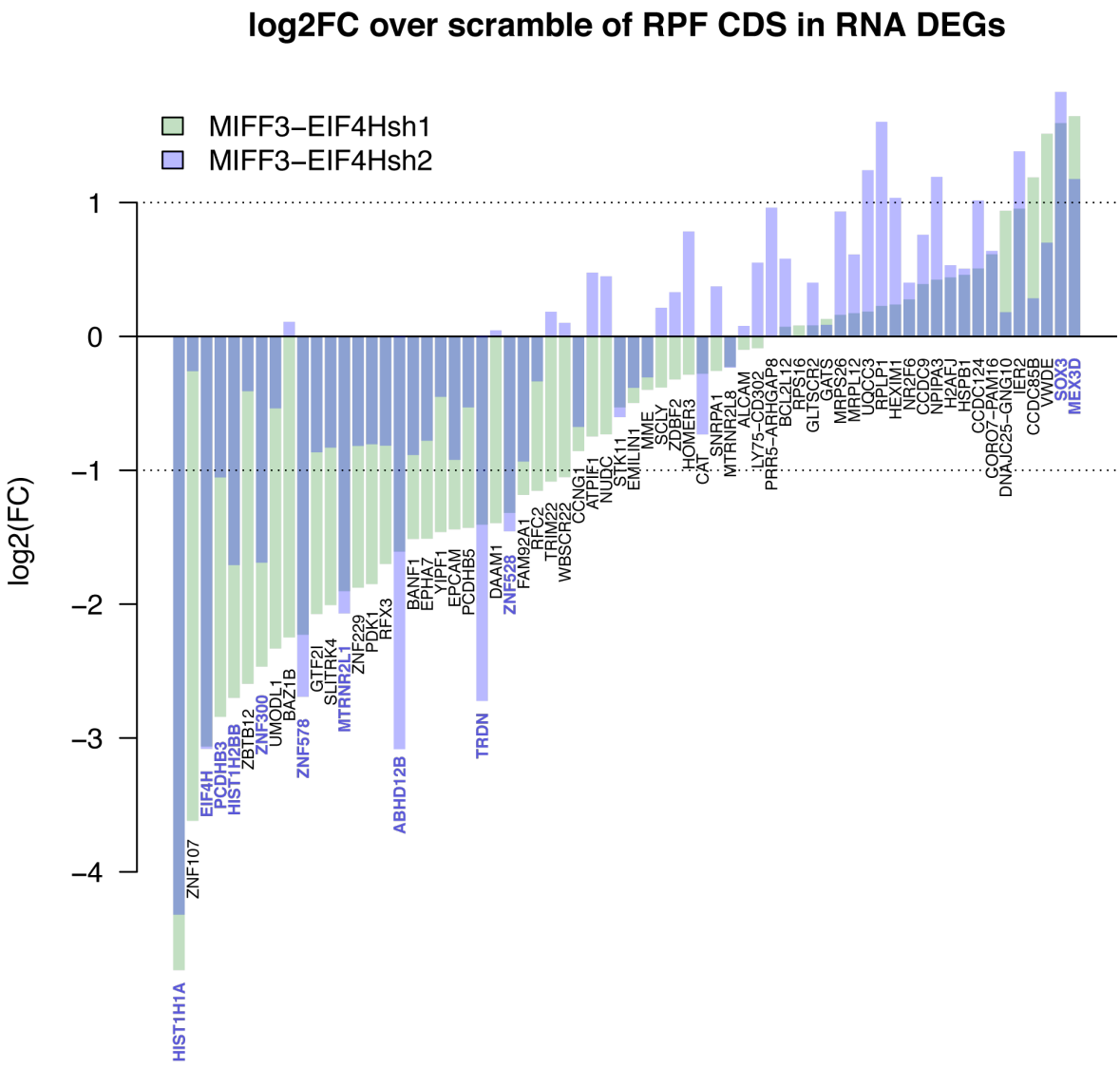

Figure 16: Barplot of log2(FC) at the RPF level of DEGs found in the RNA dataset. In blue, genes that are down- or up-regulated more than 2-fold by both EIF4H short hairpins compared to the scramble hairpin. The black dotted line represents +/- 1 log(FC), corresponding to a 2-fold change.

By plotting changes in TE with respect to scramble of genes that are called as DE only in the total RNA dataset (fig. 17), it becomes evident that more than half of the genes have a decrease in translation efficiency, with 16 genes for which TE is decreased by more than 2-fold in both hairpins, pointing to a strongly predominant effect of translation over transcription in this subset. The number of these genes increases to 22 when considering a more lenient threshold of 1.5-fold decrease in TE. This down-regulation points to a possible masking of the transcriptional effect by EIF4H,

although it does not allow us to discern whether it is a direct or indirect effect on specific mRNAs.

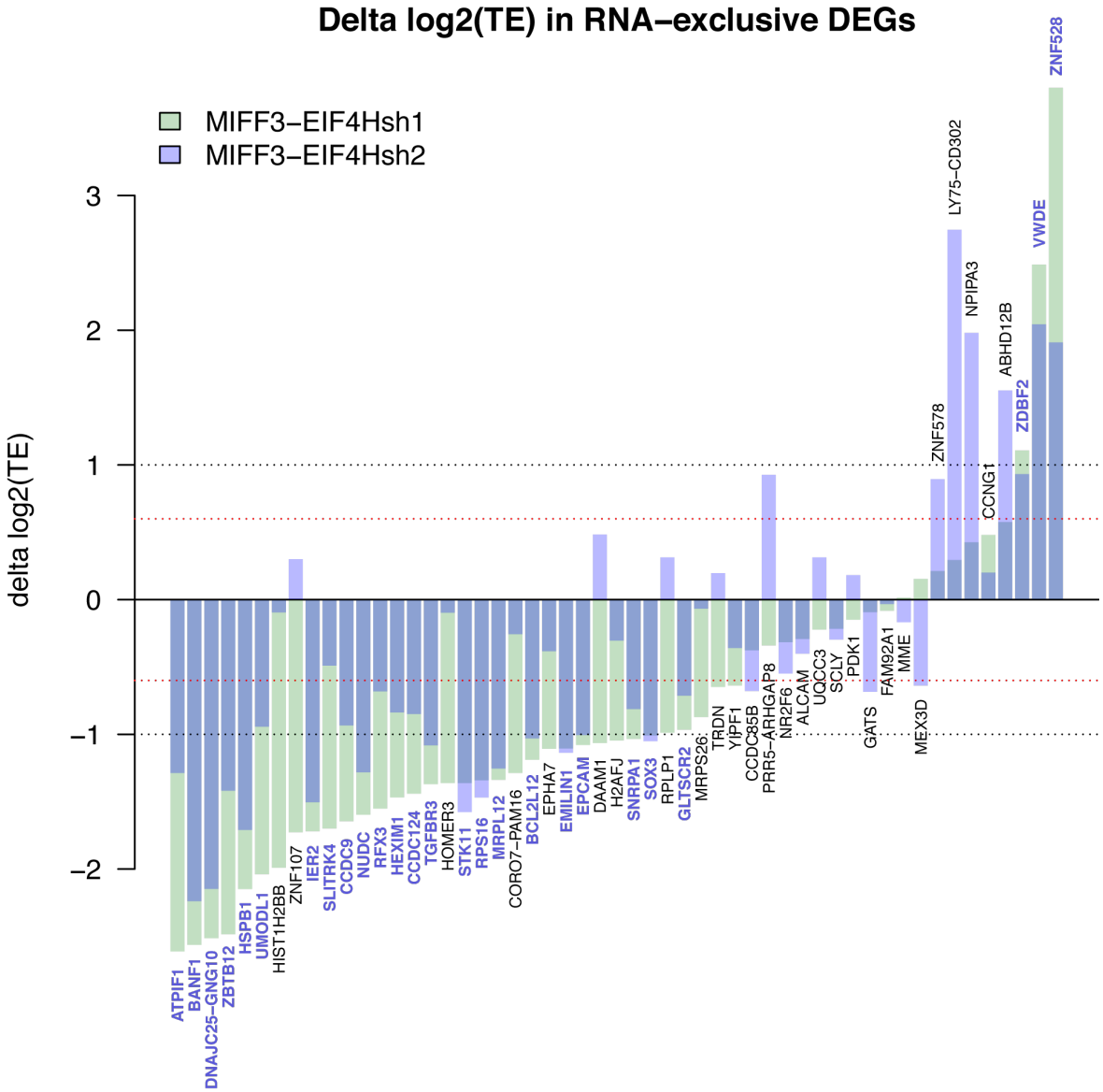

Figure 17: Barplot of differences in log2(FC) TE of DEGs found exclusively in the RNA dataset. In blue, genes whose TE is down- or up-regulated more than 1.5-fold by both EIF4H short hairpins compared to the scramble hairpin. The black dotted line represents +/- 1 log(FC), corresponding to a 2-fold change, the red dotted line represents +/- 0.6 log(FC), corresponding to a 1.5-fold change.

When applying the same TE analysis to genes that were differentially expressed in the translatome (fig. 18), we can appreciate an almost constant down-regulation of translation efficiency of DE genes, with 6 of them being down-regulated by 2-fold in 2 hairpins (9 genes by 1.5 fold). With the exception of ANKRD1, the drop in TE of these genes upon EIF4H knock-down is consistent with their trend of expression in the RPF

dataset, which correlates with EIF4H levels. These 9 genes may be directly regulated by EIF4H, although it is not possible to discern between direct and indirect effects at this stage of the analysis.

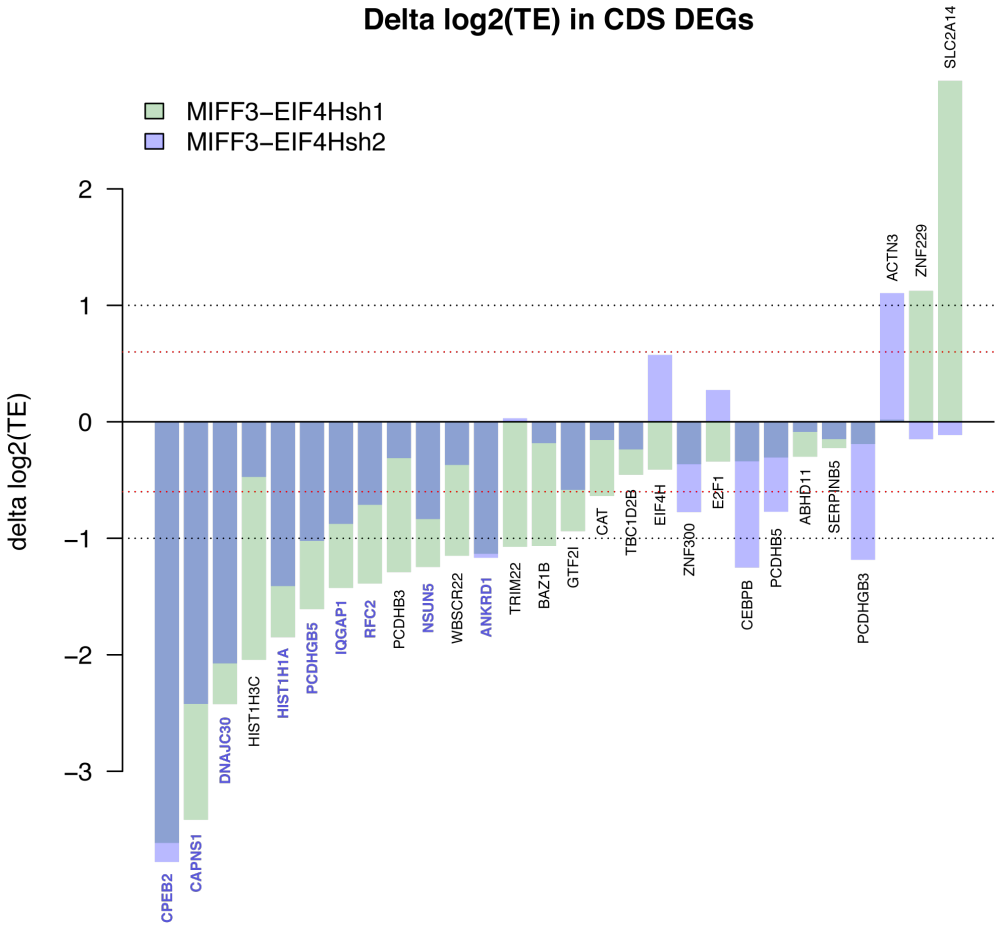

Figure 18: Barplot of differences in log2(FC) TE of DEGs found in the RPF dataset. In blue, genes whose TE is down- or up-regulated more than 1.5-fold by both EIF4H short hairpins compared to the scramble hairpin. The black dotted line represents +/- 1 log(FC), corresponding to a 2-fold change, the red dotted line represents +/- 0.6 log(FC), corresponding to a 1.5-fold change.

8. A regression-based approach reveals the modes of propagation of changes through molecular layers

Pairwise comparisons between samples are informative on the type and extent of dysregulation occurring between a particular disorder and a control, but provide

limited insight into how the dosage of genes in the WBSCR affects gene expression. The presence of key regulators of gene expression at the level of transcription and translation, and the availability of genome-wide measurements of three different expression layers, allows to: 1) make specific hypotheses on the relationship between the abundance of WBSCR genes and differences in gene expression across samples, and 2) measure the extent to which differences in gene expression in one layer are propagated to another.

In order to understand the effect on differential expression caused by changes in dosage and abundance of key regulators of gene expression, we adopted a novel strategy that uses linear regression on the abundance of single proteins (measured by SWATH-MS) to infer the magnitude and direction of changes in DEGs and DEPs. Briefly, we used the log-normalized intensity of "query" proteins as covariates in a generalized linear model (based on the negative binomial distribution implemented by edgeR) for total RNA and RPF, and in the standardized major axis (SMA) linear regression for the proteome. In other words, we checked whether there was a linear relationship between the log-transformed values of RNA or RPF abundance in each sample, and the log-transformed intensity values of a query protein in the same samples, and whether this relationship was statistically significant. These linear relationships were constructed using not only the samples from patients, but also the EIF4H knock-down and scramble samples. For every gene in each layer we obtained a slope (corresponding to the log-fold change for RNA and RPF, and the slope of the SMA linear model for the proteome) whose sign and value serve as a response coefficient to the query protein, and whose p-value describes the goodness of fit in the linear model. A gene with a statistically significant slope of +1 is perfectly correlating with the query protein, meaning that changes in the independent variable (query) result in proportional changes in the response variable. A gene with a

negative slope is anticorrelated. Genes with slopes that deviate from 1 are potentially targeted by one or more mechanisms that exacerbate or attenuate their correlation with the query protein. Genes with no statistically significant slope are not correlated to the query protein in that layer, possibly because they are not responsive to changes in query protein abundance. Although this method does not allow to draw precise causal links, it still allows to generate hypotheses in a mechanistic framework which will be validated with more targeted approaches. The use of slopes becomes particularly relevant when they are compared across layers. We computed slopes on RNA, RPF and proteome datasets using EIF4H as query. Even though read counts in total RNA and RPF have remarkable differences in their magnitude compared to protein intensities, slopes represent relative changes in abundance, thus making RNA and protein measurements numerically comparable (fig. 19).

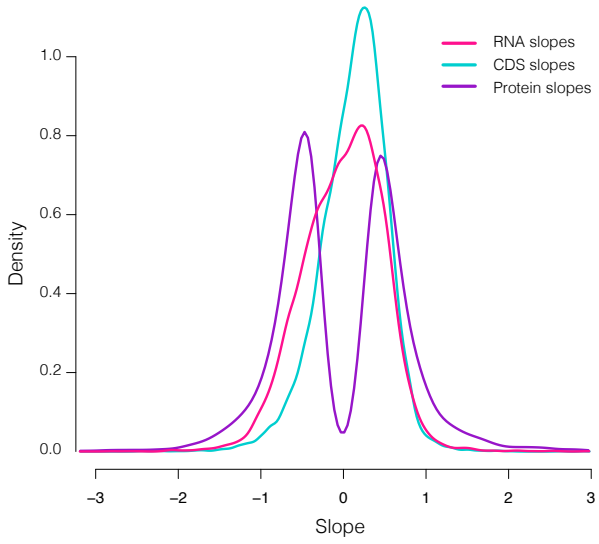

Figure 19: Total RNA, RPF and protein slopes are numerically comparable. Density plots showing the distribution of slopes by regressing on EIF4H (x axis) for all three layers. The bimodal distribution of protein slopes is due to the SMA regression test that assigns a non-zero value to each slope regardless of its statistical significance.

In principle, this method can be applied to all genes for which we can identify a statistically significant slope, with the *proviso* that many of them are unlikely to pass multiple test correction due to the high heterogeneity of samples, thus allowing us to perform only a descriptive analysis of distribution of slopes. For each analysis we can

analyse the universe of statistically significant slopes (Fig. 20A), which can then be deconvoluted into three components: intersection of genes with statistically significant slopes in both total RNA and RPFs (Fig. 20B), genes with statistically significant slopes only in total RNA (Fig. 20C) and genes with statistically significant slopes only in RPF (Fig. 20D). The fact that RNA and RPF slopes occupy almost exclusively quadrants I and III (fig. 20A) shows that most of the changes that are correlated, positively or negatively with EIF4H levels, are carried forward from the transcriptome to the translatome. As expected, genes that have significant slopes in both datasets lie close to the diagonal (fig. 20B), meaning that their changes in transcription are substantially promoted to translation. However, two interesting subsets of genes can be appreciated by looking at slopes that are significant only in one of the two layers (fig. 20C,D). These subsets, which will be referred to as "RNA-only" (fig. 20C) and "RPF-only" (fig. 20D), are delimited by including only those genes which have a statistically significant slope in one layer and not in the other, pointing to a regulation (or, at least, a correlation) that is pre-eminently acting in one of the two layers. Genes with significant slopes only in total RNA (fig. 20C) may be subjected to buffering events that dampen differences at the level of translation, while genes with significant slopes only in RPF (fig. 20D) are affected by post-transcriptional regulatory processes that are not detectable in the transcriptome. When performing GO enrichment analysis on genes that have significant slopes in only one of the two layers, we find that genes exclusive to the RNA layer are enriched in processes involved in cytoskeleton reorganization, protein localization and, strikingly, axonogenesis (fig. 21), thus confirming that iPSCs present, at the transcriptional level, perturbations that affect disease-relevant pathways.

The RPF-exclusive layer genes have different GO enrichments for processes more related to RNA metabolism and post-transcriptional regulation, and in particular are

enriched for processes involved in cell cycle regulation and genes localized in the nucleus (fig. 22). These findings point again at layer-specific modes of regulation of different cellular processes that can be appreciated only by integrating and comparing different datasets.

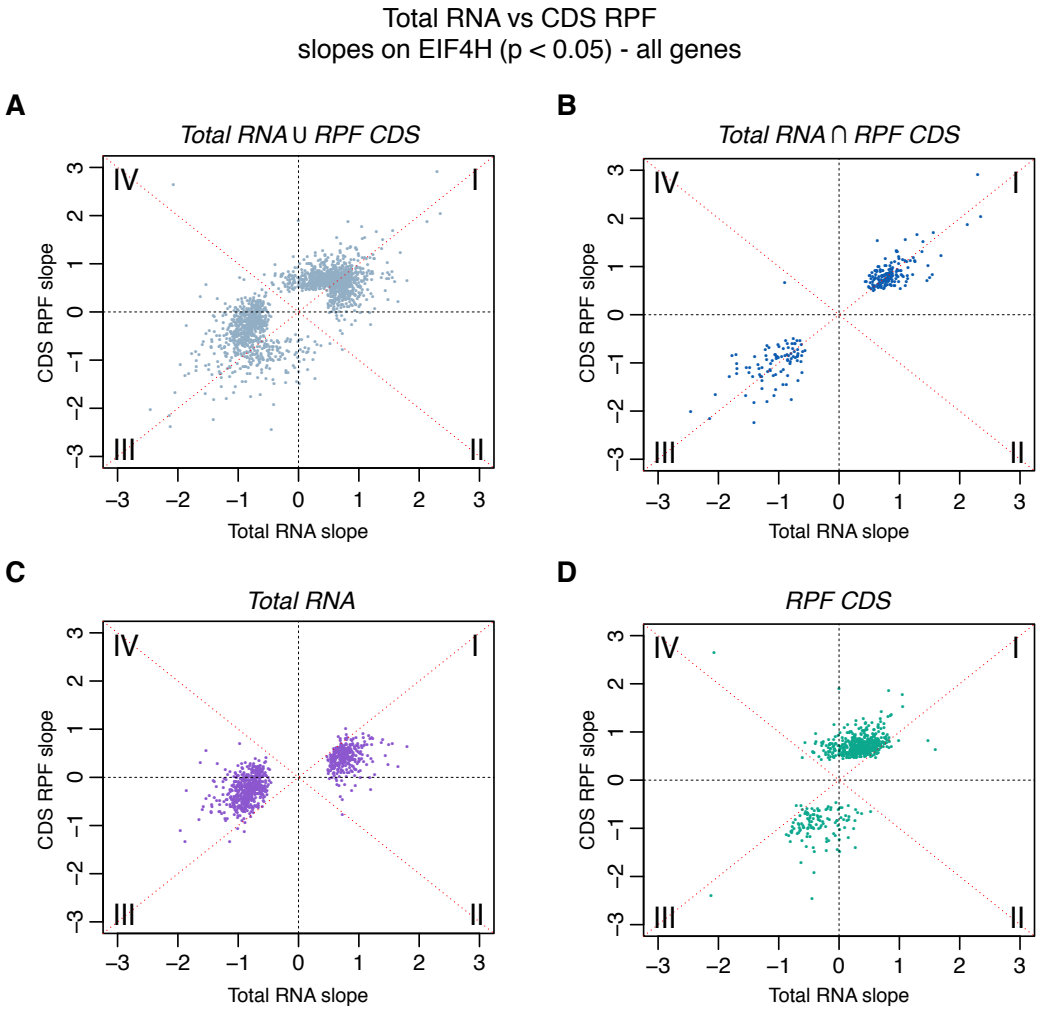

Figure 20: Quadrant graphs of statistically significant ($p < 0.05$) slopes in total RNA and RPF allow to readily visualize the type and extent of correlation with EIF4H, and its propagation across molecular layers. Each gene is represented by a coloured dot. A) Union of all genes with statistically significant slopes in either layer (light blue). B) Intersection of genes with statistically significant slopes in both layers (dark blue). C) Genes with statistically significant slopes only in total RNA (purple). D) Genes with statistically significant slopes only in RPF (green). Red dotted lines indicate diagonals, black dotted lines delimit quadrants (roman numbers displayed in each corner).

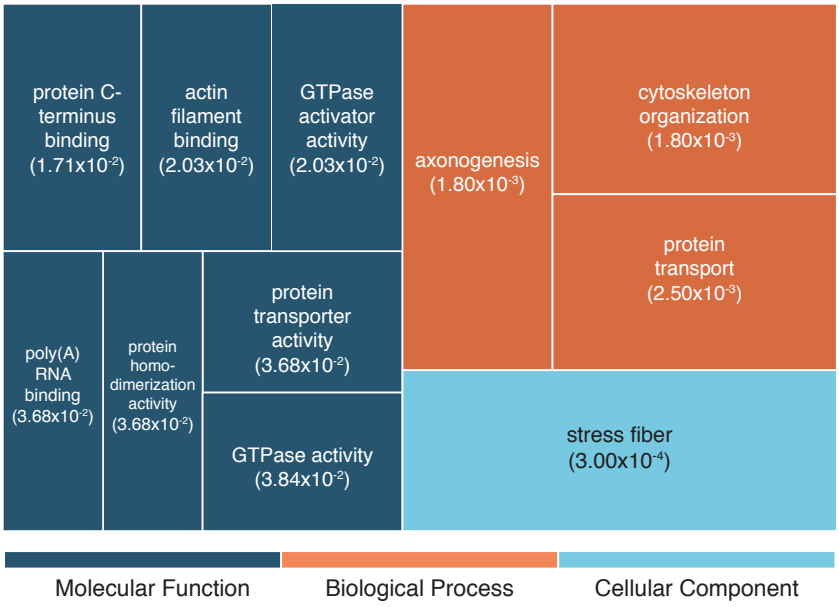

Figure 21: Treemap representing Gene Ontology terms for which there is a statistically significant enrichment among RNA-exclusive genes in the RNA vs RPF comparison. Parent categories with enriched children were removed. The size of each box is proportional to the statistical significance of the enrichment. Between brackets the p-value adjusted using Benjamini-Hochberg correction.

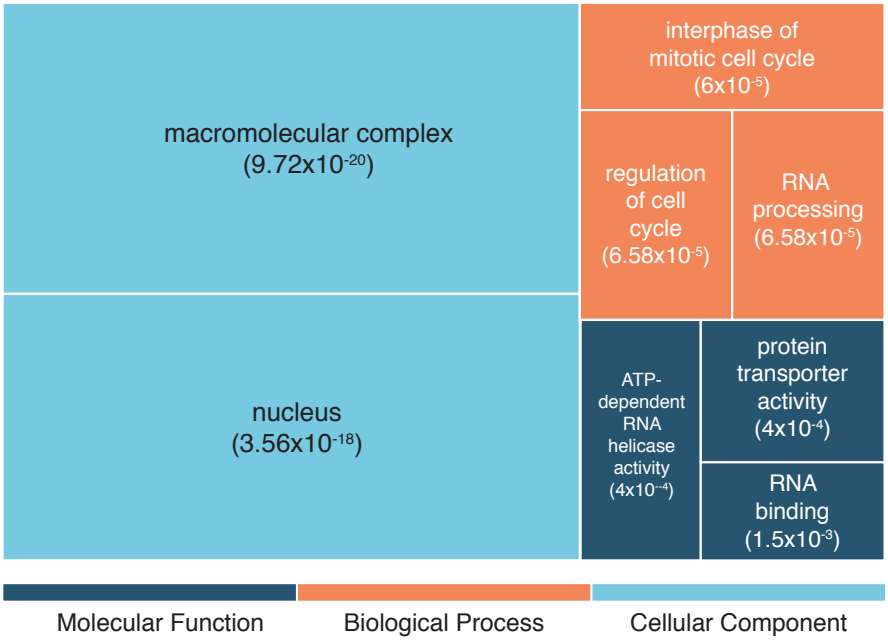

Figure 22: Treemap representing Gene Ontology terms for which there is a statistically significant enrichment among RPF-exclusive genes in the RNA vs RPF comparison. Parent categories with enriched children were removed. The size of each box is proportional to the statistical significance of the enrichment. Between brackets the p-value adjusted using Benjamini-Hochberg correction.

Taking in consideration that EIF4H is a translation initiation factor, we expect to see that most of the genes respond, at the level of translation, to increases in EIF4H abundance in a positive and dosage-sensitive fashion. And indeed our slope analysis shows the expected trend (Fig. 20D). However, since other genes of the WBSCR necessarily follow a trend that closely mirrors that of EIF4H, this evidence alone does not warrant a specific causal implication of EIF4H in the regulation of these genes. Such a link can be further probed in depth by looking at this subset of genes in the context of the EIF4H knock-down. We therefore asked whether genes clustering in the first and fourth quadrant of the RPF-only graph were significantly less translated in both RPF EIF4H knock-down datasets when compared to their scramble counterpart. Indeed, log-transformed TE distributions for genes in quadrants I and IV have a statistically significant ($p < 2.2x10^{-16}$) difference when compared to the scramble (fig. 23), showing a median reduction of 0.63 log (MIFF3 scramble vs MIFF3 EIF4H sh1) and 0.39 log (MIFF3 scramble vs MIFF3 EIF4H sh2), corresponding to a ~1.8 fold and ~1.5 fold reduction respectively.

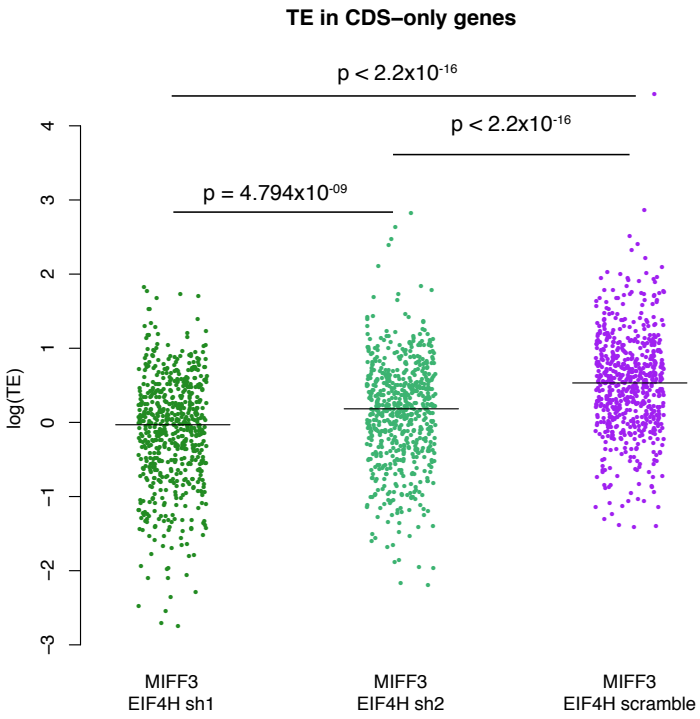

Figure 23: Effect of EIF4H knock-down on RPF-exclusive genes. Distributions of log-transformed FPKM of RPF reads for RPF-exclusive, I+IV quadrant genes in the context of EIF4H knock-down. Both short hairpins against EIF4H reduce translation levels of these genes in a statistically significant way ($p < 2.2x10^{-16}$ according to a two-sample t-test). The black bar indicates the median of each distribution.

By applying the same analysis of slopes to the comparison of translatome and proteome (fig. 24) we can immediately visualize a remarkable shift in the distribution of slopes, with a large portion of genes clustering in the second quadrant (fig. 244A). A large subset of genes in this quadrant has significant slopes in the RPF dataset only (fig. 24C), pointing to a specific compensation of changes in translation at the level of co- or post-translational regulation. Genes in the second quadrant with a significant slope in the protein dataset only (fig. 24D) are likely being dysregulated in the absence of significant changes in translation. The presence of a sizeable proportion of genes in the first quadrant indicates that the increases in translation that correlate with EIF4H levels mostly result in increases in protein abundance, again as expected given the role of EIF4H in translation. GO enrichments for RPF-exclusive genes in this analysis partially overlap with those found in the transcriptome vs translatome

comparison (fig. 25), pointing to a strongly translation-specific regulation of cell cycle and RNA metabolism.

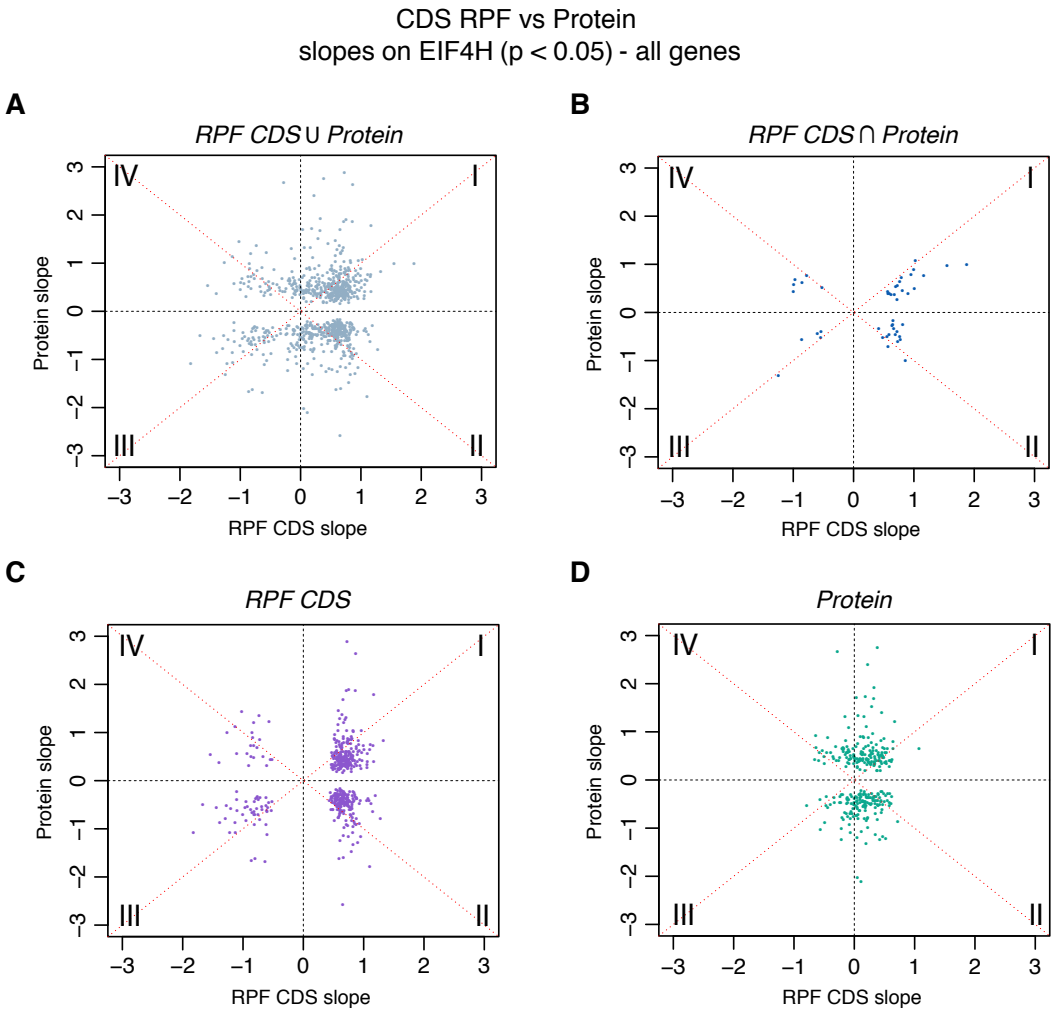

Figure 24: Quadrant graphs of statistically significant ($p < 0.05$) slopes in RPF and proteins. A) Union of all genes with statistically significant slopes in either layer (light blue). B) Intersection of genes with statistically significant slopes in both layers (dark blue). C) Genes with statistically significant slopes only in total RNA (purple). D) Genes with statistically significant slopes only in RPF (green). Red dotted lines indicate diagonals, black dotted lines delimit quadrants (roman numbers displayed in each corner).

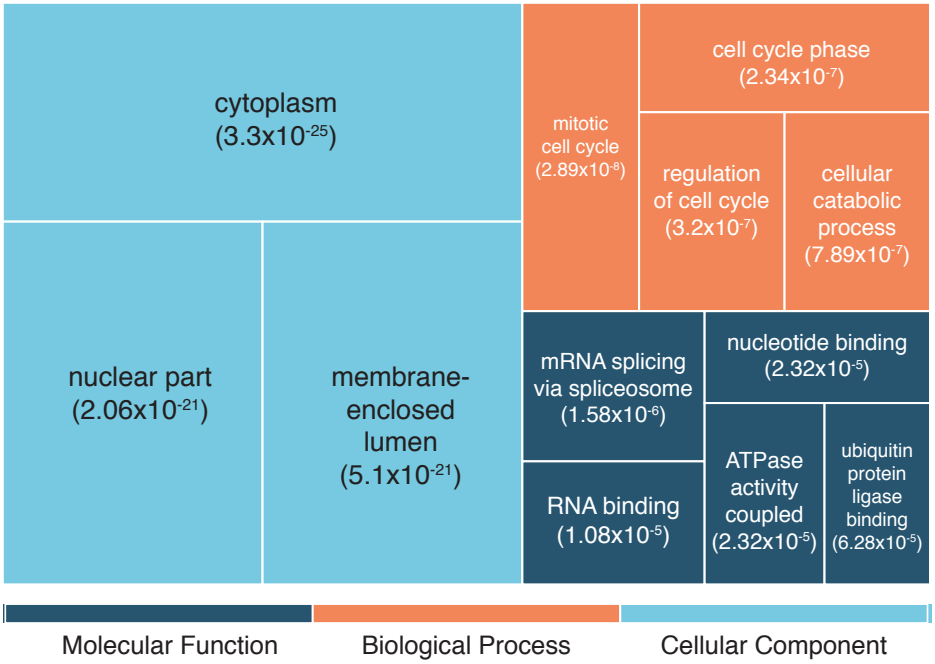

Figure 25: Treemap representing Gene Ontology terms for which there is a statistically significant enrichment among RPF-exclusive genes in the RPF vs protein comparison. Parent categories with enriched children were removed. The size of each box is proportional to the statistical significance of the enrichment. Between brackets the p-value adjusted using Benjamini-Hochberg correction.

On the other hand, protein-exclusive genes (fig. 26) are mainly related to energy metabolism and protein modification in the endoplasmic reticulum, albeit with less significant enrichments and broadly specified parent categories. Taken together, these results hint at a widespread remodeling of the proteome in the pluripotent state that, in many cases, reverts the effect of transcriptional and translational imbalances.

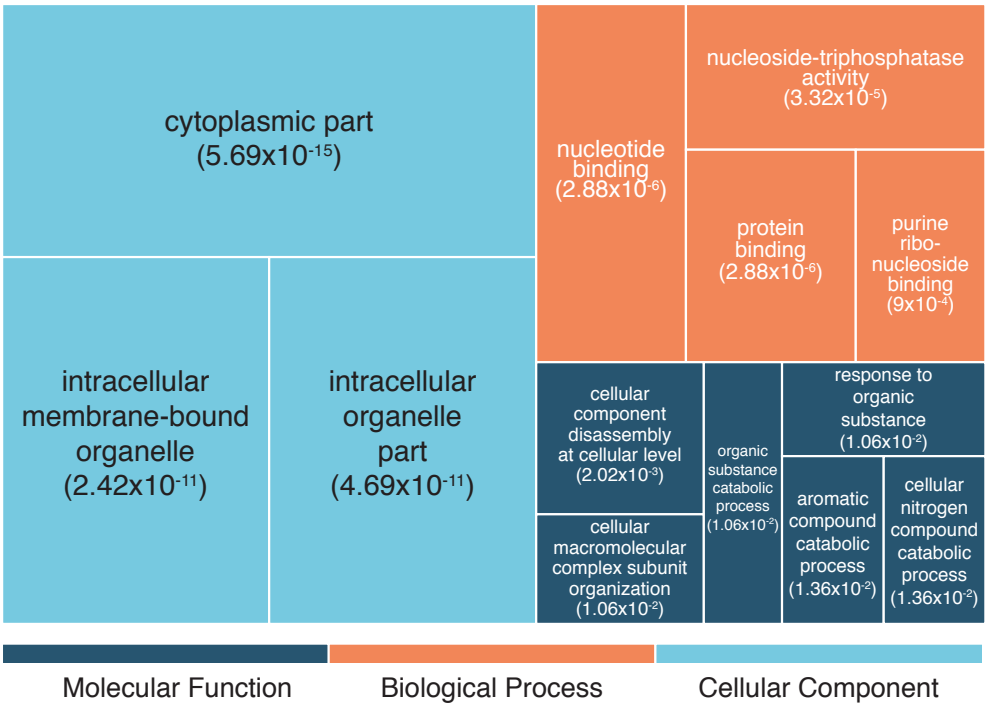

Figure 26: Treemap representing Gene Ontology terms for which there is a statistically significant enrichment among protein-exclusive genes in the RPF vs protein comparison. Parent categories with enriched children were removed. The size of each box is proportional to the statistical significance of the enrichment. Between brackets the p-value adjusted using Benjamini-Hochberg correction.

9. Regression-based approach on DEGs

Yet, if we limit our analysis to the set of DEGs for which we have good statistical confidence, i.e. the union of DEGs found in the initial RNA-seq experiment, and the DEGs and DEPs found in this study, we can query them without losing statistical power, thus making more robust claims about types and level of dysregulation for single genes (fig. 27, 28). Therefore, besides observing the global distributions of genes, we can infer how differences between layers are propagated for specific genes. We observed again that changes in the transcriptome are substantially promoted to the translatome (fig. 27A), even more so when considering genes with significant slopes in both RNA and RPF datasets (fig. 26B), with the exception of BANF1, that lies in quadrant IV. RPF-only DEGs (fig. 27D) cluster more in the positive RPF quadrants (I and IV) than in the negative quadrants (II and III), whereas RNA-only DEGs (fig.

27C) tend to be more down-regulated according to EIF4H levels. Comparing slopes in RPF and proteome (fig. 28) it can still be appreciated that a fraction of protein slopes changes direction in the II quadrant (fig. 28C, 28D), underscoring the compensatory effect on genes and proteins that have a statistically significant difference in expression. However, it must be pointed out that when comparing RPF and proteome we are limited by the number of confidently identified proteins included in the union of DEGs, which considerably lowers the number of available slopes to 96.

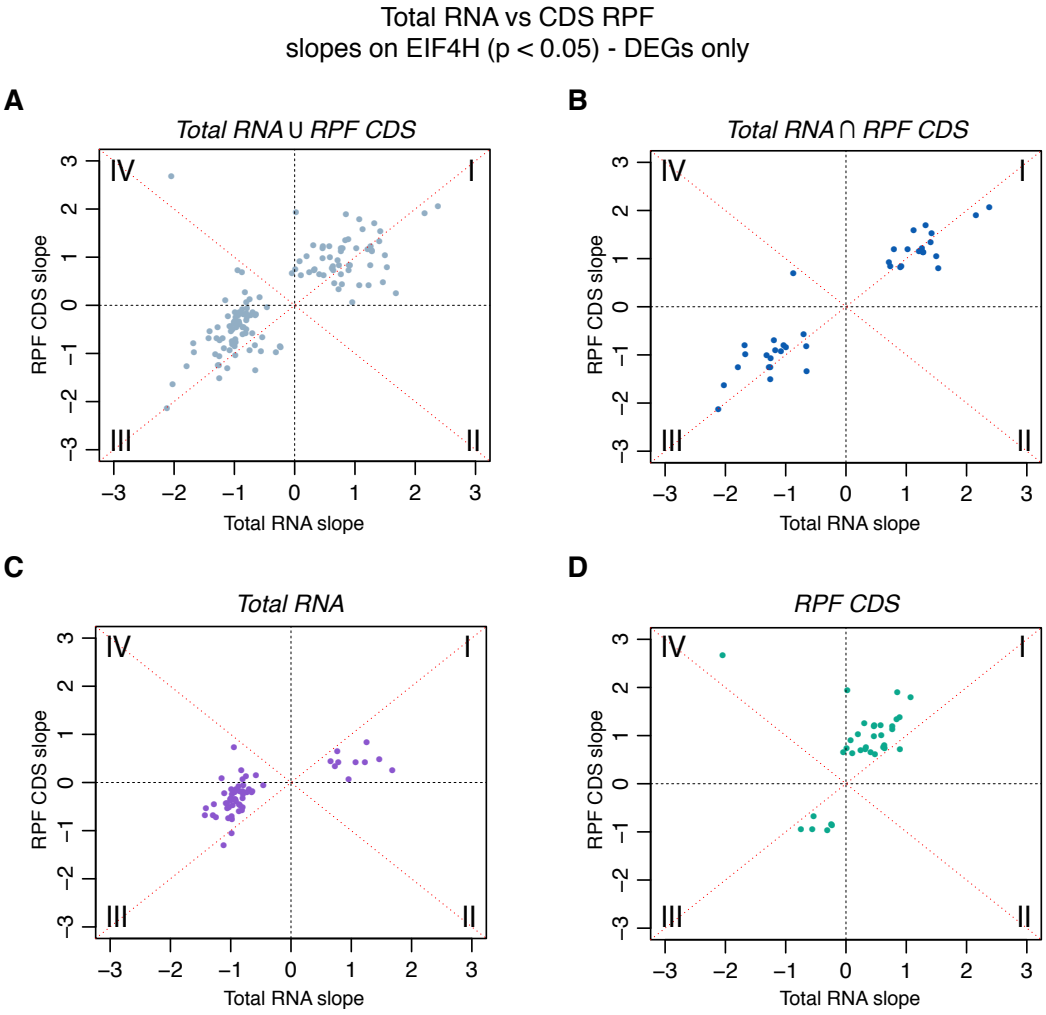

Figure 27: Quadrant graphs of statistically significant ($p < 0.05$) slopes for DEGs in total RNA and RPF. A) Union of all genes with statistically significant slopes in either layer (light blue). B) Intersection of genes with statistically significant slopes in both layers (dark blue). C) Genes with statistically significant slopes only in total RNA (purple). D) Genes with statistically significant slopes only in RPF (green). Red dotted lines indicate diagonals, black dotted lines delimit quadrants (roman numbers displayed in each corner).

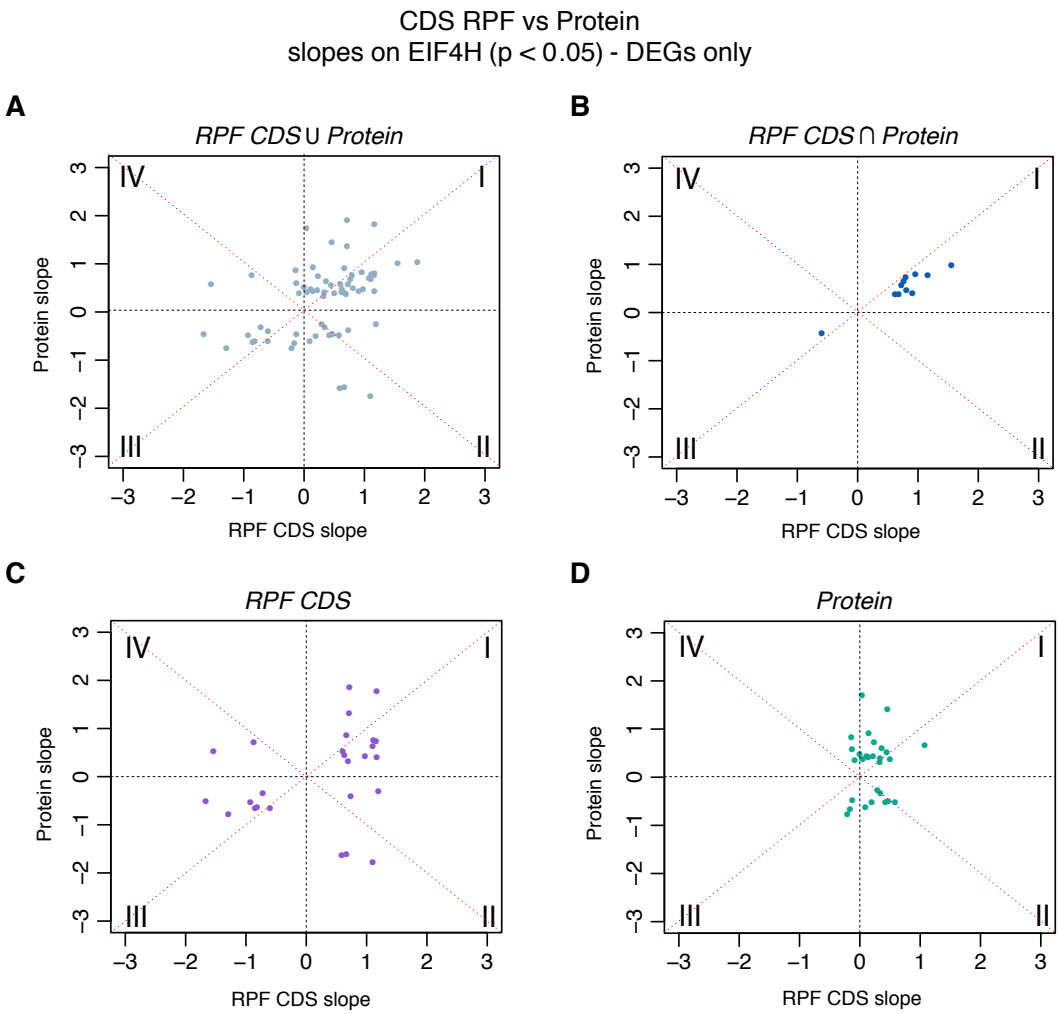

Figure 28: Quadrant graphs of statistically significant ($p < 0.05$) slopes for DEGs and DEPs in RPF and proteome. A) Union of all genes with statistically significant slopes in either layer (light blue). B) Intersection of genes with statistically significant slopes in both layers (dark blue). C) Genes with statistically significant slopes only in total RNA (purple). D) Genes with statistically significant slopes only in RPF (green). Red dotted lines indicate diagonals, black dotted lines delimit quadrants (roman numbers displayed in each corner).

Having obtained slopes for all three layers of DEGs for which we have statistical confidence, it is interesting to observe their trajectories from transcriptome to proteome (fig. 29). Hierarchical clustering of slopes identifies up to 5 different groups that follow similar trajectories, which we term *archetypes*. An expression archetype is an ideal shape drawn by the line connecting slopes in the three layers. Borrowing a simple metaphor from geography, we can identify several theoretical archetypes that resemble topographical elements, e.g. genes whose slope peaks on

translation resemble *hills*, genes whose slopes increase constantly are *creeks*, and so on. Out of all the possible shapes we identify 5 archetypes that we term *cliff* (fig. 30A), *shore* (fig. 30B), *hill* (fig. 30C), *plateau* (fig. 30D) and *creek* (fig. 30E) corresponding to the 5 clusters in the heatmap (fig. 29).

Strikingly, 3 of these archetypes are enriched for different gene ontology categories, with the remarkable example of morphogenesis and neuronal differentiation in the *creek* archetype. This division and representation serves as a functional classification of genes according to their gene expression patterns, in a way that takes into account the way they are transferred from one layer to the other, and captures novel patterns of information that result from the integration of all three layers of expression.

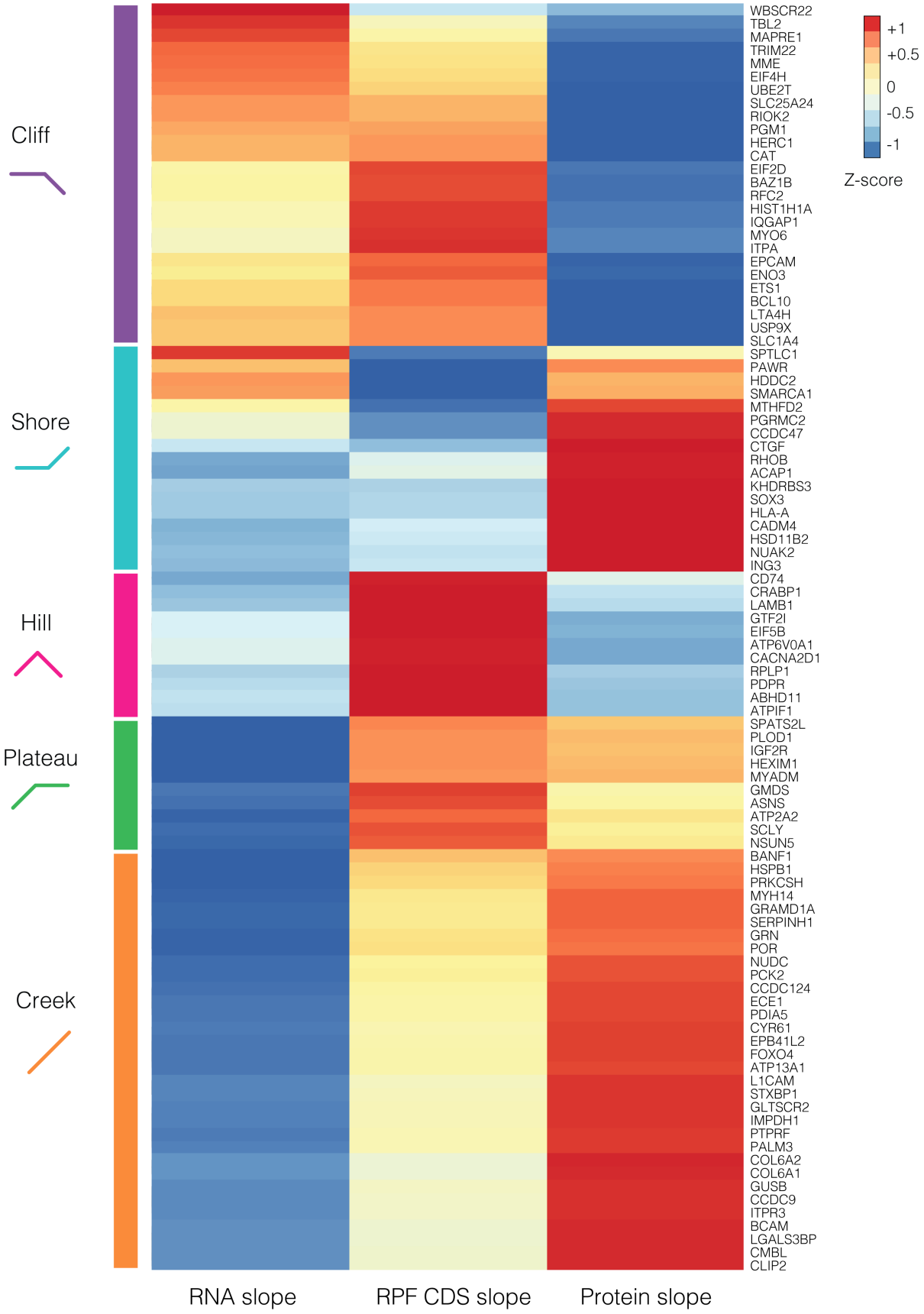

Figure 29: Trajectories of changes in expression across layers. Heatmap of Z-score for slopes in each layer. On the left, the schematic representation of archetypes.

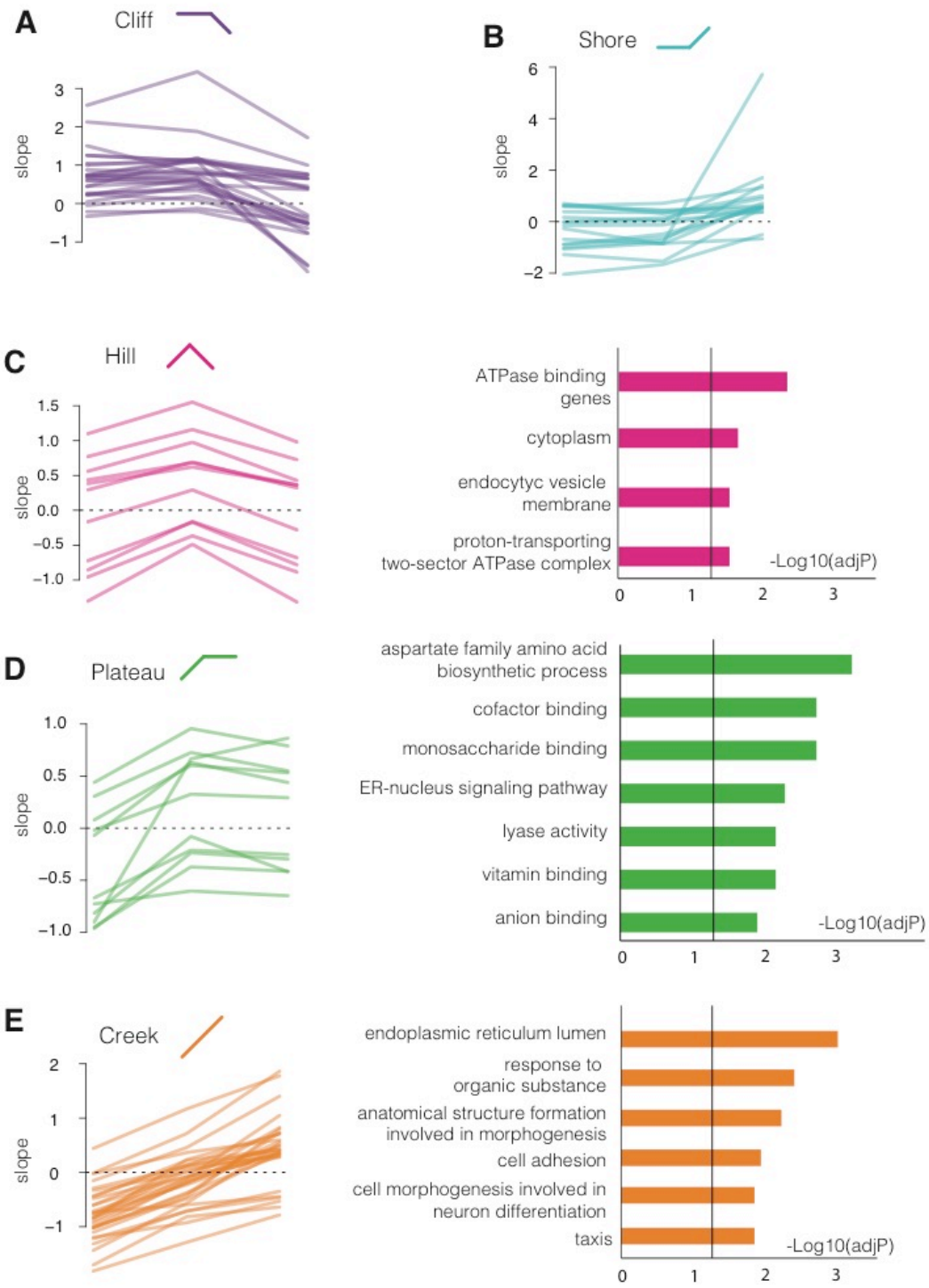

Figure 30: Representation and division of trajectories in their respective archetypes, and GO enrichments for *hill, plateau* and *creek* archetypes. The black line in the enrichment plots represents adjP = 0.05.

10. Determination of protein degradation rates and gene expression modeling

Our combination of datasets probing different molecular layers allows to pose more general questions regarding the regulation of gene expression in pluripotency. By integrating these datasets with a proteome-wide measurement of protein degradation kinetics, two goals can be achieved: gaining a deeper understanding on how differentially expressed genes are regulated at different steps, and verify a mathematical model of gene expression with our experimental data.

To perform a proteome-wide estimation of degradation rates we applied pulsed-Stable Isotope Labeling of Aminoacids in Culture (pSILAC) to a subset of our iPSC lines. Briefly, in a pSILAC design cells grow in a normal medium until a certain time t0. At t0, the culture medium is completely swapped with a "heavy" medium that contains arginine and lysine aminoacids labeled with heavy, non-radioactive nitrogen and carbon isotopes. Upon incorporation in newly synthesized proteins, these heavy aminoacids confer to proteins a heavier molecular weight. This mass shift is visible by mass spectrometry, thus allowing to make quantitative comparisons between labeled and unlabeled samples. Proteins are harvested during a time-course, and for each time point, a mixture of "light" or "old" proteins, i.e. synthesized in the medium prior to t0, and "heavy" or "new" proteins, synthesized during the pulse, will be present in the lysate. Processing these proteins by mass-spectrometry yields the simultaneous identification and quantification of both heavy and light proteins. For each protein at each time point the intensity of "heavy" proteins can be divided by the intensity of their respective "light" counterpart, with the advantage that, since each of these amounts is relative, no normalization is needed. In order to use a data-independent approach such as SWATH-MS, it is first necessary to perform a more traditional Shotgun-MS measurement, so that a spectral library for pSILAC pairs can be generated and used to instruct the SWATH-MS identification and quantification

routine. Log-transformed heavy-over-light (log(H/L+1)) values can be drawn along the time course in order to infer the turnover rate of each protein, which is the slope coefficient of the linear model that describes the relationship between log(H/L+1) and time. In other words, we assume that there is a linear relationship between the increases in the log(H/L + 1), measured at each time point, and the increases in time. The coefficient of this linear relationship (slope) will be the turnover rate.

However, protein turnover is the combined effect of synthesis and degradation of proteins. To infer degradation rates, we need to take into account the differential equation known for first-order kinetics:

$$\frac{dP}{dt} = K_{syn} + P \cdot K_{deg}$$

where in each interval *dt*, an amount of protein is gained, determined by its synthesis rate K_{syn} and a different amount of protein is lost, determined by its degradation rate K_{deg} and the amount of protein at the beginning of the interval. This equation can be easily solved with the assumption that, at steady state, the protein amount does not change:

$$\frac{dP_{SS}}{dt} = 0$$

$$0 = K_{syn} + P_{SS} \cdot K_{deg}$$

which brings us to a simple determination of both constants

$$P_{SS} = \frac{K_{syn}}{K_{deg}}$$

An important implication of the mass action law is that differences in steady-state protein amounts are directly influenced by degradation rather than synthesis.

There are several ways to determine K_{deg} starting from heavy and light protein abundances in this experimental design. We use a slightly modified version of Pratt's Relative Isotope Abundance (R.I.A) (Pratt et al., 2002), which assumes that the sum of heavy and light proteins is equivalent to the steady state in each sample:

$$R.I.A. = \frac{L}{H + L}$$

Since our experimental design entails a total replacement of L aminoacids with H ones, the L abundance is bound to decrease in time, and from the change of R.I.A. of each protein in time we can infer the degradation rate:

$$R.I.A.(t) = \frac{L_t}{H_t + L_t}$$

At each time point, R.I.A. will be influenced by the degradation of the light protein in the previous time point, and the presence of newly synthesized heavy proteins. We will assume that the total amount of proteins, represented by the sum of H and L proteins, is constant and equivalent to the steady state.

Solving this equation in continuous time leads to the expression that links R.I.A. and time by means of an exponential decay:

$$R.I.A.(t) = R.I.A.(t0) \cdot e^{-K_{deg} \cdot t}$$

Where R.I.A.(t0) is R.I.A. at the starting point.

It is reasonable to assume that, at the beginning of the time series, there are no heavy isotopes, so that R.I.A.(t0) is equal to 1. At an infinite time, R.I.A. decays exponentially from 1 to 0. Therefore, K_{deg} can be simply obtained by fitting an exponential model to experimental values of R.I.A. over time, or fitting a linear model to log(R.I.A.) values in time. An initial pilot experiment was performed on a single control iPSC line (fig. 31A), harvesting proteins at 1.5, 4.5 and 13.5 hours as done by Schwannhausser and colleagues (Schwanhausser et al., 2011), in order to perform a preliminary assessment of how heavy aminoacids were impacting cell culture, and the resolution of data that can be obtained. Each sample was prepared in duplicate.

Indeed, the number of MS2 spectra (i.e. spectra of the fragment ions, identified in the second part of the analysis upon precursor selection in MS1) with a SILAC pattern that can be identified and quantified increases with time and shows consistency between duplicates, suggesting a good dynamic range and a regular heavy aminoacid incorporation rate in the analyzed timeframe (fig. 31B). Interestingly, we found that proteins related to cell adhesion had a high turnover when compared to the rest of the proteins (fig. 32) in iPSCs, an observation that dovetails with the regulatory role of the cell-extrinsic, contact-dependent signals that regulate pluripotency. This dataset was used to validate a novel computational tool, TRIC (Röst et al., 2016), that addresses the need to reliably quantify several SWATH-MS datasets simultaneously, especially when measurements of the same protein must be performed across many samples. This pSILAC dataset represented a valuable opportunity to verify the performance of TRIC compared with more traditional alignment-based methods.

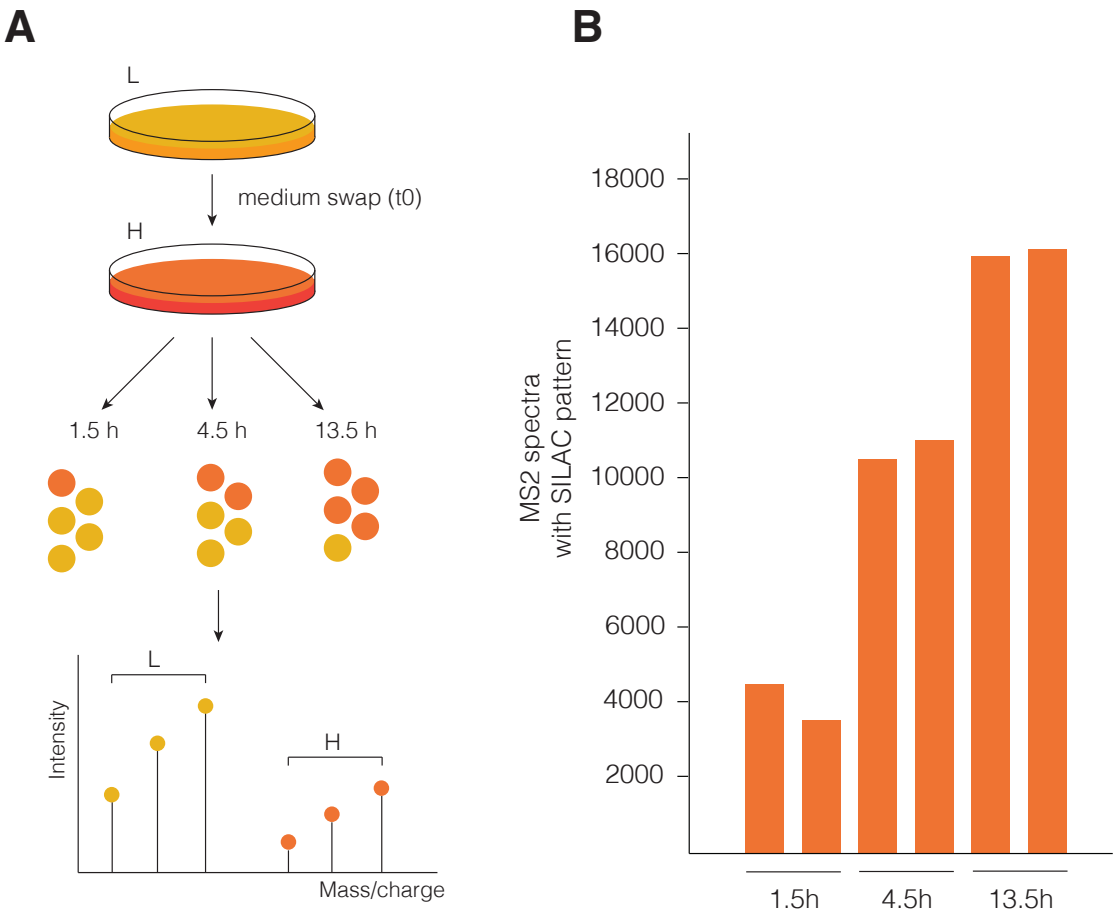

Figure 31: Experimental design and performance of the pSILAC pilot experiment. A) scheme describing the experimental setup: cells growing in light (L) medium receive a complete medium exchange with heavy (H) medium at time 0. Proteins are harvested at 1.5, 4.5 and 13.5 hours, with each time point containing an increasing number of heavy-labeled proteins (orange circles). Mass spectra with pSILAC pairs show the same patterns, with a positive shift in mass caused by heavy aminoacids. B) MS2 spectra with a SILAC pattern measured by Shotgun-MS, displaying the expected pattern of increase in aminoacid incorporation.

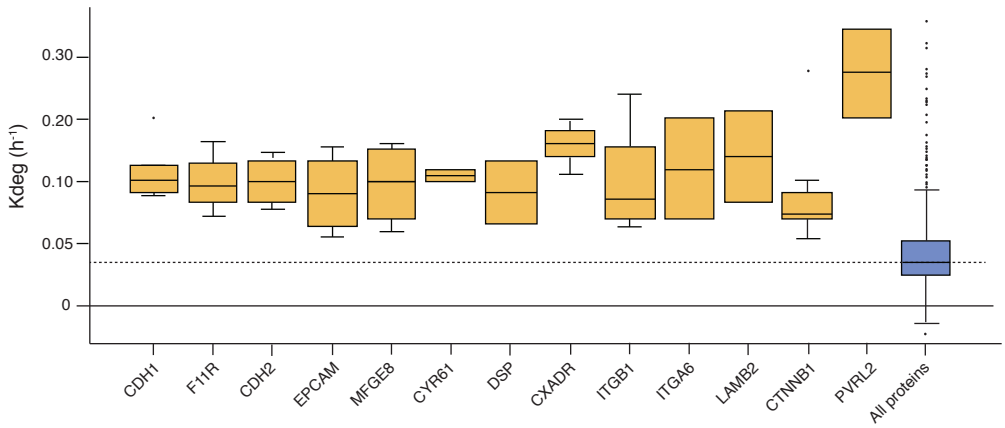

Figure 32: Degradation rates of adhesion molecules are higher than the proteome-wide average. Redrawn from (Röst et al., 2016).

We then chose a subset of iPSC lines (3 WBS lines, 3 controls and 2 7dup lines) to profile extensively the effect of protein degradation caused by the 7q11.23 CNV.

In order to avoid confounding effects imputable to non-linearity of cell duplication reaching confluence, we decided to sample a narrower time frame (2 to 8 hours), which is below half of the duplication time for iPSCs (~17 hours) and should render the effect of cell duplication negligible. Moreover, we decided to increase the precision of our measurements by preparing each sample in triplicate, resulting in 96 pSILAC proteomes. We reliably identified 3434 proteins with a pSILAC pattern in at least 3 samples, out of which 2535 proteins have enough data points to fit the R.I.A. exponential model and determine K_{deg} in at least 3 samples.

Globally, degradation rates show a small but significant ($p = 0.023$ for an ANOVA on median K_{deg}) trend that is inversely correlated with CNV dosage (fig. 33A). If we limit our observations to proteins for which we can construct an exponential decay model with a good fit, *i.e.* with a minimum of 6 data points and a relative error of the model below 0.1, we obtain a subset of 698 proteins. This subset, however, leads to a similar conclusion on the global distribution (fig. 33B), with significant differences in median K_{deg} (ANOVA $p = 0.019$).

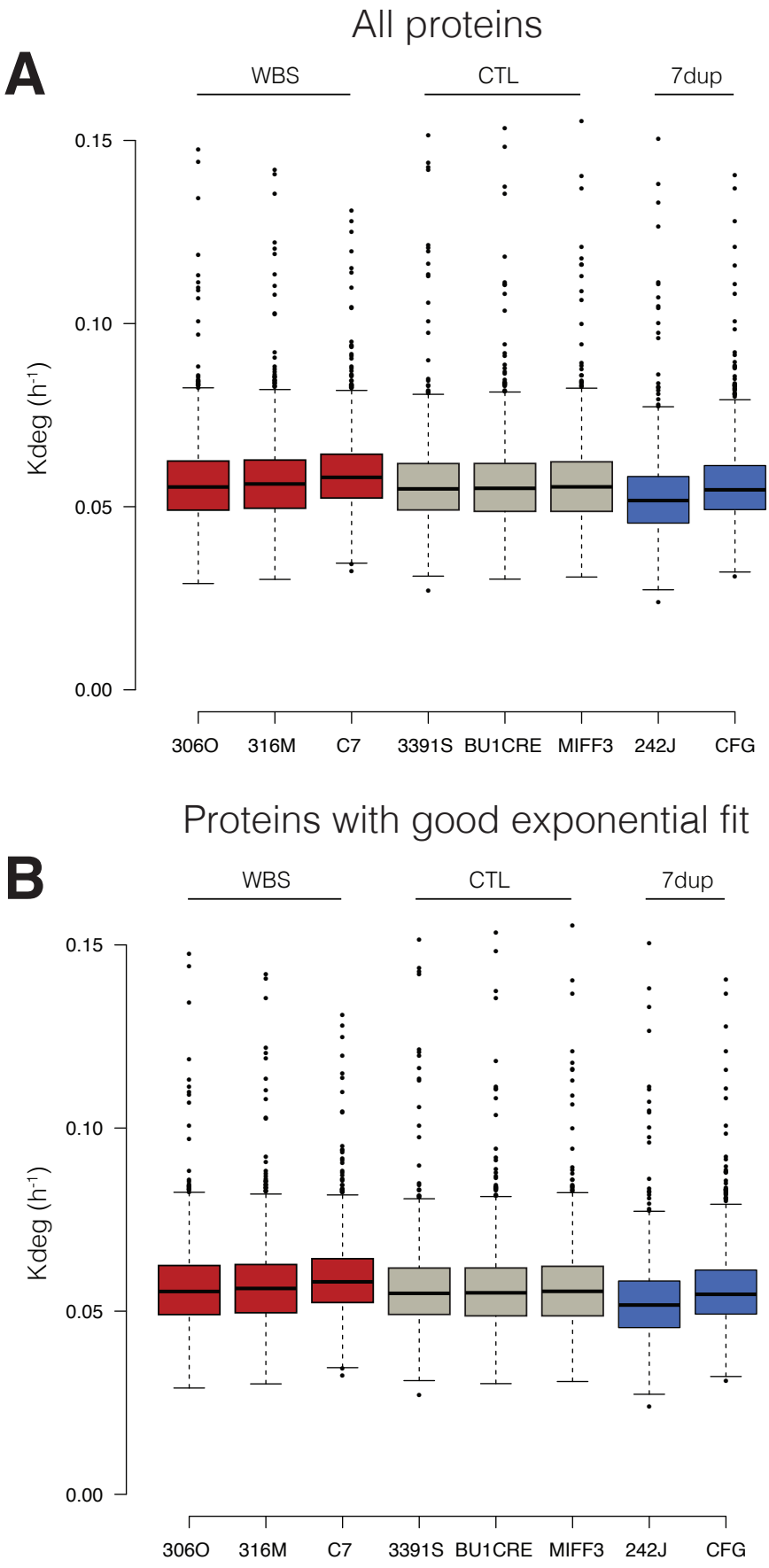

Figure 33: Boxplot of distributions of K_{deg} values per sample. A) Distribution of K_{deg} for all proteins. B) Distribution of K_{deg} for proteins with a good exponential fit. In both cases, the trend is inversely correlated with CNV dosage.

When subsetting our K_{deg} dataset with all the DEGs, we reproduce again the inverse trend in degradation (fig. 33).

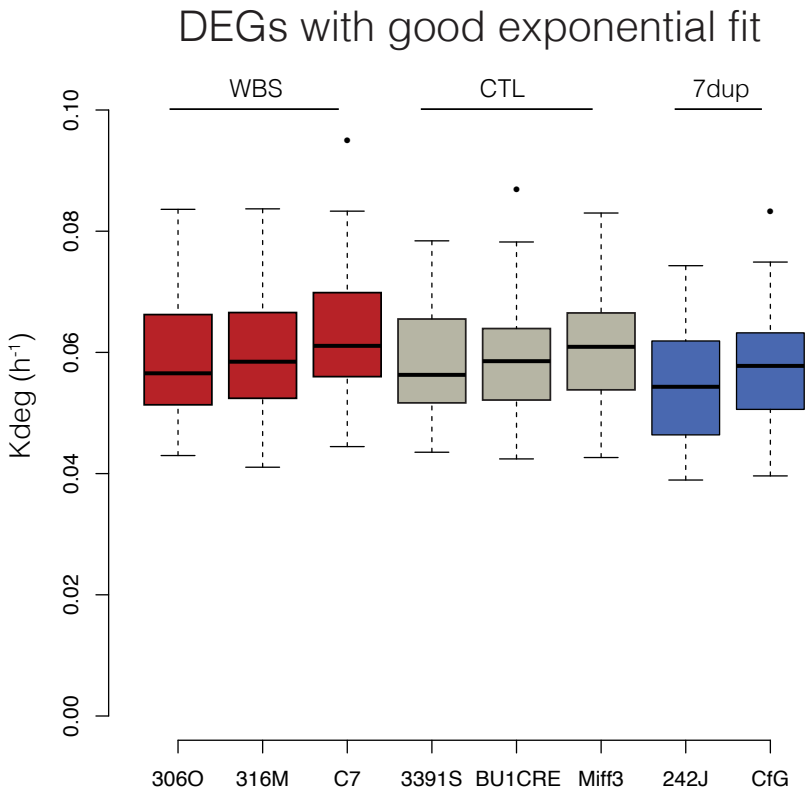

Figure 34: Distribution of K_{deg} values for DEGs with a good exponential fit.

Unfortunately, no protein located in the second quadrant nor in the fourth quadrant of the proteome comparisons (fig. 28C, 28D) was reliably identified in this dataset, leaving open the question whether changes in degradation can explain inverse trends in translation and protein abundance. Moreover, degradation rates for proteins that are differentially expressed and follow gene dosage still maintain the inverse trend (fig. 35A, 35B), which suggests that the final protein abundance is not buffered by degradation. Indeed, the fact that the expression trend persists independently of the global effect on degradation implies that it is slightly affected by it in a manner that escapes straightforward explanations. The variability in Kdeg estimation between samples is again evident, albeit smaller than the variability in steady state protein

quantification, still posing the question on what portion of the missing link between one dataset and the other can be ascribed to technical issues.

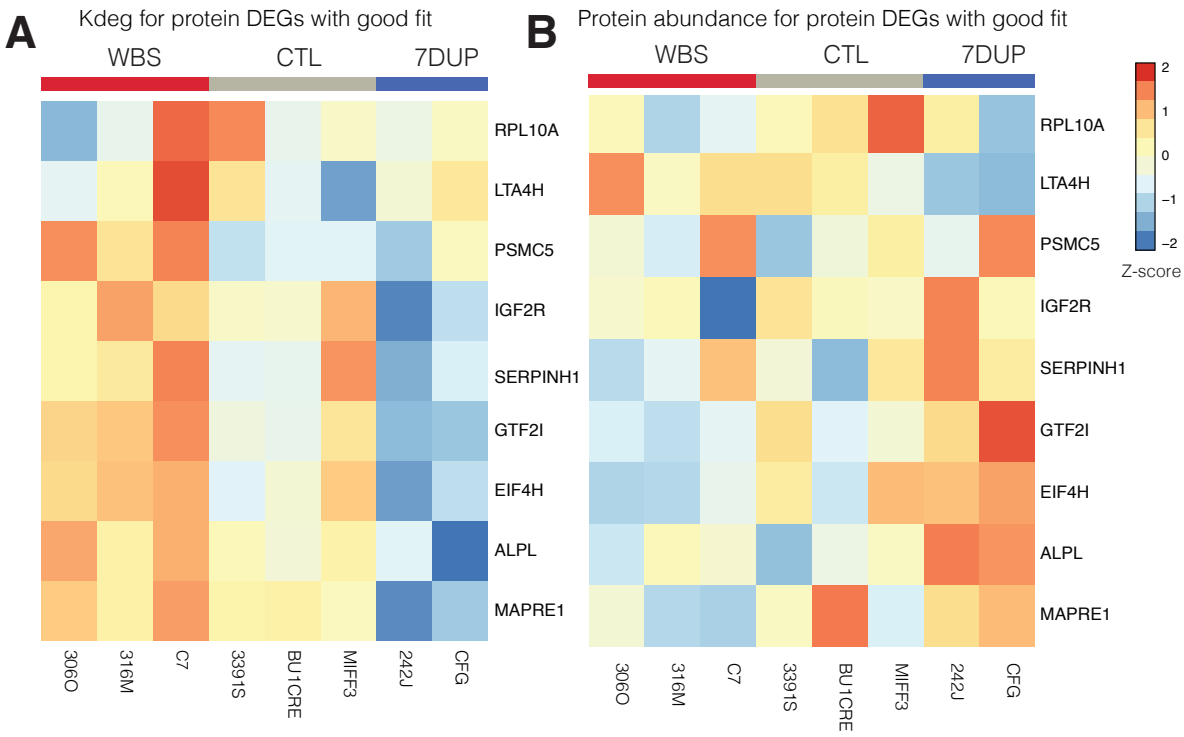

Figure 35: Trends in protein degradation and protein abundance. A) Heatmap of Z-scores of Kdeg for protein DEGs. Only K_{deg} models with a good exponential fit were selected. B) Heatmap of Z-scores of protein intensity for the same subset of proteins.

Although K_{deg} values do not seem to explain changes in protein abundance of differentially expressed genes, we probed their explanatory value when including them in a simple gene expression model derived from the aforementioned mass action law. In fact, we can integrate data from the protein dataset, the RPF dataset and the Kdeg dataset in a simple equation derived from our steady-state assumption:

$$P_{SS} = \frac{K_{syn}}{K_{deg}}$$

Where K_{syn} can be represented by RPF abundance values, which serve as a proxy for the rate of protein synthesis, while Pss is derived from the label-free quantification used for differential expression and slope computation:

$$LFQ\ intensity = \frac{RPF}{K_{deg}}$$

The ratio of RPF and Kdeg is, in other words, the "estimated" amount of protein that takes into account experimental values of translation and degradation.

In fact, RPF do provide a better correlation with protein abundance than RNA (fig. 36A, 36B), consistently with what is reported in literature (Ingolia et al., 2011).

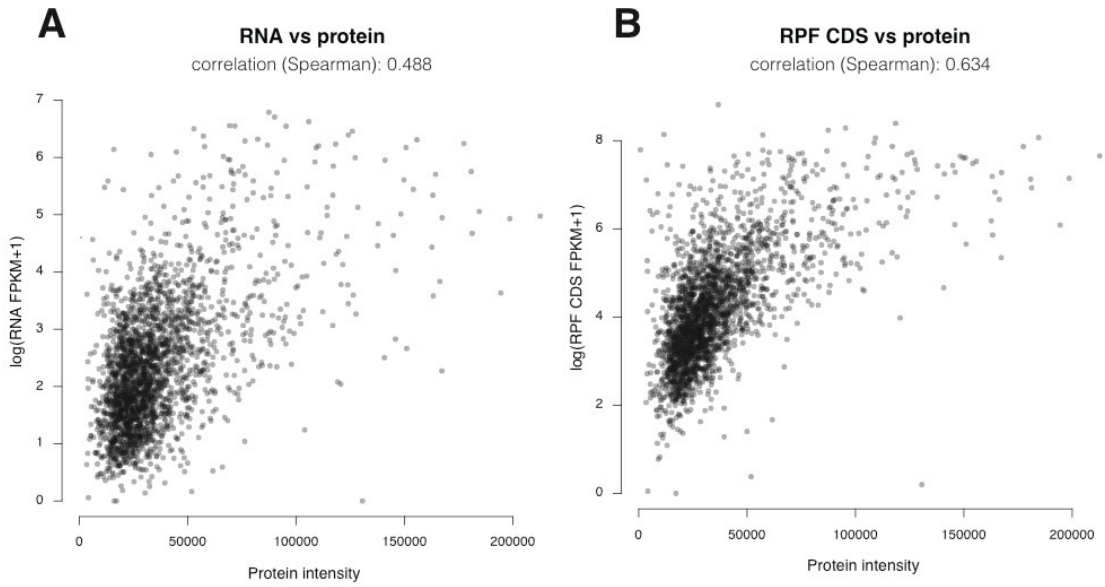

Figure 36: Global correlation between proteome and RNA or RPF for a representative sample. A) Scatterplot of log(RNA FPKM) and protein intensity. B) Scatterplot of log(RPF FPKM) and protein intensity.

A good model of gene expression should be able to minimize the distance between estimated and measured protein abundances. We calculated correlation, mean of squared residuals and mean relative error for a linear model constructed, in each sample, within the log2 of estimated protein abundance and the experimental protein abundance as measured by SWATH-MS. Each model was constructed both with and without K_{deg}. Contrary to our expectations, by including experimentally-derived Kdeg in the model, we do not observe an increase in correlation, but rather a remarkable drop when using the whole proteome (fig. 37). Correlation tends to increase upon restricting the model by increasing the minimum number of peptides used to quantify

each protein, and reaches a plateau when using proteins quantified by at least 24 peptides.

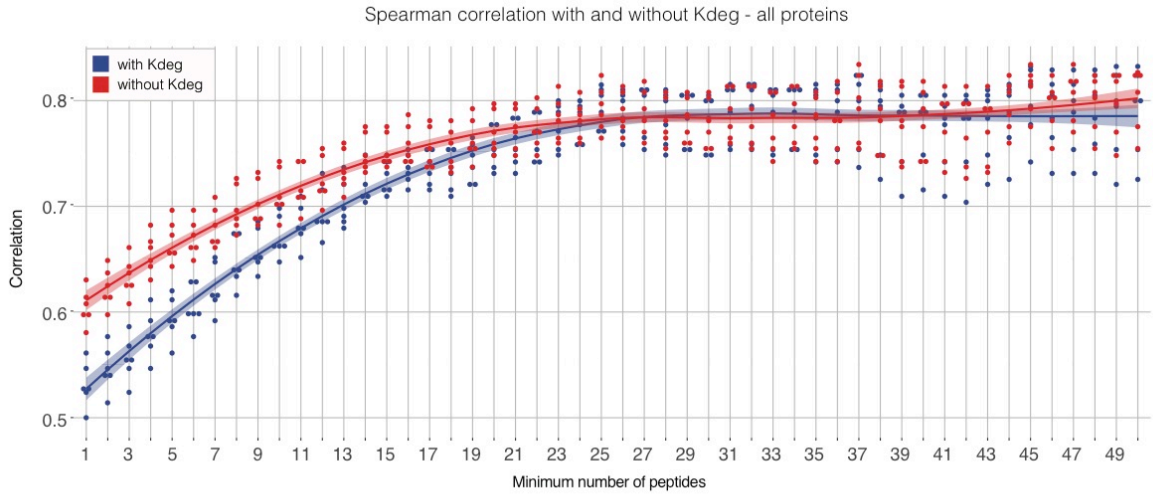

Figure 37: Spearman correlation plot of each model using all proteins. Each point represents a sample. Blue dots: model with K_{deg}. Red dots: model without K_{deg}. Solid lines are smoothing splines calculated by local regression (LOESS). Transparent areas represent the 95% confidence interval of the spline.

A similar difference can be observed for the mean relative error (fig. 38A) and the median of squared residuals (fig. 38B) in all proteins: when including K_{deg} in the model, its precision decreases, although when restricting the model with more than 20-21 peptides the trend is reversed.

To exclude that this was due to the presence of K_{deg} values obtained by suboptimal fitting, we restricted our model to the aforementioned set of proteins with a good exponential fit (fig. 39). Interestingly, both correlation values for the whole well-fitted proteome (*i.e.* minimum 1 peptide) increase compared to the model built on all proteins, but the inclusion of Kdeg does not seem to improve it, rather slightly decrease it. Mean relative error (fig. 40A) is substantially identical, but starts decreasing when filtering from a minimum of 13 peptides. The only parameter for which K_{deg} seems to improve the precision of the model is the median of squared residuals (fig. 39B), for which a consistent trend is maintained along all the thresholds.

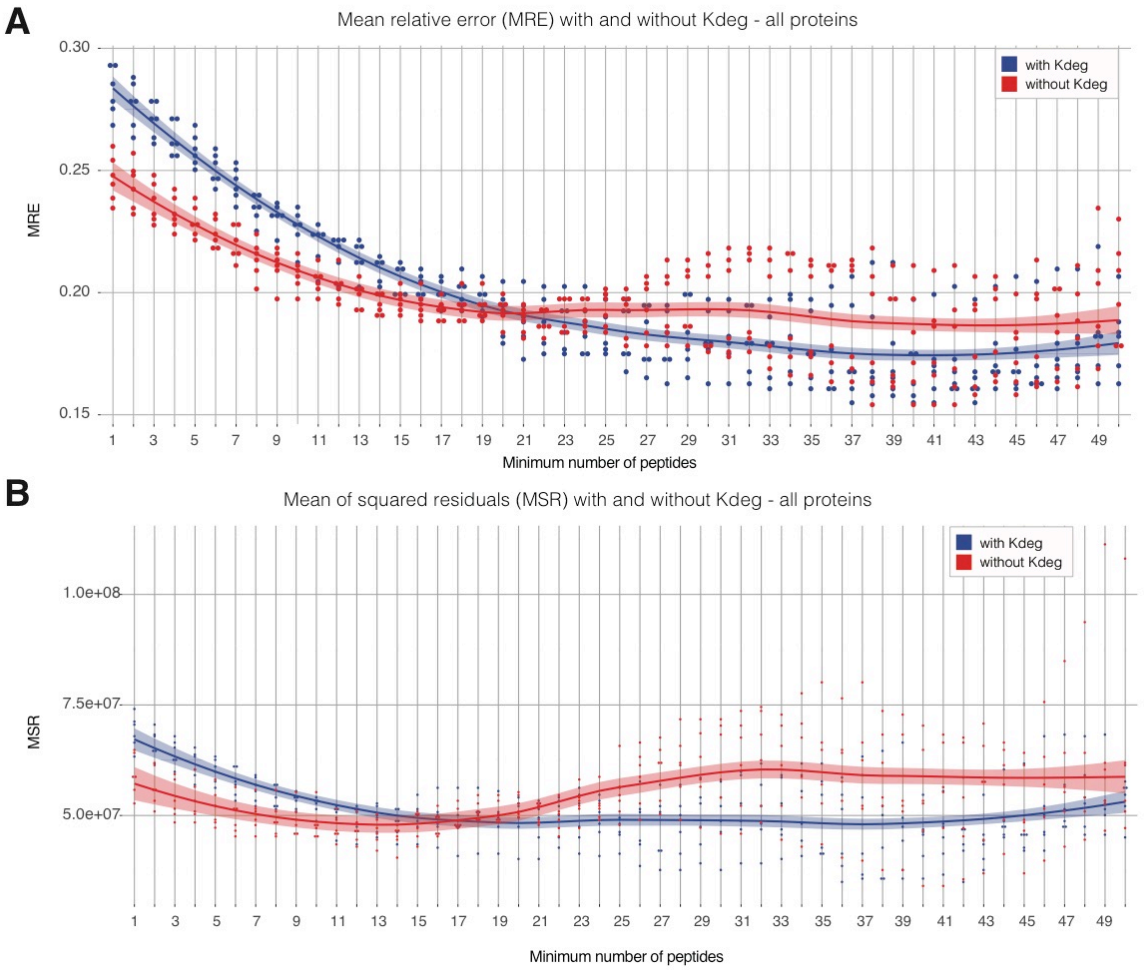

Figure 38: Error plots of each model with and without K_{deg}. A) Mean relative error plot. B) Median of squared residuals plot. Blue dots: model with K_{deg}. Red dots: model without K_{deg}. Solid lines are smoothing splines calculated by local regression (LOESS). Transparent areas represent the 95% confidence interval of the spline.

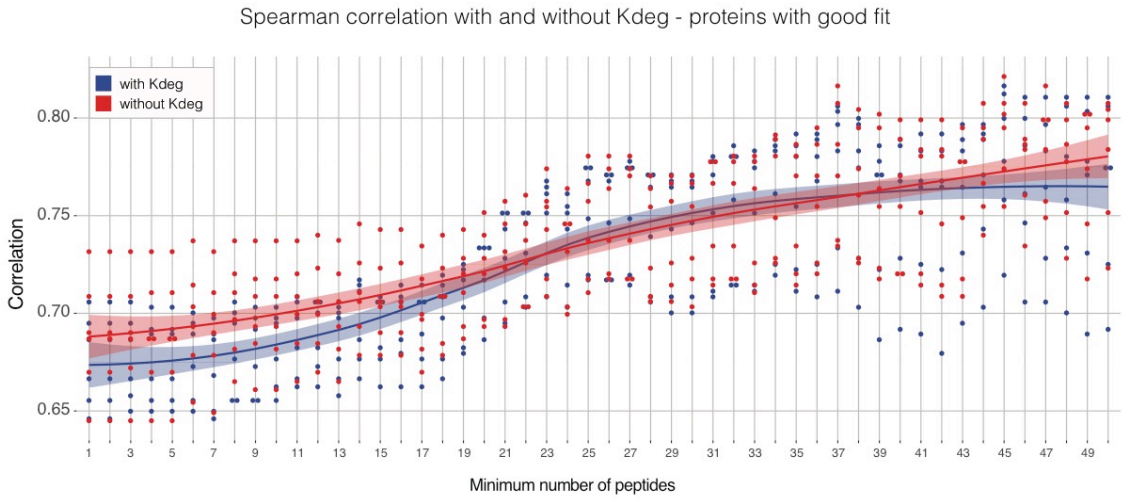

Figure 39: Spearman correlation plot of each model using only proteins with a good exponential fit. Each point represents a sample. Blue dots: model with K_{deg}. Red dots:

model without K_{deg}. Solid lines are smoothing splines calculated by local regression (LOESS). Transparent areas represent the 95% confidence interval of the spline.

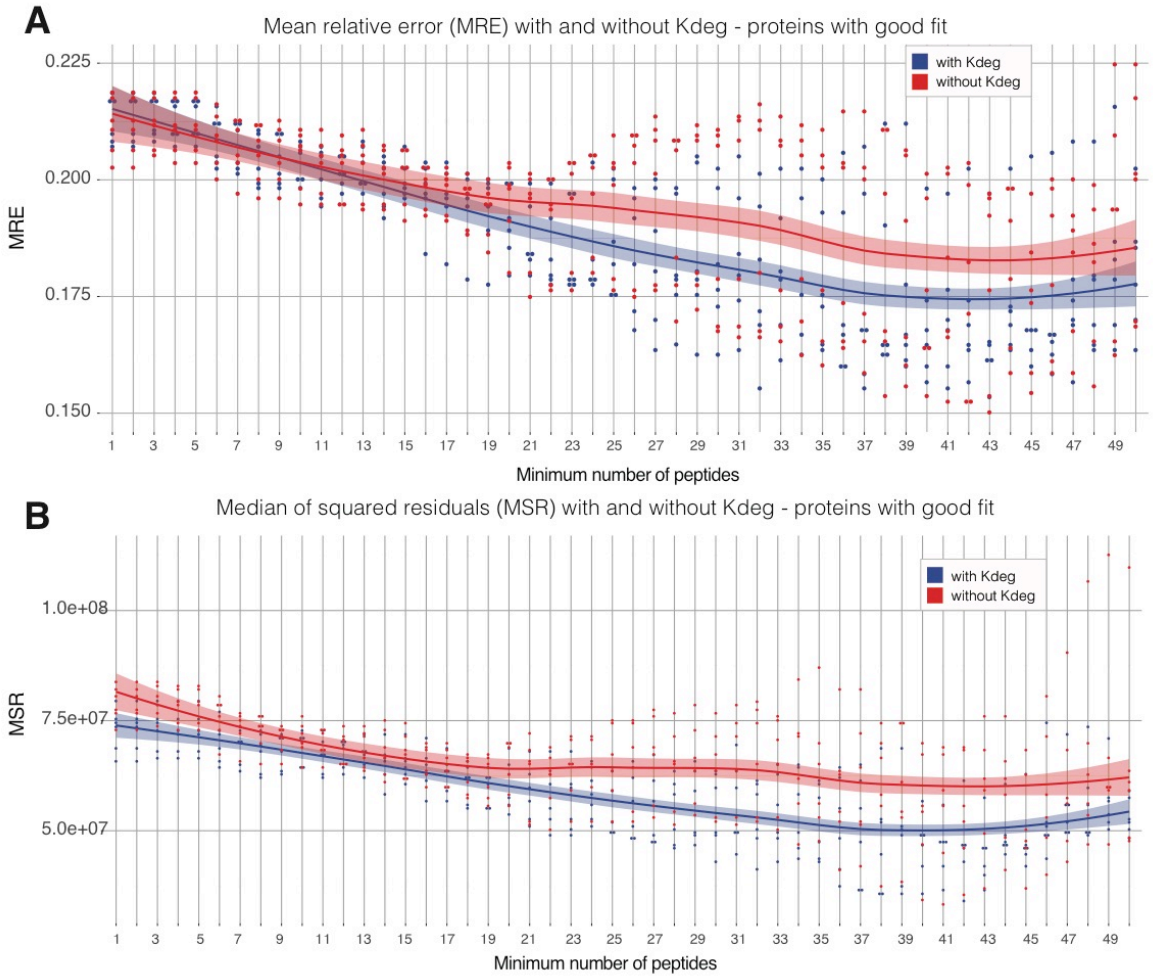

Figure 40: Error plots of each model with and without Kdeg using only proteins with a good exponential fit. A) Mean relative error plot. B) Median of squared residuals plot. Blue dots: model with K_{deg}. Red dots: model without K_{deg}. Solid lines are smoothing splines calculated by local regression (LOESS). Transparent areas represent the 95% confidence interval of the spline.

These results indicate that, in this system, the role of protein degradation as assessed by pSILAC does not have an impact on the relationship between measured amounts of proteins and their estimation by ribosome profiling, hinting at other possible explanations for the ~0.4 of missing correlation.

11. Generation of monoclonal lines for neuronal differentiation

Given the intriguing neuro-cognitive phenotypes of WBS and 7dup patients, and the roles of translation regulation in neuronal development, we established a scalable and reproducible platform for neuronal differentiation of our iPSC lines. Starting from the NGN2 neuronal differentiation protocol (Zhang et al., 2013), that makes use of two lentiviral vectors to over-express an inducible transcription factor that drives cortical neuronal differentiation, we subcloned 37 iPSC lines from 9 patients (3 WBS, 1 atypical WBS, 3 controls, 2 7dup) by single-cell cloning. The original NGN2 protocol uses an activator lentivirus, containing the reverse tetracycline transactivator (rtTA) constitutively expressed under the control of the UbC promoter, and an effector lentivirus, containing an NGN2-P2A-EGFP-T2A-Puro cDNA under the control of the tetracycline responsive element. In this protocol cells are infected with both lentiviruses, induced one day after infection and selected with puromycin the day after. The combination of both NGN2 and rtTA allows to effectively select only cells in which both viruses are properly integrated and expressed by antibiotic selection. Upon 21 days of induction of NGN2, infected iPSCs become cortical glutamatergic neurons. However, this approach has the important limitation that a viral infection should be carried out for each experiment, possibly introducing biases due to differences in batches of viral particles, and requiring cumbersome amounts of viral particles for large-scale experiments.

To circumvent this limitations, we infected 9 lines with both lentiviruses but did not induce them, letting them grow until they were ~$5x10^6$ (fig. 41). Upon reaching this number, infected iPSCs were sorted as single cells in 96-well plates, selected based on the round morphology of colonies and gradually expanded. Selected lines were then induced for one day adding doxycyline to the medium: if cells express GFP (fig. 42A), they have received both lentiviral constructs and very likely originate from the same

clone. 3 to 5 GFP-positive lines were selected and expanded, further being stabilized and characterized. These lines have the advantage of being already infected with the NGN2 dual system in a homogeneous way, so that they can be expanded to a virtually unlimited extent in order to perform a wide array of high-throughput experiments.

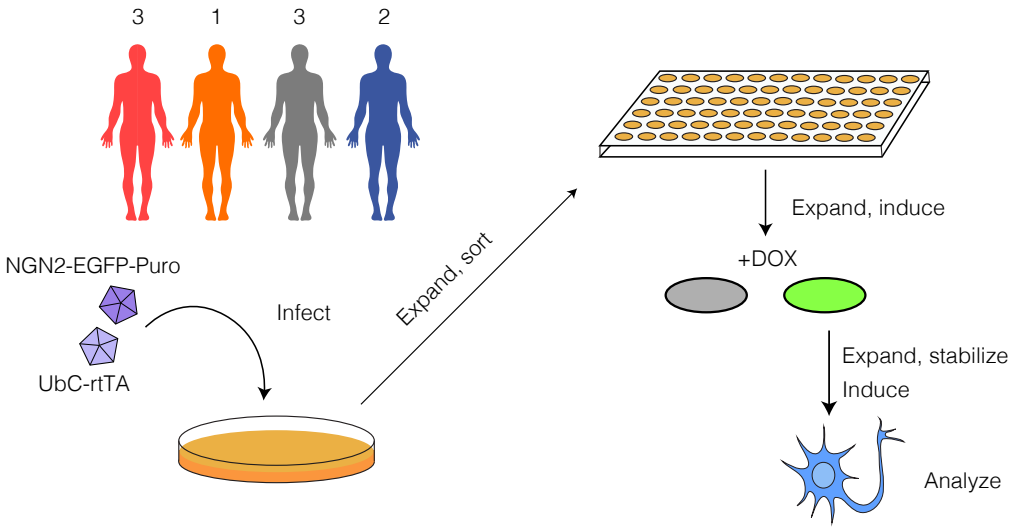

Figure 41: Derivation of monoclonal NGN2 lines. Schematic representation of the procedure. iPSC lines deriving from 9 patients are infected with both lentiviral particles, expanded and sorted as single cells. Single cell-deriving lines are expanded and induced; only GFP-positive lines are kept for further analysis. Red: WBS, orange: atypical WBS, gray: control, blue: 7dup.

As a confirmation that selected monoclonal lines are able to produce cortical, glumatergic neurons, they were induced for 21 days and stained for a panel of neuronal markers that include general neuronal antigens (TUJ1, MAP2), glutamatergic markers (VGLUT), and cortical markers (SATB2, TBR1).

This system has already been implemented in the lab for a series of different projects, all entailing high-throughput experimental setups such as RNA-seq, ChIP-seq, drug screenings. An important caveat of this protocol is that, in order to allow synapse maturation, NGN2-induced neurons must be co-cultured with replication-deficient mouse astrocytes. However, as it was the case for iPSCs, murine cells add an important confounding variable in high-throughput experiments. A concerted effort

in the lab is being carried out to replace astrocyte co-culture by using an astrocyte-conditioned medium, that contains the astrocyte secretome, necessary for neuronal trophic functions, but does not include murine cells. These monoclonal lines, and the technical advancements that will be built upon them, constitute a fundamental platform for all our future efforts to study gene expression in depth in a cell type that is highly relevant to understand the molecular mechanisms of WBS and 7dup.

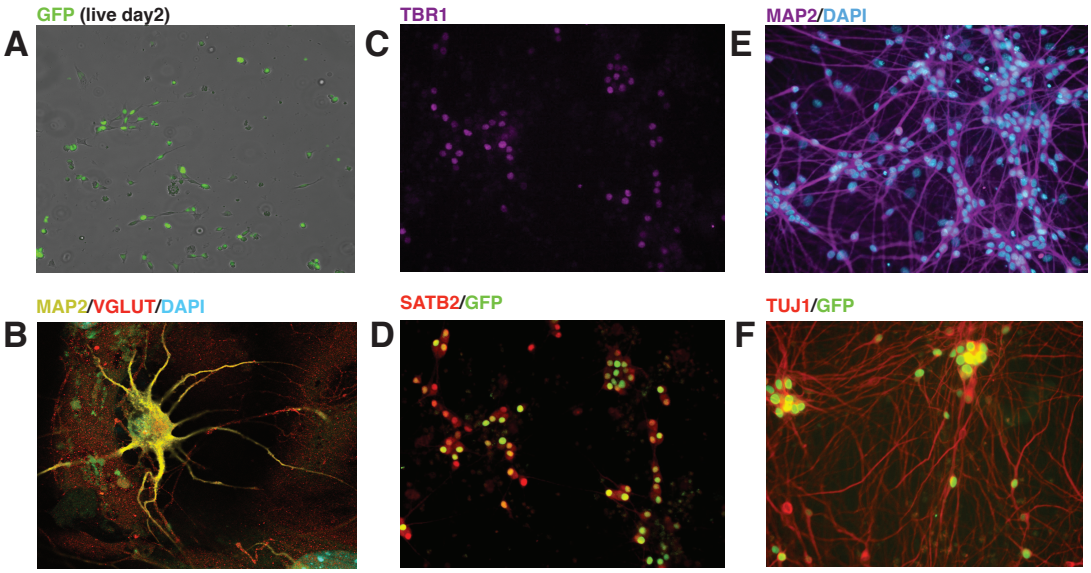

Figure 42: Characterization of NGN2 monoclonal cell lines. A) Induced monoclonal iPSCs express GFP. Magnification 20X in a live brightfield videomicroscope. B) expression pattern of the mature neuronal marker MAP2 and glutamatergic marker VGLUT. Magnification 40X. C) expression pattern of the cortical markers TBR1 and D) SATB2. Magnification 20X. E) Expression pattern of mature neuronal markers MAP2 and F) TUJ1. Magnification 20X. All pictures of stainings have been acquired at the confocal microscope.

Discussion and future directions

1. The analysis of different layers of gene expression reveals differences in gene regulation at the pluripotent state

To our knowledge, this is the first attempt at measuring, in the context of a complex set of developmental disorders at the pluripotent state, three fundamental steps in gene expression regulation: transcription, translation and protein degradation.

By profiling the transcriptome, translatome and proteome of patient-derived iPSCs, we managed to greatly expand our view on how differences in gene expression originating at the RNA level are propagated towards the protein level in pluripotency.

A first interesting finding pertains to the exclusivity of some DEGs to a specific layer. All DEGs found in the total RNA and in the RPF dataset were consistently mapped in both libraries, meaning their exclusivity is not due to technical artifacts. The proteome dataset, however, has a much smaller coverage of the repertoire of proteins that correspond to translated mRNAs, making this comparison more difficult.

The regulation at different levels of gene expression is being widely studied at a genome-wide scale since the diffusion of precise and reproducible high-throughput technologies for transcriptome, translatome and proteome quantification (Ingolia et al., 2011; Vogel and Marcotte, 2012). Different groups, in the laste decade, have attempted to give a definitive answer to the dominance of one mode of regulation over the other (Jovanovic et al., 2015; Li et al., 2014; Schwanhausser et al., 2011), by proposing different comparisons of cellular abundances of transcripts and proteins and using different statistical models and error-calibration strategies.

Although a consensus has not yet been reached, transcription is considered to shape the proteome at steady state (reviewed in Liu et al., 2016). However, changes in translation rate can be quite rapid and serve as an adaptive response to changes in

the environment, such as nutrient starvation or external stimulation (Jovanovic et al., 2015). Importantly, transcription factors are mostly under translation control upon stimulation (Jovanovic et al., 2015). While it is true that our iPSC lines grow at steady state, they represent a very early stage of development that is bound to change dramatically in a relatively short time frame, by giving rise to all three germ layers upon progressive fate acquisition. The fact that iPSC represent a cellular snapshot of a highly dynamic process that is artificially suppressed by culture conditions puts in perspective the notion of "steady state".

In this thesis I have tried to show that, already by performing differential expression analysis in iPSCs, we find that some genes are targeted mostly, if not exclusively, by different regulatory mechanisms at specific layers. These genes are particularly interesting in that they have been linked to molecular pathways involved in the clinical manifestations of WBS or 7dup.

For instance, among RNA-exclusive genes we found that *SOX3* is differentially expressed, being more highly expressed in WBS samples and lower in 7dup.

SOX3, a member of the SOX (SRY-related Homeobox) family of genes situated on the X chromosome, encodes a SRY-box transcription factor that has been linked to hormonal deficiencies (Woods et al., 2005), growth defects and intellectual disability (Jourdy et al., 2016).

SOX3 is one of the earliest markes of neuronal differentiation, and studies in animal models have shown that it acts in a tightly regulated temporal fashion. More specifically, in pluripotent cells SOX2, another member of the SOX family, binds proneural genes marked by a bivalent chromatin signature, and pluripotency genes marked by an activatory signature. During differentiation, SOX3 recognizes bivalent proneural genes bound by SOX2, and converts the signature to a monovalent one, thus activating genes that drive the transition from pluripotency to a neuronal

precursor However, SOX2 is simultaneously binding another set of bivalently marked genes, which in turn promote neuronal maturation, and keeps them repressed. Further into differentiation, SOX3 is bound by SOX11, which in turn converts bivalent, SOX3-bound marks to monovalent, activatory ones, thus allowing differentiation to continue (Bergsland et al., 2011; Bylund et al., 2003). Interestingly, SOX3 is antagonized by NGN2 (Bylund et al., 2003), the same transcription factor whose ectopic expression drives neuronal differentiation in our system. Studies on mice models have demonstrated that SOX3 is necessary to regulate the formation, during development, of the hypothalamo-pituitary axis and the dorsal telecephalon (Rizzoti et al., 2004). Tandem duplications of the locus encompassing *SOX3* have been associated, in humans, to combined hormone pituitary deficiencies (CPHD), craniofacial dysmorphisms, growth defects and mild intellectual disability (Laumonnier et al., 2002), which are clinical manifestations that partially overlap with those of WBS patients.

Another example of RNA-exclusive differences across samples affects *RPS16*, which encodes a protein of the 40S ribosomal subunit, and *MRPL12*, which encodes a protein of the 39S mitochondrial ribosome subunit. The translation efficiency of both *RPS16* and *MRPL12* decreases upon EIF4H knockdown, again suggesting a role for EIF4H in the regulation of their translation. Interestingly, *MRPL12* mutations have been associated to growth retardation (Serre et al., 2013) and nonsyndromic, autosomal deafness (Li et al., 2002), which partially overlaps with another clinical feature of WBS patients, sensorineural hearing loss.

Among DEGs that were found exclusively in the RPF dataset we found Cytoplasmic Binding Element – 2 (*CPEB2*), the human homolog of *Drosophila* Orb2, an RNA-binding protein whose expression is important for polarity formation in early embryogenesis and nervous system development(Hafer et al., 2011). Mammalian

CPEB2 has been shown to act as a repressor of translation of the hypoxia-inducible factor HIF-1a during normoxya conditions (Chen and Huang, 2012) and is highly expressed in the CNS, where it binds the mRNAs of beta-catenin and CAMKIIa (Turimella et al., 2015).

In our samples, *CPEB2* is lowly expressed at the RNA level, and it is interestingly almost exclusively translated in 7dup iPSCs. Moreover, its translation efficiency remarkably and consistently decreases upon EIF4H knockdown, possibly hinting at a role for CPEB2 in keeping the translation of selected targets at a constant level.

Another gene, ankyrin-repeat domain containing 1 (*ANKRD1*), also known as Cardiac ankyrin repeat protein (*CARP*) has been involved in the aetiology of dilated cardiomyopathy (Moulik et al., 2009), a cardiovascular disease with a strong genetic component in which the left ventricle becomes dilated and causes heart failure. It has also been observed that ANKRD1 is up-regulated in left ventricles of patients with cardiac failure (Zolk et al., 2002) compared to healthy heart biopsies. ANKRD1 binds sarcomeric proteins such as Titin and Myopalladin in human myocytes and it has a dual role: in the Z-disc of sarcomeres, it functions as a stretch sensor , whereas upon interaction with phosphorylated ERK1/2 and GATA4, it translocates to the nucleus, where it acts as a transcriptional repressor (Zhong et al., 2015). As described in the introduction, cardiac failure due to heart malformations is a shared burden of both WBS and 7dup patients and one of the most frequent causes of death for WBS patients (Pober, 2010).

Among DEPs, RPL10 is of particular interest. RPL10 encodes the ribosomal protein L10, which is a highly conserved component of the large ribosomal subunit. RPL10 is necessary for the joining of 60S at initiation (Eisinger et al., 1997).

Five distinct mutations within RPL10 have been described in families with X-linked intellectual disability (XLID) so far: three in the N-terminal (p.K78E, p.G161S and

p.A64V) and two in the C-terminal region (p.L206M and p.H213Q). Both p.K78E and p.G161S mutations are suggested as a primary cause of X-linked syndromic disorder hallmarked by microcephaly, growth retardation, seizures and minor facial anomalies (Brooks et al., 2014). The C-terminal mutations were identified in families with autistic features and moderate to severe ID, or normal cognitive development (Klauck et al., 2006). RPL10 follows a trend that is inversely correlated with WBSCR dosage.

Another protein related to translation, the eukaryotic initiation factor 2D (EIF2D) is expressed with a trend that is inversely correlated between RPF and protein,

EIF2D can perform the GTP-independent delivery of the initiator tRNAMet (Dmitriev et al., 2010) a key step of translation initiation. Its opposing levels with EIF4H dosage raise the interesting hypothesis that changes in a translation initiation factor may be counteracted by opposing changes in other initiation factors, so as to balance the initiation kinetics.

All these findings will have to be adequately validated by independent orthogonal techniques which can further expand the understanding of these modes of regulation. While differences at the RNA level can be easily validated by RT-qPCR or by Nanostring, and differences at the protein level can be validated by targeted mass spectrometry approaches such as Selected Reaction Monitoring (SRM) or antibody-based techniques, the validation of differences in translation activity requires additional steps. We have already started measuring polysome profiles obtained from the same cells that were used in this study. With polysome profiling we can infer global changes in polysome or monosome density and, more importantly, quantify the RNA present in each fraction of the gradient. Changes in the abundance of specific mRNAs across fractions and across samples will not only indicate whether the mRNA is differentially translated, but also what phase of translation is being preferentially regulated for that mRNA.

Although we can find differentially expressed genes in every layer, their expression profiles are admittedly variable among samples. This is especially true when looking at differences in gene expression at the total RNA level (see Results). Every differential gene expression analysis has then to be evaluated carefully and even more carefully validated in order to make well-grounded claims about the extent and type of dysregulation.

Besides technical issues that can be accounted for with the inclusion of additional technical replicates, the intrinsic variability of different human genetic and epigenetic backgrounds creates confounding effects, such as masking differential expression by means of inflating variance, which dramatically reduces the confidence with which genes can be identified as differentially expressed. There are two ways to circumvent this problem. The first one is to increase the statistical power of the analysis by processing more samples and including biological replicates. Two human studies aimed at integrating transcriptome, translatome and proteome (Battle et al., 2015; Cenik et al., 2015) make use of a large cohort of well characterized lymbphoblastoid cell lines (LCL) to study the regulatory consequences of human variation, with one study making use of at least 2 replicates per sample (Cenik et al., 2015). Although their biological question is remarkably different from ours, and requires more power to observe smaller differences, they are able to robustly identify differences in regulatory variation in all three layers. That being said, since our work involves the use of patient-derived iPSCs, the recruitment of new WBS or 7dup patients and the reprogramming and establishment of their iPSC lines requires a remarkable effort, whereas LCLs are readily available. As for biological replicates, a possibility would be to include different clones of the same patient; however, it is still unclear whether the analysis of different clones would lead to the discovery of differentially expressed

genes that arise due to different genomic backgrounds rather than different conditions.

Alternatively, general confounding effects in the data could be addressed computationally, making use of differential gene analysis pipelines tailored to the analysis of heterogeneous samples. A very recently published statistical tool, ELTseq (Xu and Chen, 2016), has been proposed for the analysis of typically diverse specimens such as cancer primary samples. As opposed to more traditional, negative binomial distribution-based tools such as edgeR (used in this study) and DEseq, which work under the assumption that each group of samples pertains to a series of biological replicates, ELTseq treats each sample individually without making assumptions on the distribution of read counts. The types of analysis presented in this work therefore take at face value the differential expression analysis that was performed, with the assumption that human variability is indeed a major confounding effect, and we will try in the near future to assess the extent to which the variability we observe is due to technical artifacts on top of human diversity.

Another important improvement in the analysis of ribosome profiling experiments is being carried forward by the recent emergence of tools that perform differential translation analysis, starting from count data, such as Xtail (Xiao et al., 2016), Anota (Larsson et al., 2011) or RiboDiff (Zhong et al., 2016). In fact, while count-based methods for differential expression analysis use discrete statistics, metrics such as TE or log2(FC) represent a fraction of two integers and should be dealth with using continuous probability distributions. This, in turn, makes methods based on negative binomial distributions not suitable. In the near future, we will benchmark the performance of these tools on our datasets and integrate the results with those discussed in this thesis.

2. A regression-based approach allows to readily visualize and classify genes according to the way their differences are propagated

By computing regression slopes on the three layers, we can produce a series of exploratory data visualizations that give us a "bird's eye" view on the relative propagation of gene expression in pluripotency. An alternative way of looking at how differences are propagated would be to plot, for each comparison, the fold change in one layer against the fold change in another. However, besides making the number of representations larger and more difficult to compare among each other, it would not reveal the sensitivity of gene expression to the dosage of specific proteins. The regression-based representations was preferred as it offers a synthetic and quickly understandable view of gene expression regulation.

The use of EIF4H as a query protein was dictated, in our case, by a specific hypothesis based on the disease pathophysiology and its involvement in the development of animal models (Capossela et al., 2012). By including datasets generated from knock-down samples it is possible to expand the dynamic range, thus enhancing the detection of differences that correlate with the query. However, we could have chosen any other well-quantified protein in our dataset to observe other trends, and draw a different series of slopes in the four quadrants.

Moreover, even by just relying on patient-derived samples without perturbations of protein abundance, one can imagine an evolution of this system in which all differentially expressed proteins (or all reliably quantified proteins) are used as queries, and new relationships between gene expression levels can be inferred by looking at how their positions in the quadrants change.

There are, though, some caveats worth discussing. First, the slope obtained by including a knock-down sample may be increased by the exacerbation of an effect that is not physiologically as intense. Moreover, given the high variability of samples,

slopes that are statistically significant may still not pass FDR correction, thus necessarily restricting our claims to the distribution of slopes rather than the features of single genes.

An interesting consequence of the use of slope comparisons is their partition into layer-exclusive regulation, which is a function of the statistical significance of the slopes. As shown in the results, RPF-only slopes show a pattern that greatly reflects the function of EIF4H as an enhancer of translation initiation. When computing GO enrichments for each of these partitions, we find different, partition-specific categories that point to a differential regulation, at each layer, of functional categories. Considering that iPSCs are a static snapshot of an actually dynamic cell type, it is intriguing to see that genes related to cell cycle regulation are preferentially (dys)regulated at the level of translation, as already suggested in literature (Tanenbaum et al., 2015).

The dynamicity of gene expression in iPSCs can be further appreciated by looking at how expression patterns of genes for which we have statistical confidence are propagated from RNA to protein. By clustering together these patterns, which we termed *archetypes*, another functional, data-driven classification can be made. The notable example of the *creek* archetype, for instance, includes a series of genes involved in morphogenesis and neuronal differentiation whose pattern is remarkably inverted from RNA to protein, possibly at multiple regulatory steps in each layer.

While several attempts have been made to reconstruct gene regulatory networks by looking at a single layer (mostly RNA, Basso et al., 2005) and reverse engineering RNA expression patterns by integrating it with ChIP-seq data (Qin et al., 2014), there are few studies, to my knowledge, that attempt at integrating gene expression patterns in transcriptome, translatome and proteome. The work on human variation performed by Cenik and colleagues (Cenik et al., 2015) makes use of two machine-

learning algorithms, Kohonen maps (also called self-organizing maps) and affinity propagation clustering, which are able to weigh the amount of information transferred from one layer to the others and identify clusters of expression patterns. As already stated while discussing technical limitations of the setup for differential expression analysis, an implementation of these more sophisticated techniques on our datasets may yield an additional insight into propagation of dysregulation and possibly validate our archetype-based classification.

3. Degradation rates do not improve the correlation between translatome and proteome

We leveraged the unique combination of our datasets to better understand not only to which extent each layer of gene expression is able to buffer or introduce changes in the system, but also gene expression as a process in itself.

In our experimental setup, differences in degradation across samples are much smaller than differences in transcription, translation or protein abundance. Interestingly, degradation constants do not explain changes in protein abundance across samples, and they do not increase the precision of the simple gene expression model in which we can include experimental data for all the terms of the equation. At a first glance, it could be concluded that - in this system and with this experimental setup - degradation is not a determinant of changes in protein abundance at the pluripotent steady state. However, other reasons for this lack of correlation can be proposed. In a very recent publication (McShane et al., 2016), the group of Matthias Selbach showed how a small (~10%) subset of mammalian proteins is degraded in a non-exponential decay (NED) fashion, which they term NEDs. The authors propose that NEDs are synthesized in a super-stoichiometric fashion relative to exponentially degraded (ED) proteins with which they form complexes. While all ED proteins will

be able to engage in complexes with NEDs, the remaining - free - NEDs will be degraded, thus undergoing a faster turnover.

This observation stems from theoretical work done on the mathematical modeling of pulse-chase experiments for the determination of RNA (Sin et al., 2015) and protein (Sin et al., 2016) degradation rates. In these reports, it is postulated that some proteins follow degradation kinetics which cannot be well described by the exponential decay function, but rather by a Markov process in which newly synthesized proteins (state A) can be either immediately degraded (state 0) or pass onto another state (state B), representing the interaction within a complex, a stabilizing post-translational modification, subcellular localization, and so on. Proteins in state B will then be degraded (i.e. go to state 0) with a different decay rate. Three kinetic constants should be then derived: from a A to 0, from A to B and from B to 0. This can be accomplished by using optimization algorithms that minimize the sum of squared residuals between experimental data (such as the RIA of the heavy protein) and the equation for the decay of a pulse with no chase (eq. 18 in Sin et al., 2016).

The work on NEDs becomes particularly relevant for our study since it has been proposed that NEDs are particularly enriched in aneuploidies, as super-stoichiometric amounts of proteins tend to be attenuated. The slight flattening of the slope of WBSCR proteins indeed points to a similar phenomenon, although not all WBSCR proteins are reported to be in a complex. Notably, the authors also report that 50% of the assayed proteins follow neither exponential nor non-exponential degradation, thus pointing to a cluster of “dark degradation matter”. However, as reported in the results section, the lack of correlation in our samples also affects the subset of proteins for which there is a good exponential fit, thus pointing to a problem

that does not necessarily originate from the wrong mathematical model of degradation.

Besides the important technical differences in sample preparation and analyte acquisition between count-based NGS methods and spectra and fragmentation-based proteomics methods, another possible explanation lies in the way both proteins and RPF reads are quantified.

Protein abundances are calculated, both in the LFQ and the pSILAC dataset, by compounding the measurements of peptides assigned to the same protein. Peptide spectra assignment (i.e. identification) and quantification are, therefore, fundamental for the correct measurement of protein intensity. A recent report (Bogdanow et al., 2016) has shown how the existence of unforeseen modified peptides in MS experiments can affect the assignment of spectra. A modified peptide that is not taken in consideration as such in the reference spectral library can be easily misassigned to other proteins, thus creating false positives. The authors estimate that up to 50% of proteins can be misassigned if modified peptides are not accounted for. Although we have used for LFQ a summarization method that only takes into account the most correlated peptides, for the pSILAC dataset we calculated the RIA on each consistently quantified peptide separately, so as to have as many data points as possible. A reasonable option would be to requantify RIA values by excluding potentially modified peptides and check whether there are changes in Kdeg determination. Nevertheless, since we are assessing relative differences in degradation rates rather than absolute degradation rates, this correction would make sense only under the assumption that the impact of modified peptides is CNV-dependent. On the other hand, RPF reads in our analysis were assigned to transcripts which are aligned to collapsed gene models, without taking into account alternative splicing events that greatly enhance the diversity of the proteomic repertoire (see (de Klerk and ‘t Hoen,

2015) or a review). This can in turn be further enhanced by the translation of different parts of the transcripts: upstream or alternative ORFs, alternative translation start sites, read-through translation events can all potentially give rise to different proteoforms (Floor et al., 2016). We have adopted a simplified model that works under the assumption of one-to-one correspondeces of transcribed DNA, mature RNA, translated RNA and protein, meaning that some precision will be inevitably lost to the complexity of each of these steps. Other possible explanations involving mis-quantification of ribosome occupancy profiles may lie in the bias introduced by pausing ribosomes. In fact, when translating a poly-proline tract, ribosomes tend to slow down and stall. This may artificially inflate the number of reads at pausing sites and introduce spurious differences in quantification. Ribosome profiling is a relatively new technology and many analytical tools and error-correction strategies are being increasingly proposed. There is, therefore, much room for improvement on the analytical side in the near future, in order to refine our understanding of gene expression regulation in pluripotency and in the context of 7q11.23 CNVs.

4. Adding a third dimension: disease-relevant cell types

While at the iPSC stage the transcriptional dysregulation might seem to be irrelevant, not being reflected in protein abundance, it is plausible that this equilibrium between transcriptional and translation control is fragilized in a different cellular state.

The ground-breaking potential of iPSCs to give rise to virtually all cell types in the body is what makes them especially relevant for disease modelling. We have demonstrated earlier that pluripotency is already an informative state for the molecular consequences of WBS and 7dup CNVs, and in this study I have shown that

the analysis of layers beside the transcriptome can bring a wealth of additional information with pathophysiologically meaningful implications. However, changes in cell fate are much more sensitive to differences in protein synthesis and turnover (Kristensen et al., 2014; Lu et al., 2009; Werner et al., 2015), and represent critical points in which the CNV may start to contribute to a clinical phenotype. For this reason, we have assembled a panel of NGN2-inducible lines (described in the Results section) and a panel of neural crest stem cell lines derived *in vitro* from the same iPSC lines used in this study (partially already published in Adamo et al., 2014), with the aim of performing the same set of analyses in two tissues affected by the CNVs. In fact, NGN2-induced neurons allow us to probe the type and extent of dysregulation in cortical glutamatergic neurons, which are the best candidates to study the molecular and cellular features that underlie sociality and language processing (Hutsler and Zhang, 2010). Neural crest stem cells are instead the early progenitors of a variety of cell types that give rise to cranial bones and muscles, smooth muscle cells, aortic arches, cartilage and connective tissue, thus offering us an important entry point into the study of craniofacial and cardiovascular manifestations of both syndromes.

Besides measuring the propagation of information through layers in each cell type, we will be able to measure the propagation of information through cell fates, effectively adding a third dimension (the first two being conditions and layers) to the regulation of gene expression in these syndromes.

Bibliography

Abaza, I. (2006). Drosophila UNR is required for translational repression of male-specific lethal 2 mRNA during regulation of X-chromosome dosage compensation. Genes Dev. *20*, 380–389.

Adamo, A., Atashpaz, S., Germain, P.-L., Zanella, M., D'Agostino, G., Albertin, V., Chenoweth, J., Micale, L., Fusco, C., Unger, C., et al. (2014). 7q11.23 dosage-dependent dysregulation in human pluripotent stem cells affects transcriptional programs in disease-relevant lineages. Nat. Genet. *47*, 132–141.

Altmann, H.M., Tester, D.J., Will, M.L., Middha, S., Evans, J.M., Eckloff, B.W., and Ackerman, M.J. (2015). Homozygous/Compound Heterozygous Triadin Mutations Associated With Autosomal-Recessive Long-QT Syndrome and Pediatric Sudden Cardiac Arrest: Elucidation of the Triadin Knockout Syndrome. Circulation *131*, 2051–2060.

Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., and Zoghbi, H.Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat. Genet. *23*, 185–188.

Andreu-Vieyra, C. V., Chen, R., Agno, J.E., Glaser, S., Anastassiadis, K., Stewart, A.F., and Matzuk, M.M. (2010). MLL2 Is Required in Oocytes for Bulk Histone 3 Lysine 4 Trimethylation and Transcriptional Silencing. PLoS Biol. *8*, e1000453.

Antonell, A., Del Campo, M., Magano, L.F., Kaufmann, L., de la Iglesia, J.M., Gallastegui, F., Flores, R., Schweigmann, U., Fauth, C., Kotzot, D., et al. (2010a). Partial 7q11.23 deletions further implicate GTF2I and GTF2IRD1 as the main genes responsible for the Williams-Beuren syndrome neurocognitive profile. J Med Genet *47*, 312–320.

Antonell, A., Vilardell, M., and Perez Jurado, L.A. (2010b). Transcriptome profile in Williams-Beuren syndrome lymphoblast cells reveals gene pathways implicated in glucose intolerance and visuospatial construction deficits. Hum Genet *128*, 27–37.

Arribere, J.A., Cenik, E.S., Jain, N., Hess, G.T., Lee, C.H., Bassik, M.C., and Fire, A.Z. (2016). Translation readthrough mitigation. Nature *534*, 719–723.

Ashe, A., Morgan, D.K., Whitelaw, N.C., Bruxner, T.J., Vickaryous, N.K., Cox, L.L., Butterfield, N.C., Wicking, C., Blewitt, M.E., Wilkins, S.J., et al. (2008). A genome-wide screen for modifiers of transgene variegation identifies genes with critical roles in development. Genome Biol. *9*, R182.

Auer-Grumbach, M., Bode, H., Pieber, T.R., Schabhüttl, M., Fischer, D., Seidl, R., Graf, E., Wieland, T., Schuh, R., Vacariu, G., et al. (2013). Mutations at Ser331 in the HSN type I gene SPTLC1 are associated with a distinct syndromic phenotype. Eur. J. Med. Genet. *56*, 266–269.

Avilion, A.A., Nicolis, S.K., Pevny, L.H., Perez, L., Vivian, N., and Lovell-Badge, R. (2003). Multipotent cell lineages in early mouse development depend on SOX2 function. Genes Dev. *17*, 126–140.

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jørgensen, H.F., John, R.M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., et al. (2006). Chromatin signatures of pluripotent cell lines. Nat. Cell Biol. *8*, 532–538.

Ban, N., Nissen, P., Hansen, J., Moore, P.B., and Steitz, T.A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science *289*, 905–920.

Barak, O., Lazzaro, M.A., Cooch, N.S., Picketts, D.J., and Shiekhattar, R. (2004). A Tissue-specific, Naturally Occurring Human SNF2L Variant Inactivates Chromatin

Remodeling. J. Biol. Chem. *279*, 45130–45138.

Barbosa, C., Peixeiro, I., Romão, L., Morris, D., Geballe, A., Calvo, S., Pagliarini, D., Mootha, V., Mendell, J., Sharifi, N., et al. (2013). Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. PLoS Genet. *9*, e1003529.

Basso, K., Margolin, A. a, Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. Nat Genet *37*, 382–390.

Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., and Gilad, Y. (2015). Impact of regulatory variation from RNA to protein. Science (80-.). *347*, 664–667.

Bayés, M., Magano, L.F., Rivera, N., Flores, R., and Pérez Jurado, L.A. (2003). Mutational mechanisms of Williams-Beuren syndrome deletions. Am. J. Hum. Genet. *73*, 131–151.

Bear, M.F., Huber, K.M., and Warren, S.T. (2004). The mGluR theory of fragile X mental retardation. Trends Neurosci. *27*, 370–377.

Bellugi, U., Lichtenberger, L., Mills, D., Galaburda, A., and Korenberg, J.R. (1999). Bridging cognition, the brain and molecular genetics: evidence from Williams syndrome. Trends Neurosci. *22*, 197–207.

Bergen, S.E., O'Dushlaine, C.T., Ripke, S., Lee, P.H., Ruderfer, D.M., Akterin, S., Moran, J.L., Chambert, K.D., Handsaker, R.E., Backlund, L., et al. (2012). Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. Mol. Psychiatry *17*, 880–886.

Bergsland, M., Ramskold, D., Zaouter, C., Klum, S., Sandberg, R., and Muhr, J. (2011). Sequentially acting Sox transcription factors in neural lineage development. Genes Dev. *25*, 2453–2464.

Bernier, R., Golzio, C., Xiong, B., Stessman, H.A., Coe, B.P., Penn, O., Witherspoon, K., Gerdts, J., Baker, C., Vulto-van Silfhout, A.T., et al. (2014). Disruptive CHD8 mutations define a subtype of autism early in development. Cell *158*, 263–276.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bertola, D.R., Pereira, A.C., Passetti, F., de Oliveira, P.S.L., Messiaen, L., Gelb, B.D., Kim, C.A., and Krieger, J.E. (2005). Neurofibromatosis-Noonan syndrome: Molecular evidence of the concurrence of both disorders in a patient. Am. J. Med. Genet. Part A *136A*, 242–245.

Beunders, G., van de Kamp, J.M., Veenhoven, R.H., van Hagen, J.M., Nieuwint, A.W.M., and Sistermans, E.A. (2010). A triplication of the Williams-Beuren syndrome region in a patient with mental retardation, a severe expressive language delay, behavioural problems and dysmorphisms. J. Med. Genet. *47*, 271–275.

BEUREN, A.J., APITZ, J., and HARMJANZ, D. (1962). Supravalvular aortic stenosis in association with mental retardation and a certain facial appearance. Circulation *26*, 1235–1240.

Bhat, M., Robichaud, N., Hulea, L., Sonenberg, N., Pelletier, J., and Topisirovic, I. (2015). Targeting the translation machinery in cancer. Nat. Rev. Drug Discov. *14*, 261–278.

Bogdanow, B., Zauber, H., and Selbach, M. (2016). Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides. Mol. Cell. Proteomics

15, 2791–2801.

Boocock, G.R.B., Morrison, J.A., Popovic, M., Richards, N., Ellis, L., Durie, P.R., and Rommens, J.M. (2002). Mutations in SBDS are associated with Shwachman–Diamond syndrome. Nat. Genet. *33*, 97–101.

Borralleras, C., Sahun, I., Pérez-Jurado, L.A., and Campuzano, V. (2015). Intracisternal Gtf2i Gene Therapy Ameliorates Deficits in Cognition and Synaptic Plasticity of a Mouse Model of Williams-Beuren Syndrome. Mol. Ther. *23*, 1691–1699.

Bowman, M., Oldridge, M., Archer, C., O'Rourke, A., McParland, J., Brekelmans, R., Seller, A., and Lester, T. (2012). Gross deletions in TCOF1 are a cause of Treacher–Collins–Franceschetti syndrome. Eur. J. Hum. Genet. *20*, 769–777.

Brennand, K.J., Simone, A., Jou, J., Gelboin-Burkhart, C., Tran, N., Sangar, S., Li, Y., Mu, Y.L., Chen, G., Yu, D., et al. (2011). Modelling schizophrenia using human induced pluripotent stem cells. Nature *473*, 221–+.

Broadbent, H., Farran, E.K., Chin, E., Metcalfe, K., Tassabehji, M., Turnpenny, P., Sansbury, F., Meaburn, E., and Karmiloff-Smith, A. (2014). Genetic contributions to visuospatial cognition in Williams syndrome: insights from two contrasting partial deletion patients. J. Neurodev. Disord. *6*, 18.

Brooks, S.S., Wall, A.L., Golzio, C., Reid, D.W., Kondyles, A., Willer, J.R., Botti, C., Nicchitta, C. V., Katsanis, N., and Davis, E.E. (2014). A Novel Ribosomopathy Caused by Dysfunction of RPL10 Disrupts Neurodevelopment and Causes X-Linked Microcephaly in Humans. Genetics *198*, 723–733.

Brykczynska, U., Hisano, M., Erkek, S., Ramos, L., Oakeley, E.J., Roloff, T.C., Beisel, C., Schübeler, D., Stadler, M.B., and Peters, A.H.F.M. (2010). Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa. Nat. Struct. Mol. Biol. *17*, 679–687.

Buchan, J.R., and Parker, R. (2009). Eukaryotic Stress Granules: The Ins and Outs of Translation. Mol. Cell *36*, 932–941.

Burgold, T., Spreafico, F., De Santa, F., Totaro, M.G., Prosperini, E., Natoli, G., and Testa, G. (2008). The Histone H3 Lysine 27-Specific Demethylase Jmjd3 Is Required for Neural Commitment. PLoS One *3*, e3034.

Burgold, T., Voituron, N., Caganova, M., Tripathi, P.P., Menuet, C., Tusi, B.K., Spreafico, F., Bevengut, M., Gestreau, C., Buontempo, S., et al. (2012). The H3K27 Demethylase JMJD3 Is Required for Maintenance of the Embryonic Respiratory Neuronal Network, Neonatal Breathing, and Survival. Cell Rep *2*, 1244–1258.

Burn, J. (1986). Williams syndrome. J. Med. Genet. *23*, 389–395.

Burns, F.R., von Kannen, S., Guy, L., Raper, J.A., Kamholz, J., and Chang, S. (1991). DM-GRASP, a novel immunoglobulin superfamily axonal surface protein that supports neurite extension. Neuron *7*, 209–220.

Bykhovskaya, Y., Casas, K., Mengesha, E., Inbal, A., and Fischel-Ghodsian, N. (2004). Missense Mutation in Pseudouridine Synthase 1 (PUS1) Causes Mitochondrial Myopathy and Sideroblastic Anemia (MLASA). Am. J. Hum. Genet. *74*, 1303–1308.

Bylund, M., Andersson, E., Novitch, B.G., and Muhr, J. (2003). Vertebrate neurogenesis is counteracted by Sox1–3 activity. Nat. Neurosci. *6*, 1162–1168.

Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc. Natl. Acad. Sci. U. S. A. *106*, 7507–7512.

Del Campo, M., Antonell, A., Magano, L.F., Muñoz, F.J., Flores, R., Bayés, M., and Pérez Jurado, L.A. (2006). Hemizygosity at the NCF1 Gene in Patients with Williams-Beuren Syndrome Decreases Their Risk of Hypertension. Am. J. Hum. Genet. *78*, 533–542.

Canales, C.P., Wong, A.C.Y., Gunning, P.W., Housley, G.D., Hardeman, E.C., and Palmer, S.J. (2015). The role of GTF2IRD1 in the auditory pathology of Williams–Beuren Syndrome. Eur. J. Hum. Genet. *23*, 774–780.

Capitão, L., Sampaio, A., Sampaio, C., Vasconcelos, C., Férnandez, M., Garayzábal, E., Shenton, M.E., and Gonçalves, Ó.F. (2011). MRI amygdala volume in Williams Syndrome. Res. Dev. Disabil. *32*, 2767–2772.

Capossela, S., Muzio, L., Bertolo, A., Bianchi, V., Dati, G., Chaabane, L., Godi, C., Politi, L.S., Biffo, S., D'Adamo, P., et al. (2012). Growth defects and impaired cognitive-behavioral abilities in mice with knockout for Eif4h, a gene located in the mouse homolog of the Williams-Beuren syndrome critical region. Am J Pathol *180*, 1121–1135.

Cartwright, P., McLean, C., Sheppard, A., Rivett, D., Jones, K., and Dalton, S. (2005). LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. Development *132*, 885–896.

Cenik, C., Cenik, E.S., Byeon, G.W., Grubert, F., Candille, S.I., Spacek, D., Alsallakh, B., Tilgner, H., Araya, C.L., Tang, H., et al. (2015). Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. Genome Res. *25*, 1610–1621.

Chailangkarn, T., Trujillo, C.A., Freitas, B.C., Hrvoj-Mihic, B., Herai, R.H., Yu, D.X., Brown, T.T., Marchetto, M.C., Bardy, C., McHenry, L., et al. (2016). A human neurodevelopmental model for Williams syndrome. Nature *536*, 338–343.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell *113*, 643–655.

Chambers, S.M., Qi, Y., Mica, Y., Lee, G., Zhang, X.J., Niu, L., Bilsland, J., Cao, L., Stevens, E., Whiting, P., et al. (2012). Combined small-molecule inhibition accelerates developmental timing and converts human pluripotent stem cells into nociceptors. Nat Biotechnol *30*, 715–720.

Chau, V., Tobias, J.W., Bachmair, A., Marriott, D., Ecker, D.J., Gonda, D.K., and Varshavsky, A. (1989). A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein. Science *243*, 1576–1583.

Chen, P.-J., and Huang, Y.-S. (2012). CPEB2-eEF2 interaction impedes HIF-1α RNA translation. EMBO J. *31*, 959–971.

Christiansen, H.E., Schwarze, U., Pyott, S.M., AlSwaid, A., Al Balwi, M., Alrasheed, S., Pepin, M.G., Weis, M.A., Eyre, D.R., and Byers, P.H. (2010). Homozygosity for a Missense Mutation in SERPINH1, which Encodes the Collagen Chaperone Protein HSP47, Results in Severe Recessive Osteogenesis Imperfecta. Am. J. Hum. Genet. *86*, 389–398.

Chu, D., Kazana, E., Bellanger, N., Singh, T., Tuite, M.F., and von der Haar, T. (2014). Translation elongation can control translation initiation on eukaryotic mRNAs. EMBO J. *33*, 21–34.

Clayton-Smith, J., and Laan, L. (2003). Angelman syndrome: a review of the clinical and genetic aspects. J. Med. Genet. *40*, 87–95.

Cohen, A.S.A., Tuysuz, B., Shen, Y., Bhalla, S.K., Jones, S.J.M., and Gibson, W.T. (2015). A novel mutation in EED associated with overgrowth. J. Hum. Genet. *60*, 339–342.

Collins, B.C., Gillet, L.C., Rosenberger, G., Röst, H.L., Vichalkovski, A., Gstaiger, M., and Aebersold, R. (2013). Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. Nat. Methods *10*, 1246–1253.

Cowan, C.A. (2005). Nuclear Reprogramming of Somatic Cells After Fusion with Human Embryonic Stem Cells. Science (80-.). *309*, 1369–1373.

Crespi, B.J., Hurd, P.L., Mervis, C., Klein-Tasman, B., Dykens, E., Martens, M., Wilson, S., Reutens, D., Schubert, C., Morris, C., et al. (2014). Cognitive-behavioral phenotypes of Williams syndrome are associated with genetic variation in the GTF2I gene, in a healthy population. BMC Neurosci. *15*, 127.

CRICK, F. (1970). Central Dogma of Molecular Biology. Nature *227*, 561–563.

Dai, L., Bellugi, U., Chen, X.N., Pulst-Korenberg, A.M., Jarvinen-Pasley, A., Tirosh-Wagner, T., Eis, P.S., Graham, J., Mills, D., Searcy, Y., et al. (2009). Is it Williams syndrome? GTF2IRD1 implicated in visual-spatial construction and GTF2I in sociability revealed by high resolution arrays. Am J Med Genet A *149A*, 302–314.

Davis, L.J. (1997). The disability studies reader. Routledge *2nd*, x, 454 .

Deaton, A.M., Webb, S., Kerr, A.R.W., Illingworth, R.S., Guy, J., Andrews, R., and Bird, A. (2011). Cell type-specific DNA methylation at intragenic CpG islands in the immune system. Genome Res. *21*, 1074–1086.

Dennis, M.D., Jefferson, L.S., and Kimball, S.R. (2012a). Role of p70S6K1-mediated phosphorylation of eIF4B and PDCD4 proteins in the regulation of protein synthesis. J. Biol. Chem. *287*, 42890–42899.

Dennis, M.D., Jefferson, L.S., and Kimball, S.R. (2012b). Role of p70S6K1-mediated Phosphorylation of eIF4B and PDCD4 Proteins in the Regulation of Protein Synthesis. J. Biol. Chem. *287*, 42890–42899.

Dheedene, A., Maes, M., Vergult, S., and Menten, B. (2014). A de novo POU3F3 Deletion in a Boy with Intellectual Disability and Dysmorphic Features. Mol. Syndromol. *5*, 32–35.

Didiot, M.-C., Subramanian, M., Flatter, E., Mandel, J.-L., and Moine, H. (2009). Cells lacking the fragile X mental retardation protein (FMRP) have normal RISC activity but exhibit altered stress granule assembly. Mol. Biol. Cell *20*, 428–437.

Dmitriev, S.E., Terenin, I.M., Andreev, D.E., Ivanov, P.A., Dunaevsky, J.E., Merrick, W.C., and Shatsky, I.N. (2010). GTP-independent tRNA Delivery to the Ribosomal P-site by a Novel Eukaryotic Translation Factor. J. Biol. Chem. *285*, 26779–26787.

Dominguez, M.H., Ayoub, A.E., and Rakic, P. (2013). POU-III transcription factors (Brn1, Brn2, and Oct6) influence neurogenesis, molecular identity, and migratory destination of upper-layer cells of the cerebral cortex. Cereb. Cortex *23*, 2632–2643.

Edelmann, L., Prosnitz, A., Pardo, S., Bhatt, J., Cohen, N., Lauriat, T., Ouchanov, L., Gonzalez, P.J., Manghi, E.R., Bondy, P., et al. (2007). An atypical deletion of the Williams-Beuren syndrome interval implicates genes associated with defective visuospatial processing and autism. J Med Genet *44*, 136–143.

Eisinger, D.P., Dick, F.A., and Trumpower, B.L. (1997). Qsr1p, a 60S ribosomal subunit protein, is required for joining of 40S and 60S subunits. Mol. Cell. Biol. *17*, 5136–5145.

Eulalio, A., Mano, M., Ferro, M.D., Zentilin, L., Sinagra, G., Zacchigna, S., and Giacca, M.

(2012). Functional screening identifies miRNAs inducing cardiac regeneration. Nature *492*, 376–381.

Evans, M.J., and Kaufman, M.H. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature *292*, 154–156.

Evgrafov, O. V, Mersiyanova, I., Irobi, J., Van Den Bosch, L., Dierick, I., Leung, C.L., Schagina, O., Verpoorten, N., Van Impe, K., Fedotov, V., et al. (2004). Mutant small heat-shock protein 27 causes axonal Charcot-Marie-Tooth disease and distal hereditary motor neuropathy. Nat. Genet. *36*, 602–606.

Felsenfeld, G., and Bell, A.C. (2000). Methylation of a CTCF-dependent boundary controls imprinted expressionof the Igf2 gene. Nature *405*, 482–485.

Ferrero, G.B., Howald, C., Micale, L., Biamino, E., Augello, B., Fusco, C., Turturo, M.G., Forzano, S., Reymond, A., and Merla, G. (2010). An atypical 7q11.23 deletion in a normal IQ Williams-Beuren syndrome patient. Eur J Hum Genet *18*, 33–38.

Floor, S.N., Doudna, J.A., Amrani, N., Ghosh, S., Mangus, D., Jacobson, A., Arava, Y., Wang, Y., Storey, J., Liu, C., et al. (2016). Tunable protein synthesis by transcript isoforms in human cells. Elife *5*, 1276–1280.

Fodor, J.A. (1983). The modularity of mind : an essay on faculty psychology (MIT Press).

Francelle, L., Galvan, L., Gaillard, M.-C., Petit, F., Bernay, B., Guillermier, M., Bonvento, G., Dufour, N., Elalouf, J.-M., Hantraye, P., et al. (2015). The striatal long noncoding RNA Abhd11os is neuroprotective against an N-terminal fragment of mutant huntingtin in vivo. Neurobiol. Aging *36*, 1601.e7-1601.e16.

Frangiskakis, J.M., Ewart, A.K., Morris, C.A., Mervis, C.B., Bertrand, J., Robinson, B.F., Klein, B.P., Ensing, G.J., Everett, L.A., Green, E.D., et al. (1996). LIM-kinase1 hemizygosity implicated in impaired visuospatial constructive cognition. Cell *86*, 59–69.

Fusco, C., Micale, L., Egorov, M., Monti, M., D'Addetta, E. V, Augello, B., Cozzolino, F., Calcagni, A., Fontana, A., Polishchuk, R.S., et al. (2012). The E3-ubiquitin ligase TRIM50 interacts with HDAC6 and p62, and promotes the sequestration and clearance of ubiquitinated proteins into the aggresome. PLoS One *7*, e40440.

Fusco, C., Micale, L., Augello, B., Teresa Pellico, M., Menghini, D., Alfieri, P., Cristina Digilio, M., Mandriani, B., Carella, M., Palumbo, O., et al. (2014). Smaller and larger deletions of the Williams Beuren syndrome region implicate genes involved in mild Fusco, C., Micale, L., Augello, B., Teresa Pellico, M., Menghini, D., Alfieri, P., … Merla, G. (2014). Smaller and larger deletions of the Williams Beuren. Eur J Hum Genet *22*, 64–70.

Gao, M.C., Bellugi, U., Dai, L., Mills, D.L., Sobel, E.M., Lange, K., and Korenberg, J.R. (2010). Intelligence in Williams Syndrome is related to STX1A, which encodes a component of the presynaptic SNARE complex. PLoS One *5*, e10292.

Germain, P.-L., Ratti, E., and Boem, F. (2014). Junk or functional DNA? ENCODE and the function controversy. Biol. Philos. *29*, 807–831.

Germain, P.L., Vitriolo, A., Adamo, A., Laise, P., Das, V., and Testa, G. (2016). RNAontheBENCH: Computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. Nucleic Acids Res. *44*, 5054–5067.

Gessert, S., Maurus, D., Brade, T., Walther, P., Pandur, P., and Kühl, M. (2008). DM-GRASP/ALCAM/CD166 is required for cardiac morphogenesis and maintenance of

cardiac identity in first heart field derived cells. Dev. Biol. *321*, 150–161.

Gibson, W.T., Hood, R.L., Zhan, S.H., Bulman, D.E., Fejes, A.P., Moore, R., Mungall, A.J., Eydoux, P., Babul-Hirji, R., An, J., et al. (2012). Mutations in EZH2 Cause Weaver Syndrome. Am. J. Hum. Genet. *90*, 110–118.

Gillet, L.C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. Mol. Cell. Proteomics *11*, O111.016717-O111.016717.

Gkogkas, C.G., Khoutorsky, A., Ran, I., Rampakakis, E., Nevarko, T., Weatherill, D.B., Vasuta, C., Yee, S., Truitt, M., Dallaire, P., et al. (2013). Autism-related deficits via dysregulated eIF4E-dependent translational control. Nature *493*, 371–377.

Glaser, S., Lubitz, S., Loveland, K.L., Ohbo, K., Robb, L., Schwenk, F., Seibler, J., Roellig, D., Kranz, A., Anastassiadis, K., et al. (2009). The histone 3 lysine 4 methyltransferase, Mll2, is only required briefly in development and spermatogenesis. Epigenetics Chromatin *2*, 5.

Glickman, M.H., and Ciechanover, A. (2002). The Ubiquitin-Proteasome Proteolytic Pathway: Destruction for the Sake of Construction. Physiol. Rev. *82*, 373–428.

Goes, F.S., McGrath, J., Avramopoulos, D., Wolyniec, P., Pirooznia, M., Ruczinski, I., Nestadt, G., Kenny, E.E., Vacic, V., Peters, I., et al. (2015). Genome-wide association study of schizophrenia in Ashkenazi Jews. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. *168*, 649–659.

Gray, V., Karmiloff-Smith, A., Funnell, E., and Tassabehji, M. (2006). In-depth analysis of spatial cognition in Williams syndrome: A critical assessment of the role of the LIMK1 gene. Neuropsychologia *44*, 679–685.

Gupte, R.S., Piotr, P., Grabarek, J., Traganos, F., Darzynkiewicz, Z., and Lee, M.Y.W.T. (2005). RIα influences cellular proliferation in cancer cells by transporting RFC40 into the nucleus. Cancer Biol. Ther. *4*, 435–443.

GURDON, J.B. (1962). The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. J. Embryol. Exp. Morphol. *10*, 622–640.

Gurdon, J.B., Laskey, R.A., and Reeves, O.R. (1975). The developmental capacity of nuclei transplanted from keratinized skin cells of adult frogs. J. Embryol. Exp. Morphol. *34*, 93–112.

Hafer, N., Xu, S., Bhat, K.M., and Schedl, P. (2011). The Drosophila CPEB Protein Orb2 Has a Novel Expression Pattern and Is Important for Asymmetric Cell Division and Nervous System Function. Genetics *189*.

van Hagen, J.M., van der Geest, J.N., van der Giessen, R.S., Lagers-van Haselen, G.C., Eussen, H.J.F.M.M., Gille, J.J.P., Govaerts, L.C.P., Wouters, C.H., de Coo, I.F.M., Hoogenraad, C.C., et al. (2007). Contribution of CYLN2 and GTF2IRD1 to neurological and cognitive symptoms in Williams Syndrome. Neurobiol. Dis. *26*, 112–124.

Hakre, S., Tussie-Luna, M.I., Ashworth, T., Novina, C.D., Settleman, J., Sharp, P.A., and Roy, A.L. (2006). Opposing functions of TFII-I spliced isoforms in growth factor-induced gene expression. Mol Cell *24*, 301–308.

Hammoud, S.S., Nix, D.A., Zhang, H., Purwar, J., Carrell, D.T., and Cairns, B.R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. Nature.

Han, J.R., Gu, W., and Hecht, N.B. (1995). Testis-brain RNA-binding protein, a

testicular translational regulatory RNA-binding protein, is present in the brain and binds to the 3' untranslated regions of transported brain mRNAs. Biol. Reprod. *53*, 707–717.

Han, P., Hang, C.T., Yang, J., and Chang, C.-P. (2011). Chromatin remodeling in cardiovascular development and physiology. Circ. Res. *108*, 378–396.

Hay, N., and Sonenberg, N. (2004). Upstream and downstream of mTOR. Genes Dev. *18*, 1926–1945.

Helsmoortel, C., Vulto-van Silfhout, A.T., Coe, B.P., Vandeweyer, G., Rooms, L., van den Ende, J., Schuurs-Hoeijmakers, J.H.M., Marcelis, C.L., Willemsen, M.H., Vissers, L.E.L.M., et al. (2014). A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. Nat. Genet. *46*, 380–384.

Henrichsen, C.N., Csardi, G., Zabot, M.T., Fusco, C., Bergmann, S., Merla, G., and Reymond, A. (2011). Using transcription modules to identify expression clusters perturbed in Williams-Beuren syndrome. PLoS Comput Biol *7*, e1001054.

Hirota, H., Matsuoka, R., Chen, X.N., Salandanan, L.S., Lincoln, A., Rose, F.E., Sunahara, M., Osawa, M., Bellugi, U., and Korenberg, J.R. (2003). Williams syndrome deficits in visual spatial processing linked to GTF2IRD1 and GTF2I on chromosome 7q11.23. Genet Med *5*, 311–321.

Hobart, H.H., Morris, C.A., Mervis, C.B., Pani, A.M., Kistler, D.J., Rios, C.M., Kimberley, K.W., Gregg, R.G., and Bray-Ward, P. (2010). Inversion of the Williams syndrome region is a common polymorphism found more frequently in parents of children with Williams syndrome. Am. J. Med. Genet. Part C Semin. Med. Genet. *154C*, 220–228.

Hoeft, F., Dai, L., Haas, B.W., Sheau, K., Mimura, M., Mills, D., Galaburda, A., Bellugi, U., Korenberg, J.R., and Reiss, A.L. (2014). Mapping Genetically Controlled Neural Circuits of Social Behavior and Visuo-Motor Integration by a Preliminary Examination of Atypical Deletions with Williams Syndrome. PLoS One *9*, e104088.

Höglund, P., Kurotaki, N., Kytölä, S., Miyake, N., Somer, M., and Matsumoto, N. (2003). Familial Sotos syndrome is caused by a novel 1 bp deletion of the NSD1 gene. J. Med. Genet. *40*, 51–54.

Hong, S.-K., and Dawid, I.B. (2009). FGF-dependent left-right asymmetry patterning in zebrafish is mediated by Ier2 and Fibp1. Proc. Natl. Acad. Sci. *106*, 2230–2235.

van Hoof, A. (2002). Exosome-Mediated Recognition and Degradation of mRNAs Lacking a Termination Codon. Science (80-.). *295*, 2262–2264.

Hutsler, J.J., and Zhang, H. (2010). Increased dendritic spine densities on cortical projection neurons in autism spectrum disorders. Brain Res. *1309*, 83–94.

Illingworth, R., Kerr, A., DeSousa, D., Jørgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., et al. (2008). A Novel CpG Island Set Identifies Tissue-Specific Methylation at Developmental Gene Loci. PLoS Biol. *6*, e22.

Ingham, P.W. (1983). Differential expression of bithorax complex genes in the absence of the extra sex combs and trithorax genes. Nature *306*, 591–593.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science (80-.). *324*, 218–223.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. *3*, 318–356.

Ji, Z., Song, R., Regev, A., and Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife *4*, e08890.

Jones, W., Bellugi, U., Lai, Z., Chiles, M., Reilly, J., Lincoln, A., and Adolphs, R. (2000). II. Hypersociability in Williams Syndrome. J. Cogn. Neurosci. *12 Suppl 1*, 30–46.

Jourdy, Y., Chatron, N., Carage, M.-L., Fretigny, M., Meunier, S., Zawadzki, C., Gay, V., Negrier, C., Sanlaville, D., and Vinciguerra, C. (2016). Study of six patients with complete *F9* deletion characterized by cytogenetic microarray: role of the *SOX3* gene in intellectual disability. J. Thromb. Haemost. *14*, 1988–1993.

Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R., et al. (2015). Dynamic profiling of the protein life cycle in response to pathogens. Science (80-.). *347*, 1259038–1259038.

Kadoch, C., Hargreaves, D.C., Hodges, C., Elias, L., Ho, L., Ranish, J., and Crabtree, G.R. (2013). Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. Nat. Genet. *45*, 592–601.

Kahan, B.W., and Ephrussi, B. (1970). Developmental potentialities of clonal in vitro cultures of mouse testicular teratoma. J. Natl. Cancer Inst. *44*, 1015–1036.

Kaul, G., Pattan, G., and Rafeequi, T. (2011). Eukaryotic elongation factor-2 (eEF2): its regulation and peptide chain elongation. Cell Biochem. Funct. *29*, 227–234.

Kelava, I., Lewitus, E., and Huttner, W.B. (2013). The secondary loss of gyrencephaly as an example of evolutionary phenotypical reversal. Front. Neuroanat. *7*, 16.

Kevelam, S.H., Bierau, J., Salvarinova, R., Agrawal, S., Honzik, T., Visser, D., Weiss, M.M., Salomons, G.S., Abbink, T.E.M., Waisfisz, Q., et al. (2015). Recessive *ITPA* mutations cause an early infantile encephalopathy. Ann. Neurol. *78*, 649–658.

Khan, Z., Ford, M.J., Cusanovich, D.A., Mitrano, A., Pritchard, J.K., and Gilad, Y. (2013). Primate Transcript and Protein Expression Levels Evolve Under Compensatory Selection Pressures. Science (80-.). *342*.

Kirchhoff, M., Bisgaard, A.-M., Bryndorf, T., and Gerdes, T. (2007). MLPA analysis for a panel of syndromes with mental retardation reveals imbalances in 5.8% of patients with mental retardation and dysmorphic features, including duplications of the Sotos syndrome and Williams–Beuren syndrome regions. Eur. J. Med. Genet. *50*, 33–42.

Klauck, S.M., Felder, B., Kolb-Kokocinski, A., Schuster, C., Chiocchetti, A., Schupp, I., Wellenreuther, R., Schmötzer, G., Poustka, F., Breitenbach-Koller, L., et al. (2006). Mutations in the ribosomal protein gene RPL10 suggest a novel modulating disease mechanism for autism. Mol. Psychiatry *11*, 1073–1084.

Kleefstra, T., Kramer, J.M., Neveling, K., Willemsen, M.H., Koemans, T.S., Vissers, L.E.L.M., Wissink-Lindhout, W., Fenckova, M., van den Akker, W.M.R., Kasri, N.N., et al. (2012). Disruption of an EHMT1-Associated Chromatin-Modification Module Causes Intellectual Disability. Am. J. Hum. Genet. *91*, 73–82.

Kleinsmith, L.J., and Pierce, G.B. (1964). Multipotentiality of Single Embryonal Carcinoma Cells. Cancer Res. *24*.

de Klerk, E., and 't Hoen, P.A.C. (2015). Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. Trends Genet. *31*, 128–139.

Knight, S.J., Flannery, A. V, Hirst, M.C., Campbell, L., Christodoulou, Z., Phelps, S.R., Pointon, J., Middleton-Price, H.R., Barnicoat, A., and Pembrey, M.E. (1993). Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. Cell *74*, 127–134.

Knudtzon, J., Aksnes, L., Akslen, L.A., and Aarskog, D. (2008). Elevated 1,25-dihydroxyvitamin D and normocalcaemia in presumed familial Williams syndrome. Clin. Genet. *32*, 369–374.

Kondo, T., Asai, M., Tsukita, K., Kutoku, Y., Ohsawa, Y., Sunada, Y., Imamura, K., Egawa, N., Yahata, N., Okita, K., et al. (2013). Modeling Alzheimer's Disease with iPSCs Reveals Stress Phenotypes Associated with Intracellular Aβ and Differential Drug Responsiveness. Cell Stem Cell *12*, 487–496.

Koolen, D.A., Kramer, J.M., Neveling, K., Nillesen, W.M., Moore-Barton, H.L., Elmslie, F. V, Toutain, A., Amiel, J., Malan, V., Tsai, A.C.-H., et al. (2012). Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. Nat. Genet. *44*, 639–641.

Kriek, M., White, S.J., Szuhai, K., Knijnenburg, J., van Ommen, G.-J.B., den Dunnen, J.T., and Breuning, M.H. (2006). Copy number variation in regions flanked (or unflanked) by duplicons among patients with developmental delay and/or congenital malformations; detection of reciprocal and partial Williams-Beuren duplications. Eur. J. Hum. Genet. *14*, 180–189.

Kriks, S., Shim, J.W., Piao, J., Ganat, Y.M., Wakeman, D.R., Xie, Z., Carrillo-Reid, L., Auyeung, G., Antonacci, C., Buch, A., et al. (2011). Dopamine neurons derived from human ES cells efficiently engraft in animal models of Parkinson's disease. Nature *480*, 547–551.

Kristensen, A.R., Gsponer, J., and Foster, L.J. (2014). Protein synthesis rate is the predominant regulator of protein expression during differentiation. Mol. Syst. Biol. *9*, 689–689.

Kumar, A., Wolpert, C., Kandt, R.S., Segal, J., Pufky, J., Roses, A.D., Pericak-Vance, M.A., and Gilbert, J.R. (1995). A de novo frame-shift mutation in the tuberin gene. Hum. Mol. Genet. *4*, 1471–1472.

Kunszt, P., Blum, L., Hullár, B., Schmid, E., Srebniak, A., Wolski, W., Rinn, B., Elmer, F.-J., Ramakrishnan, C., Quandt, A., et al. (2015). iPortal: the swiss grid proteomics portal: Requirements and new features based on experience and usability considerations. Concurr. Comput. Pract. Exp. *27*, 433–445.

Van Laarhoven, P.M., Neitzel, L.R., Quintana, A.M., Geiger, E.A., Zackai, E.H., Clouthier, D.E., Artinger, K.B., Ming, J.E., and Shaikh, T.H. (2015). Kabuki syndrome genes KMT2D and KDM6A: functional analyses demonstrate critical roles in craniofacial, heart and brain development. Hum. Mol. Genet. *24*, 4443–4453.

Lalani, S.R., Safiullah, A.M., Fernbach, S.D., Harutyunyan, K.G., Thaller, C., Peterson, L.E., McPherson, J.D., Gibbs, R.A., White, L.D., Hefner, M., et al. (2006). Spectrum of CHD7 Mutations in 110 Individuals with CHARGE Syndrome and Genotype-Phenotype Correlation. Am. J. Hum. Genet. *78*, 303–314.

Lalli, M.A., Jang, J., Park, J.-H.C., Wang, Y., Guzman, E., Zhou, H., Audouard, M., Bridges, D., Tovar, K.R., Papuc, S.M., et al. (2016). Haploinsufficiency of BAZ1B contributes to Williams syndrome through transcriptional dysregulation of neurodevelopmental pathways. Hum. Mol. Genet. *25*, 1294–1306.

Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., King, N., Stein, S.E., and Aebersold, R.

(2007). Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics *7*, 655–667.

Lam, P.P.L., Leung, Y.-M., Sheu, L., Ellis, J., Tsushima, R.G., Osborne, L.R., and Gaisano, H.Y. (2005). Transgenic mouse overexpressing syntaxin-1A as a diabetes model. Diabetes *54*, 2744–2754.

Larsson, O., Sonenberg, N., and Nadon, R. (2011). anota: Analysis of differential translation in genome-wide studies. Bioinformatics *27*, 1440–1441.

Laumonnier, F., Ronce, N., Hamel, B.C.J., Thomas, P., Lespinasse, J., Raynaud, M., Paringaux, C., Van Bokhoven, H., Kalscheuer, V., Fryns, J.-P., et al. (2002). Transcription factor SOX3 is involved in X-linked mental retardation with growth hormone deficiency. Am. J. Hum. Genet. *71*, 1450–1455.

Laurent, B., Ruitu, L., Murn, J., Hempel, K., Ferrao, R., Xiang, Y., Liu, S., Garcia, B.A., Wu, H., Wu, F., et al. (2015). A specific LSD1/KDM1A isoform regulates neuronal differentiation through H3K9 demethylation. Mol. Cell *57*, 957–970.

Lawlor, M.A. (2002). Essential role of PDK1 in regulating cell size and development in mice. EMBO J. *21*, 3728–3738.

Lee, T.I., and Young, R.A. (2013). Transcriptional regulation and its misregulation in disease. Cell *152*, 1237–1251.

Lee, A.S.Y., Kranzusch, P.J., Doudna, J.A., and Cate, J.H.D. (2016). eIF3d is an mRNA cap-binding protein that is required for specialized translation initiation. Nature *536*, 96–99.

Lewis, P.H. (1949). Pc: Polycomb. Drosoph. Inf. Serv. *21*, 69.

Li, D.Y., Toland, A.E., Boak, B.B., Atkinson, D.L., Ensing, G.J., Morris, C.A., and Keating, M.T. (1997). Elastin point mutations cause an obstructive vascular disease, supravalvular aortic stenosis. Hum. Mol. Genet. *6*, 1021–1028.

Li, J.J., Bickel, P.J., and Biggin, M.D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ *2*, e270.

Li, X., Li, R., Lin, X., and Guan, M.-X. (2002). Isolation and characterization of the putative nuclear modifier gene MTO1 involved in the pathogenesis of deafness-associated mitochondrial 12 S rRNA A1555G mutation. J. Biol. Chem. *277*, 27256–27264.

Li, Y., McClintick, J., Zhong, L., Edenberg, H.J., Yoder, M.C., and Chan, R.J. (2005). Murine embryonic stem cell differentiation is promoted by SOCS-3 and inhibited by the zinc finger transcription factor Klf4. Blood *105*, 635–637.

Liao, B.-Y., and Zhang, J. (2008). Null mutations in human and mouse orthologs frequently result in different phenotypes. Proc. Natl. Acad. Sci. *105*, 6987–6992.

Lindqvist, L., Imataka, H., and Pelletier, J. (2008). Cap-dependent eukaryotic initiation factor-mRNA interactions probed by cross-linking. RNA *14*, 960–969.

Liu, T., Nie, F., Yang, X., Wang, X., Yuan, Y., Lv, Z., Zhou, L., Peng, R., Ni, D., Gu, Y., et al. (2015). MicroRNA-590 is an EMT-suppressive microRNA involved in the TGFβ signaling pathway. Mol. Med. Rep.

Liu, Y., Hüttenhain, R., Surinova, S., Gillet, L.C.J., Mouritsen, J., Brunner, R., Navarro, P., and Aebersold, R. (2013). Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS. Proteomics *13*, 1247–1256.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein

Levels on mRNA Abundance. Cell *165*, 535–550.

Longerich, S., San Filippo, J., Liu, D., and Sung, P. (2009). FANCI Binds Branched DNA and Is Monoubiquitinated by UBE2T-FANCL. J. Biol. Chem. *284*, 23182–23186.

Lu, R., Markowetz, F., Unwin, R.D., Leek, J.T., Airoldi, E.M., MacArthur, B.D., Lachmann, A., Rozov, R., Ma'ayan, A., Boyer, L.A., et al. (2009). Systems-level dynamic analyses of fate change in murine embryonic stem cells. Nature *462*, 358–362.

Macrae, T., Sargeant, T., Lemieux, S., Hébert, J., Deneault, E., and Sauvageau, G. (2013). RNA-Seq reveals spliceosome and proteasome genes as most consistent transcripts in human cancer cells. PLoS One *8*, e72884.

Magnani, D., Morle, L., Hasenpusch-Theil, K., Paschaki, M., Jacoby, M., Schurmans, S., Durand, B., and Theil, T. (2015). The ciliogenic transcription factor Rfx3 is required for the formation of the thalamocortical tract by regulating the patterning of prethalamus and ventral telencephalon. Hum. Mol. Genet. *24*, 2578–2593.

Makino, Y., Yamano, K., Kanemaki, M., Morikawa, K., Kishimoto, T., Shimbara, N., Tanaka, K., and Tamura, T. (1997). SUG1, a component of the 26 S proteasome, is an ATPase stimulated by specific RNAs. J. Biol. Chem. *272*, 23201–23205.

Malenfant, P., Liu, X., Hudson, M.L., Qiao, Y., Hrynchak, M., Riendeau, N., Hildebrand, M.J., Cohen, I.L., Chudley, A.E., Forster-Gibson, C., et al. (2011). Association of GTF2i in the Williams-Beuren Syndrome Critical Region with Autism Spectrum Disorders. J Autism Dev Disord.

Malenfant, P., Liu, X., Hudson, M.L., Qiao, Y., Hrynchak, M., Riendeau, N., Hildebrand, M.J., Cohen, I.L., Chudley, A.E., Forster-Gibson, C., et al. (2012). Association of GTF2i in the Williams-Beuren syndrome critical region with autism spectrum disorders. J Autism Dev Disord *42*, 1459–1469.

Marchetto, M.C.N., Carromeu, C., Acab, A., Yu, D., Yeo, G.W., Mu, Y., Chen, G., Gage, F.H., and Muotri, A.R. (2010). A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells. Cell *143*, 527–539.

Marintchev, A., Edmonds, K.A., Marintcheva, B., Hendrickson, E., Oberer, M., Suzuki, C., Herdy, B., Sonenberg, N., and Wagner, G. (2009). Topology and regulation of the human eIF4A/4G/4H helicase complex in translation initiation. Cell *136*, 447–460.

Marler, J.A., Elfenbein, J.L., Ryals, B.M., Urban, Z., and Netzloff, M.L. (2005). Sensorineural hearing loss in children and adults with Williams syndrome. Am. J. Med. Genet. Part A *138A*, 318–327.

McShane, E., Sin, C., Zauber, H., Wells, J.N., Donnelly, N., Wang, X., Hou, J., Chen, W., Storchova, Z., Marsh, J.A., et al. (2016). Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. Cell *167*, 803–815.e21.

Meijer, H.A., Kong, Y.W., Lu, W.T., Wilczynska, A., Spriggs, R. V., Robinson, S.W., Godfrey, J.D., Willis, A.E., and Bushell, M. (2013). Translational Repression and eIF4A2 Activity Are Critical for MicroRNA-Mediated Gene Regulation. Science (80-.). *340*.

Meloni, M., and Testa, G. (2014). Scrutinizing the epigenetics revolution. Biosocieties *9*, 431–456.

Mervis, C.B., Dida, J., Lam, E., Crawford-Zelli, N.A., Young, E.J., Henderson, D.R., Onay, T., Morris, C.A., Woodruff-Borden, J., Yeomans, J., et al. (2012). Duplication of GTF2I results in separation anxiety in mice and humans. Am J Hum Genet *90*, 1064–1070.

Mervis, C.B., Klein-Tasman, B.P., Huffman, M.J., Velleman, S.L., Pitts, C.H., Henderson, D.R., Woodruff-Borden, J., Morris, C.A., and Osborne, L.R. (2015). Children with

7q11.23 duplication syndrome: Psychological characteristics. Am. J. Med. Genet. Part A *167*, 1436–1450.

Meyer, H.-J., and Rape, M. (2014). Enhanced Protein Degradation by Branched Ubiquitin Chains. Cell *157*, 910–921.

Meyer-Lindenberg, A., Kohn, P., Mervis, C.B., Kippenhan, J.S., Olsen, R.K., Morris, C.A., and Berman, K.F. (2004). Neural Basis of Genetically Determined Visuospatial Construction Deficit in Williams Syndrome. Neuron *43*, 623–631.

Meyer-Lindenberg, A., Hariri, A.R., Munoz, K.E., Mervis, C.B., Mattay, V.S., Morris, C.A., and Berman, K.F. (2005). Neural correlates of genetically abnormal social cognition in Williams syndrome. Nat. Neurosci. *8*, 991–993.

Meyer-Lindenberg, A., Mervis, C.B., and Berman, K.F. (2006). Neural mechanisms in Williams syndrome: a unique window to genetic influences on cognition and behaviour. Nat. Rev. Neurosci. *7*, 380–393.

Micale, L., Turturo, M.G., Fusco, C., Augello, B., Jurado, L.A., Izzi, C., Digilio, M.C., Milani, D., Lapi, E., Zelante, L., et al. (2010). Identification and characterization of seven novel mutations of elastin gene in a cohort of patients affected by supravalvular aortic stenosis. Eur J Hum Genet *18*, 317–323.

Middeljans, E., Wan, X., Jansen, P.W., Sharma, V., Stunnenberg, H.G., and Logie, C. (2012). SS18 Together with Animal-Specific Factors Defines Human BAF-Type SWI/SNF Complexes. PLoS One *7*, e33834.

Miyake, N., Koshimizu, E., Okamoto, N., Mizuno, S., Ogata, T., Nagai, T., Kosho, T., Ohashi, H., Kato, M., Sasaki, G., et al. (2013). *MLL2* and *KDM6A* mutations in patients with Kabuki syndrome. Am. J. Med. Genet. Part A *161*, 2234–2243.

Morris, C.A., Mervis, C.B., and Osborne, L.R. (2011). Frequency of the 7q11.23 inversion polymorphism in transmitting parents of children with Williams syndrome and in the general population does not differ between North America and Europe. Mol. Cytogenet. *4*, 7.

Mortimer, S.E., and Hedstrom, L. (2005). Autosomal dominant retinitis pigmentosa mutations in inosine 5′-monophosphate dehydrogenase type I disrupt nucleic acid binding. Biochem. J. *390*, 41–47.

Moulik, M., Vatta, M., Witt, S.H., Arola, A.M., Murphy, R.T., McKenna, W.J., Boriek, A.M., Oka, K., Labeit, S., Bowles, N.E., et al. (2009). ANKRD1, the Gene Encoding Cardiac Ankyrin Repeat Protein, Is a Novel Dilated Cardiomyopathy Gene. J. Am. Coll. Cardiol. *54*, 325–333.

Murata, Y., and Wharton, R.P. (1995). Binding of pumilio to maternal hunchback mRNA is required for posterior patterning in Drosophila embryos. Cell *80*, 747–756.

Napoli, I., Mercaldo, V., Boyl, P.P., Eleuteri, B., Zalfa, F., De Rubeis, S., Di Marino, D., Mohr, E., Massimi, M., Falconi, M., et al. (2008). The Fragile X Syndrome Protein Represses Activity-Dependent Translation through CYFIP1, a New 4E-BP. Cell *134*, 1042–1054.

Navia-Paldanius, D., Patel, J.Z., López Navarro, M., Jakupović, H., Goffart, S., Pasonen-Seppänen, S., Nevalainen, T.J., Jääskeläinen, T., Laitinen, T., Laitinen, J.T., et al. (2016). Chemoproteomic, biochemical and pharmacological approaches in the discovery of inhibitors targeting human α/β-hydrolase domain containing 11 (ABHD11). Eur. J. Pharm. Sci. *93*, 253–263.

Nicholas, C.R., Chen, J., Tang, Y., Southwell, D.G., Chalmers, N., Vogt, D., Arnold, C.M.,

Chen, Y.J., Stanley, E.G., Elefanty, A.G., et al. (2013). Functional Maturation of hPSC-Derived Forebrain Interneurons Requires an Extended Timeline and Mimics Human Neural Development. Cell Stem Cell *12*, 573–586.

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Scholer, H., and Smith, A. (1998). Formation of pluripotent stem cells in the mammalian embryo dependes on the POU transcription factor Oct4. Cell *95*, 379–391.

Noble, J.A., and Valdes, A.M. (2011). Genetics of the HLA region in the prediction of type 1 diabetes. Curr. Diab. Rep. *11*, 533–542.

Nowotny, H., and Testa, G. Naked genes : reinventing the human in the molecular age.

Orr, S.J., and McVicar, D.W. (2011). LAB/NTAL/Lat2: a force to be reckoned with in all leukocytes? J. Leukoc. Biol. *89*, 11–19.

Osborne, L.R. (2010). Animal Models of Williams Syndrome. Am. J. Med. Genet. Part C-Seminars Med. Genet. *154C*, 209–219.

Õunap, K., Leetsi, L., Matsoo, M., and Kurg, R. (2015). The Stability of Ribosome Biogenesis Factor WBSCR22 Is Regulated by Interaction with TRMT112 via Ubiquitin-Proteasome Pathway. PLoS One *10*, e0133841.

Palacios-Verdú, M.G., Segura-Puimedon, M., Borralleras, C., Flores, R., Del Campo, M., Campuzano, V., and Pérez-Jurado, L.A. (2015). Metabolic abnormalities in Williams–Beuren syndrome. J. Med. Genet. *52*, 248–255.

Parsyan, A., Svitkin, Y., Shahbazian, D., Gkogkas, C., Lasko, P., Merrick, W.C., and Sonenberg, N. (2011). mRNA helicases: the tacticians of translational control. Nat Rev Mol Cell Biol *12*, 235–245.

Pasca, S.P., Portmann, T., Voineagu, I., Yazawa, M., Shcheglovitov, A., Pasca, A.M., Cord, B., Palmer, T.D., Chikahisa, S., Nishino, S., et al. (2011). Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy syndrome. Nat Med *17*, 1657–1662.

Paterson, S.J., Girelli, L., Butterworth, B., and Karmiloff-Smith, A. (2006). Are numerical impairments syndrome specific? Evidence from Williams syndrome and Down's syndrome. J. Child Psychol. Psychiatry *47*, 190–204.

Patil, S.J., Salian, S., Bhat, V., Girisha, K.M., Shrivastava, Y., VS, K., and Sapare, A. (2015). Familial 7q11.23 duplication with variable phenotype. Am. J. Med. Genet. Part A *167*, 2727–2730.

Pérez Jurado, L.A., Peoples, R., Kaplan, P., Hamel, B.C., and Francke, U. (1996). Molecular definition of the chromosome 7 deletion in Williams syndrome and parent-of-origin effects on growth. Am. J. Hum. Genet. *59*, 781–792.

Pestova, T. V, and Kolupaeva, V.G. (2002). The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. Genes Dev. *16*, 2906–2922.

Pober, B.R. (2010a). MEDICAL PROGRESS Williams-Beuren Syndrome (vol 362, pg 239, 2010). N. Engl. J. Med. *362*, 2142.

Pober, B.R., Wang, E., Caprio, S., Petersen, K.F., Brandt, C., Stanley, T., Osborne, L.R., Dzuria, J., and Gulanski, B. (2010). High prevalence of diabetes and pre-diabetes in adults with Williams syndrome. Am. J. Med. Genet. C. Semin. Med. Genet. *154C*, 291–298.

Pratt, J.M., Petty, J., Riba-Garcia, I., Robertson, D.H.L., Gaskell, S.J., Oliver, S.G., and

Beynon, R.J. (2002). Dynamics of Protein Turnover, a Missing Dimension in Proteomics. Mol. Cell. Proteomics *1*, 579–591.

Prescott, J., Jariwala, U., Jia, L., Cogan, J.P., Barski, A., Pregizer, S., Shen, H.C., Arasheben, A., Neilson, J.J., Frenkel, B., et al. (2007). Androgen receptor-mediated repression of novel target genes. Prostate *67*, 1371–1383.

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R., et al. (2015). Codon optimality is a major determinant of mRNA stability. Cell *160*, 1111–1124.

Prontera, P., Serino, D., Caldini, B., Scarponi, L., Merla, G., Testa, G., Muti, M., Napolioni, V., Mazzotta, G., Piccirilli, M., et al. (2014). Brief Report: Functional MRI of a Patient with 7q11.23 Duplication Syndrome and Autism Spectrum Disorder. J Autism Dev Disord.

Puente, X.S., Quesada, V., Osorio, F.G., Cabanillas, R., Cadiñanos, J., Fraile, J.M., Ordóñez, G.R., Puente, D.A., Gutiérrez-Fernández, A., Fanjul-Fernández, M., et al. (2011). Exome Sequencing and Functional Analysis Identifies BANF1 Mutation as the Cause of a Hereditary Progeroid Syndrome. Am. J. Hum. Genet. *88*, 650–656.

Qin, J., Hu, Y., Xu, F., Yalamanchili, H.K., and Wang, J. (2014). Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. Methods *67*, 294–303.

Quadrato, G., Brown, J., and Arlotta, P. (2016). The promises and challenges of human brain organoids as models of neuropsychiatric disease. Nat. Med. *22*, 1220–1228.

Richter, N.J., Rogers Jr., G.W., Hensold, J.O., and Merrick, W.C. (1999). Further biochemical and kinetic characterization of human eukaryotic initiation factor 4H. J Biol Chem *274*, 35415–35424.

Rivera-Pomar, R., Niessing, D., Schmidt-Ott, U., Gehring, W.J., and Jacklë, H. (1996). RNA binding and translational suppression by bicoid. Nature *379*, 746–749.

Rizzoti, K., Brunelli, S., Carmignac, D., Thomas, P.Q., Robinson, I.C., and Lovell-Badge, R. (2004). SOX3 is required during the formation of the hypothalamo-pituitary axis. Nat. Genet. *36*, 247–255.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rosenberger, G., Koh, C.C., Guo, T., Röst, H.L., Kouvonen, P., Collins, B.C., Heusel, M., Liu, Y., Caron, E., Vichalkovski, A., et al. (2014). A repository of assays to quantify 10,000 human proteins by SWATH-MS. Sci. Data *1*, 140031.

Röst, H.L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S.M., Schubert, O.T., Wolski, W., Collins, B.C., Malmström, J., Malmström, L., et al. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat. Biotechnol. *32*, 219–223.

Röst, H.L., Liu, Y., D'Agostino, G., Zanella, M., Navarro, P., Rosenberger, G., Collins, B.C., Gillet, L., Testa, G., Malmström, L., et al. (2016). TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. Nat. Methods 1–14.

Rougvie, A.E., and Lis, J.T. (1988). The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. Cell *54*, 795–804.

Roy, A.L. (2001). Biochemistry and biology of the inducible multifunctional transcription factor TFII-I. Gene *274*, 1–13.

Rozovsky, N., Butterworth, A.C., and Moore, M.J. (2008). Interactions between eIF4AI and its accessory factors eIF4B and eIF4H. RNA *14*, 2136–2148.

Sakurai, T., Dorr, N.P., Takahashi, N., McInnes, L.A., Elder, G.A., and Buxbaum, J.D. (2011). Haploinsufficiency of Gtf2i, a gene deleted in Williams Syndrome, leads to increases in social interactions. Autism Res *4*, 28–39.

Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A., et al. (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron *70*, 863–885.

Santen, G.W.E., Aten, E., Sun, Y., Almomani, R., Gilissen, C., Nielsen, M., Kant, S.G., Snoeck, I.N., Peeters, E.A.J., Hilhorst-Hofstee, Y., et al. (2012). Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome. Nat. Genet. *44*, 379–380.

Santini, E., Huynh, T., Macaskill, A., Carter, A., Pierre, P., Ruggero, D., Kaphzan, H., and Klann, E. (2012). Exaggerated translation causes synaptic and behavioural aberrations associated with autism. Nature.

Schepens, B., Tinton, S.A., Bruynooghe, Y., Beyaert, R., and Cornelis, S. (2005). The polypyrimidine tract-binding protein stimulates HIF-1alpha IRES-mediated translation during hypoxia. Nucleic Acids Res. *33*, 6884–6894.

Schosserer, M., Minois, N., Angerer, T.B., Amring, M., Dellago, H., Harreither, E., Calle-Perez, A., Pircher, A., Gerstl, M.P., Pfeifenberger, S., et al. (2015). Methylation of ribosomal RNA by NSUN5 is a conserved mechanism modulating organismal lifespan. Nat. Commun. *6*, 6158.

Schuh, A.L., and Audhya, A. (2014). The ESCRT machinery: from the plasma membrane to endosomes and back again. Crit. Rev. Biochem. Mol. Biol. *49*, 242–261.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Schwartz, C.E., Tarpey, P.S., Lubs, H.A., Verloes, A., May, M.M., Risheg, H., Friez, M.J., Futreal, P.A., Edkins, S., Teague, J., et al. (2007). The original Lujan syndrome family has a novel missense mutation (p.N1007S) in the MED12 gene. J. Med. Genet. *44*, 472–477.

Seifert, W., Beninde, J., Hoffmann, K., Lindner, T.H., Bassir, C., Aksu, F., Hübner, C., Verbeek, N.E., Mundlos, S., and Horn, D. (2009). HPGD mutations cause cranioosteoarthropathy but not autosomal dominant digital clubbing. Eur. J. Hum. Genet. *17*, 1570–1576.

Serre, V., Rozanska, A., Beinat, M., Chretien, D., Boddaert, N., Munnich, A., Rötig, A., and Chrzanowska-Lightowlers, Z.M. (2013). Mutations in mitochondrial ribosomal protein MRPL12 leads to growth retardation, neurological deterioration and mitochondrial translation deficiency. Biochim. Biophys. Acta *1832*, 1304–1312.

Shantz, L.M., and Pegg, A.E. (1999). Translational regulation of ornithine decarboxylase and other enzymes of the polyamine pathway. Int. J. Biochem. Cell Biol. *31*, 107–122.

Shi, Y., Kirwan, P., Smith, J., Robinson, H.P., and Livesey, F.J. (2012a). Human cerebral

cortex development from pluripotent stem cells to functional excitatory synapses. Nat Neurosci *15*, 477–486, S1.

Shi, Y., Kirwan, P., and Livesey, F.J. (2012b). Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. Nat Protoc *7*, 1836–1846.

Shteynberg, D., Deutsch, E.W., Lam, H., Eng, J.K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R.L., Aebersold, R., and Nesvizhskii, a. I. (2011). iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates. Mol. Cell. Proteomics *10*, M111.007690-M111.007690.

Sin, C., Chiarugi, D., and Valleriani, A. (2015). Single-molecule modeling of mRNA degradation by miRNA: Lessons from data. BMC Syst. Biol. *9*, S2.

Sin, C., Chiarugi, D., and Valleriani, A. (2016). Degradation Parameters from Pulse-Chase Experiments. PLoS One *11*, e0155028.

van Slegtenhorst, M., de Hoogt, R., Hermans, C., Nellist, M., Janssen, B., Verhoef, S., Lindhout, D., van den Ouweland, A., Halley, D., Young, J., et al. (1997). Identification of the tuberous sclerosis gene TSC1 on chromosome 9q34. Science *277*, 805–808.

Soldi, M., and Bonaldi, T. (2013). The Proteomic Investigation of Chromatin Functional Domains Reveals Novel Synergisms among Distinct Heterochromatin Components. Mol. Cell. Proteomics *12*, 764–780.

Somerville, M.J., Mervis, C.B., Young, E.J., Seo, E.-J., del Campo, M., Bamforth, S., Peregrine, E., Loo, W., Lilley, M., Pérez-Jurado, L.A., et al. (2005). Severe Expressive-Language Delay Related to Duplication of the Williams–Beuren Locus. N. Engl. J. Med. *353*, 1694–1701.

Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell *136*, 731–745.

Spirin, A.S. (2009). How does a scanning ribosomal particle move along the 5'-untranslated region of eukaryotic mRNA? Brownian Ratchet model. Biochemistry *48*, 10688–10692.

Sterneck, E., and Johnson, P.F. (1998). CCAAT/enhancer binding protein beta is a neuronal transcriptional regulator activated by nerve growth factor receptor signaling. J. Neurochem. *70*, 2424–2433.

Stessman, H.A., Bernier, R., and Eichler, E.E. (2014). A Genotype-First Approach to Defining the Subtypes of a Complex Disease. Cell *156*, 872–877.

Sun, Y., Atas, E., Lindqvist, L., Sonenberg, N., Pelletier, J., and Meller, A. (2012). The eukaryotic initiation factor eIF4H facilitates loop-binding, repetitive RNA unwinding by the eIF4A DEAD-box helicase. Nucleic Acids Res *40*, 6199–6207.

Swartz, J.R., Waller, R., Bogdan, R., Knodt, A.R., Sabhlok, A., Hyde, L.W., and Hariri, A.R. (2015). A Common Polymorphism in a Williams Syndrome Gene Predicts Amygdala Reactivity and Extraversion in Healthy Adults. Biol. Psychiatry.

Tada, M., Takahama, Y., Abe, K., Nakatsuji, N., and Tada, T. (2001). Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells. Curr. Biol. *11*, 1553–1558.

Taipale, M., Tucker, G., Peng, J., Krykbaeva, I., Lin, Z.-Y., Larsen, B., Choi, H., Berger, B., Gingras, A.-C., and Lindquist, S. (2014). A Quantitative Chaperone Interaction Network Reveals the Architecture of Cellular Protein Homeostasis Pathways. Cell *158*,

434–448.

Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell *126*, 663–676.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872.

Tanaka, K., Suzuki, T., Hattori, N., and Mizuno, Y. (2004). Ubiquitin, proteasome and parkin. Biochim. Biophys. Acta - Mol. Cell Res. *1695*, 235–247.

Tanenbaum, M.E., Stern-Ginossar, N., Weissman, J.S., Vale, R.D., Aviner, R., Geiger, T., Elroy-Stein, O., Belloc, E., Mendez, R., Bonneau, A., et al. (2015). Regulation of mRNA translation during mitosis. Elife *4*, 1834–1844.

Tassabehji, M. (2005). GTF2IRD1 in Craniofacial Development of Humans and Mice. Science (80-.). *310*, 1184–1187.

Tassabehji, M., Metcalfe, K., Donnai, D., Hurst, J., Reardon, W., Burch, M., and Read, A.P. (1997). Elastin: genomic structure and point mutations in patients with supravalvular aortic stenosis. Hum. Mol. Genet. *6*, 1029–1036.

Testa, G. (2011). The time of timing: how Polycomb proteins regulate neurogenesis. Bioessays *33*, 519–528.

Thomas, N.S., Durkie, M., Potts, G., Sandford, R., Van Zyl, B., Youings, S., Dennis, N.R., and Jacobs, P.A. (2006). Parental and chromosomal origins of microdeletion and duplication syndromes involving 7q11.23, 15q11-q13 and 22q11. Eur. J. Hum. Genet. *14*, 831–837.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. Science *282*, 1145–1147.

Tipney, H.J., Hinsley, T.A., Brass, A., Metcalfe, K., Donnai, D., and Tassabehji, M. (2004). Isolation and characterisation of GTF2IRD2, a novel fusion gene and member of the TFII-I family of transcription factors, deleted in Williams-Beuren syndrome. Eur. J. Hum. Genet. *12*, 551–560.

Tordjman, S., Anderson, G.M., Botbol, M., Toutain, A., Sarda, P., Carlier, M., Saugier-Veber, P., Baumann, C., Cohen, D., Lagneaux, C., et al. (2012). Autistic Disorder in Patients with Williams-Beuren Syndrome: A Reconsideration of the Williams-Beuren Syndrome Phenotype. PLoS One *7*, e30778.

Torniero, C., dalla Bernardina, B., Novara, F., Vetro, A., Ricca, I., Darra, F., Pramparo, T., Guerrini, R., and Zuffardi, O. (2007). Cortical dysplasia of the left temporal lobe might explain severe expressive-language delay in patients with duplication of the Williams-Beuren locus. Eur J Hum Genet *15*, 62–67.

Torniero, C., Dalla Bernardina, B., Novara, F., Cerini, R., Bonaglia, C., Pramparo, T., Ciccone, R., Guerrini, R., and Zuffardi, O. (2008). Dysmorphic features, simplified gyral pattern and 7q11.23 duplication reciprocal to the Williams-Beuren deletion. Eur. J. Hum. Genet. *16*, 880–887.

Tsukumo, Y., Tsukahara, S., Furuno, A., Iemura, S., Natsume, T., and Tomida, A. (2015). The endoplasmic reticulum-localized protein TBL2 interacts with the 60S ribosomal subunit. Biochem. Biophys. Res. Commun. *462*, 383–388.

Tsukumo, Y., Tsukahara, S., Furuno, A., Iemura, S., Natsume, T., and Tomida, A. (2016).

TBL2 Associates With *ATF4* mRNA Via Its WD40 Domain and Regulates Its Translation During ER Stress. J. Cell. Biochem. *117*, 500–509.

Tsurusaki, Y., Okamoto, N., Ohashi, H., Kosho, T., Imai, Y., Hibi-Ko, Y., Kaname, T., Naritomi, K., Kawame, H., Wakui, K., et al. (2012). Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. Nat. Genet. *44*, 376–378.

Turimella, S.L., Bedner, P., Skubal, M., Vangoor, V.R., Kaczmarczyk, L., Karl, K., Zoidl, G., Gieselmann, V., Seifert, G., Steinhäuser, C., et al. (2015). Characterization of cytoplasmic polyadenylation element binding 2 protein expression and its RNA binding activity. Hippocampus *25*, 630–642.

Turner, D.J., Miretti, M., Rajan, D., Fiegler, H., Carter, N.P., Blayney, M.L., Beck, S., and Hurles, M.E. (2008). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. Nat. Genet. *40*, 90–95.

Tuttle, A.M., Hoffman, T.L., and Schilling, T.F. (2014). Rabconnectin-3a Regulates Vesicle Endocytosis and Canonical Wnt Signaling in Zebrafish Neural Crest Migration. PLoS Biol. *12*, e1001852.

Uehara, T., Kage-Nakadai, E., Yoshina, S., Imae, R., and Mitani, S. (2015). The Tumor Suppressor BCL7B Functions in the Wnt Signaling Pathway. PLoS Genet. *11*, e1004921.

Vandeweyer, G., Van der Aa, N., Reyniers, E., and Kooy, R.F. (2012). The Contribution of CLIP2 Haploinsufficiency to the Clinical Manifestations of the Williams-Beuren Syndrome. Am. J. Hum. Genet. *90*, 1071–1078.

Varga, E.A., Pastore, M., Prior, T., Herman, G.E., and McBride, K.L. (2009). The prevalence of PTEN mutations in a clinical pediatric cohort with autism spectrum disorders, developmental delay, and macrocephaly. Genet. Med. *11*, 111–117.

Vaysse, C., Philippe, C., Martineau, Y., Quelen, C., Hieblot, C., Renaud, C., Nicaise, Y., Desquesnes, A., Pannese, M., Filleron, T., et al. (2015). Key contribution of eIF4H-mediated translational control in tumor promotion. Oncotarget *6*, 39924–39940.

Vincent, M., Geneviève, D., Ostertag, A., Marlin, S., Lacombe, D., Martin-Coignard, D., Coubes, C., David, A., Lyonnet, S., Vilain, C., et al. (2016). Treacher Collins syndrome: a clinical and molecular study based on a large series of patients. Genet. Med. *18*, 49–56.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet.

Wan, M., Lee, S.S.J., Zhang, X., Houwink-Manville, I., Song, H.-R., Amir, R.E., Budden, S., Naidu, S., Pereira, J.L.P., Lo, I.F.M., et al. (1999). Rett Syndrome and Beyond: Recurrent Spontaneous and Familial MECP2 Mutations at CpG Hotspots. Am. J. Hum. Genet. *65*, 1520–1529.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M.Q., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet. *40*, 897–903.

Watatani, K., Hirabayashi, Y., Itoh, Y., and Gotoh, Y. (2012). PDK1 regulates the generation of oligodendrocyte precursor cells at an early stage of mouse telencephalic development. Genes to Cells *17*, 326–335.

Weiss, M.J., Cole, D.E., Ray, K., Whyte, M.P., Lafferty, M.A., Mulivor, R.A., and Harris, H. (1988). A missense mutation in the human liver/bone/kidney alkaline phosphatase gene causing a lethal form of hypophosphatasia. Proc. Natl. Acad. Sci. U. S. A. *85*,

7666–7669.

Wells, S.E., Hillner, P.E., Vale, R.D., and Sachs, A.B. (1998). Circularization of mRNA by Eukaryotic Translation Initiation Factors. Mol. Cell *2*, 135–140.

Wendell, S. (1996). The rejected body : feminist philosophical reflections on disability. In Routledge, p. 206.

Werner, A., Iwasaki, S., McGourty, C.A., Medina-Ruiz, S., Teerikorpi, N., Fedrigo, I., Ingolia, N.T., and Rape, M. (2015). Cell-fate determination by ubiquitin-dependent regulation of translation. Nature *525*, 523–527.

WILLIAMS, J.C., BARRATT-BOYES, B.G., and LOWE, J.B. (1961). Supravalvular aortic stenosis. Circulation *24*, 1311–1318.

Wilmut, I., Schnieke, A.E., McWhir, J., Kind, A.J., and Campbell, K.H. (1997). Viable offspring derived from fetal and adult mammalian cells. Nature *385*, 810–813.

Woods, K.S., Cundall, M., Turton, J., Rizotti, K., Mehta, A., Palmer, R., Wong, J., Chong, W.K., Al-Zyoud, M., El-Ali, M., et al. (2005). Over- and underdosage of SOX3 is associated with infundibular hypoplasia and hypopituitarism. Am. J. Hum. Genet. *76*, 833–849.

Wu, X.Q., and Hecht, N.B. (2000). Mouse testis brain ribonucleic acid-binding protein/translin colocalizes with microtubules and is immunoprecipitated with messenger ribonucleic acids encoding myelin basic protein, alpha calmodulin kinase II, and protamines 1 and 2. Biol. Reprod. *62*, 720–725.

Wu, D., Matsushita, K., Matsubara, H., Nomura, F., and Tomonaga, T. (2011). An alternative splicing isoform of eukaryotic initiation factor 4H promotes tumorigenesis in vivo and is a potential therapeutic target for human cancer. Int. J. Cancer *128*, 1018–1030.

Xiao, Z., Zou, Q., Liu, Y., Yang, X., Schwanhausser, B., Brar, G.A., Ingolia, N.T., Lareau, L.F., Weissman, J.S., Bazzini, A.A., et al. (2016). Genome-wide assessment of differential translations with ribosome profiling data. Nat. Commun. *7*, 11194.

Xiol, J., Cora, E., Koglgruber, R., Chuma, S., Subramanian, S., Hosokawa, M., Reuter, M., Yang, Z., Berninger, P., Palencia, A., et al. (2012). A Role for Fkbp6 and the Chaperone Machinery in piRNA Amplification and Transposon Silencing. Mol. Cell *47*, 970–979.

Xu, M., and Chen, L. (2016). An empirical likelihood ratio test robust to individual heterogeneity for differential expression analysis of RNA-seq. Brief. Bioinform. bbw103.

Xu, G.L., Bestor, T.H., Bourc'his, D., Hsieh, C.L., Tommerup, N., Bugge, M., Hulten, M., Qu, X., Russo, J.J., and Viegas-Péquignot, E. (1999). Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature *402*, 187–191.

Yau, R., and Rape, M. (2016). The increasing complexity of the ubiquitin code. Nat. Cell Biol. *18*, 579–586.

Zhang, B., Kirov, S., and Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res. *33*, W741-8.

Zhang, Y., Pak, C., Han, Y., Ahlenius, H., Zhang, Z., Chanda, S., Marro, S., Patzke, C., Acuna, C., Covy, J., et al. (2013). Rapid single-step induction of functional neurons from human pluripotent stem cells. Neuron *78*, 785–798.

Zhong, L., Chiusa, M., Cadar, A.G., Lin, A., Samaras, S., Davidson, J.M., and Lim, C.C.

(2015). Targeted inhibition of ANKRD1 disrupts sarcomeric ERK-GATA4 signal transduction and abrogates phenylephrine-induced cardiomyocyte hypertrophy. Cardiovasc. Res. *106*, 261–271.

Zhong, Y., Karaletsos, T., Drewe, P., Sreedharan, V., Kuo, D., Singh, K., Wendel, H.-G., and Rätsch, G. (2016). RiboDiff: Detecting Changes of mRNA Translation Efficiency from Ribosome Footprints. Bioinformatics btw585.

Zipursky, S.L., and Sanes, J.R. (2010). Chemoaffinity revisited: dscams, protocadherins, and neural circuit assembly. Cell *143*, 343–353.

Zolk, O., Frohme, M., Maurer, A., Kluxen, F.-W., Hentsch, B., Zubakov, D., Hoheisel, J.D., Zucker, I.H., Pepe, S., and Eschenhagen, T. (2002). Cardiac ankyrin repeat protein, a negative regulator of cardiac gene expression, is augmented in human heart failure. Biochem. Biophys. Res. Commun. *293*, 1377–1382.

Zorbas, C., Nicolas, E., Wacheul, L., Huvelle, E., Heurgue-Hamard, V., and Lafontaine, D.L.J. (2015). The human 18S rRNA base methyltransferases DIMT1L and WBSCR22-TRMT112 but not rRNA modification are required for ribosome biogenesis. Mol. Biol. Cell *26*, 2080–2095.

Zucchelli, S., Fasolo, F., Russo, R., Cimatti, L., Patrucco, L., Takahashi, H., Jones, M.H., Santoro, C., Sblattero, D., Cotella, D., et al. (2015). SINEUPs are modular antisense long non-coding RNAs that increase synthesis of target proteins in cells. Front. Cell. Neurosci. *9*.

Appendices

Appendix 1: DEGs found in the total RNA dataset, divided by comparison and ranked by log2(FC)

WBS vs 7dup

Gene	log2(FC) WBS vs 7DUP
CORO7-PAM16	7.965545609
ANKHD1-EIF4EBP3	2.316202082
EPCAM	-1.299165539
TRIM22	-1.452383771
EIF4H	-1.50731784
RFC2	-1.659907188
WBSCR22	-1.669503698
GTF2I	-1.730595533
FAM92A1	-1.922358746
ZNF772	-1.948471795
BAZ1B	-2.015771021
PCDHA3	-2.219030747
PCDHB5	-2.678254054
ZNF528	-2.777897495
HIST1H1A	-3.079409806
FAM218A	-3.316002196
CAT	-3.713503487
ZNF726	-3.779889179
TRDN	-3.805645334
HIST1H2BB	-3.843365858
COL22A1	-4.250411012
ZNF229	-4.373393706
ZNF300	-5.01540476
GOLGA6L9	-6.971521788
ZNF560	-6.980700052
ZNF728	-8.564891818

WBS vs CTL

Gene	log2(FC) WBS vs CTL
PRR5-ARHGAP8	7.021821998
MAGEA6	-7.177349673
GATS	-7.864540299
NPIPA3	-8.453526352

CTL vs 7dup

Gene	log2(FC) CTL vs 7DUP
C7orf55-LUC7L2	9.67004816
YIPF1	9.593671534
TGFBR3	8.909501057
LY75-CD302	7.94028697
TAS2R19	3.944479728
ZNF528	2.176261753
MME	2.023307779
SLITRK4	1.858640026
ZNF37A	1.830231458
ZNF107	1.700412421
DAAM1	1.691160536
PDK1	1.68748495
PCDHB3	1.655572502
EPHA7	1.531901415
TMEM212	1.470566477
VWDE	1.399316049
CCNG1	1.186953281
ZDBF2	1.135681944
GLTSCR2	-1.066231509
UQCC3	-1.15809764
ATPIF1	-1.17656681
SCLY	-1.181441119
MRPS26	-1.201324597
BANF1	-1.203710336
STK11	-1.212246089
HEXIM1	-1.218319377
ZBTB12	-1.283237923
NUDC	-1.406838347
BCL2L12	-1.50092379
RPLP1	-1.522696959
CCDC9	-1.61648646
MEX3D	-1.651273109
EMILIN1	-1.669217957
CCDC124	-1.681376167
ERICH1	-1.750113792
MRPL12	-1.827099067
SOX3	-1.854492361
CCDC85B	-1.862113079
IER2	-1.88152359
HSPB1	-2.071171356
RPS16	-2.103611664
MTRNR2L5	-2.554641907
DNAJC25-GNG10	-3.06331733
MTRNR2L4	-3.679767912
NPIPB4	-4.149696614

MTRNR2L3	-4.491985162
MTRNR2L6	-4.545419681
MTRNR2L1	-4.729834655
MTRNR2L8	-4.941854179
MTRNR2L12	-4.975125751
MTRNR2L10	-5.012305496
UMODL1	-6.380614689
CORO7-PAM16	-7.503891435

Appendix 2: DEGs found in the RPF dataset, divided by comparison and ranked by log2(FC)

WBS vs 7dup

Gene	log2(FC) WBS vs 7DUP
TRIM22	-1.394273481
EIF4H	-1.525684575
RFC2	-1.525697254
WBSCR22	-1.650549701
ABHD11	-1.652178076
GTF2I	-1.71696254
ACTN3	-1.719656287
NSUN5	-1.775139352
PCDHGB1	-2.043856112
PCDHB3	-2.067585376
BAZ1B	-2.071233546
DNAJC30	-2.108031991
PCDHGB5	-2.485500987
ZNF300	-3.376391537
PCDHB5	-3.536538934
HIST1H1A	-3.915754195
CPEB2	-3.938420519
CES1	-4.042795098
HIST1H3C	-4.266643498
CAT	-4.391741678
ZNF560	-4.777739089
COL22A1	-4.782345235
ZNF229	-5.273198484
PCDHA3	-5.369211071
NKX2-3	-5.655229804
TRIM67	-6.144417428
ZNF835	-7.767511261
ZNF726	-7.899961189

WBS vs CTL

Gene	log2(FC) WBS vs CTL
HSPE1-MOB4	9.01499279
PCDHGB3	4.021343045
PCDHB3	2.180861705
IQGAP1	1.939059859
PCDHGB5	1.775102165

CTL vs 7dup

Gene	log2(FC) CTL vs 7DUP
HSPE1-MOB4	8.782546337
CXorf40A	6.690733608
ANKRD1	3.135558081
SLC2A14	2.477482582
E2F1	-3.045628871
CNTNAP3B	-4.690827255
HIST1H3C	-4.924954956
CAPNS1	-5.253214866
SERPINB5	-5.370477795
TBC1D2B	-6.456628803
ISY1-RAB43	-7.252935923
CEBPB	-8.500903868
CPEB2	-8.827313798
TRIM67	-10.02779981

Appendix 3: DEPs found in the protein dataset, divided by comparison and ranked by log2(FC)

WBS vs 7dup

Gene	lof2(FC) WBS vs 7DUP
HPGD	0.338711903
BCL10	0.225354189
IMPDH1	0.181847797
LTA4H	0.17914281
UBE2T	0.174876545
LRRC16A	0.147508649
ENO2	0.132426323
PITHD1	0.100725452
ALPL	0.08948509
EIF2D	0.06525309
ITPA	0.024563487
RIOK2	-0.023686976
HLA-A	-0.055653366
ABHD11	-0.061739094
ATP6V0A1	-0.073414655
IGF2R	-0.080751009
MAPRE1	-0.113814095
CLIP2	-0.124566678
SMARCA1	-0.13587009
SERPINH1	-0.143026047
RFC2	-0.145857385
HSD11B2	-0.175041966
RCN1	-0.213467758
EPCAM	-0.24199208
KHDRBS3	-0.253512939
CAT	-0.295973553
SPTLC1	-0.35717112
BAZ1B	-0.360190419
TBL2	-0.37886345
WBSCR22	-0.415526022
GTF2I	-0.447215704
NSUN5	-0.507731336
EIF4H	-0.587631406

WBS vs CTL

Gene	log2(FC) WBS vs CTL
HPGD	0.338711903
BCL10	0.225354189
LTA4H	0.17914281
SMARCA1	-0.13587009
RFC2	-0.145857385
EPCAM	-0.24199208
SPTLC1	-0.35717112
BAZ1B	-0.360190419
TBL2	-0.37886345
WBSCR22	-0.415526022
GTF2I	-0.447215704
NSUN5	-0.507731336
EIF4H	-0.587631406

7dup vs CTL

Gene	log2(FC) DUP vs CTL
HLA-A	1.555228768
CLIP2	0.829262291
ABHD11	0.621673719
GTF2I	0.530783645
ENO2	0.482921748
BAZ1B	0.464556662
ALPL	0.377140845
TBL2	0.28352144
NSUN5	0.278563753
RFC2	0.217115069
IGF2R	0.189178131
RIOK2	-0.417987554

Appendix 4: Code

Differential gene expression on total RNA and RPF

```
degClinical <- function(rna, rpfcds, rpfutr, filter=T,
atyp="exclude", nf.rna, nf.rpf){

     require(edgeR)
rna <- round(rna)
rpfcds <- round(rpfcds)
rpfutr <- round(rpfutr)

     samples <- c("316M", "306O", "C7", "192B", "MIFF3",
"BU1CRE", "3391S", "242J", "242K", "CFG", "V548MB")

if(filter == T){
     rna <- rna[which(apply(rna,1,FUN=function(x){
sum(x>10)>3 })),]
     rpfcds <- rpfcds[which(apply(rpfcds,1,FUN=function(x){
sum(x>10)>3 })),]
     rpfutr <- rpfutr[which(apply(rpfutr,1,FUN=function(x){
sum(x>10)>3 })),]
     }

     rna <- rna[,samples]
     rpfcds <- rpfcds[,samples]
     rpfutr <- rpfutr[,samples]
     nf.rna <- nf.rna[samples,1]
     nf.rpf <- nf.rpf[samples,1]

     if (atyp == "control"){
          nc <- 4
          ns <- 4
          nw <- 3
               }else
     if (atyp == "wbs"){
          nc <- 3
          ns <- 5
          nw <- 4
               }else
     if (atyp == "exclude"){
          nc <- 3
          ns <- 5
          nw <- 3
          }

     d.dup.ctl <- data.frame(samples[c(ns:11)],
condition=c(rep("CTRL", nc), rep("DUP", 4)))
     d.wbs.ctl <- data.frame(samples[c(1:(nc+nw))],
condition=c(rep("WBS", nw), rep("CTL", nc)))
     d.wbs.dup <- data.frame(samples[c(1:nw, 8:11)],
condition=c(rep("WBS", nw), rep("DUP", 4)))

     rnadds <- DGEList(rna, norm.factors=nf.rna)
     rpfcdsdds <- DGEList(rpfcds, norm.factors=nf.rpf)
```

```
     rpfutrdds <- DGEList(rpfutr, norm.factors=nf.rpf)


#SUBSET NORMALIZATION FACTORS SPECIFIC FOR THE COMPARISONS

     rnaDCnf <- rnadds$samples$norm.factors[c(ns:11)]
     rnaWCnf <- rnadds$samples$norm.factors[c(1:(nc+nw))]
     rnaWDnf <- rnadds$samples$norm.factors[c(1:nw, 8:11)]

     rpfcdsDCnf <- rpfcdsdds$samples$norm.factors[c(ns:11)]
     rpfcdsWCnf <-
rpfcdsdds$samples$norm.factors[c(1:(nc+nw))]
     rpfcdsWDnf <- rpfcdsdds$samples$norm.factors[c(1:nw,
8:11)]

     rpfutrDCnf <- rpfutrdds$samples$norm.factors[c(ns:11)]
     rpfutrWCnf <-
rpfutrdds$samples$norm.factors[c(1:(nc+nw))]
     rpfutrWDnf <- rpfutrdds$samples$norm.factors[c(1:nw,
8:11)]


#MAKE MODEL MATRICES

     mmDC <- model.matrix(~condition, data=d.dup.ctl)
     mmWC <- model.matrix(~condition, data=d.wbs.ctl)
     mmWD <- model.matrix(~condition, data=d.wbs.dup)

#ESTIMATE DISPERSION

     rnaddsDC <- DGEList(rna[,c(ns:11)],
norm.factors=rnaDCnf)
     rnaddsWC <- DGEList(rna[,c(1:(nc+nw))],
norm.factors=rnaWCnf)
     rnaddsWD <- DGEList(rna[,c(1:nw, 8:11)],
norm.factors=rnaWDnf)

     rnaddsDC <- estimateDisp(rnaddsDC, mmDC)
     rnaddsWC <- estimateDisp(rnaddsWC, mmWC)
     rnaddsWD <- estimateDisp(rnaddsWD, mmWD)

     rpfcdsddsDC <- DGEList(rpfcds[,c(ns:11)],
norm.factors=rpfcdsDCnf)
     rpfcdsddsWC <- DGEList(rpfcds[,c(1:(nc+nw))],
norm.factors=rpfcdsWCnf)
     rpfcdsddsWD <- DGEList(rpfcds[,c(1:nw, 8:11)],
norm.factors=rpfcdsWDnf)

     rpfcdsddsDC <- estimateDisp(rpfcdsddsDC, mmDC)
     rpfcdsddsWC <- estimateDisp(rpfcdsddsWC, mmWC)
     rpfcdsddsWD <- estimateDisp(rpfcdsddsWD, mmWD)

     rpfutrddsDC <- DGEList(rpfutr[,c(ns:11)],
norm.factors=rpfutrDCnf)
     rpfutrddsWC <- DGEList(rpfutr[,c(1:(nc+nw))],
norm.factors=rpfutrWCnf)
```

```
     rpfutrddsWD <- DGEList(rpfutr[,c(1:nw, 8:11)],
norm.factors=rpfutrWDnf)

     rpfutrddsDC <- estimateDisp(rpfutrddsDC, mmDC)
     rpfutrddsWC <- estimateDisp(rpfutrddsWC, mmWC)
     rpfutrddsWD <- estimateDisp(rpfutrddsWD, mmWD)


#FIT

     rnaddsDCfit <- glmFit(rnaddsDC, mmDC)
     rnaddsWCfit <- glmFit(rnaddsWC, mmWC)
     rnaddsWDfit <- glmFit(rnaddsWD, mmWD)

     rpfcdsddsDCfit <- glmFit(rpfcdsddsDC, mmDC)
     rpfcdsddsWCfit <- glmFit(rpfcdsddsWC, mmWC)
     rpfcdsddsWDfit <- glmFit(rpfcdsddsWD, mmWD)

     rpfutrddsDCfit <- glmFit(rpfutrddsDC, mmDC)
     rpfutrddsWCfit <- glmFit(rpfutrddsWC, mmWC)
     rpfutrddsWDfit <- glmFit(rpfutrddsWD, mmWD)

#LOG LIKELIHOOD


     rnaddsDCLRT <- glmLRT(rnaddsDCfit, "conditionDUP")
     rnaddsWCLRT <- glmLRT(rnaddsWCfit, "conditionWBS")
     rnaddsWDLRT <- glmLRT(rnaddsWDfit, "conditionWBS")

     rpfcdsddsDCLRT <- glmLRT(rpfcdsddsDCfit,
"conditionDUP")
     rpfcdsddsWCLRT <- glmLRT(rpfcdsddsWCfit,
"conditionWBS")
     rpfcdsddsWDLRT <- glmLRT(rpfcdsddsWDfit,
"conditionWBS")

     rpfutrddsDCLRT <- glmLRT(rpfutrddsDCfit,
"conditionDUP")
     rpfutrddsWCLRT <- glmLRT(rpfutrddsWCfit,
"conditionWBS")
     rpfutrddsWDLRT <- glmLRT(rpfutrddsWDfit,
"conditionWBS")

#SIG RESULTS

     rna.res.DC <- as.data.frame(topTags(rnaddsDCLRT,
30000))
     rna.res.WC <- as.data.frame(topTags(rnaddsWCLRT,
30000))
     rna.res.WD <- as.data.frame(topTags(rnaddsWDLRT,
30000))

     rpfcds.res.DC <- as.data.frame(topTags(rpfcdsddsDCLRT,
30000))
     rpfcds.res.WC <- as.data.frame(topTags(rpfcdsddsWCLRT,
30000))
```

```
     rpfcds.res.WD <- as.data.frame(topTags(rpfcdsddsWDLRT,
30000))

     rpfutr.res.DC <- as.data.frame(topTags(rpfutrddsDCLRT,
30000))
     rpfutr.res.WC <- as.data.frame(topTags(rpfutrddsWCLRT,
30000))
     rpfutr.res.WD <- as.data.frame(topTags(rpfutrddsWDLRT,
30000))

degs_sig_dataframes <- list(rna.res.DC = rna.res.DC,
rna.res.WC = rna.res.WC, rna.res.WD = rna.res.WD,
rpfcds.res.DC = rpfcds.res.DC, rpfcds.res.WC =
rpfcds.res.WC, rpfcds.res.WD = rpfcds.res.WD)

#SIG GENES THRESHOLDED

     rna_DC_sig_005 <-
unique(row.names(rna.res.DC)[which(rna.res.DC$FDR < 0.05)])
     rna_DC_sig_01 <-
unique(row.names(rna.res.DC)[which(rna.res.DC$FDR < 0.1)])

     rna_WC_sig_005 <-
unique(row.names(rna.res.WC)[which(rna.res.WC$FDR < 0.05)])
     rna_WC_sig_01 <-
unique(row.names(rna.res.WC)[which(rna.res.WC$FDR < 0.1)])

     rna_WD_sig_005 <-
unique(row.names(rna.res.WD)[which(rna.res.WD$FDR < 0.05)])
     rna_WD_sig_01 <-
unique(row.names(rna.res.WD)[which(rna.res.WD$FDR < 0.1)])

     rpfcds_DC_sig_005 <-
unique(row.names(rpfcds.res.DC)[which(rpfcds.res.DC$FDR <
0.05)])
     rpfcds_DC_sig_01 <-
unique(row.names(rpfcds.res.DC)[which(rpfcds.res.DC$FDR <
0.1)])

     rpfcds_WC_sig_005 <-
unique(row.names(rpfcds.res.WC)[which(rpfcds.res.WC$FDR <
0.05)])
     rpfcds_WC_sig_01 <-
unique(row.names(rpfcds.res.WC)[which(rpfcds.res.WC$FDR <
0.1)])

     rpfcds_WD_sig_005 <-
unique(row.names(rpfcds.res.WD)[which(rpfcds.res.WD$FDR <
0.05)])
     rpfcds_WD_sig_01 <-
unique(row.names(rpfcds.res.WD)[which(rpfcds.res.WD$FDR <
0.1)])

     rpfutr_DC_sig_005 <-
unique(row.names(rpfutr.res.DC)[which(rpfutr.res.DC$FDR <
0.05)])
```

```
     rpfutr_DC_sig_01 <-
unique(row.names(rpfutr.res.DC)[which(rpfutr.res.DC$FDR <
0.1)])

     rpfutr_WC_sig_005 <-
unique(row.names(rpfutr.res.WC)[which(rpfutr.res.WC$FDR <
0.05)])
     rpfutr_WC_sig_01 <-
unique(row.names(rpfutr.res.WC)[which(rpfutr.res.WC$FDR <
0.1)])

     rpfutr_WD_sig_005 <-
unique(row.names(rpfutr.res.WD)[which(rpfutr.res.WD$FDR <
0.05)])
     rpfutr_WD_sig_01 <-
unique(row.names(rpfutr.res.WD)[which(rpfutr.res.WD$FDR <
0.1)])


#SIG GENES TABLES: ONLY DEGS

     deg_DC <- list(deg_RNA_01 = rna_DC_sig_01, deg_RNA_005
= rna_DC_sig_01, deg_RPF_CDS_01 = rpfcds_DC_sig_01,
deg_RPF_CDS_005 = rpfcds_DC_sig_01, deg_RPF_UTR_005 =
rpfutr_DC_sig_005, deg_RPF_UTR_01 = rpfutr_DC_sig_01)
     deg_WC <- list(deg_RNA_01 = rna_WC_sig_01, deg_RNA_005
= rna_WC_sig_01, deg_RPF_CDS_01 = rpfcds_WC_sig_01,
deg_RPF_CDS_005 = rpfcds_WC_sig_01, deg_RPF_UTR_005 =
rpfutr_WC_sig_005, deg_RPF_UTR_01 = rpfutr_WC_sig_01)
     deg_WD <- list(deg_RNA_01 = rna_WD_sig_01, deg_RNA_005
= rna_WD_sig_01, deg_RPF_CDS_01 = rpfcds_WD_sig_01,
deg_RPF_CDS_005 = rpfcds_WD_sig_01, deg_RPF_UTR_005 =
rpfutr_WD_sig_005, deg_RPF_UTR_01 = rpfutr_WD_sig_01)

#SIG GENES: 1st INTERSECTION

     deg_common_RNA_sig_01 <- intersect(rna_WD_sig_01,
(intersect(rna_DC_sig_01, rna_WC_sig_01)))
     deg_common_RPF_CDS_sig_01 <-
intersect(rpfcds_WD_sig_01, (intersect(rpfcds_DC_sig_01,
rpfcds_WC_sig_01)))
     deg_common_RPF_UTR_sig_01 <-
intersect(rpfutr_WD_sig_01, (intersect(rpfutr_DC_sig_01,
rpfutr_WC_sig_01)))

     deg_common_RNA_sig_005 <- intersect(rna_WD_sig_005,
(intersect(rna_DC_sig_005, rna_WC_sig_005)))
     deg_common_RPF_CDS_sig_005 <-
intersect(rpfcds_WD_sig_005, (intersect(rpfcds_DC_sig_005,
rpfcds_WC_sig_005)))
     deg_common_RPF_UTR_sig_005 <-
intersect(rpfutr_WD_sig_005, (intersect(rpfutr_DC_sig_005,
rpfutr_WC_sig_005)))

deg_common_layers <- list(RNA_sig_01 =
deg_common_RNA_sig_01, RNA_sig_005 = deg_common_RNA_sig_005,
```

```
RPF_CDS_sig_01 = deg_common_RPF_CDS_sig_01, RPF_CDS_sig_005
= deg_common_RPF_CDS_sig_005, RPF_UTR_sig_01 =
deg_common_RPF_UTR_sig_01, RPF_UTR_sig_005 =
deg_common_RPF_UTR_sig_005)

#SIG GENES: ALL INTERSECTIONS

#DEGs common to all layers
     deg_common_all_01 <-
intersect(intersect(deg_common_RNA_sig_01,
deg_common_RPF_CDS_sig_01), deg_common_RPF_UTR_sig_01)
     deg_common_all_005 <-
intersect(intersect(deg_common_RNA_sig_005,
deg_common_RPF_CDS_sig_005), deg_common_RPF_UTR_sig_005)

#DEGs common between RNA and CDS RPF
     deg_common_RNA_RPF_CDS_01 <-
intersect(deg_common_RNA_sig_01, deg_common_RPF_CDS_sig_01)
     deg_common_RNA_RPF_CDS_005 <-
intersect(deg_common_RNA_sig_005,
deg_common_RPF_CDS_sig_005)

#DEGs common between CDS RPF and UTR RPF
     deg_common_RPF_CDS_UTR_01 <-
intersect(deg_common_RPF_CDS_sig_01,
deg_common_RPF_UTR_sig_01)
     deg_common_RPF_CDS_UTR_005 <-
intersect(deg_common_RPF_CDS_sig_005,
deg_common_RPF_UTR_sig_005)


#SIG GENES: 1st UNION

     deg_union_RNA_sig_01 <- union(rna_WD_sig_01,
(union(rna_DC_sig_01, rna_WC_sig_01)))
     deg_union_RPF_CDS_sig_01 <- union(rpfcds_WD_sig_01,
(union(rpfcds_DC_sig_01, rpfcds_WC_sig_01)))
     deg_union_RPF_UTR_sig_01 <- union(rpfutr_WD_sig_01,
(union(rpfutr_DC_sig_01, rpfutr_WC_sig_01)))

     deg_union_RNA_sig_005 <- union(rna_WD_sig_005,
(union(rna_DC_sig_005, rna_WC_sig_005)))
     deg_union_RPF_CDS_sig_005 <- union(rpfcds_WD_sig_005,
(union(rpfcds_DC_sig_005, rpfcds_WC_sig_005)))
     deg_union_RPF_UTR_sig_005 <- union(rpfutr_WD_sig_005,
(union(rpfutr_DC_sig_005, rpfutr_WC_sig_005)))


#SIG GENES: ALL UNIONS

#DEGs union to all layers
     deg_union_all_01 <- union(union(deg_union_RNA_sig_01,
deg_union_RPF_CDS_sig_01), deg_union_RPF_UTR_sig_01)
     deg_union_all_005 <- union(union(deg_union_RNA_sig_005,
deg_union_RPF_CDS_sig_005), deg_union_RPF_UTR_sig_005)
```

```
#DEGs union between RNA and CDS RPF
     deg_union_RNA_RPF_CDS_01 <- union(deg_union_RNA_sig_01,
deg_union_RPF_CDS_sig_01)
     deg_union_RNA_RPF_CDS_005 <-
union(deg_union_RNA_sig_005, deg_union_RPF_CDS_sig_005)

#DEGs union between CDS RPF and UTR RPF
     deg_union_RPF_CDS_UTR_01 <-
union(deg_union_RPF_CDS_sig_01, deg_union_RPF_UTR_sig_01)
     deg_union_RPF_CDS_UTR_005 <-
union(deg_union_RPF_CDS_sig_005, deg_union_RPF_UTR_sig_005)

#SIG GENES: EXCLUSIONS OF INTERSECTION

     deg_union_RNA_ex_CDS_01 <-
setdiff(deg_union_RNA_sig_01, deg_union_RPF_CDS_sig_01)
     deg_union_CDS_ex_RNA_01 <-
setdiff(deg_union_RPF_CDS_sig_01, deg_union_RNA_sig_01)
     deg_union_CDS_ex_UTR_01 <-
setdiff(deg_union_RPF_CDS_sig_01, deg_union_RPF_UTR_sig_01)
     deg_union_UTR_ex_CDS_01 <-
setdiff(deg_union_RPF_UTR_sig_01, deg_union_RPF_CDS_sig_01)

     deg_union_RNA_ex_CDS_005 <-
setdiff(deg_union_RNA_sig_005, deg_union_RPF_CDS_sig_005)
     deg_union_CDS_ex_RNA_005 <-
setdiff(deg_union_RPF_CDS_sig_005, deg_union_RNA_sig_005)
     deg_union_CDS_ex_UTR_005 <-
setdiff(deg_union_RPF_CDS_sig_005,
deg_union_RPF_UTR_sig_005)
     deg_union_UTR_ex_CDS_005 <-
setdiff(deg_union_RPF_UTR_sig_005,
deg_union_RPF_CDS_sig_005)


#DATAFRAME GENERATION

     clin_deg_common_bylayer <- list(RNA_sig_01 =
deg_common_RNA_sig_01, RNA_sig_005 = deg_common_RNA_sig_005,
RPF_CDS_sig_01 = deg_common_RPF_CDS_sig_01, RPF_CDS_sig_005
= deg_common_RPF_CDS_sig_005, RPF_UTR_sig_01 =
deg_common_RPF_UTR_sig_01, RPF_UTR_sig_005 =
deg_common_RPF_UTR_sig_005)
     clin_deg_union_bylayer <- list(RNA_sig_01 =
deg_union_RNA_sig_01, RNA_sig_005 = deg_union_RNA_sig_005,
RPF_CDS_sig_01 = deg_union_RPF_CDS_sig_01, RPF_CDS_sig_005 =
deg_union_RPF_CDS_sig_005, RPF_UTR_sig_01 =
deg_union_RPF_UTR_sig_01, RPF_UTR_sig_005 =
deg_union_RPF_UTR_sig_005)

     clin_deg_common_acrosslayer <- list(all_common_01 =
deg_common_all_01,  all_common_005 = deg_common_all_005,
RNA_RPF_common_01 = deg_common_RNA_RPF_CDS_01,
RNA_RPF_common_005 = deg_common_RNA_RPF_CDS_005,
RPF_CDS_UTR_common_01 = deg_common_RPF_CDS_UTR_01,
RPF_CDS_UTR_common_005 = deg_common_RPF_CDS_UTR_005)
```

```
     clin_deg_union_acrosslayer <- list(all_union_01 =
deg_union_all_01,  all_union_005 = deg_union_all_005,
RNA_RPF_union_01 = deg_union_RNA_RPF_CDS_01,
RNA_RPF_union_005 = deg_union_RNA_RPF_CDS_005,
RPF_CDS_UTR_union_01 = deg_union_RPF_CDS_UTR_01,
RPF_CDS_UTR_union_005 = deg_union_RPF_CDS_UTR_005)

     clin_deg_excluded_bylayer <- list(union_RNA_ex_CDS_01 =
deg_union_RNA_ex_CDS_01,union_RNA_ex_CDS_005 =
deg_union_RNA_ex_CDS_005, union_CDS_ex_RNA_01 =
deg_union_CDS_ex_RNA_01, union_CDS_ex_RNA_005 =
deg_union_CDS_ex_RNA_005,  union_CDS_ex_UTR_01 =
deg_union_CDS_ex_UTR_01,union_CDS_ex_UTR_005 =
deg_union_CDS_ex_UTR_005, union_UTR_ex_CDS_01 =
deg_union_UTR_ex_CDS_01, union_UTR_ex_CDS_005 =
deg_union_UTR_ex_CDS_005)

     clin_deg_comparisons <- list(deg_WC=deg_WC,
deg_WD=deg_WD, deg_DC=deg_DC)

     table_out <- list(clin_deg_common_bylayer =
clin_deg_common_bylayer, clin_deg_union_bylayer =
clin_deg_union_bylayer, clin_deg_common_acrosslayer =
clin_deg_common_acrosslayer, clin_deg_union_acrosslayer =
clin_deg_union_acrosslayer, clin_deg_excluded_bylayer =
clin_deg_excluded_bylayer, clin_deg_comparisons =
clin_deg_comparisons, degs_sig_dataframes =
degs_sig_dataframes)

return(table_out)

}

# example call with data provided
# clin_degs <- degClinical(rna.total$rna.total.count.agg,
rpf.aggregated$cds.agg, rpf.aggregated$utr5.agg, filter=T,
atyp="exclude", nf.rna=rna.nf, nf.rpf=rpf.nf)
```

Slope computation and plotting

```
# SMA and return slope
smab <- function(y,x,robust=F, ...){
         y <- as.numeric(y)
         x <- as.numeric(x)
         t <- try(sma(y~x, robust=robust,
slope.test=0),silent=T)
         if("try-error" %in% class(t)) return(NA)
         return(as.numeric(unlist(t$coef[2]))
    }

# SMA and return p-value

    smap <- function(y,x,robust=F, ...){
         y <- as.numeric(y)
         x <- as.numeric(x)
```

```
        t <- try(sma(y~x, robust=robust,
slope.test=0),silent=T)
        if("try-error" %in% class(t)) return(NA)
        return(as.numeric(t$p))
    }

# generate protein SMA slope dataframe

proteinSMAslope <- function(o, levels, robust=F, ...){
require(smatr)
slopes <- apply(log(o+1), x=log(levels+1), robust=robust,
MARGIN=1, FUN=smab)
slopes.pval <- apply(log(o+1),x=log(levels+1),
robust=robust, method = method, MARGIN=1, FUN=smap)
res <- as.data.frame(cbind(slopes, slopes.pval),
row.names=row.names(o))
res$FDR <- p.adjust(res[,2], method="fdr")
res$gene <- sapply(row.names(res),FUN=function(x){
     x <-
unique(as.character(uniprotconv[which(uniprotconv[,1]==x),2]
))
     x <- x[which(!is.na(x) & x!="")]
     paste(x,collapse=", ")
})
colnames(res)[c(1,2)] <- c("slope", "p.value")
return(res)
}

# example call: res <- proteinMechDegs(prot.ok,
eif4h.prot.levels$int, robust=F, method="SMA")

# aggregate RNA/RPF

aggregateRna <- function(o){

     ag <- aggregate(o, by=list(gene=conv[row.names(o),1]),
FUN=sum)
          row.names(ag) <- ag$gene
          ag$gene <- NULL

     return(ag)
}

aggregateRPF <- function(o){

          cds <- o[grep("utr", row.names(o), invert=T),]

          utr3 <- o[grep("utr3",row.names(o)),]
          utr5 <- o[grep("utr5",row.names(o)),]

          rownames(utr5) <- gsub(".utr5", "",
row.names(utr5), fixed=T)
          rownames(utr3) <- gsub(".utr3", "",
row.names(utr3), fixed=T)

          cds.agg <- aggregateRna(cds)
```

```
          utr5.agg <- aggregateRna(utr5)
          utr3.agg <- aggregateRna(utr3)

          ag <- list(cds.agg = cds.agg, utr5.agg = utr5.agg,
utr3.agg = utr3.agg)

     return(ag)
}

# get norm. factors for slopped transcripts
getNF <- function(o){
     require(edgeR)
          #elements <- aggregateRPF(o)

     dds <- DGEList(rbind(o$cds.agg, o$utr3.agg,
o$utr5.agg))

     dds <- calcNormFactors(dds)

     nf <- as.data.frame(dds$samples$norm.factors)
     row.names(nf) <- colnames(o$cds.agg)

return(nf)
}

# get norm. factors for RNA

getRNANF <- function(o){
require(edgeR)
dds <- DGEList(o)
dds <- calcNormFactors(dds)

     nf <- as.data.frame(dds$samples$norm.factors)
     row.names(nf) <- colnames(o)

return(nf)
}

filterTables <- function(o, thresh){
               f <- o[apply(o, MARGIN = 1, function(x) all(x
> thresh)), ]
               return(f)}

# get log(FC) for RPF

rpfLogFC <- function(o, nf=rpf.nf, filter=F, mincounts=50,
minsamples=3){
     require(edgeR)

          if(filter == T){
               o <- o[which(apply(o,1,FUN=function(x){
sum(x>mincounts)>minsamples })),]
          }

          dds <- DGEList(o, norm.factors=nf)
```

```
              d <- data.frame(row.names=colnames(o),
condition=c("WBS", "WBS", "WBS","AtWBS", "CTRL.4Hsh",
"CTRL.4Hsh", "CTRL.Baz1B.sh", "SCR", "CTRL", "CTRL", "CTRL",
"DUP", "DUP", "DUP", "DUP"))
              d$EIF4H.CN <- c(1, 1, 1, 1, 1, 1, 2, 2, 2, 2,
2, 3, 3, 3, 3)
              d$EIF4H.prot.levels <- c(0.7146365,
0.7046330, 0.9349140, 1.0790609, 0.6714633, 0.3808189,
0.8121058, 1.5301023, 1.5301023, 0.7957576, 1.3108425,
1.5211638, 1.6774462, 1.6282297, 1.4295153)
              d$EIF4H.rpf.levels <- c(1, 1, 1, 1, 0.2, 0.2,
2, 2, 2, 2, 2, 3, 3, 3, 3)
              d$EIF4H.rna.levels <- c(0.6566165, 0.8570071,
0.7647101, 0.8001617, 0.2467174, 0.1675075, 1.9995914,
1.3707070, 1.5250787, 0.8444991, 1.3021972, 1.9022625,
1.8728631, 2.2141652, 1.9380043)

          mm <- model.matrix(~EIF4H.prot.levels, data=d)

          dds <- estimateDisp(dds, mm)

          ddsfit <- glmFit(dds, mm)

          ddsLRT <- glmLRT(ddsfit, "EIF4H.prot.levels")

          rpfLRT <- as.data.frame(topTags(ddsLRT, 30000))

          rpflogfc <- as.data.frame(rpfLRT$logFC)
          row.names(rpflogfc) <- row.names(rpfLRT)
          rpflogfc$p.value <- rpfLRT$PVal
          rpflogfc$gene <- row.names(rpflogfc)
          rpflogfc$FDR <- p.adjust(rpflogfc$p.value,
method="fdr")
          colnames(rpflogfc)[c(1,2)] <- c("slope",
"p.value")

     return(rpflogfc)
}

# get log(FC) for RNA

rnaLogFC <- function(o, nf=rna.nf, filter=F, mincounts=50,
minsamples=3){

     require(edgeR)

          if(filter == T){
               o <- o[which(apply(o,1,FUN=function(x){
sum(x>mincounts)>minsamples })),]
          }
          dds <- DGEList(o, norm.factors=nf)
          d <- data.frame(row.names=colnames(o),
condition=c("WBS", "WBS", "WBS","AtWBS", "CTRL.4Hsh",
"CTRL.4Hsh", "CTRL.Baz1B.sh", "SCR", "CTRL", "CTRL", "CTRL",
"DUP", "DUP", "DUP", "DUP"))
```

```
               d$EIF4H.CN <- c(1, 1, 1, 1, 1, 1, 2, 2, 2, 2,
2, 3, 3, 3, 3)
               d$EIF4H.prot.levels <- c(0.7146365,
0.7046330, 0.9349140, 1.0790609, 0.6714633, 0.3808189,
0.8121058, 1.5301023, 1.5301023, 0.7957576, 1.3108425,
1.5211638, 1.6774462, 1.6282297, 1.4295153)
               d$EIF4H.rna.levels <- c(0.6566165, 0.8570071,
0.7647101, 0.8001617, 0.2467174, 0.1675075, 1.9995914,
1.3707070, 1.5250787, 0.8444991, 1.3021972, 1.9022625,
1.8728631, 2.2141652, 1.9380043)

          mm <- model.matrix(~EIF4H.prot.levels, data=d)

          dds <- estimateDisp(dds, mm)

          ddsfit <- glmFit(dds, mm)

          ddsLRT <- glmLRT(ddsfit, "EIF4H.prot.levels")
          rnaLRT <- as.data.frame(topTags(ddsLRT, 30000))

          rnalogfc <- as.data.frame(rnaLRT$logFC)
          row.names(rnalogfc) <- row.names(rnaLRT)
          rnalogfc$p.value <- rnaLRT$PVal
          rnalogfc$gene <- row.names(rnalogfc)
          rnalogfc$FDR <- p.adjust(rnalogfc$p.value,
method="fdr")
          colnames(rnalogfc)[c(1,2)] <- c("slope",
"p.value")

     return(rnalogfc)
}


# Plot genes in quadrants
plotCompSlopes <- function(o1, o2, alpha, lim=2, sig="both",
text=T, axes=T, fillzero=F, subset=F, ...){

     in_common <- intersect(o1$gene, o2$gene)

     if(subset == T){
     o1 <- o1[intersect(rownames(o1),alldegs),]
     o2 <- o2[intersect(rownames(o2),alldegs),]}


          switch(sig,
        both = {in_pval <-
intersect(row.names(o1[which(o1$p.value < alpha),]),
row.names(o2[which(o2$p.val < alpha),])); col="red"},
        first = {in_pval <- row.names(o1[which(o1$p.value <
alpha),]); col="green"},
        second = {in_pval <- row.names(o2[which(o2$p.value <
alpha),]); col="green"},
        union = {in_pval <-
union(row.names(o1[which(o1$p.value < alpha),]),
row.names(o2[which(o2$p.value < alpha),])); col="blue"},
```

```
        onlyfirst = {in_pval <-
setdiff(row.names(o1[which(o1$p.value < alpha),]),
row.names(o2[which(o2$p.value < alpha),])); col="orange"},
        onlysecond = {in_pval <-
setdiff(row.names(o2[which(o2$p.value < alpha),]),
row.names(o1[which(o1$p.value < alpha),])); col="purple"},
          stop(paste(sig,": unknown selection",sep=""))
          )

    if(fillzero == T){
     o1$slope[which(o1$p.value >= alpha)] <- 0
     o2$slope[which(o2$p.value >= alpha)] <- 0
     }

     plot(o1[in_pval,1], o2[in_pval,1], xlim=c(-lim,+lim),
ylim=c(-lim,+lim), col=col, pch=16, cex=0.6)
     abline(h=0,v=0, col="black", lwd=0.4, lty="dashed")
     text(y=lim, x=lim-lim/5, label="1st quadrant")
     text(y=-lim, x=lim-lim/5, label="2nd quadrant")
     text(y=-lim, x=-(lim-lim/5), label="3rd quadrant")
     text(y=lim, x=-(lim-lim/5), label="4th quadrant")

     if(text == T){
          text(o1[in_pval,1], o2[in_pval,1],
labels=o1[in_pval,3], cex=0.5, pos=3)
                    }
     if(axes == T){
          abline(0,1, col="red", lwd=0.7, lty="dotted")
          abline(0,-1, col="red", lwd=0.7, lty="dotted")
          }


}

# Generate list of genes divided by quadrants

listCompSlopes <- function(o1, o2, alpha, sig="both",
subset=F){

          require(GTscripts)

          name.o1 <- deparse(substitute(o1))
          name.o2 <- deparse(substitute(o2))

          in_common <- intersect(o1$gene, o2$gene)
          o1 <- o1[in_common,]
          o2 <- o2[in_common,]

          if(subset ==T){
          o1 <- o1[intersect(rownames(o1),alldegs),]
          o2 <- o2[intersect(rownames(o2),alldegs),]}

       switch(sig,
          both = {in_pval <-
intersect(row.names(o1[which(o1$p.value < alpha),]),
row.names(o2[which(o2$p.val < alpha),]))},
```

```
          first = {in_pval <- row.names(o1[which(o1$p.value
< alpha),])},
                second = {in_pval <-
row.names(o2[which(o2$p.value < alpha),])},
                union = {in_pval <-
union(row.names(o1[which(o1$p.value < alpha),]),
row.names(o2[which(o2$p.value < alpha),]))},
          onlyfirst = {in_pval <-
setdiff(row.names(o1[which(o1$p.value < alpha),]),
row.names(o2[which(o2$p.value < alpha),]))},
             onlysecond = {in_pval <-
setdiff(row.names(o2[which(o2$p.value < alpha),]),
row.names(o1[which(o1$p.value < alpha),]))},
          stop(paste(sig,": unknown selection",sep=""))
          )

        o1 <- o1[in_pval,]
        o2 <- o2[in_pval,]
        res <- as.data.frame(cbind(o1, o2))
     res <- res[,-c(3,4,7,8)]


     first_quadrant_genes <-
intersect(row.names(o1[which(o1$slope > 0),]),
row.names(o2[which(o2$slope > 0),]))
          second_quadrant_genes <-
intersect(row.names(o1[which(o1$slope > 0),]),
row.names(o2[which(o2$slope < 0),]))
          third_quadrant_genes <-
intersect(row.names(o1[which(o1$slope < 0),]),
row.names(o2[which(o2$slope < 0),]))
          fourth_quadrant_genes <-
intersect(row.names(o1[which(o1$slope < 0),]),
row.names(o2[which(o2$slope > 0),]))

          colnames(res) <- c(paste(name.o1,".slope",
sep=""), paste(name.o1,".p.value", sep=""), paste(name.o2,
".slope", sep=""), paste(name.o2, ".p.value", sep=""))

          first_quadrant <-
as.data.frame(res[first_quadrant_genes,])
          second_quadrant <-
as.data.frame(res[second_quadrant_genes,])
          third_quadrant <-
as.data.frame(res[third_quadrant_genes,])
          fourth_quadrant <-
as.data.frame(res[fourth_quadrant_genes,])

     quadrants <- list(first_quadrant=first_quadrant,
second_quadrant=second_quadrant,
third_quadrant=third_quadrant,
fourth_quadrant=fourth_quadrant)

     return(quadrants)
}
```

```
# Make a slopematrix using slope data frames

buildMatrix <- function(rna, rpf, prot, alpha=0.05,
stringent=F){

     in_common <- intersect(intersect(row.names(rna),
row.names(rpf)), row.names(prot))

     rna <- rna[in_common,]
     rpf <- rpf[in_common,]
     prot <- prot[in_common,]

     rna2 <- rna[which(rna$p.value < alpha),]
     rpf2 <- rpf[which(rpf$p.value < alpha),]
     prot2 <- prot[which(prot$p.value < alpha),]

     if(stringent == T){
     in_common_2 <- intersect(intersect(rownames(rna2),
rownames(rpf2)), rownames(prot2))
     }else{
     in_common_2 <- union(union(rownames(rna2),
rownames(rpf2)), rownames(prot2))}


     rna <- rna[in_common_2,]
     rpf <- rpf[in_common_2,]
     prot <- prot[in_common_2,]

     res <- cbind(rna[,c(1,2)], rpf[,c(1,2)], prot[,c(1,2)])

     matrix <- res[,c(1,3,5)]
     colnames(matrix) <- c("RNA", "RPF", "PROT")

     return(matrix)

}
```

Gene expression modeling with degadation parameters

```
# Formats and fetches data to be used in all modeling steps.
# Allows to choose the minimum number of peptides (x)

getData <- function(o,x){

     o$protein$nbp <-
sapply(o$protein$peptides,FUN=function(x){
length(strsplit(as.character(x),", ",fixed=T)[[1]])})
     protein_cut <- o$protein[which(o$protein$nbp >= x),]
     colnames(protein_cut)[3:(2+nrow(o$design))] <-
as.character(o$design$sample)
     row.names(protein_cut) <- protein_cut$protein

     rows_used <- intersect(rownames(o$kloss),
rownames(o$RPF))
     rows_used <- intersect(rows_used,
rownames(protein_cut))
```

```
     samples <- intersect(colnames(o$RPF),colnames(o$kloss))

     prot_gf <- protein_cut[rows_used, samples]
     rpf_gf <- o$RPF[rownames(prot_gf),samples]
     pk_gf <- o$kloss[rownames(prot_gf),samples]
     rna_gf <- o$RNA[rownames(prot_gf),samples]
     elements <- list(prot_gf=prot_gf, rpf_gf=rpf_gf,
pk_gf=pk_gf, rna_gf=rna_gf)
     return(elements)
}


# Plots estimatedP vs measuredP with their linear model and
# G value by calling recursively estimateGlobalFactor with
# doplot=T

plotGlobalFactor <- function(o, x, doplot=T, kl=T, rnacor=F,
ylim=c(0,200000)){

          layout(matrix(1:6, nrow=2))
          samples <- c("306O", "316M", "BU1CRE", "MIFF3",
"CFG", "242J")


               elements <- getData(o=o, x=x)
                    for(j in 1:length(samples)){

     estimateGlobalFactor(elements$rpf_gf[,j],
elements$prot_gf[,j], elements$pk_gf[,j],
elements$rna_gf[,j], main=paste("",samples[j]), doplot=T,
kl=kl, ylim=ylim)
                                                        }

}

# Estimates the global factor with the possibility to choose
# whether to plot and whether to use K_loss in the model.

estimateGlobalFactor <- function(rpf, prot, kloss, rna,
rnacor=F, doplot=T, main="", ylim=c(0,200000), kl=TRUE,
protwisefactors=NULL){
     if(is.matrix(rpf) | is.data.frame(rpf)) stop("Your
input is a matrix or dataframe! You should work on one
sample at a time.")
     if(length(unique(c(length(rpf),length(prot),length(klos
s))))!=1) stop("Each vector (rpf, prot, kloss) should
contain the values for the same proteins, in the same
order.")
     library(MASS)
     library(LSD)
     if(kl == TRUE){
     estimatedP = (log2(rpf+0.1))/kloss}
     else{
     estimatedP = log2(rpf+0.1)}
     if(rnacor == TRUE){
     estimatedP = (log2(rna+0.1))}
```

```
     else{
     estimatedP = log2(rpf+0.1)}
          if(!is.null(protwisefactors)) estimatedP <-
estimatedP*protwisefactors
     if(is.null(ylim)) ylim <- c(0, max(prot))
     correlation <- cor(log2(prot+0.1), estimatedP,
use="pairwise", method="spearman")
     if(doplot) heatscatter(estimatedP,prot,ylab=c("Measured
P"),xlab=c("Estimated P"),main=main, ylim=ylim)
     mod <- try(rlm(prot~estimatedP+0),silent=T)
     if("try-error" %in% class(mod)){
          warning("Could not fit robust model... trying
normal lm")
          mod <- try(lm(prot~estimatedP+0),silent=T)
          if("try-error" %in% class(mod))    stop("Could not
fit model")
     }
     if(doplot) abline(mod,lwd=2, col="blue", lty="dashed")
     #modelmad <- median(abs(mod$residuals))
     if(doplot) legend("topleft", bty="n",
legend=paste(c("G:","MSR:","Cor
(spearman):"),c(round(coefficients(mod)[[1]]),round(median(m
od$residuals^2)),round(correlation, digits=3))))

     gvalue <- coefficients(mod)[[1]]
     msr <- round(median(mod$residuals^2))
     mre <- median(abs(mod$residuals)/mod$fitted.values)
     meansr <- round(mean(mod$residuals^2))
     features <- list(gvalue=gvalue, msr=msr, meansr=meansr,
correlation=correlation, mre=mre)


     return(features)
}

# Gets the features of a model: global factor, median
# squared residuals, median relative error, spearman
# correlation, mean squared residuals.

getModelFeatures <- function(o, x, kl=T,
protwisefactors=NULL, doplot=F){

     require(reshape)

     samples <- c("306O", "316M", "BU1CRE", "MIFF3", "CFG",
"242J")
     gf <- matrix(0, ncol=6, nrow=x,
dimnames=list(1:x,samples))
     msr <- matrix(0, ncol=6, nrow=x,
dimnames=list(1:x,samples))
     cors <- matrix(0, ncol=6, nrow=x,
dimnames=list(1:x,samples))
     mre <- matrix(0, ncol=6, nrow=x,
dimnames=list(1:x,samples))
     meansr <- matrix(0, ncol=6, nrow=x,
dimnames=list(1:x,samples))
```

```
     for(i in 1:x){
          elements <- getData(o=o, x=i)
          if(is.null(protwisefactors)){
               pwf <- NULL
          }else{
               pwf <-
protwisefactors[row.names(elements$prot_gf)]
          }
          for(j in 1:length(samples)){
               features <-
estimateGlobalFactor(elements$rpf_gf[,j],
elements$prot_gf[,j], elements$pk_gf[,j], doplot=doplot,
kl=kl, protwisefactors=pwf)
               gf[i,j] <- features$gvalue
               msr[i,j] <- features$msr
               cors[i,j] <- features$correlation
               mre[i,j] <- features$mre
               meansr[i,j] <- features$meansr
          }
     }

     list(gf = gf, msr = msr, cors = cors, mre = mre, meansr
= meansr, name=deparse(substitute(o)), kl=kl)
}


# Plots values for features of the model, i.e. correlation,
# median relative error, mean squared residuals along an
# increasing number of minimum peptide thresholds

plot_Cor <- function(o, x, method="correlation", ylim=NULL){

          fa <- switch(method,
               gvalue=o$gf[x,],
               correlation=o$cors[x,],
               mre=o$mre[x,],
               msr=o$msr[x,],
               stop("unknown method!")
               )


          cols <- c("red", "red", "gray", "gray", "blue",
"blue")

          plot(x=rep(x,ncol(fa)), y=as.numeric(fa),
col=rep(cols, each=nrow(fa)), pch=16, ylim=ylim,
ylab=paste(method), xlab="Peptide threshold",
main=paste(method, "for", deparse(substitute(o))))
}


# Plots values for features of two models in the same plot
# for comparisons along an increasing number of minimum
# peptide thresholds
```

```
compare_features <- function(o1, o2, x,
method="correlation"){

     fa <- switch(method,
               gvalue=o1$gf[x,],
               correlation=o1$cors[x,],
               mre=o1$mre[x,],
               msr=o1$msr[x,],
               meansr=o1$meansr[x,],
               stop("unknown method!")
               )

     fa2 <- switch(method,
               gvalue=o2$gf[x,],
               correlation=o2$cors[x,],
               mre=o2$mre[x,],
               msr=o2$msr[x,],
               meansr=o2$meansr[x,],
               stop("unknown method!")
               )

          fam <- melt(fa)
          fam$thresh <- x
          fa2m <- melt (fa2)
          fa2m$thresh <- x

          plot <- ggplot() +
     geom_dotplot(data=fam, aes(x=factor(thresh), y=value),
binaxis = "y", stackdir = "center", binpositions="all",
colour = "NA", fill="blue", binwidth = 0.015*(max(fa)-
min(fa))) +
     geom_smooth(data=data.frame(x=rep(1:nrow(fa),6),y=as.nu
meric(fa)), aes(x=x,y=y), fill="blue", color="blue")   +
     geom_dotplot(data=fa2m, aes(x=factor(thresh), y=value),
binaxis = "y", stackdir = "center", binpositions="all",
colour = "NA", fill="red", binwidth = 0.015*(max(fa)-
min(fa))) +
     geom_smooth(data=data.frame(x=rep(1:nrow(fa2),6),y=as.n
umeric(fa2)), aes(x=x,y=y), fill="red", color="red")   +
     ggtitle(paste(method, "for minimum 1 to",x[length(x)],
"peptides in",o1$name,"and", o2$name))

     return(plot)

}
```

Acknowledgments

This project has been made possible thanks to the support, guidance and mentorship of my supervisor, Giuseppe Testa. I am really grateful to him also for creating a vibrant intellectual environment that never made me regret my choice to do a PhD.

I also want to thank my external supervisor, Claudia Bagni, and my internal supervisor, Pier Paolo Di Fiore, for their advice and engaging scientific conversations.

Yansheng Liu and Ruedi Aebersold at ETH Zürich were instrumental in processing and interpreting all the proteomic data and shaping the project together with us.

A heartfelt thanks goes to all the facilities in the campus that make our life easier: the cell culture facility, the sequencing facility, the imaging facility and the warehouse.

I also want to acknowledge our funding agencies, who made this work possible: Fondazione Umberto Veronesi, Telethon, the European Research Council, EPIGEN, Regione Lombardia and Fondation Jerome Lejeune.

I need to reserve a special thanks to people in the lab, past and present.
Philip K. Dick, in the afterword of *A Scanner Darkly*, described his friends in a simple but beautiful way:

> *These were comrades whom I had; there are no better.*

First of all, I need to thank Marija and Pierre-Luc for their unbelievable support and the dedication they showed to making this project solid, interesting and innovative. Also for their help in times of dire need and for every kind word or gesture I have received from them. I owe them a large part of my scientific and personal growth.

I also want to thank my colleagues, inside and outisde of the lab, past and present members of the campus, with whom I've shared great moments: Giulia, Matteo, Jacopo, Michele, Luca C, Mattia, Mariaelena, Pietro, Virginia, Federico, James, Alessandro, Luca M, Alejandro, Agnese, Berta, Maddalena, Nicolò, Sina, Elena, Patricio, Veronica, Emanuele, Giulia F., Italia, Danila, and I'm probably forgetting someone, so I'll just play it cheap and say: you know who you are, and I won't forget you.

My achievements would not have been even thinkable without the support of my parents, my friends from before, and the person I'm proud to share my life with, who somehow managed to keep me afloat, give me new ideas, and help me in every possible situation.