Establishing a transcriptome-based drug discovery paradigm for neurodevelopmental disorders

Ryan S. Dhindsa[1,2*], Anthony W. Zoghbi[1,3], Daniel K. Krizay[1,4], Chirag Vasavda[5], David B. Goldstein[1,4*]

[1] Institute for Genomic Medicine, Columbia University Irving Medical Center, New York, USA

[2] Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University Irving Medical Center, New York, NY 10032, USA

[3] Department of Psychiatry, Columbia University Irving Medical Center, New York, USA; New York State Psychiatric Institute, New York, USA

[4] Department of Genetics & Development, Columbia University Irving Medical Center, New York, USA

[5] The Solomon H. Snyder Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

*Correspondence: rsd2135@cumc.columbia.edu (R.S.D.), dg2875@cumc.columbia.edu (D.B.G.)

Abstract

Advances in genetic discoveries have created substantial opportunities for precision medicine in neurodevelopmental disorders. Many of the genes implicated in these diseases encode proteins that regulate gene expression, such as chromatin associated proteins, transcription factors, and RNA-binding proteins. The identification of targeted therapeutics for individuals carrying mutations in these genes remains a challenge, as the encoded proteins can theoretically regulate thousands of downstream targets in a considerable number of cell types. Here, we propose the application of a drug discovery approach called "transcriptome reversal" for these disorders. This approach, originally developed for cancer, attempts to identify compounds that reverse gene-expression signatures associated with disease states.

Introduction

The introduction of whole-exome sequencing has led to genetic discoveries that have dramatically improved our understanding of mechanisms underlying neurodevelopmental disorders[1]. Some of these discoveries have translated to the development of targeted treatments[2], including enzyme replacement therapies in deficiency disorders[3], antisense oligonucleotide therapies for spinal muscle atrophy[4,5], and channel modulators for gain-of-function channelopathies[6]. Despite these important achievements, the development of precision therapeutics has lagged far behind the rapid progress in gene discovery.

One of the major challenges facing precision medicine in neurodevelopmental disorders—including autism spectrum disorder (ASD), developmental epileptic encephalopathy (EE), developmental delay with cognitive manifestations (DD), and schizophrenia—is the tremendous genetic heterogeneity underlying these conditions. Implicated genes have revealed a wide range of etiologies for each of these diseases, such as abnormal synaptic transmission, chromatin remodeling, transcription regulation, and ion channel function[7-11]. It is likely that this heterogeneity will require different drug discovery paradigms for each functional class of genes.

We argue that neurodevelopmental disorder associated genes that directly influence the transcriptome—namely chromatin modifiers, transcription factors and co-factors, and RNA-binding proteins—represent one particular group of genes amenable to drug discovery efforts that are well-developed in cancer but largely unexplored in neurology. This strategy, termed transcriptome reversal, posits that if gene expression changes underlie the pathophysiology of a particular disease, then correcting this transcriptomic signature toward a normal state may have therapeutic potential[12,13].

Here, we outline a strategy for the development of a systematic program for transcriptome-guided drug discovery in neurodevelopmental disease. We first provide an overview of the transcriptomic signature reversal approach and its successes in other disease areas. We next discuss opportunities for analogous approaches in neurodevelopmental disorders by focusing on transcriptional regulators that have been implicated in EE, ASD, DD, and schizophrenia. Finally, we propose immediate next steps that would be required to systematically implement this drug discovery strategy.

The transcriptome reversal paradigm

In theory, many different classes of molecules can reverse transcriptomic signatures toward a healthy state, including small-interfering RNAs, antisense oligonucleotides, peptides, and classical small molecules. Here we focus on small molecules because they are especially

pharmacologically versatile and have been the focus of most transcriptome reversal efforts to date. Broadly speaking, this approach requires three steps: identification of the disease gene expression signature, *in silico* or experimental screening to prioritize compounds most likely to reverse this disease signature, and targeted experimental validation of candidate compounds (**Figure 1**).
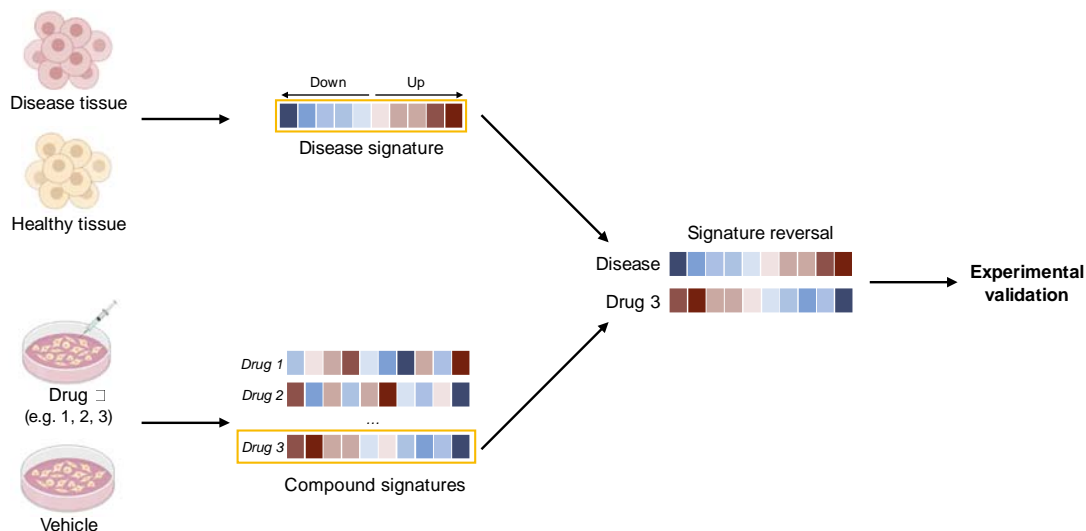

Figure 1. The transcriptome reversal approach. Gene expression profiling (e.g. RNA-seq or single-cell RNA-seq) is used to derive disease expression signatures. This signature is then compared to the signatures of cells treated with small molecules in order to identify compounds most likely to reverse disease-associated expression changes. Candidate compounds are then screened experimentally to determine whether they reverse disease phenotypes.

Generating disease gene expression signatures

Generating disease expression signatures requires the unbiased assessment of gene expression changes in tissue or cells derived from patients or a disease model. By performing RNA-sequencing on disease samples and healthy samples, researchers can then use differential gene expression analysis (DGE) to identify quantitative gene expression level changes between sample groups. This strategy reveals the sets of genes that are up- and down-regulated in the disease state compared to a normal state (i.e. the disease expression signature). In addition to being useful for transcriptome reversal, these expression signatures can reveal perturbed biological pathways and provide insight into disease mechanisms[14].

Importantly, advances in single-cell RNA-sequencing (scRNA-seq) allow interrogation of these expression changes at an unprecedented resolution. Until recently, RNA-seq was predominantly performed on bulk tissues. While these studies led to transformative discoveries, bulk RNA-sequencing methods cannot resolve specific cell types and only provide an average

expression signal for an ensemble of cells. scRNA-seq, on the other hand, can resolve cell type-specific expression signatures. The most common scRNA-seq methods involve microfluidic chips that encapsulate single cells into reagent-filled oil droplets[15]. Reverse transcription occurs within each oil droplet, and the resulting cDNAs receive barcodes that allow for the assignment of each resulting sequencing read back to its own single cell of origin. It is important to note that performing scRNA-seq necessitates particularly careful experimental design and analyzing the resulting data requires statistical rigor[14,16].

In silico screening of small molecules

Once a disease gene expression signature has been identified, the next step is to identify compounds that are predicted to reverse these expression changes. This process requires two components: (1) a database of gene expression signatures of cells or tissues treated with small molecules; and (2) a statistical method to compare the disease signature to these small molecule signatures.

There have been considerable efforts to create publicly available compendia of small molecule expression signatures. In 2006, Lamb and colleagues introduced the "Connectivity Map" (CMap), which included microarray-derived signatures for roughly 1,300 small molecules applied to human cell lines[13]. More recently, a new version of the Connectivity Map was introduced that leverages a novel gene expression profiling technique called L1000[17]. This assay measures the expression level of 978 "landmark genes" that were selected to capture a large proportion of genome-wide variation in gene expression. The measurements of these genes are then used to infer the expression of 11,350 other genes in the transcriptome. Because the L1000 assay is far more affordable than typical RNA-sequencing, CMap investigators were able to use this platform to profile roughly 20,000 small molecules in a variable number of human cell lines. This resource has not only facilitated drug discovery in cancer[18-20], but also in other non-neurological diseases too, including diabetes[21], inflammatory bowel disease[22], and osteoporosis[23] (reviewed by Keenan et al.[24]).

Investigators have introduced several statistical methods that compare disease and compound signatures to predict which compounds are most likely to invert the disease signature. For example, the CMap uses the Connectivity Score: a normalized similarity metric based on the weighted Kolmogorov-Smirnov enrichment statistic. Other methods rely on the organization of expression data into networks to infer disease- and drug-induced changes in master regulator activity[18,25]. Computational algorithms, such as OncoTreat[19], can be used to identify individual or combinations of compounds that are expected to reverse master regulator

activity in the disease state. Importantly, as discussed below, the success of this *in silico* screening step relies on the availability of cell type-specific perturbation data.

Experimental validation

Once small molecules are prioritized via *in silico* screening, the next step is to verify that the top-scoring compounds in fact reverse the disease profile. In cases where the small molecule signatures were generated in cell types not related to the disease of interest, it is critical to appreciate that the compounds identified may not be the best compounds in cell types relevant to the disease. Therefore, validation requires the administration of candidate compounds to patient cells or cells derived from the disease model followed by gene expression profiling. The most promising compounds should reverse disease-related expression changes, such that the signature more closely resembles the wildtype signature. Furthermore, candidate compounds should be administered at multiple doses and timepoints to pinpoint the biological conditions that lead to the strongest restoration of the transcriptome. When possible, these compounds should then be evaluated for their ability to rescue disease phenotypes in a relevant disease model.

Transcriptome reversal opportunities in neurodevelopmental disorders

Despite the promise of this approach, there are relatively few examples of transcriptome reversal studies in neurodevelopmental disorders. Most attempts have focused on non-genetic disease models or tissue derived from patients who lack a clear Mendelian diagnosis. For example, Readhead and colleagues recently used L1000 to identify compounds that reverse transcriptomic signatures of post-mortem tissue derived from 12 individuals with schizophrenia[26]. Other studies have identified compounds that reverse disease signatures identified in mouse models of acquired epilepsy[27,28] and in surgical tissue derived from patients with unilateral mesial temporal lobe epilepsy[29].

These studies have provided preliminary proofs of concept, but when the underlying cause of disease is unknown, it can be challenging to decipher whether the observed disease signatures reflect causal versus compensatory gene expression changes. Thus, we argue that an important path forward is to focus on genetic models in which gene expression dysregulation represents the primary mechanism of pathogenicity. Here, we define primary transcriptomic regulators as transcription factors/co-factors, chromatin-associated proteins, and RNA-binding proteins. Strikingly, these genes encompass a substantial proportion of the genes implicated in

developmental delay with cognitive impairment (13%), autism spectrum disorder (45%), developmental and epileptic encephalopathy (18%), and schizophrenia (20%) (**Figure 2**). We briefly review the functions of these gene expression regulators below and provide select examples of disease-related genes that fall into each of these classes.
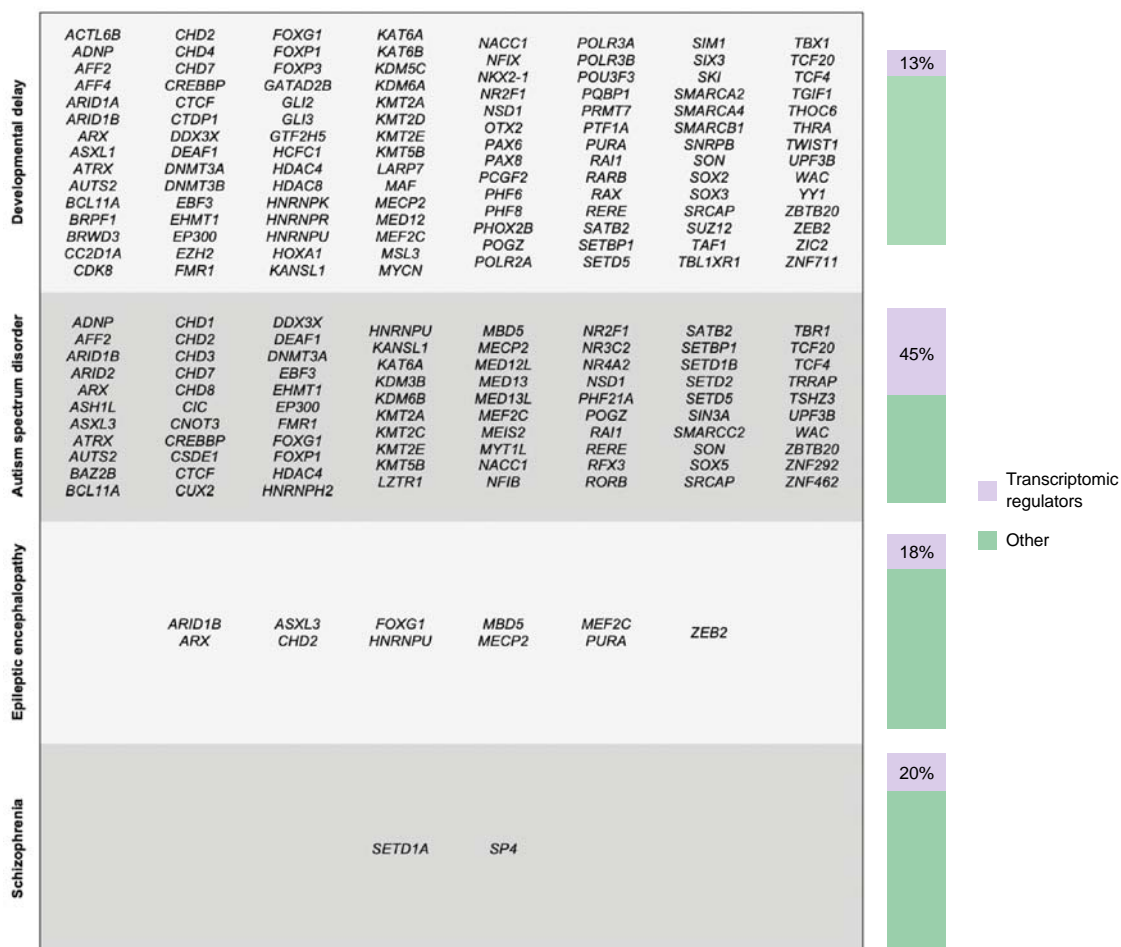
Developmental delay — 13%

ACTL6B	CHD2	FOXG1	KAT6A	NACC1	POLR3A	SIM1	TBX1
ADNP	CHD4	FOXP1	KAT6B	NFIX	POLR3B	SIX3	TCF20
AFF2	CHD7	FOXP3	KDM5C	NKX2-1	POU3F3	SKI	TCF4
AFF4	CREBBP	GATAD2B	KDM6A	NR2F1	PQBP1	SMARCA2	TGIF1
ARID1A	CTCF	GLI2	KMT2A	NSD1	PRMT7	SMARCA4	THOC6
ARID1B	CTDP1	GLI3	KMT2D	OTX2	PTF1A	SMARCB1	THRA
ARX	DDX3X	GTF2H5	KMT2E	PAX6	PURA	SNRPB	TWIST1
ASXL1	DEAF1	HCFC1	KMT5B	PAX8	RAI1	SON	UPF3B
ATRX	DNMT3A	HDAC4	LARP7	PCGF2	RARB	SOX2	WAC
AUTS2	DNMT3B	HDAC8	MAF	PHF6	RAX	SOX3	YY1
BCL11A	EBF3	HNRNPK	MECP2	PHF8	RERE	SRCAP	ZBTB20
BRPF1	EHMT1	HNRNPR	MED12	PHOX2B	SATB2	SUZ12	ZEB2
BRWD3	EP300	HNRNPU	MEF2C	POGZ	SETBP1	TAF1	ZIC2
CC2D1A	EZH2	HOXA1	MSL3	POLR2A	SETD5	TBL1XR1	ZNF711
CDK8	FMR1	KANSL1	MYCN				

Autism spectrum disorder — 45%

ADNP	CHD1	DDX3X	HNRNPU	MBD5	NR2F1	SATB2	TBR1
AFF2	CHD2	DEAF1	KANSL1	MECP2	NR3C2	SETBP1	TCF20
ARID1B	CHD3	DNMT3A	KAT6A	MED12L	NR4A2	SETD1B	TCF4
ARID2	CHD7	EBF3	KDM3B	MED13	NSD1	SETD2	TRRAP
ARX	CHD8	EHMT1	KDM6B	MED13L	PHF21A	SETD5	TSHZ3
ASH1L	CIC	EP300	KMT2A	MEF2C	POGZ	SIN3A	UPF3B
ASXL3	CNOT3	FMR1	KMT2C	MEIS2	RAI1	SMARCC2	WAC
ATRX	CREBBP	FOXG1	KMT2E	MYT1L	RERE	SON	ZBTB20
AUTS2	CSDE1	FOXP1	KMT5B	NACC1	RFX3	SOX5	ZNF292
BAZ2B	CTCF	HDAC4	LZTR1	NFIB	RORB	SRCAP	ZNF462
BCL11A	CUX2	HNRNPH2					

Epileptic encephalopathy — 18%

ARID1B	ASXL3	FOXG1	MBD5	MEF2C	ZEB2
ARX	CHD2	HNRNPU	MECP2	PURA	

Schizophrenia — 20%

SETD1A	SP4

Legend: Transcriptomic regulators / Other

Figure 2. Transcriptomic regulators implicated in neurodevelopmental disorders. Table of genes encoding transcription factors, chromatin-associated proteins, and RNA-binding proteins in developmental delay with intellectual disability, autism spectrum disorder, epileptic encephalopathy, and schizophrenia. Bar charts depict the percentage of transcriptomic regulators amongst the total number of genes implicated in each disease area.

Transcription factors

Transcription factors (TFs) bind to specific DNA motifs to recruit and regulate the transcription machinery. These proteins can be broadly separated into two classes based on their regulatory function: control of transcription initiation and control of transcription elongation[30]. This distinction is not absolute, however, as some TFs mediate both of these processes.

Furthermore, TFs can work alone or with other proteins (e.g. co-factors) in a complex to activate or repress the recruitment of RNA polymerase. Unsurprisingly, pathogenic mutations in TFs associated with neurodevelopmental disorders cause widespread transcriptomic defects in the brain.

Mutations in the transcription factor-encoding gene *MEF2C* are associated with autosomal dominant intellectual disability, autism spectrum disorder, and epilepsy[31,32]. Specifically deleting *MEF2C* in forebrain excitatory neurons in mice leads to broad changes in genes that influence neuronal specification and synapse development, leading to a dramatic increase in inhibitory synaptic transmission[33]. The FOXO (forkhead box) family of transcription factors also influences a wide network of genes important in neurodevelopment, and patients with mutations in *FOXG1* present with autism spectrum disorder[34]. Sequencing chromatin immunoprecipitation of Foxg1 identified thousands of loci in the mouse genome where this protein binds, suggesting that Foxg1 regulates hundreds of targets *in vivo*[35].

Chromatin-associated proteins

In eukaryotes, DNA resides in the nucleus where it is packaged into a highly condensed structure called chromatin. The primary unit of chromatin is the nucleosome, which consists of DNA wrapped around eight histone proteins. Chromatin remodelers, which move histone proteins to make DNA accessible to transcription machinery, have been implicated in neurodevelopmental disorders. Additionally, each of these histone proteins has a peptide tail that is post-translationally modified (e.g. acetylated or methylated) to activate or repress transcription. Mutations in chromatin regulators that create, remove, and recognize these chemical modifications can also result in neurodevelopmental disorders. DNA nucleotides can also be chemically modified themselves in ways that influence gene expression. The most well-studied chemical modification is DNA methylation, in which cytosines in the context of CpG dinucleotides receive methyl groups to regulate gene expression.

Loss-of-function mutations in the gene *MECP2* (methyl CpG binding protein 2)*,* which encodes a methyl CpG binding protein, cause Rett Syndrome[36]. Conditionally ablating *MECP2* from the mouse brain recapitulates several findings in patients with Rett Syndrome, such as uncoordinated gait and repetitive movements[37,38]. Gene expression profiling of mouse models

as well as human brain tissue has revealed widespread changes in gene expression across brain regions[39-41], consistent with *MECP2*'s role in modulating chromatin topology.

RNA-binding proteins

RNA-binding proteins play a key role in the post-transcriptional control of RNAs[42]. They regulate nearly all aspects of RNA biogenesis, including RNA splicing, polyadenylation, mRNA localization and degradation, and translation. There are hundreds of RNA-binding proteins with diverse functions, each with different RNA-sequence specificities and affinities. Many of these proteins seem to bind to thousands of transcripts per cell. Indeed, pathogenic mutations in these genes can lead to the dysregulation of hundreds or thousands of genes. For example, mutations in *HNRNPU*, an RNA- and DNA-binding protein, is associated with epileptic encephalopathy and autism spectrum disorder[7,43,44]. We recently reported that a mouse model of *HNRNPU* haploinsufficiency exhibits neuroanatomical abnormalities, global developmental delay, and increased seizure susceptibility[45]. Single-cell RNA-sequencing of mutant hippocampal and neocortical cells revealed thousands of dysregulated genes across neuronal subtypes that converged on biological pathways important to neurodevelopment[45].

Other gene expression regulators

Although we focus here on transcription factors/co-factors, chromatin-associated proteins, and RNA-binding proteins, we note that other genes implicated in neurodevelopmental disorders may also affect gene expression. Some examples include kinases, G-protein subunits, and ubiquitin ligases. Many of these proteins, however, are also directly involved in many alternate signaling pathways, making it less clear that the primary cause of disease is the associated transcriptomic dysregulation.

A systematic transcriptomic signature reversal program for neurodevelopmental disorders

The substantial contribution of transcriptomic regulators to genetic neurodevelopmental diseases highlights the enormous potential for developing a transcriptomic reversal program for these diseases. However, this undertaking will require a concerted, community-based effort.

Here, we outline concrete steps and special considerations required to facilitate the successful implementation of this approach.

Generating mouse and organoid models of disease

Except in rare cases, it is not possible to acquire brain tissue from individuals with neurodevelopmental disorders. Therefore, generating disease expression signatures for these diseases requires robust model systems. Ideal pre-clinical models for this approach need to express the disease-associated gene in the same cell type as *in loco*, recapitulate perturbations to the regulatory network and disease phenotypes, and enable high-throughput screening of compounds[46]. No single model satisfies all of these criteria. For example, *in vitro* models offer the most potential for high-throughput screening and can be developed for human cells, but of course cannot represent developmental or behavioral consequences of neurodevelopmental disease-causing mutations[46]. We therefore suggest an integrated approach that includes both human organoids of brain regions and mouse models.

Genetically engineered mouse lines have contributed to significant advances in neurodevelopmental disease gene research. Importantly, mouse models can provide important behavioral and electrophysiological phenotypic endpoints for small molecule screening[47,48]. Mouse models of genetic epilepsies, for example, may display spontaneous seizures or altered seizure thresholds. Additionally, while mouse models cannot fully capture the entirety of symptoms associated with ASD and schizophrenia, they can model certain aspects of these disorders that may manifest in an equivalent manner in rodents[49]. For example, mouse models of schizophrenia-causing mutations in *SETD1A* and ASD-causing mutations in *CHD8*, display cognitive impairments that overlap with human pathology[50,51].

Despite the utility of mouse models, there are inherent differences between the development, architecture, and function of the mouse and human brain[52]. Therefore, mouse studies should be compared with models based on human cells that capture as much of human brain development as is currently possible. Cortical organoids, which are self-organizing three-dimensional structures that resemble features of human cortical development, have emerged as one such model[53-55]. Although reductionist in nature, this system is amenable to genetic engineering and high-throughput screening. Because organoids are derived from human cells, they also overcome potential species-specific differences in gene expression regulation. In fact, investigators have successfully used organoids to identify molecular phenotypes in various neurodevelopmental conditions that have been difficult to study in mice, such as lissencephaly[56] and periventricular nodular heterotopia[57]. Organoids can be developed directly using human

cells from individuals with the disease of interest, reprogramming those cells into induced pluripotent stem cells (iPSCs), and then differentiating them into cortical organoids. Alternatively, genome-editing can be used to introduce a disease-causing mutation in a control iPSC line. While organoids perhaps bring us closer to human biology, this model system has its own limitations. Most notably, recent work has shown that these artificial systems ectopically activate cellular stress pathways, which can impact cell type specification[58]. Moreover, organoids are now known to show substantial variability from one batch to another, making it challenging to uncover consistent transcriptomic, morphological, or functional phenotypes.

Clearly, neither model is perfect. Mice lack human-specific features of development, while organoids are subject to artificial conditions that impact cellular development. We do not yet know how often gene expression signatures in mouse models will be reflected in organoids of the same model, and vice versa. When mouse and organoid models are congruent, we can be more confident in the relevance of the disease signatures. However, when the expression changes between models differ, we must attempt to elucidate the source of this divergence. For example, if organoids are differentiated with unrelated protocols but retain the same transcriptomic signatures, those signatures are likely relevant to the underlying biology than an experimental artifact. It would be similarly worthwhile to compare transcriptomes of global knockout mice with either targeted knockouts or knock-ins. Both targeted knockouts and knock-ins should presumably recapitulate the global knockout, while providing granular insight into the functions of the gene of interest. If transcriptomes from global knockouts show important differences compared with targeted mutants however, peripheral tissues may be playing an underappreciated role in the disease biology. Unlike mice, organoids are isolated systems that would miss these effects entirely.

Deriving disease gene expression signatures via single-cell RNA-sequencing

It is becoming increasingly clear that disease-causing mutations in transcriptomic regulators have cell type-specific effects. One recent study used *in vivo* Perturb-Seq (a pooled CRISPR screening method) to introduce mutations in 35 ASD associated genes within the developing mouse brain *in utero* and then perform scRNA-seq postnatally[59]. Many of the edited genes were transcriptomic regulators, such as *Chd8, Gatad2b, and Set5d,* and perturbations of these genes led to cell type-specific gene expression changes. For example, loss of *Chd8* led to the dysregulation of gene expression programs required for oligodendrocyte differentiation and maturation. In a separate study, scRNA-seq revealed that haploinsufficiency of the schizophrenia-associated gene*, Setd1a*, led to stronger differential gene expression effects in

cortical pyramidal cells than in any other cell type[50]. We recently reported that a mouse model of *HNRNPU*-mediated EE results in an increased burden of differentially expressed genes in pyramidal cells of the subiculum[45]. Using CMap, we found that the compounds predicted to reverse the subiculum signature did not strongly overlap with the compounds predicted to reverse the bulk signature[45].

Together, these studies highlight the need for scRNA-seq in generating cell type-specific disease expression signatures. We argue that each implicated transcriptomic regulator should be modeled and profiled, and the resulting expression data should be made available publicly. Although a massive undertaking, this effort would both maximize the number of transcriptomic reversal successes and equip investigators to identify potentially convergent disease mechanisms among these genes.

Developing a compendium of compound signatures

Although the CMap provides thousands of compound signatures, most of these signatures were derived from cancer cell lines. A small proportion of compounds (n = 768) were also profiled in iPSC-derived neural progenitor cells (NPCs) and neurons. Using these data, we performed a cluster analysis to compare the differential gene expression responses between neurons, NPCs, and the two most commonly assayed cancer cell lines, MCF7 and PC3. Perhaps unsurprisingly, NPCs and neuron signatures clustered separately from the cancer cell line signatures (**Figure 3A**). Consistent with other reports, these results suggest cell type-dependent gene expression responses[26]. Therefore, compound signatures generated in cancer cell lines may have limited utility in identifying transcriptomic reversal candidates for neurological diseases.

Small molecule-induced gene expression changes not only vary between cancer cell lines and neural cells, but also between subtypes of neural cells. To demonstrate this point, we performed single-cell RNA-sequencing of cultured mouse primary cortical neurons treated with seven different FDA-approved compounds with known pharmacological effects on the central nervous system (naloxone, perphenazine, trazodone, trimipramine, risperidone, epirizole, and carbamazepine). These seven compounds were chosen because they elicited the most orthogonal signatures in the CMap NPC data, suggesting they represent the transcriptomic landscape of FDA-approved blood-brain penetrant compounds. We dosed each compound at its reported maximum serum concentration for 24 hours on the 10th day *in vitro* (DIV 10). We then

compared expression responses in four major cell types: astrocytes, oligodendrocytes, excitatory neurons, and inhibitory neurons (**Figure 3B**). Interestingly, both the number of differentially expressed genes and magnitude of expression changes varied by markedly cell type and by drug (**Figure 3C**).

The availability of cell type-specific signatures for both drugs and disease models would allow for comparisons in matched cell types. This approach should theoretically reduce the number of false positive transcriptomic reversal candidates. The current per-sample costs of conventional scRNA-seq approaches, however, preclude the generation of even a modestly sized compendium of signatures. Fortunately, emerging multiplexing techniques, such as Cell Hashing[60], MULTI-seq[61], and sci-Plex[62], dramatically reduce costs and increase the feasibility of generating cell type-specific signatures. Ideally, these signatures should be generated in both organoids and cultured primary mouse neurons. Compounds should be dosed at physiologically relevant concentrations, which can be estimated from a compound's reported maximum serum concentration or maximum therapeutic dose if known.
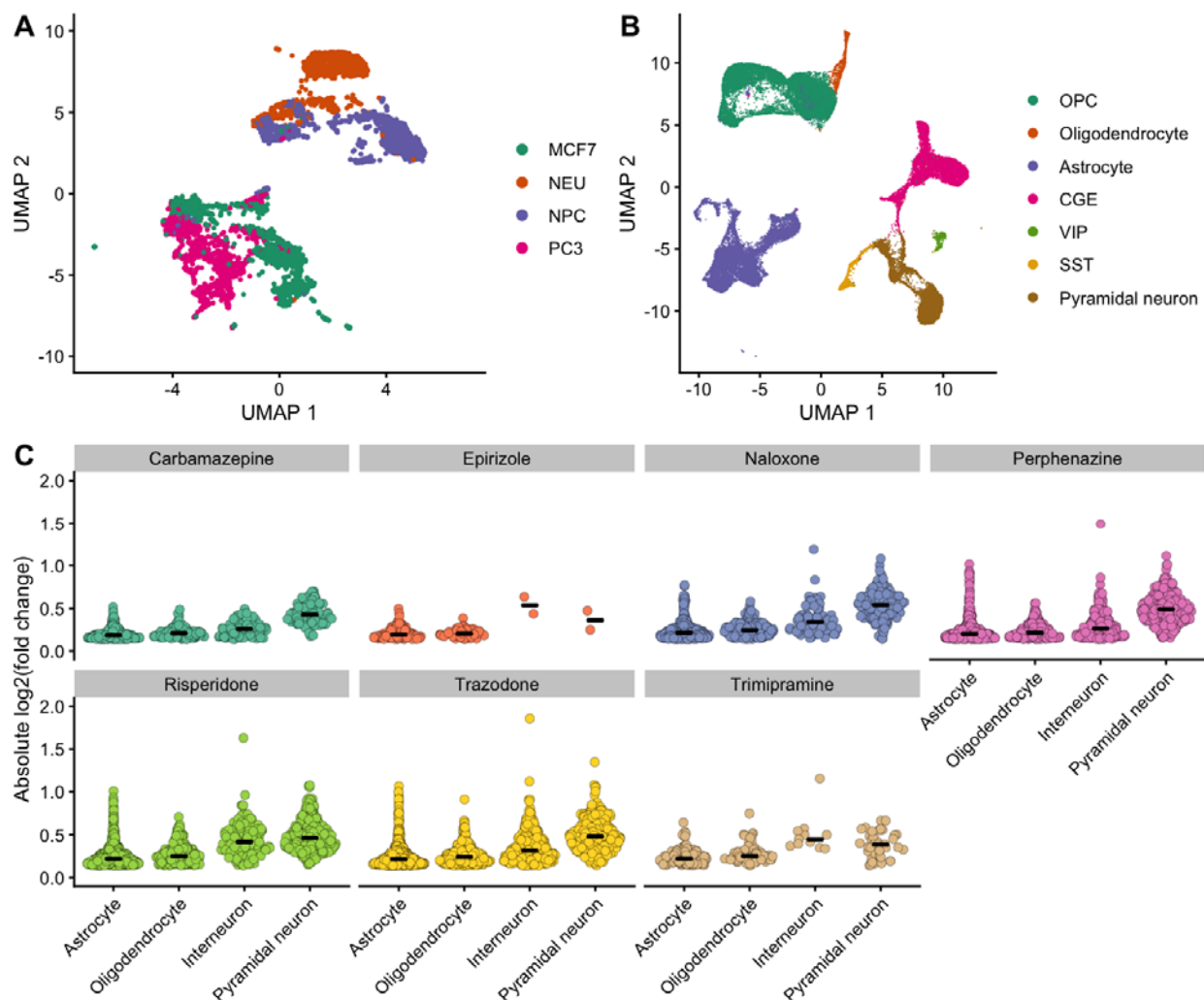
Figure 3. Cell type-specific gene expression responses to small molecules. (A) UMAP plot representing differential gene expression responses to roughly 800 small molecules in cancer cell lines (MCF7 and PC3) versus iPSC-derived neurons and neural progenitor cells (NPCs). Gene expression was measured via the L1000 assay. **(B)** UMAP plot representing clusters of major cell types detected via scRNA-seq of cultured primary mouse cortical neurons (47,494 cells). Cells were dosed with one of seven compounds or with DMSO (control). **(C)** Violin plots representing the average effect size per cell type of differentially expressed genes (FDR < .01 and expression change > 10%) in each treatment condition.

Compound prioritization and validation

In any transcriptomic reversal approach, candidate compounds need to be administered to cells containing the mutated gene to verify that they in fact restore the transcriptome toward normal. Once validated, these compounds need to be tested for the functional consequence of this restoration. We do not yet know to what extent the transcriptome must be normalized to rescue molecular and behavioral phenotypes, and this will only be learned through experience of such

studies. Furthermore, it will be crucial to determine whether compounds elicit the same transcriptomic effects *in vivo* as they do *in vitro* through experimental comparisons.

The cell type-specific transcriptomic reversal approach offers unique opportunities during the validation phase. Consider the case in which more than one cell type emerges as particularly vulnerable to a disease-causing mutation. In this instance, it is possible that drugs that reverse the signature in one cell type may not reverse the signature in another. While this scenario would require the interrogation of more compounds, it also offers the ability to probe the contribution of each cell type to the disease phenotype and to identify potential combination therapies. For example, administering compounds that revert the transcriptome in one cell type may rescue some phenotypes, whereas compounds that revert the transcriptome of another cell type could rescue others. Finally, the prioritization of compounds should also consider the safety profile of disease-signature reversing compounds. Ideally, a compound would reverse the disease signature while having minimal impact on the rest of the transcriptome. In reality, this ideal is unlikely to be often achievable. Approaches will need to be developed to assess whether expression changes beyond the disease-signature reversal are likely to be harmful. For example, one potential approach would be to consider how many genes that are genetically intolerant[63-65] and that are intolerant to expression variation[66] have been perturbed.

Conclusions

We have outlined a precision medicine paradigm that could lead to the discovery of numerous treatment options for neurodevelopmental disorders caused by mutations in transcriptomic regulators. We argue that the paradigm can be most easily refined by working on these Mendelian conditions, in which we can be confident of the causal consequences of transcriptomic alterations. Once the key aspects of this paradigm are better understood in the context of neurological disorders, such as how much restoration of the altered transcriptome is required, it will be possible to extend the paradigm to conditions with more complex genetics, such as the more common forms of the aforementioned diseases.

Establishing such a systematic program would require a community-based effort and substantial data sharing. As demonstrated by the CMap, publicly accessible drug signatures can open the floodgates for identifying novel therapies. Furthermore, meeting the goals outlined here, including the establishment of robust disease models, implementation of large-scale scRNA-seq analyses, and the evaluation of candidate compounds, will require the collaborative efforts between researchers with biological, clinical, and computational expertise. Given the remarkable importance of genes with transcriptomic effects in neurodevelopmental disorders we

believe that careful implementation of a paradigm of transcriptomic-reversal has the potential to identify effective treatments for many genetic diseases.

Methods

Connectivity Map analysis

We compared Connectivity Map[17] L1000 data for 2,103 compounds that were dosed in MCF7 cells, PC3 cells, iPSC-derived neurons, and iPSC-derived neural progenitor cells (NPCs). Data were downloaded from the Gene Expression Omnibus (accession number GSE92742). We only considered gene expression values for the 978 genes directly assessed by the L1000 assay. We then created a UMAP plot to visualize clustering patterns for gene expression signatures elicited by these 2,103 compounds.

To find compounds that elicited orthogonal gene expression signatures in neural cells, we considered the L1000 data for blood-brain barrier penetrant compounds that were assayed in NPCs. We opted to use the NPC signatures rather than neuron signatures as there were more compounds assayed in NPCs (141 versus 53). We performed clustering using affinity propagation[67], which resulted in 7 clusters with the following examplars (i.e. cluster members that are representative of the group): naloxone, perphenazine, trazodone, trimipramine, risperidone, epirizole, and carbamazepine. These compounds were chosen for dosing in mouse cortical neurons.

Mouse husbandry

Experiments were performed on the inbred background C57BL/6NJ (005304 JAX stock). All mice were maintained in ventilated cages with controlled humidity at ~60%, 12h:12h light:dark cycles (lights on at 7:00AM, off 7:00PM) and controlled temperature of 22–23°C. Mice had access to regular chow and water, ad libitum. Breeding cages were fed a high fat breeder chow. Mice were maintained and all procedures were performed within the Columbia University Institute of Comparative Medicine, which is fully accredited by the Association for Assessment and Accreditation of Laboratory Animal Care. All protocols were approved by the Columbia Institutional Animal Care and Use Committee.

Primary cortical culture

Prior to use, 24-well TC Treated plates (Corning) were coated overnight with 50 μg/mL poly-D-lysine (Sigma) in 0.1M borate buffer (pH 8.5). Cortices acquired from multiple sets of male littermates were incubated in activated 20 U/mL Papain/DNase (Worthington) for 15 minutes at 37°C, centrifuged at 300 x g for 5 min, and washed in 1X PBS. Cell pellets were suspended in NBA/B27 [consisting of Neurobasal-A (Life Technologies), 1X B27 (Life Technologies), 1X GlutaMax (Life Technologies), 1% HEPES, and 1X Penicillin/Streptomycin], supplemented with 1% fetal bovine serum (Gibco) and 5 μg/mL laminin. Five hundred thousand cells were plated per well. The day after plating, media was removed and replaced with pre-warmed NBA/B27. Cultures were maintained at 37°C in 5% CO_2. 50% of the medium was changed every other day with NBA/B27 starting on DIV3. On DIV10, seven compounds and the control media were introduced to the appropriate wells in duplicate at their reported maximum serum concentrations[68]:

Drug	Maximum serum concentration (nM)	Dose (nM)
Naloxone	91.7	100

Perphenazine	49.4	50
Trazodone hyrochloride	4302.7	4000
Trimipramine maleate salt	1018.9	1000
Risperidone	48.7	50
Epirizole	Unknown	1000
Carbamazepine	84.6	100
DMSO (negative control)	-	-

After 24 hours of compound exposure, the cells were dissociated with activated 20 U/mL Papain, strained through 40 um cell strainers (MTC Bio) to remove cell clumps and collected for single-cell RNA-sequencing (scRNA-seq) on the 10X Chromium in two batches. scRNA-seq libraries were constructed using the 10X Chromium Single Cell 3′ Reagent Kits v2 according to manufacturer descriptions, and samples were sequenced on a NovaSeq 6000. Reads were aligned to the mm10 genome using the 10X CellRanger pipeline with default parameters to generate the feature-barcode matrix.

scRNA-seq analysis
We used Seurat v3 to perform downstream QC and analyses on feature-barcode matrices[37,38]. We removed all genes that were not detected in at least 4 cells. We further removed cells with fewer than 200 genes or more than 2,500 genes detected. We further removed all cells with greater than 25% of reads mapping to mitochondrial genes. 47,494 cells remained after filtering:

Drug	Batch 1	Batch 2	Total
Naloxone	2203	3360	5563
Perphenazine	2457	4254	6711
Trazodone	2298	3482	5780
Trimipramine	3516	3299	6815
Risperidone	2367	2757	5124
Epirizole	3051	2982	6033
Carbamazepine	3189	3021	6210
DMSO (control)	1770	3488	5258

The filtered matrices were log-normalized and scaled to 10,000 transcripts per cell. We used the variance-stabilizing transformation implemented in the FindVariableFeatures function in order to identify the top 2,000 most variable genes per sample. We used Seurat's data integration method to harmonize gene expression across datasets prior to clustering. We first identified anchors between samples in each dataset using the FindIntegrationAnchors function, which uses canonical correlation analysis (CCA) to identify pairwise cell correspondences between samples. We then computed an integrated expression matrix using these anchors as input to the IntegrateData function.

Next, we used linear regression to regress out the number of UMIs per cell and percentage of mitochondrial reads using the ScaleData function on the integrated expression

matrices. We then performed dimensionality reduction using PCA. For each dataset, we selected the top 30 dimensions to compute a cellular distance matrix, which was used to generate a K-nearest neighbor graph. The KNN was used as input to the Louvain Clustering algorithm implemented in the FindClusters function. For clustering via Louvain, we chose a resolution parameter of 0.8. We visualized the cells using UMAP via the RunUMAP function. To annotate and merge clusters, we performed differential gene expression analysis on the integrated expression values between each cluster using the default parameters in the FindMarkers function, which implements a Wilcoxon test and corrects p-values via Bonferroni correction. Additionally, we visualized the expression of canonical marker genes aggregated from previous single-cell publications[71–74].

Differential gene expression analysis

We performed cell-type-specific differential gene expression analysis using MAST[75], as implemented in Seurat's FindMarkers function, in order to identify genes dysregulated between drug-treated and DMSO-treated cells. We excluded all non-coding genes, genes encoding ribosomal proteins, and pseudogenes from our analysis to reduce the multiple testing burden. For each cell type, we fit a linear mixed model that included the gene detection rate (ngeneson) and sequencing batch as latent variables. We corrected the p-values using the Benjamini-Hochberg FDR method, and considered genes with an FDR < 0.01 as differentially expressed.

Declaration of interests

D.B.G. is a founder of and holds equity in Praxis, holds equity in Q-State Biosciences, serves as a consultant to AstraZeneca, and has received research support from Janssen, Gilead, Biogen, AstraZeneca, and UCB. R.S.D. serves as a consultant to AstraZeneca. C.V., A.W.Z, and D.K. declare no competing interests.

Acknowledgments

We are grateful to Wayne Frankel for helpful discussions and his comments on the manuscript.

References

1.   Claussnitzer M, Cho JH, Collins R, et al. A brief history of human disease genetics. *Nature* 2020; **577**(7789): 179-89.

2.   Dhindsa RS, Goldstein DB. Genetic Discoveries Drive Molecular Analyses and Targeted Therapeutic Options in the Epilepsies. *Curr Neurol Neurosci Rep* 2015; **15**(10): 70.

3.   Desnick RJ, Schuchman EH. Enzyme replacement and enhancement therapies: Lessons from lysosomal disorders (vol 3, pg 954, 2002). *Nature Reviews Genetics* 2003; **4**(2): 157-.

4.   Chiriboga CA, Swoboda KJ, Darras BT, et al. Results from a phase 1 study of nusinersen (ISIS-SMN(Rx)) in children with spinal muscular atrophy. *Neurology* 2016; **86**(10): 890-7.

5.   Finkel RS, Chiriboga CA, Vajsar J, et al. Treatment of infantile-onset spinal muscular atrophy with nusinersen: a phase 2, open-label, dose-escalation study. *Lancet* 2016; **388**(10063): 3017-26.

6.   Atkin TA, Maher CM, Gerlach AC, et al. A comprehensive approach to identifying repurposed drugs to treat SCN8A epilepsy. *Epilepsia* 2018; **59**(4): 802-13.

7.   Epi4K Consortium, Epilepsy Phenome/Genome P, Allen AS, et al. De novo mutations in epileptic encephalopathies. *Nature* 2013; **501**(7466): 217-21.

8.   EuroEpinomics Consortium, Epilepsy Phenome/Genome Project, Epi4K Consortium. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am J Hum Genet* 2014; **95**(4): 360-70.

9.   Buxbaum J, Consortium AS. Synaptic, Transcriptional, and Chromatin Genes Disrupted in Autism: Findings from 13,000 Exomes. *Neuropsychopharmacol* 2014; **39**: S409-S.

10.   Deciphering Developmental Disorders S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 2015; **519**(7542): 223-8.

11.   Singh T, Kurki MI, Curtis D, et al. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci* 2016; **19**(4): 571-+.

12.   Lamb J. The Connectivity Map: a new tool for biomedical research. *Nat Rev Cancer* 2007; **7**(1): 54-60.

13.   Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006; **313**(5795): 1929-35.

14.   Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet* 2019; **20**(11): 631-56.

15.   Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 2015; **161**(5): 1202-14.

16.   Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* 2019; **15**(6).

17.   Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 2017; **171**(6): 1437-52 e17.

18.   Mitrofanova A, Aytes A, Zou M, Shen MM, Abate-Shen C, Califano A. Predicting Drug Response in Human Prostate Cancer from Preclinical Analysis of In Vivo Mouse Models. *Cell Rep* 2015; **12**(12): 2060-71.

19.   Alvarez MJ, Subramaniam PS, Tang LH, et al. A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors. *Nat Genet* 2018; **50**(7): 979-89.

20.   Chen B, Ma L, Paik H, et al. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nat Commun* 2017; **8**: 16022.

21.   Zhang M, Luo H, Xi ZR, Rogaeva E. Drug Repositioning for Diabetes Based on 'Omics' Data Mining. *Plos One* 2015; **10**(5).

22.   Dudley JT, Sirota M, Shenoy M, et al. Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease. *Sci Transl Med* 2011; **3**(96).

23.   Brum AM, van de Peppel J, van der Leije CS, et al. Connectivity Map-based discovery of parbendazole reveals targetable human osteogenic pathway. *P Natl Acad Sci USA* 2015; **112**(41): 12711-6.

24.   Keenan AB, Wojciechowicz ML, Wang ZC, et al. Connectivity Mapping: Methods and Applications. *Annu Rev Biomed Da S* 2019; **2**: 69-92.

25.   Alvarez MJ, Subramaniam PS, Tang LH, et al. A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors. *Nature Genetics* 2018; **50**(7): 979-+.

26.   Readhead B, Hartley BJ, Eastwood BJ, et al. Expression-based drug screening of neural progenitor cells from individuals with schizophrenia (vol 9, 4412, 2018). *Nature Communications* 2018; **9**.

27.   Srivastava PK, van Eyll J, Godard P, et al. A systems-level framework for drug discovery identifies Csf1R as an anti-epileptic drug target. *Nat Commun* 2018; **9**(1): 3561.

28.   Delahaye-Duriez A, Srivastava P, Shkura K, et al. Rare and common epilepsies converge on a shared gene regulatory network providing opportunities for novel antiepileptic drug discovery. *Genome Biol* 2016; **17**(1): 245.

29.   Brueggeman L, Sturgeon ML, Martin RM, et al. Drug repositioning in epilepsy reveals novel antiseizure candidates. *Ann Clin Transl Neur* 2019; **6**(2): 295-309.

30.   Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. *Cell* 2013; **152**(6): 1237-51.

31.   Le Meur N, Holder-Espinasse M, Jaillard S, et al. MEF2C haploinsufficiency caused by either microdeletion of the 5q14.3 region or mutation is responsible for severe mental retardation with stereotypic movements, epilepsy and/or cerebral malformations. *J Med Genet* 2010; **47**(1): 22-9.

32.   Zweier C, Trautmann U, Ekici A, Rauch A. A 15Mb duplication of 6q24.1-q25.3 associated with typical but milder features of the duplication 6q syndrome. *Eur J Med Genet* 2008; **51**(4): 358-61.

33.   Harrington AJ, Raissi A, Rajkovich K, et al. MEF2C regulates cortical inhibitory and excitatory synapses and behaviors relevant to neurodevelopmental disorders. *Elife* 2016; **5**.

34.   Ariani F, Hayek G, Rondinella D, et al. FOXG1 is responsible for the congenital variant of Rett syndrome. *Am J Hum Genet* 2008; **83**(1): 89-93.

35.   Kumamoto T, Toma K, Gunadi, et al. Foxg1 coordinates the switch from nonradially to radially migrating glutamatergic subtypes in the neocortex through spatiotemporal repression. *Cell Rep* 2013; **3**(3): 931-45.

36.   Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics* 1999; **23**(2): 185-8.

37.   Chen RZ, Akbarian S, Tudor M, Jaenisch R. Deficiency of methyl-CpG binding protein-2 in CNS neurons results in a Rett-like phenotype in mice. *Nat Genet* 2001; **27**(3): 327-31.

38.   Guy J, Hendrich B, Holmes M, Martin JE, Bird A. A mouse Mecp2-null mutation causes neurological symptoms that mimic Rett syndrome. *Nat Genet* 2001; **27**(3): 322-6.

39.   Shovlin S, Tropea D. Transcriptome level analysis in Rett syndrome using human samples from different tissues. *Orphanet J Rare Dis* 2018; **13**(1): 113.

40.   Ben-Shachar S, Chahrour M, Thaller C, Shaw CA, Zoghbi HY. Mouse models of MeCP2 disorders share gene expression changes in the cerebellum and hypothalamus. *Hum Mol Genet* 2009; **18**(13): 2431-42.

41.   Renthal W, Boxer LD, Hrvatin S, et al. Characterization of human mosaic Rett syndrome brain tissue by single-nucleus RNA sequencing. *Nat Neurosci* 2018; **21**(12): 1670-9.

42.   Brinegar AE, Cooper TA. Roles for RNA-binding proteins in development and disease. *Brain Res* 2016; **1647**: 1-8.

43.   Bramswig NC, Ludecke HJ, Hamdan FF, et al. Heterozygous HNRNPU variants cause early onset epilepsy and severe intellectual disability. *Hum Genet* 2017; **136**(7): 821-34.

44.   Yates TM, Vasudevan PC, Chandler KE, et al. De novo mutations in HNRNPU result in a neurodevelopmental syndrome. *Am J Med Genet A* 2017; **173**(11): 3003-12.

45.   Dugger SA, Dhindsa RS, De Almeida Sampaio G, et al. Neurodevelopmental deficits and cell-type-specific transcriptomic perturbations in a mouse model of *HNRNPU* haploinsufficiency. *bioRxiv* 2020: 2020.05.01.072512.

46.   EpiPM Consortium. A roadmap for precision medicine in the epilepsies. *Lancet Neurol* 2015; **14**(12): 1219-28.

47.   Silverman JL, Yang M, Lord C, Crawley JN. Behavioural phenotyping assays for mouse models of autism. *Nat Rev Neurosci* 2010; **11**(7): 490-502.

48.   Loscher W. Animal Models of Seizures and Epilepsy: Past, Present, and Future Role for the Discovery of Antiseizure Drugs. *Neurochem Res* 2017; **42**(7): 1873-88.

49.   Gogos JA, Crabtree G, Diamantopoulou A. The abiding relevance of mouse models of rare mutations to psychiatric neuroscience and therapeutics. *Schizophr Res* 2019.

50.   Mukai J, Cannavo E, Crabtree GW, et al. Recapitulation and Reversal of Schizophrenia-Related Phenotypes in Setd1a-Deficient Mice. *Neuron* 2019; **104**(3): 471-87 e12.

51.   Gompers AL, Su-Feher L, Ellegood J, et al. Germline Chd8 haploinsufficiency alters brain development in mouse. *Nat Neurosci* 2017; **20**(8): 1062-73.

52.   Amin ND, Pasca SP. Building Models of Brain Disorders with Three-Dimensional Organoids. *Neuron* 2018; **100**(2): 389-405.

53.   Lancaster MA, Renner M, Martin CA, et al. Cerebral organoids model human brain development and microcephaly. *Nature* 2013; **501**(7467): 373-9.

54.   Pollen AA, Bhaduri A, Andrews MG, et al. Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell* 2019; **176**(4): 743-+.

55.   Di Lullo E, Kriegstein AR. The use of brain organoids to investigate neural development and disease. *Nat Rev Neurosci* 2017; **18**(10): 573-84.

56.   Bershteyn M, Nowakowski TJ, Pollen AA, et al. Human iPSC-Derived Cerebral Organoids Model Cellular Features of Lissencephaly and Reveal Prolonged Mitosis of Outer Radial Glia. *Cell Stem Cell* 2017; **20**(4): 435-49 e4.

57.   Klaus J, Kanton S, Kyrousi C, et al. Altered neuronal migratory trajectories in human cerebral organoids derived from individuals with neuronal heterotopia. *Nat Med* 2019; **25**(4): 561-+.

58.   Bhaduri A, Andrews MG, Leon WM, et al. Cell stress in cortical organoids impairs molecular subtype specification. *Nature* 2020.

59.   Jin X, Simmons SK, Guo AX, et al. In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with Autism risk genes. *bioRxiv* 2019: 791525.

60.   Stoeckius M, Zheng S, Houck-Loomis B, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol* 2018; **19**(1): 224.

61.   McGinnis CS, Patterson DM, Winkler J, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nature Methods* 2019; **16**(7): 619-+.

62.   Srivatsan SR, McFaline-Figueroa JL, Ramani V, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science* 2020; **367**(6473): 45-+.

63.   Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 2013; **9**(8): e1003709.

64.   Petrovski S, Gussow AB, Wang Q, et al. The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLoS Genet* 2015; **11**(9): e1005492.

65.   Dhindsa RS, Copeland BR, Mustoe AM, Goldstein DB. Natural selection shapes codon usage in the human genome. *bioRxiv* 2019: 688051.

66.   Mohammadi P, Castel SE, Cummings BB, et al. Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science* 2019; **366**(6463): 351-6.

67.   Frey BJ, Dueck D. Clustering by passing messages between data points. *Science* 2007; **315**(5814): 972-6.

68.   Schulz M, Iwersen-Bergmann S, Andresen H, Schmoldt A. Therapeutic and toxic blood concentrations of nearly 1,000 drugs and other xenobiotics. *Crit Care* 2012; **16**(4).