Title

High density SNP array and reanalysis of genome sequencing uncovers CNVs associated with neurodevelopmental disorders in KOLF2.1J iPSCs

Authors

Carolina Gracia-Diaz[1,2], Jonathan E. Perdomo[1,3], Munir E. Khan[4], Brianna Disanza[1,5], Gregory G. Cajka[1,6], Sunyimeng Lei[1,2], Alyssa Gagne[1,2], Jean Ann Maguire[1,2], Thomas Roule[1,2], Ophir Shalem[1,6], Elizabeth J. Bhoj[1,4,5], Rebecca C. Ahrens-Nicklas[1,5], Deborah French[1,2], Ethan M. Goldberg[7,8], Kai Wang[1,2,*], Joseph Glessner[4,*], Naiara Akizu[1,2,10*]

Affiliations

[1]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA.
[2]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, USA.
[3]School of Biomedical Engineering, Drexel University, Philadelphia, PA 19104, USA
[4]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA.
[5]Department of Pediatrics, University of Pennsylvania, Philadelphia, PA, USA.
[6]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA.
[7]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA.
[8]Departmen of Neurology, University of Pennsylvania, Philadelphia, PA, USA.
[9]Division of Neurology, Department of Pediatrics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA.
[10]Lead contact
*Correspondance: wangk@chop.edu (K.W.), glessner@chop.edu (J.G.), aquizun@chop.edu (N.A.)

Summary

The KOLF2.1J iPSC line was recently proposed as a reference iPSC to promote the standardization of research studies in the stem cell field. Due to overall good performance differentiating to neural cell lineages, high gene editing efficiency, and absence of genetic variants associated to neurological disorders KOLF2.1J iPSC line was particularly recommended for neurodegenerative disease modeling. However, our work uncovers that KOLF2.1J hPSCs carry heterozygous small copy number variants (CNVs) that cause *DTNBP1, JARID2* and *ASTN2* haploinsufficiencies, all of which are associated with neurological disorders. We further determine that these CNVs arose *in vitro* over the course of KOLF2.1J iPSC generation from a healthy donor-derived KOLF2 iPSC line and affect the expression of DNTBP1, JARID2 and ASTN2 proteins in KOLF2.1J iPSCs and neural progenitors. Therefore, our study suggests that KOLF2.1J iPSCs carry genetic variants that may be deleterious for neural cell lineages. This data is essential for a careful interpretation of neural cell studies derived from KOLF2.1J iPSCs and highlights the need for a catalogue of iPSC lines that includes a comprehensive genome characterization analysis.

1

Introduction

The derivation of the first human embryonic stem cells (hESC)[1] followed by the generation of induced pluripotent stem cells (iPSC) from somatic cells[2] and more recent advances in genome engineering methods[3] have transformed the way we study human biology and diseases. Today, we can generate virtually every human cell type *in vitro* and we can do so in the genetic background of choice. This progress has particularly benefited research focused on the study of human brain development, mechanisms and disorders by providing a nearly universal access to human neural cells, which are otherwise of limited accessibility.

Since the differentiation of the first neurons from hESC and iPSCs (collectively known as human pluripotent stem cells, hPSCs)[4,5], protocols for 2D and 3D neural tissue generation have exponentially increased and expanded the horizon for mechanistic and therapeutic discoveries of human brain disorders. Yet, these advances come with new challenges for the field, including the multiple sources of technical and biological variables that lead to the expression of phenotypes often difficult to reconcile[6,7]. To overcome these challenges, early efforts focused on improving the reproducibility and efficiency of neural differentiation protocols[8-11]. However, even with the most reproducible protocol, the genetic background of an individual hPSC line can significantly contribute to the phenotypic heterogeneity, just like in the general population. Studies using high quality genome assemblies obtained with long read sequencing estimate that each human genome contains approximately ~4,000,000 single nucleotide variants (SNVs), ~800,000 insertion deletions (indels) and 25,000 structural variants (SV), such as copy number variants (CNVs) or inversions and insertions of >50bp[12]. Given that recapitulating the large genetic heterogeneity of the human population in a dish would limit the identification of significant genotype-phenotypes relations, recent efforts have proposed to homogenize the genetic background by adopting a widespread use of reference hPSCs[13]. The success of this idea is supported by decades of key discoveries in isogenic strains of model organisms[14,15].

In a recent effort to generate a readily available collection of isogenic iPSC carrying mutations associated with neurodegenerative disorders for distribution and collaboration across different laboratories, the iPSC Neurodegenerative Disease Initiative (iNDI) sought to identify an overall well-performing iPSC line that would serve as a reference for the research community[13,16]. To achieve this goal, nine iPSCs lines from public repositories were selected for a deep functional and genomic characterization. One of these iPSC lines, named KOLF2.1J, was generated from the KOLF2_C1 iPSC line by CRISPR/Cas9 correction of a heterozygous 19bp deletion in *ARID2*[17], a variant likely pathogenic for a neurodevelopmental disorder known as Coffin-Siris syndrome. After a deep characterization of several subclones of the nine iPSC lines, KOLF2.1J was selected as a candidate reference iPSC line for its good performance at pluripotent stage and differentiation to neural cell linages, high CRISPR-Cas9 editing efficiency and low burden of genetic variants associated with neurological disorders[13]. To further test KOLF2.1J for the presence of genetic variants that might affect the experimental interpretations, additional genomic analyses were performed, including comparative studies with the parental fibroblasts (from a 55-59yo healthy male donor) and reprogrammed iPSC clone KOLF2-C1. These analyses revealed ~3.3 million high confidence SNPs and indels affecting coding regions in KOLF2.1J. Twenty-five

2

of these were present only in KOLF2.1J and not in KOLF-C1 iPSCs and 37 were inherited from the parental fibroblasts. Only three of these variants affected genes predicted to be intolerant to loss of function or listed as haploinsufficent (*COL3A1*, *SHOX* and *DEDD*) but none were suspected to compromise neurological disease modeling. Notably, structural variants of 50 bp-5 Mbp in size were not assessed in these genomic analyses due to technical limitations of the selected methods[13].

Here we show that KOLF2.1J iPSCs carry at least 5 small CNVs (<1Mbp), two of which cause *DTNBP1, JARID2* and *ASTN2* haploinsufficiency. Given the association of these genes with neurodevelopmental disorders[18-27] and constraints for disruptive variants in human genomes[28-30] our work uncovers that KOLF2.1J iPSCs carry genetic variants that may affect the interpretation of developmental and neural cell phenotypes. Our work further highlights the need for the inclusion of structural variants in genome sequencing analyses of iPSC lines and the creation of a catalogue of reference iPSCs which shall include a comprehensive characterization of genes predicted to be intolerant to disruptive variation during human development.

Results

Genome integrity validation detects CNVs in KOLF2.1J iPSCs

Owing to their deep characterization and malleability for genome editing, we recently adopted KOLF2.1J iPSCs (The Jackson Laboratory, JIPSC1000) for neurodegenerative and neurodevelopmental studies conducted in our laboratories. Our projects originate from the identification of candidate pathogenic variants through exome and genome sequencing of children affected with progressive or non-progressive neurodevelopmental disorders. We then design *in silico* and biochemical studies to test the impact of the variants on the encoded proteins and functionally validate those with a significant molecular effect. The hPSC based neural differentiation models constitute the top tier of our studies for which we reprogram patient cells or use CRISPR-Cas9 to introduce the candidate variants in an isogenic background to compare neural cell phenotypes between variants and controls.

Following this strategy, we selected several variants to CRISPR-Cas9 edit in KOLF2.1J iPSCs and attest their high editing efficiency. After gene editing and expansion of edited KOLF2.1J clones to generate frozen stocks, we applied our routine quality control tests, including a genome integrity validation. For this purpose, we extracted genomic DNA (gDNA) from the edited KOLF2.1J iPSCs and subsequently analyzed with an array containing common and rare alleles of ~650.000 single nucleotide polymorphisms (SNPs) spanning the whole human genome. As a method for structural integrity validation we analyzed the array for CNVs causing deletions or duplications of genome fragments using PennCNV[31]. Furthermore, to rule out the possibility of undesired genetic events arising during the editing process, we compared the results with the CNV analysis of an unedited KOLF2.1J iPSC clone. Although, the comparison confirmed absence of acquired CNVs in most of the edited clones, results unexpectedly revealed recurrent small CNVs (<1Mbp) in all edited and unedited KOLF2.1J iPSCs (Table 1), suggesting the presence of these CNVs in our KOLF2.1J iPSC stock. To corroborate this finding, we conducted the same SNP array based CNV analysis in one of our KOLF2.1J iPSCs stocks at +1 passage (p3), KOLF2.1J iPSCs carrying the doxycycline inducible *NGN2* transgene (both kindly donated by Dr.

3

Skarnes) and a KOLF2.1J stock at p2 recently received from The Jackson Laboratory. Results confirmed the presence of the 5 CNVs in all the stock KOLF2.1J iPSCs (Table 1, Fig 1 and Supplementary Fig 1 and 2).

6p22 and 9p33 deletion CNVs are associated with neurodevelopmental disorders

With an estimate of ~25,000 structural variants (SV) carried per genome in average, it is not surprising that most hPSCs used for research also carry SV[32]. Furthermore, numerous studies have demonstrated that the origin of the parental cell, the reprogramming process and the culture conditions can contribute to structural genetic variation[33-38]. KOLF2.1J iPSCs were derived from a CRISPR-Cas9 mediated correction of a 19 bp deletion in one copy of *ARID2* present in the parental KOLF2-C1 iPSC line and expanded to hundreds of replicate vials to ensure their availability for distribution[13]. Although all these procedures are a known source of genetic structural variation[36,38,39], most type of structural variants were not reported in KOLF2.1J iPSCs[13]. Given the discrepancy between these and our results, we took a closer look to the SNPs at the CNV regions of KOLF2.1J iPSCs by visualizing their signal intensity (Log R Ratio (LRR)) and B allele frequency (BAF). Consistent with a heterozygous deletion, we observed a drop in LRR values of SNPs located within 3p13, 6p22 and 9q33 CNV regions relative to those in adjacent diploid regions, while the signal intensity of duplicated CNV regions in 3p14 and 18q33 was slightly above (Fig 1, Supplementary Fig 1 and 2). Furthermore, as expected by one copy loss of a DNA fragment, the regions with reduced LRRs also showed lack of heterozygous SNPs in KOLF2.1J iPSCs (Fig 1, Supplementary Fig 1 and 2) while the control iPSC line we analyzed in parallel showed LRR and BAF values of 6p22, 9q33 and adjacent SNPs that were consistent with diploid state (Fig 1).

Among the 5 CNVs identified, two span coding regions. One is a ~0.2Mbp heterozygous copy loss located in chromosome 6p22 which deletes one copy of the entire *DTNBP1* gene and 11 out of the 18 coding exons of *JARID2* (Fig 2A). *DTNBP1* encodes Dysbindin, a protein involved in lysosome-related organelle biogenesis which expression increases with neuronal differentiation in the developing and adult human brain (Supplementary Fig 3A)[40,41]. While there is no evidence of pathogenic consequences for heterozygous *DTNBP1* loss, homozygous nonsense mutations cause Hermansky-Pudlak syndrome 7 (OMIM #614076)[42-44] and the locus is associated with increased susceptibility for schizophrenia[25-27]. More concerning for studies involving neural cell types is the heterozygous loss of *JARID2*, a gene highly intolerant to loss of function (pLI=1, pLOF o/e=0.09)[28], and copy number variation (pHaplo=0.99, pTriplo=0.99 and pHI=0.59)[29,30] in the human population (Table 2). *JARID2* is highly expressed in pluripotent cells and in the developing human brain with expression slightly declining as neurons differentiate (Supplementary Fig 3B)[40,41]. As part of the Polycomb Repressive Complex 2 (PRC2) regulatory subunit, JARID2 is critical for embryonic development and tissue homeostasis[45]. Notably, similar CNVs causing partial or full deletion of one copy of *JARID2*, are the molecular basis for a neurodevelopmental delay with variable intellectual disability and dysmorphic facies syndrome (OMIM #620098)[18-21].

The other CNV affecting a coding region is a ~0.1Mbp heterozygous deletion in chromosome 9q33 overlapping with the exon 20 of *ASTN2* and its antisense gene (*ASTN2-AS1*) (Fig 2J). *ASTN2* encodes a protein involved in trafficking and degradation of neuronal cell adhesion

4

molecules important for the regulation of neuronal migration and synaptic development[46]. Similar to *JARID2*, *ASTN2* is a gene with low tolerance for loss of function and copy number variation (pLI=1, pHaplo=0.88, pTriplo=0.672 and pHI=0.95)[28-30] (Table 2) and its genetic disruption is a risk factor for multiple neurodevelopmental disorders[22-24]. The expression of *ASTN2* is low at pluripotent stage but increases in neurons and is widely expressed in the developing and adult brain (Supplementary Fig 3C)[40,41]. Although we have not experimentally investigated the consequence of this CNV for *ASTN2* expression, the loss of exon 20 is predicted to lead to a premature stop codon and nonsense mediated RNA decay.

To further validate the SNP array CNV results with an alternative method, we performed a quantitative PCR (qPCR) analysis of genomic DNA (gDNA) with a set of primers targeting the CNVs and upstream and downstream regions not predicted to be affected by the CNVs. As a positive control for sensitivity of this approach distinguishing one copy from two copies of genome fragments, we used a previously reported patient iPSC line carrying a heterozygous deletion within *EZH1* gene[47]. As expected, the level of amplification of *EZH1* region in this patient-derived iPSCs was half of the control hPSCs (Supplementary Fig 4). Similarly, the qPCR with primers targeting 6p22 and 9q33 CNVs showed half levels in KOLF2.1J iPSCs compared to control hPSCs, while the level of amplification of diploid neighboring regions were similar between all the hPSC lines (Fig 2B-I and K-P). These data confirm *JARID2* and *ASTN2* haploinsufficiency in several stocks of KOLF2.1J iPSCs that we independently received.

Reanalysis of genome sequencing data confirms that KOLF2.1J iPSCs carry CNVs associated with neurodevelopmental disorders

Given that our data shows that the stock of KOLF2.1J iPSCs being currently distributed carry CNVs associated with neurodevelopmental disorders, we wondered whether the 5 CNVs were already present in the KOLF2.1J genome sequencing data, yet missed during the analysis previously reported[13]. In support of this possibility, we found that the original structural analysis of the genome sequencing was performed with an algorithm that considers discordant read-pairs and split-reads for CNV calling[48], which can lead to false negatives when there is not enough evidence of aberrant read pairs passing quality control filtering.

The detection of structural variants from short read genome sequencing studies remains challenging. However, several algorithms that consider read depth in their analysis show improved sensitivity for CNV calling from short-read genome sequencing data[49]. To determine if the genetic background of KOLF2.1J iPSCs includes the 5 CNVs identified in our stocks, we downloaded the genome sequencing data deposited in the Alzheimer's Disease Workbench (https://doi.org/10.34688/KOLF2.1J.2021.12.14)[13] and aligned to the reference genome in paired-end short read mode. Then, we calculated the Log2 Ratio and BAF across alignments with read coverages and allele counts obtained for each CNV region using an in-house developed software. Strikingly, the five CNVs identified by the SNP array were also present in the genome sequencing data of KOLF2.1J iPSCs (Fig 3 and Supplementary Fig 5). In addition, the reanalysis allowed us to define the breakpoints of each CNV at base-resolution, which revealed that the 3p14 duplication is larger than indicated by the SNP array CNV calling (Supplementary Fig 5B). Notably, these new coordinates indicate that the CNV 3p14 duplication overlaps with coding regions of

5

FHIT and *PTPRB,* although functional consequences are uncertain because it is difficult to tell whether there is a tandem duplication or an extra copy elsewhere in the genome.

To further validate the genome sequencing CNV analysis results, we analyzed the Chr6 and Chr9 CNV breakpoint junctions in KOLF2.1J iPSCs by Sanger sequencing. Results confirmed the breakpoint coordinates for the Chr6 CNV, with only one bp discrepancy from the genome sequencing analysis (Fig 3C). For the Chr9 CNV, the Sanger sequencing result corrected the breakpoint junction by 56bps (Fig 3D). Interestingly, we noted that these 56bps were repeated almost identically (except for two nucleotides) in the sequences surrounding the 5' and 3' breakpoints of the Chr9 CNV. Thus, we had a closer look at these genomic regions and uncovered that the 56bps repeats are part of the microhomologies of two *AluS* elements flanking the Chr9 CNV. This finding indicates that the Chr9 CNV arose from an *Alu/Alu* mediated genomic rearrangement, which is a known source of genomic structural variation and estimated to cause ~0.3% of human genetic diseases[50].

Chr6 and Chr9 CNVs arose *in vitro* and affect the expression of *JARID2, DTNBP1* and *ASTN2* genes in KOLF2.1J iPSC line

Having confirmed that KOLF2.1J iPSCs carry CNVs that are associated with neurological disorders, we next wondered whether these CNVs were already present in the healthy, 55-59 year-old donor. This possibility would suggest that the CNVs are not deleterious. Of note, KOLF2.1J iPSC line was derived from a subclone of KOLF2 iPSCs reprogramed from the donor's fibroblasts, and KOLF2 genomic data is publicly accessible through the European Nucleotide Archive. Therefore, we downloaded the available KOLF2 SNP genotyping data , which contains LRR and BAF information for 511,966 SNPs, and analyzed for CNVs. CNV calling revealed that only Chr3p14.2 and Chr18q22.1 duplication CNVs were present in KOLF2 iPSCs (Table 3). Furthermore, visual analysis of SNPs within the Chr3p13 showed BAFs of ~0.3 and ~0.7 which suggests the subclonal presence of this CNV (Table 3 and Supplementary Fig 6A-C). Strikingly, Chr6 and Chr9 CNVs were not called nor detected by visual review of SNPs LRRs and BAFs in KOLF2 iPSCs (Table 3 and Fig 4A,B), indicating that they arose *in vitro* over the course of culturing, subcloning and CRISPR/Cas9 editing from KOLF2 iPSC to KOLF2.1J iPSC generation.

Given that the healthy donor does not carry Chr6 and Chr9 CNVs, we could not exclude a deletereus effect. We then reasoned that for a CNVs to cause deleterious effects, it should first alter the expression of the affected genes. Chr6 and Chr9 CNVs cause a complete deletion of *DTNBP1* and partial deletions predicting loss of expression of *JARID2* and *ASTN2.* Indeed, a recent proteomic study shows that compared to H9 hESCs, KOLF2.1J iPSCs express about half of ASTN2, JARID2 and DTNBP1 both at pluripotent and neural progenitor stages (Fig 4C,D)[51].

Altogether, our work reveals that KOLF2.1J iPSCs carry five CNVs, two of which arose *in vitro* and affect the expression of genes associated with neurological disorders. These findings are essential for the interpretation of neural phenotypes derived from studies using KOLF2.1J iPSCs and to guide the selection of iPSC lines for future studies.

Discussion

The mutational burden of an hPSC line constitutes a collection of inherited variants and variants arising during the reprograming, passaging, and genome engineering when relevant. Several studies have shown that SNVs and small insertion and deletions (indels) are largely derived from parental cells, whereas large structural alterations (chromosomal rearrangements and CNVs >1Mb) arise during the reprograming or passaging[33-38]. However, less is known about the occurrence of other structural variants (SV), such as smaller CNVs (<1Mb). A typical human genome is estimated to carry 25,000 SVs and a large fraction of these are small structural events which are rarely reported in hPSC genome characterization studies due to the limited resolution of approaches chosen for the analysis[32,49,52]. For instance, in the genomic characterization of KOLF2.1J iPSCs only large events (>5Mb) and small indels (<50bp) were confidently assessed[13]. Although the genome sequencing data performed at 30x mean coverage should, in principle, provide the resolution to detect most classes of SVs, there is no single algorithm that can comprehensively call all types of SVs. The algorithm originally chosen for KOLF2.1J genome sequencing analysis depends on the detection of split reads and discordant read-pairs, which are often filtered out due to poor quality indicators or low read coverage. Therefore, a large fraction of KOLF2.1J iPSC genome is likely still uncharacterized.

Nonetheless, even with a complete catalogue of genetic variation, their effects on hPSC performance, differentiation and interaction with other genetic variants for disease modeling would still be difficult to ascertain. Yet, evidence derived from population and disease genomic studies can provide some functional insights. Indeed, mutational constraint metrics indicate that disruptive variants in *JARID2* and *ASTN2* are rare in human genomes thus predicting that KOLF2.1J CNVs that delete *JARID2* and *ASTN2* coding exons are likely deleterious for humans. Accordingly, mounting evidence show that *ASTN2* haploinsufficiency provides risk to neurodevelopmental disorders, and heterozygous loss of *JARID2* is the cause of an autosomal dominant neurodevelopmental syndrome. Furthermore, *JARID2* acts as a hematopoietic tumor suppressor through Cyclin D1 repression and deletions drive transformation to secondary acute myeloid leukemias[53-55]. Together, genetic constraint and disease associations of *JARID2* are reminiscent of *ARID2*, in which a heterozygous 19bp deletion was corrected to generate the KOLF2.1J iPSC line from KOLF2-C1s[13]. Of note, both ARID2 and JARID2 belong to the ARID family, comprised of 15 proteins with diverse functions in development, proliferation and tissue specific gene expression[56].

JARID2, in particular, is a regulatory subunit of the Polycomb Repressive Complex 2 (PRC2)[57-60], which maintains transcriptional repression of developmental genes by fostering chromatin condensation through catalyzing histone H3 lysine 27 (H3K27) methylation[45]. Genetic disruption of PRC2 is a major cause of cancer and developmental disorders in humans[61-63] and postnatal depletion in the mouse brain leads to neurodegeneration due to the de-repression of death promoting genes[64], while full body disruption causes embryonic lethally[65-67]. Likewise, *Jarid2* deletion leads to gastrulation arrest in *Xenopus* and knock out mice embryos die at embryonic day 10.5-15.5 due to neural tube, heart and hematopoietic defects that vary in severity depending on the genetic background of the mouse strain[58,68,69]. Data from mouse ESC (mESC) cultures also support that *in vitro* differentiation is *Jarid2* dose sensitive, given that homozygous and

7

heterozygous *Jarid2* knock out mESCs show delayed differentiation to the three germ layer derivatives, including neurons[57,59,70,71]. In contrast, little is known about the consequences of *JARID2* deletion in hPSCs. The functional characterization of KOLF2.1J iPSCs, which carry only one intact copy of *JARID2,* indicates that global pluripotency and neuronal differentiation phenotypes are comparable to other hPSCs, with the exception of few parameters (i.e. reduced amplitude of evoked excitatory postsynaptic currents in derived neurons)[13]. Furthermore, a recent study shows that the proteome remodeling trajectory during neural induction is similar between KOLF2.1J iPSCs and H9 hESCs, although at steady state level the two lines show profound protein expression differences[51]. Whether KOLF2.1J iPSC-derived neurons display specific neurological disease phenotypes or higher susceptibility to genetic variants associated to neurological disorders remain unknown, but our findings urge to consider these possibilities when working with KOLF2.1J iPSCs.

The identification of CNVs affecting *JARID2*, *DTNBP1* and *ASTN2* may limit the potential of KOLF2.1Js as a "global" reference iPSC line. However, it is likely that such a "global" reference iPSC may never exist. While some users may argue that the phenotypic consequences of carrying at least three potentially deleterious variants are overturned by the overall good growth, editing and differentiation performance of the KOLF2.1J line, users studying neural phenotypes may want to search alternative genetic backgrounds "free" of deleterious variants known to be associated with neurological diseases. This search by itself raises several issues including the lack of a comprehensive catalogue of genomically characterized iPSC lines. Furthermore, although current studies suggest that ~5% of the apparently healthy human genomes carry a rare loss of function variant known to be deleterious for human development[72], it is likely that this percentage increases over time as we continue surveying "healthy" and disease associated human genomes with continuously developing sequencing and computational technologies, and therefore no perfect iPSC line exists. While we fill these gaps in knowledge, we propose that the field should work together to generate a catalogue of iPSC lines which includes a comprehensive genome sequencing analysis that enable users with limited resources and computational expertise to interrogate their genes of interest and retrieve unambiguous information about coverage and the type of genetic variants assessed and reported. A similar resource has already been launched for hESCs (https://hscgp.broadinstitute.org/hscgp) and serves as a great example. Ideally, this resource should enable the selection of iPSC lines that do not carry deleterious mutations clearly associated with neurological disorders (or disorders of interest to other studies).

Limitations of the study
Our study is focused on the five new CNVs incidentally identified by an SNP array in KOLF2.1J iPSCs but our analysis is not comprehensive due to limits in the resolution of the SNP array and targeted analysis applied on the genome sequencing data. Based on estimates for variation in human genomes we expect that KOLF2.1J iPSCs will carry many more SVs that will be revealed with future additional sequencing methods and computational algorithms. Furthermore, while our data show CNVs associated with rare neurodevelopmental disorders in KOLF2.1J iPSCs, the potential of KOLFs to acquire overall characteristic features of neural cell lines upon diverse differentiation protocols has been recapitulated by multiple independent laboratories. Thus, the impact of these CNVs for pluripotency and neural cell phenotypes, as well as their interaction with

other variants that researchers may want to introduce to generate experimental models of neurological disorders, remain unknown.

Acknowledgments

Author contribution
C.G-D., J.E.P., M.E.K., B.D., G.G.C., S.L., A.G., J.A.M., and T.R. conducted experiments. C.G-D, J.E.P., K.W., J.G. and N.A. designed experiments and wrote the paper. All the authors contributed to data interpretation and edited and reviewed the paper. K.W. supervised the genome sequencing reanalysis. J.G. supervised the SNP array CNV analysis. N.A. supervised and coordinated the project.

Declaration of Interests
The authors declare no competing interests.
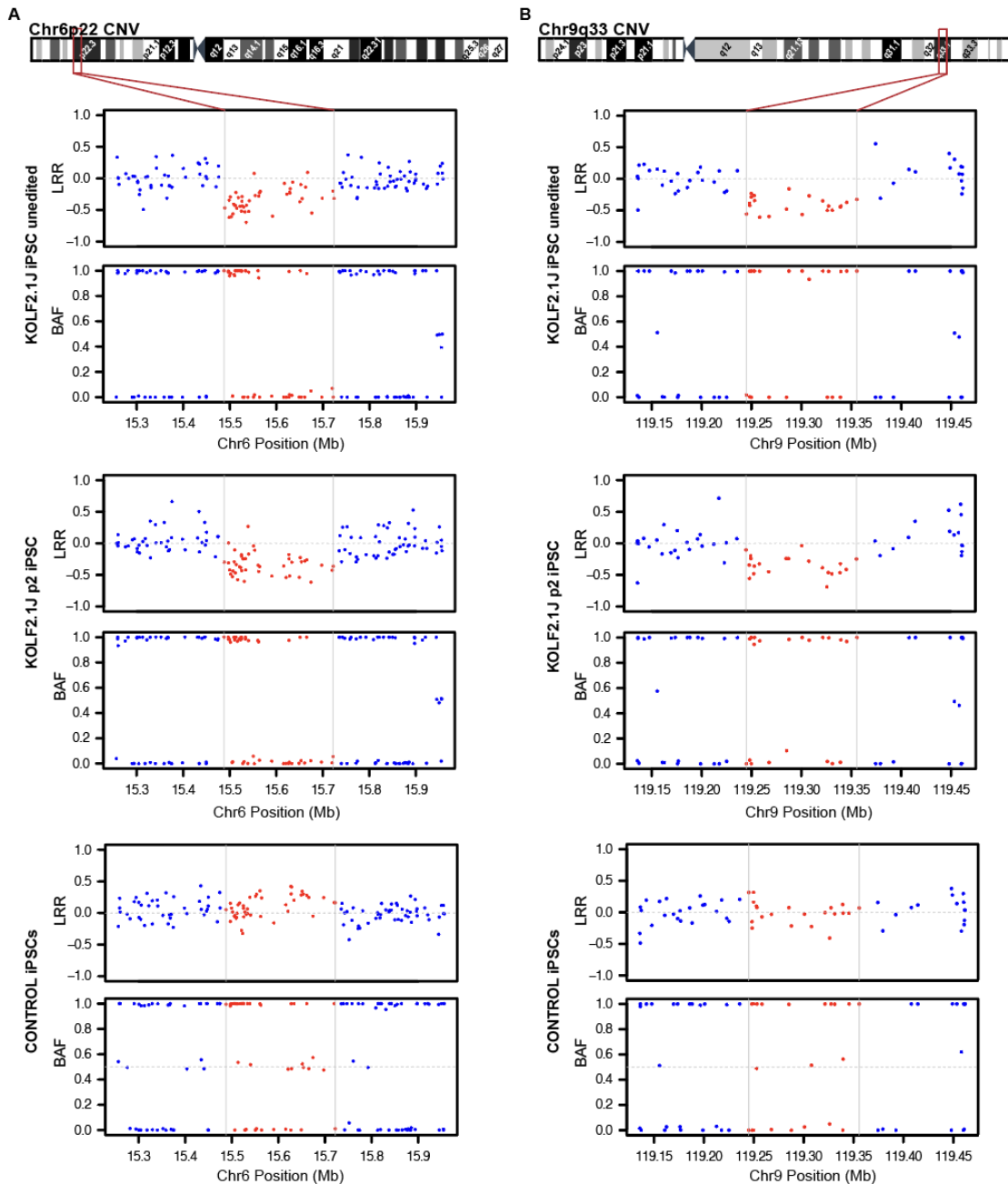
Figures

Figure 1

Figure 1: High density SNP array uncovers CNVs affecting coding genes in Chr6p22 and Chr9q33 in KOLF2.1J iPSCs. (A) Chromosome 6 cytoband schematics (top) and Log R Ratio (LRR) and B Allele frequency (BAF) plots (bottom) show reduction of signal intensity and a loss of heterozygosity in 6p22 region of unedited and p2 KOLF2.1J iPSCs compared to a control iPSC line. (**B)** Chromosome 9 cytoband schematics (top) and LRR and BAF plots (bottom) show reduction of signal intensity and a loss of heterozygosity in 9q33 region of KOLF2.1J iPSCs unedited and at p2 compared to a control iPSC line.
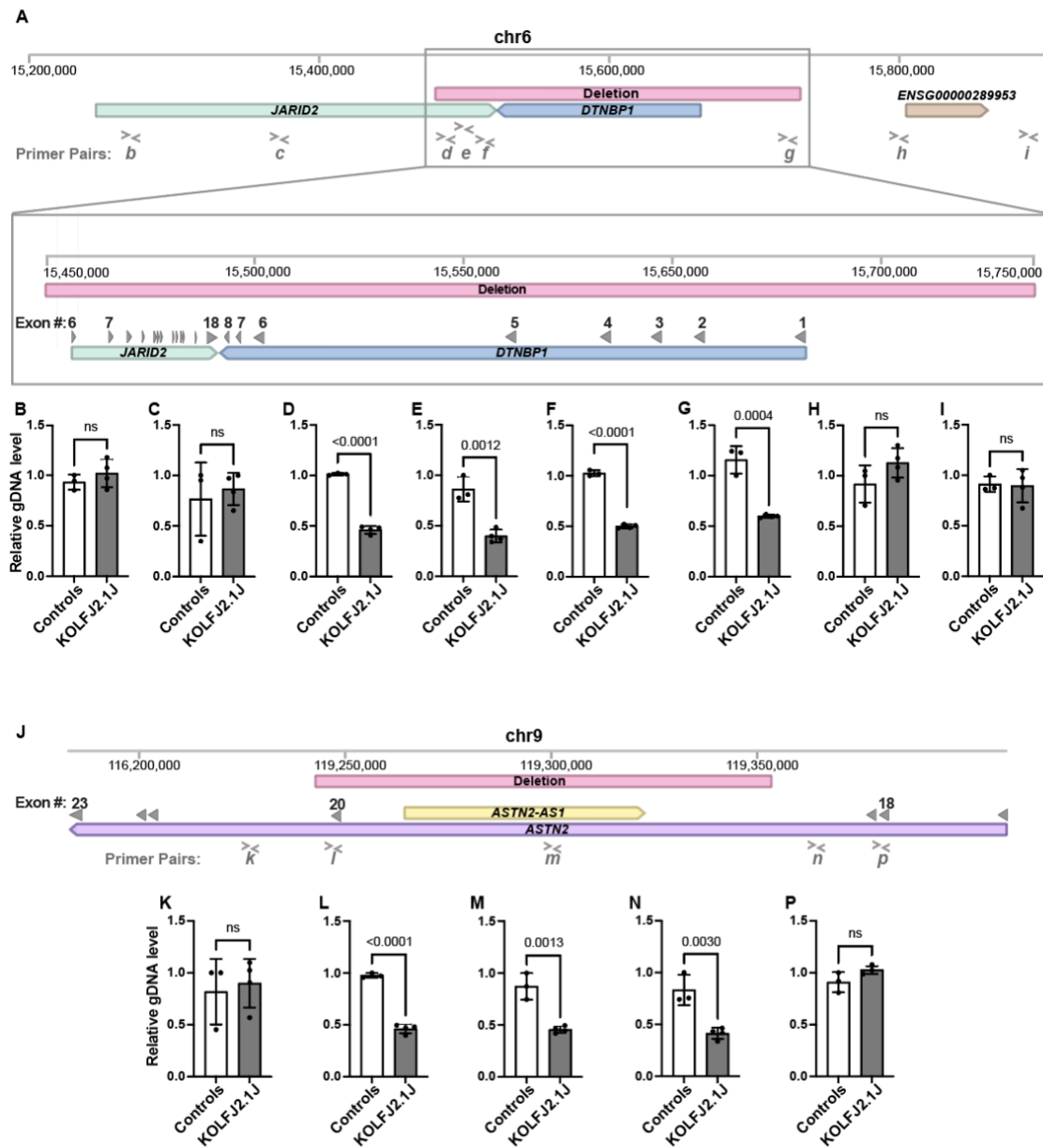
Figure 2

Figure 2: gDNA qPCR analysis confirms the presence of chr6p22 and chr9q33 CNV in KOLF2.1J iPSCs. (A) Schematic representation of Chromosome 6 (Chr6:15,200,000-15,900,000). Window illustrates the 0.2 Mb deleted region (red bar) and affected genes, *JARID2* (green bar) and *DTBNP1* (blue bar). **(B-I)** Genomic DNA qPCR results showing half levels of amplification within the Chr6 deleted CNV region in KOLF2.1J compared to control iPSCs (D-G) and regions upstream and downstream of the deletion (A-B and H-I). **(J)** Schematic representation of Chromosome 9 (Chr9:116,200,000-119,400,000) with the 0.1 Mb deleted region (red bar) and affected genes, *ASTN2* (purple bar) and *ASTN-AS1* (yellow bar). **(K-P)** Genomic DNA qPCR results showing half levels of amplification within the Chr9 deleted CNV region in KOLF2.1J compared to control iPSCs (L-N) and regions upstream and downstream of the deletion (K and P). Graphs show mean +/-SD of n=3 control hPSC lines and n=4 KOLF2.1J iPSC independent stocks. p-values were calculated with a two-sided unpaired t-test. ns=non-significant. Grey filled arrowheads indicate exons. > and < symbols indicate qPCR primer pair positions.
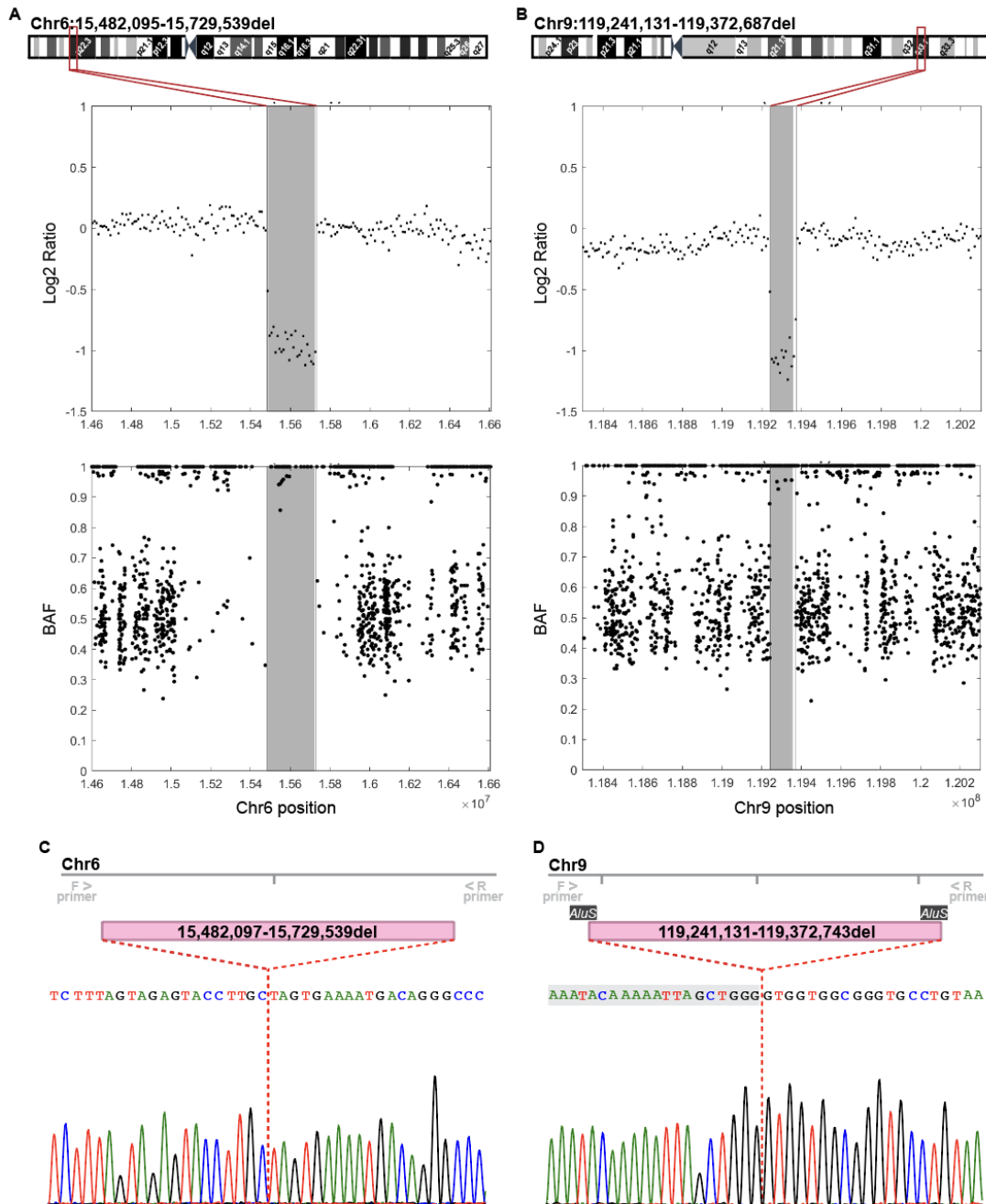
11

Figure 3

Figure 3: Genome sequencing reanalysis confirms CNVs at Chr6p22 and Chr9q33. (A, B) Log2 ratio and BAF plots show reduction of signal intensity (Log2 ratio) and a loss of heterozygosity (BAF) in Chr6p22 (A) and Chr9q33 (B) CNV regions compared to diploid flanking regions. Shadowed area represents the CNV defined by the SNP array and vertical lines the base-resolution breakpoints determined from the genome sequencing data. Chromosome 6 and 9 cytoband schematics with the CNV breakpoints at base resolution are shown on the top. **(C, D)** Sanger sequencing of PCR amplicons including ~200bp up- and downstream of the predicted breakpoints confirm deletion CNVs in Chr6(C) and Chr9 (D) and adjust the breakpoint positions. Nucleotides shadowed in grey (D) are part of the 56 bp long microhomology domain of two *AluS* repetitive elements localized in Chr9 CNV breakpoints.
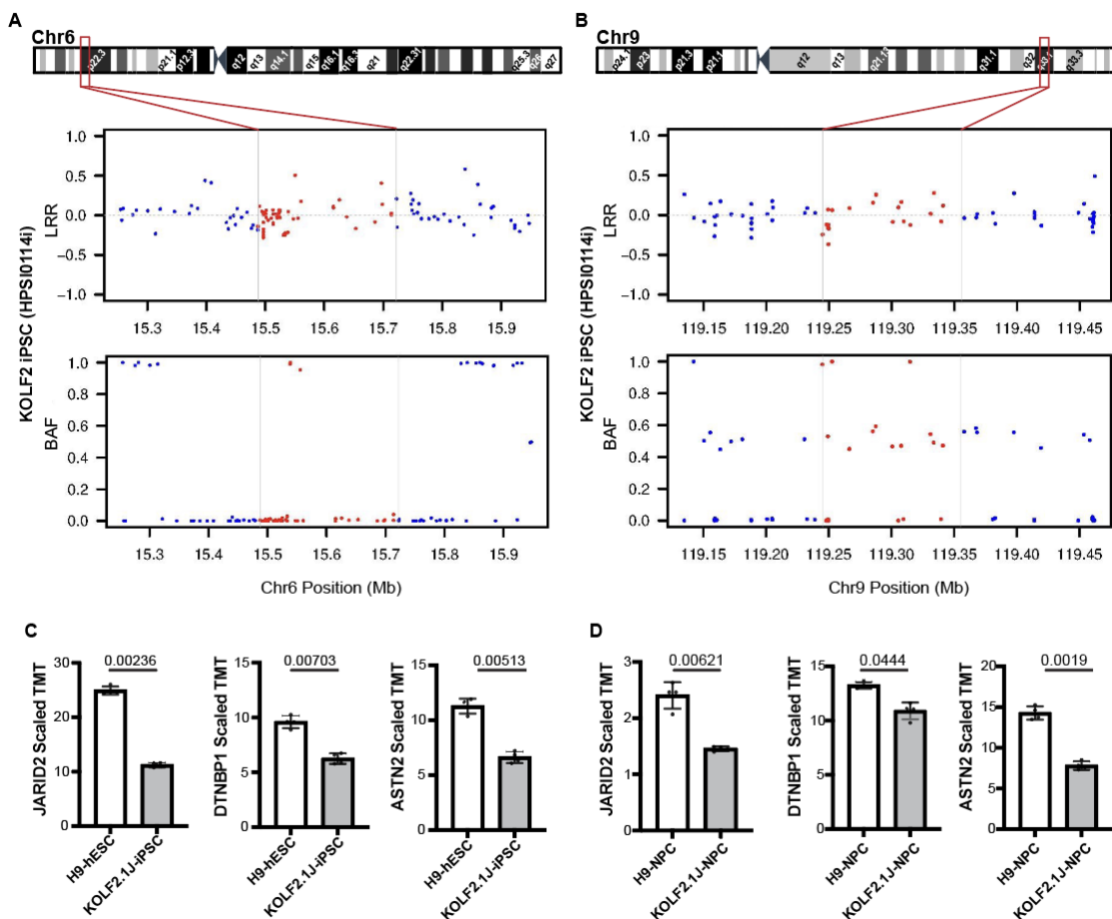
12

Figure 4

Figure 4: Chr6p22 and Chr9q33 CNVs arose *in vitro* and affect the expression of *JARID2*, *DTNBP1* and *ASTN2* genes. (A, B) LRR and BAF plots of the KOLF2 iPSC line SNP array show diploid Chr6 (A) and Chr9 (B) regions, indicating that the CNVs arose in culture between KOLF2 to KOLF2.1J generation. Horizontal lines and red dots depict SNPs that fall within KOLF2.1J CNV regions. **(C, D)** Quantitative proteomic data retrieved from Nam et al[51] show that JARID2, DTNBP1 and ASTN2 protein levels are ~half in KOLF2.1J compared to the reference H9 line at both pluripotency (C) and neural progenitor stage (D). Graphs show mean +/-SD of n=4 cultures and q-values as calculated in the reference proteomic analysis with Welsch's t-test and FDR correction for multiple comparisons[51].

13

Tables

Table 1: CNVs detected in KOLF2.1J iPSC lines by high density SNP array.

Sample	Cytoband	Aprox. lengh	Aprox. Coordinates	CNV type	Affected Genes
KOLF2.1J Unedited	3p13	0.1 Mb	Chr3:72,289,657-72,404,542	Deletion	No coding genes
	3q14	0.3 Mb	Chr3:61,242,795-61,528,270	Duplication	No coding genes
	6p22	0.2 Mb	Chr6:15,487,649-15,722,102	Deletion	*JARID2* and *DTNBP1*
	9p22	0.1 Mb	Chr9:119,244,942-119,355,584	Deletion	*ASTN2* and *ASTN2-AS1*
	18q22	0.1 Mb	Chr18:62,115,075-62,264,706	Duplication	No coding genes
KOLF2.1J Edited 1	3p13	0.1 Mb	Chr3:72,289,657-72,404,542	Deletion	No coding genes
	3q14	0.3 Mb	Chr3:61,242,795-61,528,270	Duplication	No coding genes
	6p22	0.2 Mb	Chr6:15,487,649-15,722,102	Deletion	*JARID2* and *DTNBP1*
	9p22	0.1 Mb	Chr9:119,244,942-119,355,584	Deletion	*ASTN2* and *ASTN2-AS1*
	18q22	0.1 Mb	Chr18:62,115,075-62,264,706	Duplication	No coding genes
KOLF2.1J p2	3p13	0.1 Mb	Chr3:72,289,657-72,404,542	Deletion	No coding genes
	3q14	0.3 Mb	Chr3:61,242,795-61,528,270	Duplication	No coding genes
	6p22	0.2 Mb	Chr6:15,487,649-15,722,102	Deletion	*JARID2* and *DTNBP1*
	9p22	0.1 Mb	Chr9:119,244,942-119,355,584	Deletion	*ASTN2* and *ASTN2-AS1*
	18q22	0.1 Mb	Chr18:62,115,075-62,264,706	Duplication	No coding genes
KOLF2.1J p3	3p13	0.1 Mb	Chr3:72,289,657-72,404,542	Deletion	No coding genes
	3q14	0.3 Mb	Chr3:61,242,795-61,528,270	Duplication	No coding genes
	6p22	0.2 Mb	Chr6:15,487,649-15,722,102	Deletion	*JARID2* and *DTNBP1*
	9p22	0.1 Mb	Chr9:119,244,942-119,355,584	Deletion	*ASTN2* and *ASTN2-AS1*
	18q22	0.1 Mb	Chr18:62,115,075-62,264,706	Duplication	No coding genes
KOLF2.1J-NGN2 p10	3p13	0.1 Mb	Chr3:72,289,657-72,404,542	Deletion	No coding genes
	3q14	0.3 Mb	Chr3:61,242,795-61,528,270	Duplication	No coding genes
	6p22	0.2 Mb	Chr6:15,487,649-15,722,102	Deletion	*JARID2* and *DTNBP1*
	9p22	0.1 Mb	Chr9:119,244,942-119,355,584	Deletion	*ASTN2* and *ASTN2-AS1*
	18q22	0.1 Mb	Chr18:62,115,075-62,264,706	Duplication	No coding genes

Table 2: Constriction metrics and OMIM phenotypes of genes affected by KOLF2.1J CNVs compared to *ARID2*.

Gene	OMIM #	pHaplo[29]	pTriplo[29]	Decipher p(HI)[30]	GenomAD[28] pLI	GnomAD[28] pLOF o/e
JARID2	620098	0.997	0.992	0.58982626|12.14%	1	0.09 (0.05 - 0.19)
DTNBP1	614076	0.769	0.272	0.148024092|43.28%	0	0.59 (0.37 - 0.98)
ASTN2		0.884	0.672	0.952118309|2.19%	1	0.14 (0.08 - 0.25)
ARID2	617808	0.997	0.981	0.622938783|11.02%	1	0.04 (0.02 - 0.1)

Table 3: CNVs detected in KOLF2 iPSCs by the analysis of publicly available HumanCoreExome-12_v1 SNP array data.

Sample	CNVs called in KOLF2.1J	CNV called by PennCNV	CNV called by visual review
KOLF2 iPSCs	3p13del	No	Possible mosaic deletion
	3q14dup	Yes	Yes
	6p22del	No	No
	9p22del	No	No
	18q22dup	Yes	Yes

15

Methods

hPSC research

iPSCs and ESCs used in this study were previously generated from material obtained under informed consent and appropriate ethical approvals. All the work was performed under approval from the Children's Hospital of Philadelphia Institutional Review Board (IRB) and Institutional Biosafety Committee (IBC) and following the ISSCR standards for human stem cell use in research.

hPSC culturing

KOLF2.1J human pluripotent stem cells (JAX, JIPSC1000) were cultured on Matrigel (Corning, #354277) coated plates in mTeSR1 media (STEMCELL Technologies, #85850). The medium was changed every day and cells were passaged when 70% confluent using ReleSR (STEMCELL Technologies, #100-0484) at a 1:20 ratio. When thawing KOLF2.1Js 10 μM Rho-associated protein kinase (ROCK) inhibitor (Y-27632) (Tocris, #1254) was added into the culture media overnight.

Genomic DNA extraction and quantitative PCR

hPSC pellets were processed for gDNA extraction using the GeneJET Genomic DNA Purification Kit (Thermo Scientific, #K0721). Eluted gDNA was quantified using NanoDrop. gDNA qPCR was performed in a 10 μL reaction volume containing 12ng of gDNA, 0.5 μM of forward primers, 0.5 μM of reverse primers (Supplementary Table) and 1X Power SYBR Green PCR Master Mix (Thermo Fisher, #4367659) in a 384-well plate (Thermo Fisher, #AB1384W) and ran using the CFX384 Touch Real-Time PCR Detection System (Bio-Rad, #1855484). Primers were designed with Primer3[73].

Genomic DNA PCR and Sanger Sequencing

Primers targeting sequences upstream and downstream the breakpoints of Chr6 and Chr9 deletions CNVs were designed for PCR and Sanger Sequencing (Supplementary Table). PCR was performed in a 30 μL reaction volume containing 1X Colorless GoTaq Reaction Buffer (Promega, #M792B), 200 μM dNTPs (Invitrogen, #18427013), 0.5 μM forward and 0.5 μM reverse primers, GoTaq DNA Polymerase (Promega, #M3008) and 1μL of template gDNA. To prepare PCR products for Sanger Sequencing we enzymatically purified the products using ExoSAP-IT PCR Product Cleanup Reagent (Applied Biosystems, #78201.1.ML). Purified PCR products were mixed with the forward primer at 1.67 μM and Sanger sequenced in Genewiz.

SNP genotyping array

Sample gDNA was hybridized to an Illumina Infinium GSA-24v3 BeadChip, consisting of ~650,000 short invariant 50 mer oligonucleotide probes conjugated to silica beads. After hybridization a single-base, hybridization-dependent extension reaction was performed at the target SNP which label alternate alleles (herein denoted A and B) with different fluorophores. Arrays were loaded onto an iScan System and scanned to extract data. DMAP files enabled the identification of bead locations on the BeadChip and quantification of the signal associated with each bead. Raw fluorescence intensity from the two-color channels was processed into a discrete genotype call (normalized to continuous value 0-1 B-Allele Frequency (BAF)) and the total intensity from both channels (normalized to continuous value with median=0 Log R Ratio (LRR)) at each SNP informed for copy number. The resulting raw Intensity Data files (*.idat) were converted to Genotype Call files (*.gtc) and then into BAF and LRR signal files.

CNV detection from SNP genotyping array
PennCNV was used as the main CNV detection algorithm. PennCNV calls were filtered to include CNVs with number of SNPs supporting >=20 and length >=100,000 and Segmental Duplication track coverage < 0.5. Related cell line clones CNV calls were compared to ensure consistency in CNV calling. All genomic coordinates are in GRCh37/hg19 human build version. For KOLF2 CNV analysis, genotyping data containing SNP based BAF and LRR values was downloaded from European Nucleotide Archive (HPSI0114i-kolf_2.wec.gtarray.HumanCoreExome-12_v1_0.20141111.genotypes.vcf.gz), which contains data for 511,966.

Structural variant analysis of genome sequencing
KOLF2.1J genome sequencing data was downloaded from ADWB (https://doi.org/10.34688/KOLF2.1J.2021.12.14). Illumina paired-end whole genome sequencing reads were aligned to the hg19 reference using minimap2[74] version 2.17-r941 in paired-end short read mode. To validate CNV regions identified from SNP arrays, we calculate coverage log2 ratios and B-allele frequencies. For the log2 ratio, we use an in-house software tool which employs a 10kb non-overlapping sliding window spanning the CNV and calculates the log2 ratio between mean window and mean chromosome coverage. This value represents fold change in coverage, with positive values indicating high coverage with potential duplications, and negative values low coverage with potential deletions. B-allele frequency (BAF) is calculated as the fraction of reads supporting non-reference alleles at SNP locations. In regions with normal diploid copy number, BAF values cluster around 0.5 for heterozygous variants, and cluster around 1 for homozygous variants. An increased dispersion around 0.67 or 0.33 indicates allelic imbalance corresponding to a duplication, and the lack of clustering around 0.5 indicates allele loss corresponding to a deletion. We use DeepVariant[75] to call SNPs and obtain BAFs. We then use BCFTools view (version 1.9)[76] to filter SNP locations by variant quality score >30 and read depth >10. Finally, to obtain base-resolution breakpoints for each CNV, we run the DELLY[77] SV caller on the alignment file and filter results by region, SV type and a minimum length of 100kb to obtain a single SV candidate for each region.
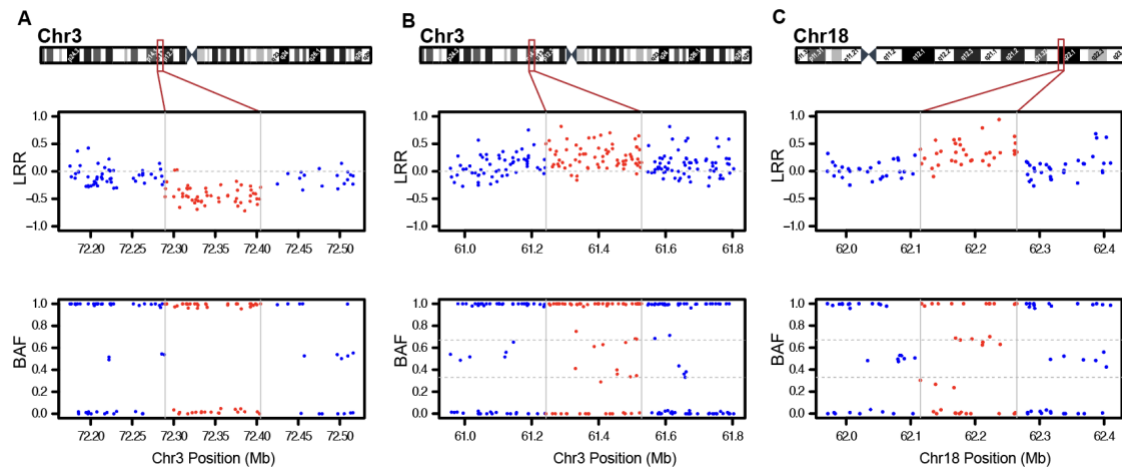
Supplemental Information
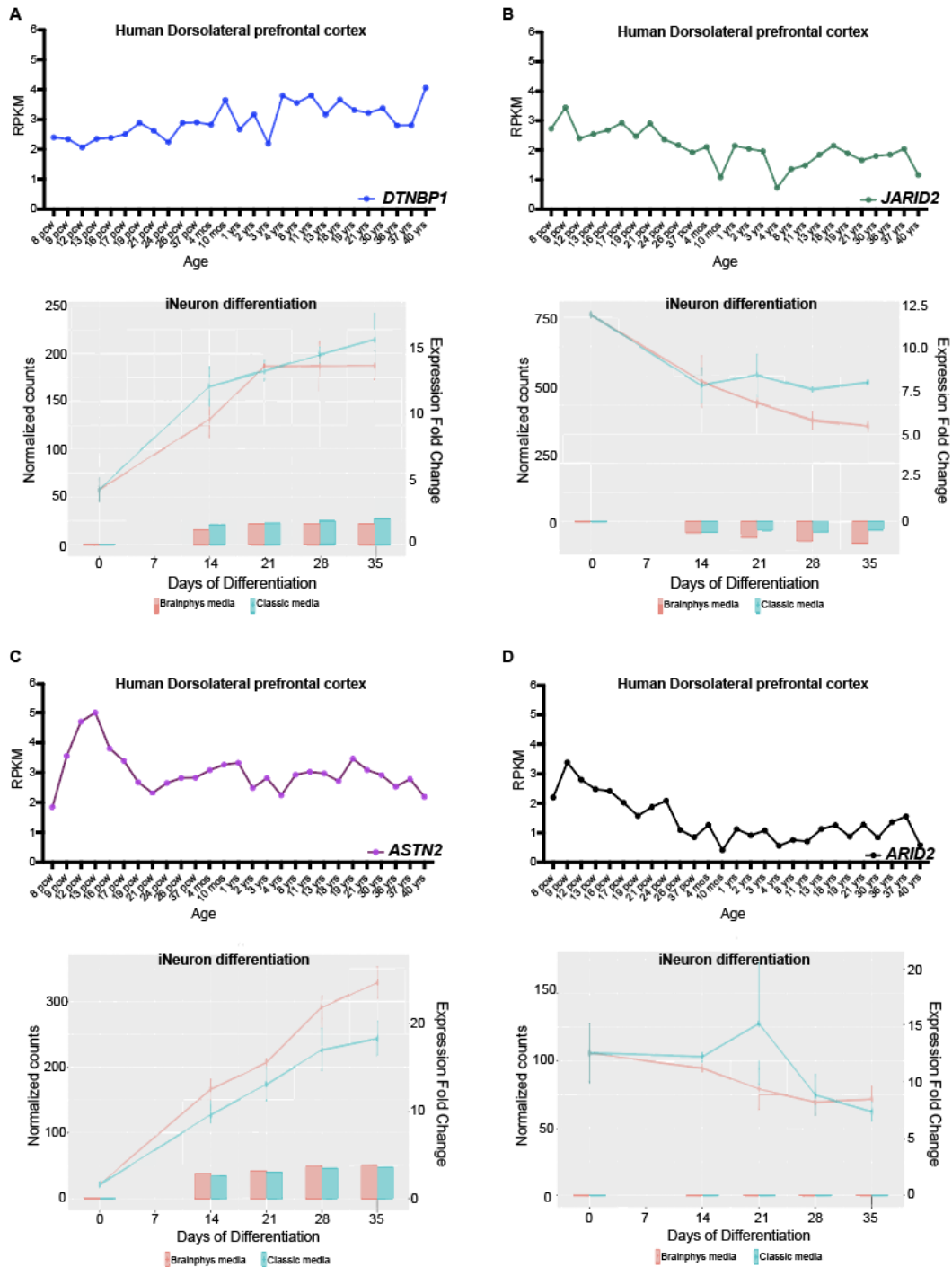
Supplemental Figure 1
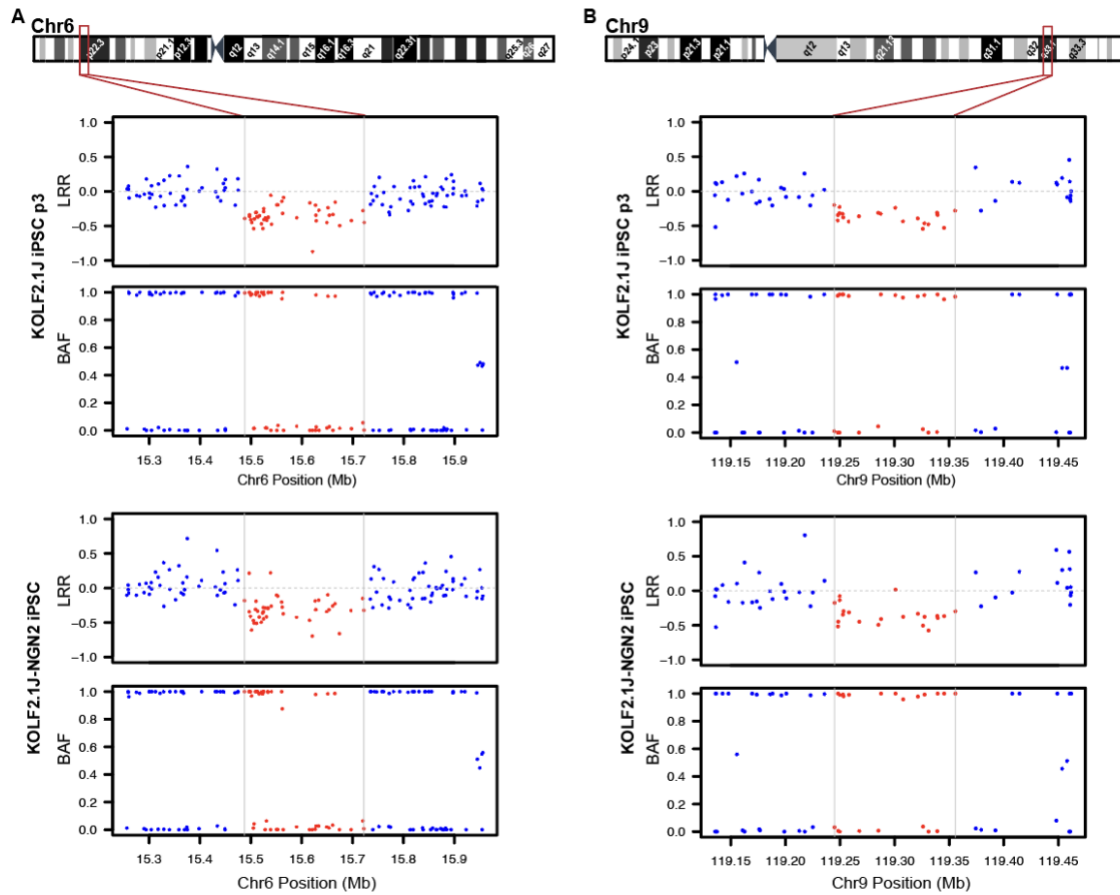
Supplemental Figure 1: High density SNP array uncovers CNVs affecting coding genes in Chr6p22 and Chr9q33 in a stock of KOLF2.1J iPSCs and KOLF2.1J with doxycycline inducible *NGN2* transgene. (A) Chromosome 6 cytoband schematics (top) and Log R Ratio (LRR) and B Allele frequency (BAF) plots (bottom) show reduction of signal intensity and a loss of heterozygosity in 6p22 region of KOLF2.1J iPSCs p3 and KOLF2.1J-NGN2. (**B**) Chromosome 9 cytoband schematics (top) and LRR and BAF plots (bottom) show reduction of signal intensity and a loss of heterozygosity in 9q33 region of KOLF2.1J iPSCs p3 and KOLF2.1J-NGN2.

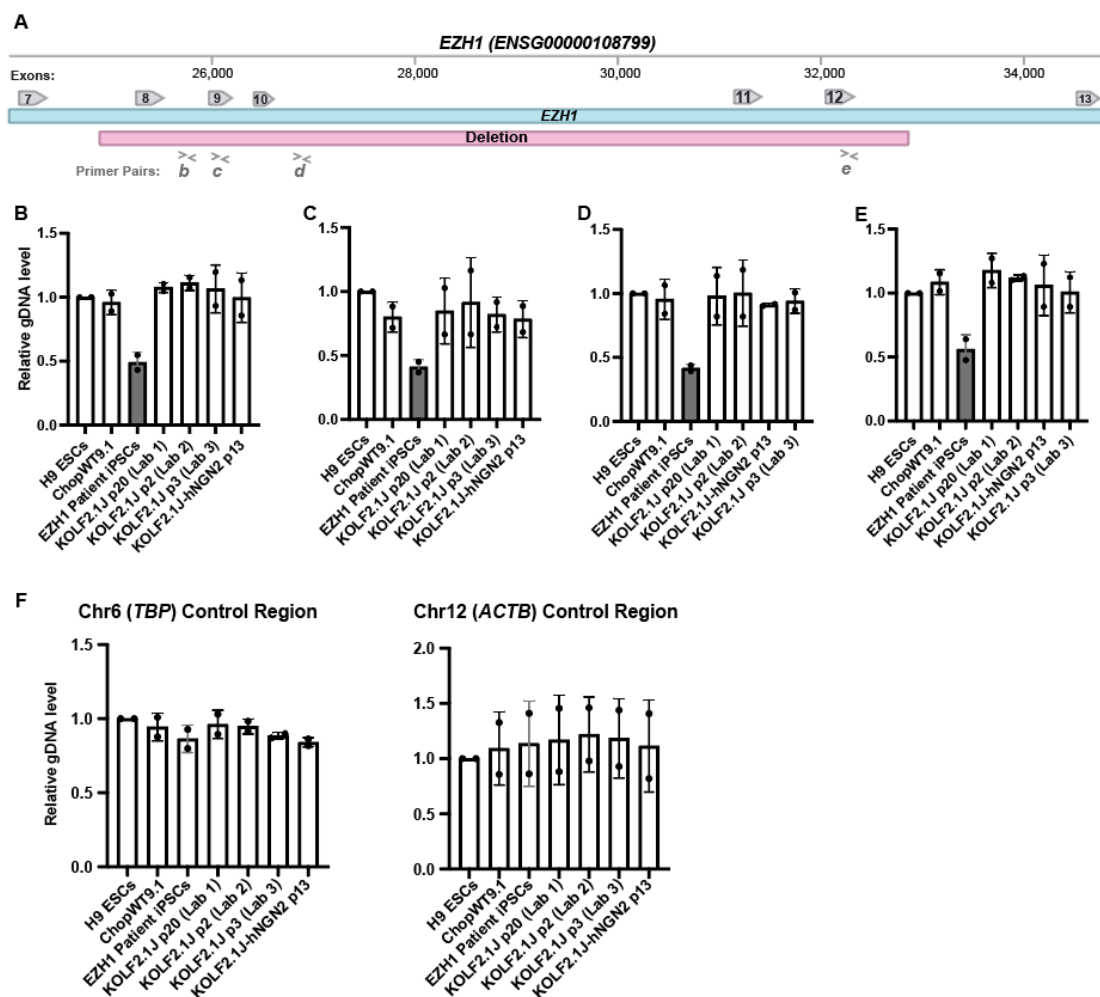**Supplemental Figure 2**

Supplemental Figure 2: CNVs detected by SNP array in Chr3 and Chr18 do not overlap with coding regions. (A-B) Chromosome 3 cytoband schematics (top) and LRR and BAF plots (bottom) showing reduction of signal intensity and a loss of heterozygosity in 3p13 deleted region (A) and increased signal intensity and altered BAF in 3p14 duplicated region (B) of KOLF2.1J iPSCs compared to a control iPSC line. (**C**) Chromosome 18 cytoband schematics (top) and LRR and BAF plots (bottom) showing increased signal intensity and altered BAF in 18q22 duplicated region of KOLF2.1J iPSCs.

19

Supplemental Figure 3: Genes affected by Chr6p22 and Chr9q33 are expressed in neurons in the developing and adult brain. (A-D) DTNBP1 (A), JARID2 (B), ASTN2 (C) and ARID2 (D) expression in developing and adult human prefrontal cortex and in hPSC to iNeuron differentiation. Human prefrontal cortex expression data is retrieved from Brainspan (https://www.brainspan.org/rnaseq/search/index.html) and iNeuron differentiation expression plots from the web app created by Connor Ludwig, Kampmann Lab (https://kampmannlab.ucsf.edu/ineuron-rna-seq). Pcw; postconception week. Yrs; years.
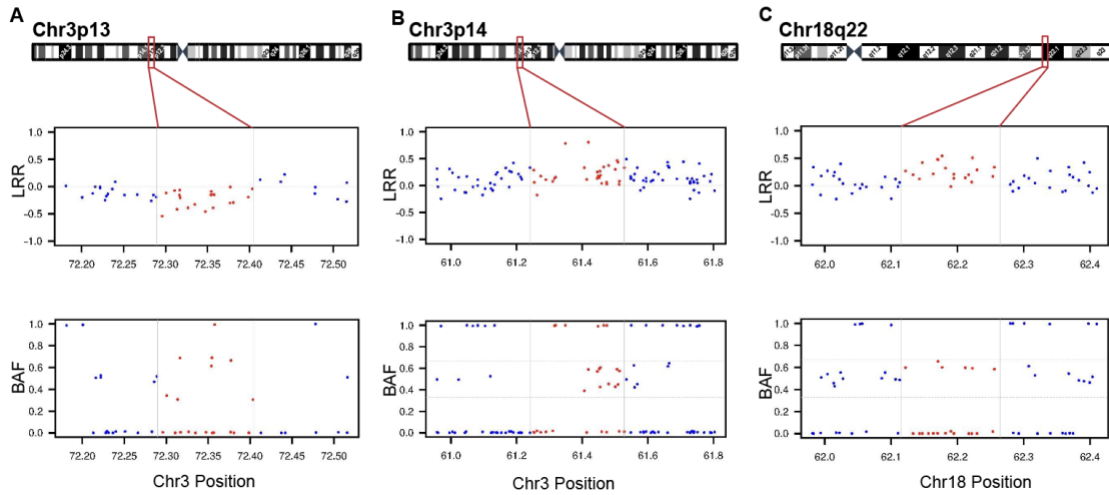
Supplemental Figure 4

Supplemental Figure 4: gDNA qPCR on an iPSC line with one copy of EZH1 exon 8-12 deletion demonstrates sensitivity of the technique for detection of hemizygous regions. **(A)** Schematic representation of *EZH1* exon 7-13. Red line illustrates the region deleted in a EZH1 patient iPSC line. Arrowheads (>, <) labeled with b-e indicate position of the primers used for gDNA qPCR. **(B-E)** gDNA qPCR results showing half levels amplification in regions deleted in EZH1 patient iPSCs compared to control and KOLF2.1Js that are expected to be diploid. **(F)** Randomly selected regions of the genome show similar amplification levels across all the hPSC lines.

21

Supplemental Figure 5
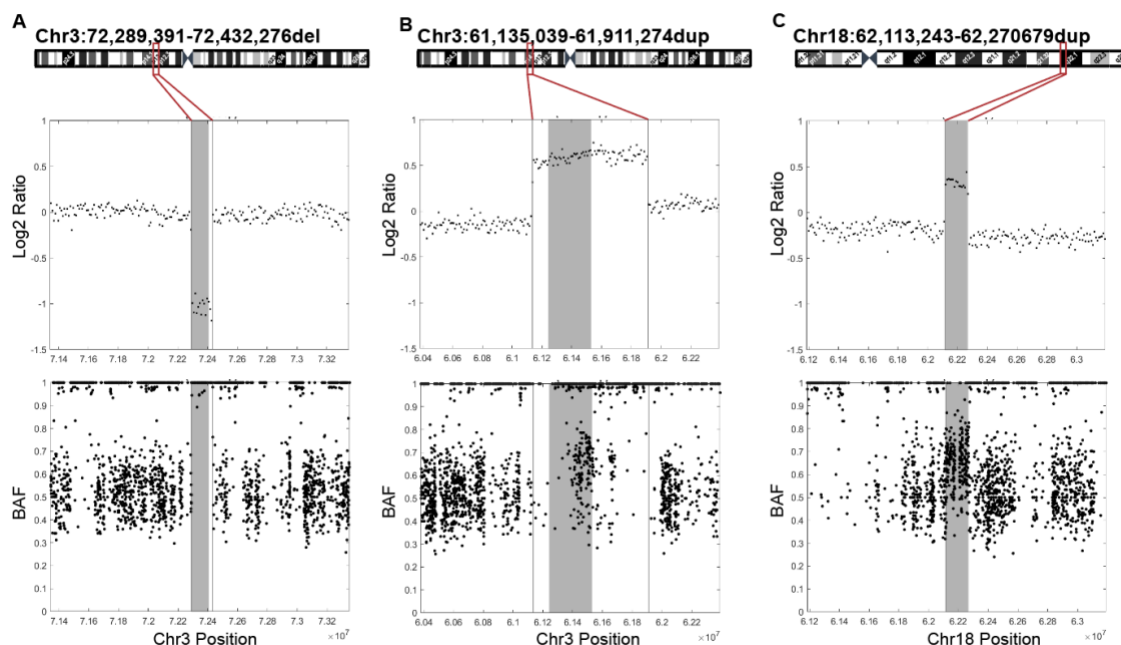
Supplemental Figure 5: Genome sequencing reanalysis confirms CNVs at Chr3p13, Chr3p14 and Chr18q22. (A-C). Chromosome cytoband schematics with base resolution break points of the CNVs (top) and LRR and BAF plots (bottom) obtained from the KOLF2.1J iPSC genome sequencing reanalysis. LRR plots show reduction of signal intensity and BAF plots show loss of heterozygosity in Chr3p13 (A) and gain of signal intensity and altered BAF in Chr3p14 (B) and Chr18q22 (C) CNV regions compared to diploid up and downstream regions. Shadowed area represents the CNV defined by the SNP array and vertical lines the base-resolution breakpoints determined from the genome sequencing data.

Supplemental Figure 6

Supplemental Figure 6: Chr3p13, Chr3p14 and Chr18q22 CNVs in KOLF2.1J were inherited from KOLF2 iPSC line. (A, B) LRR and BAF plots of the KOLF2 iPSC line SNP array deposited in HipSci, show that KOLF2 iPSC line carries two of the Chr3p14 (B) and Chr18q22 (C) CNVs in heterozygosity and is likely mosaic for Chr3p13 (A) deletion.

23

Supplementary Table 1: Description and sequence of primers used across the study

Primer Pair	Targeted Region (GRCh37/hg19)	Notes	Primer Orientation	Sequence (5'->3')	Product Length (bp)
1	chr6:15372726+15372836	Upstream deleted Chr6 region	Forward	gactacaggcgtgcaccaccac	111
			Reverse	aaggcggacggatcacaaggga	
2	chr6:15270358+15270456	Upstream deleted Chr6 region	Forward	tctgtccgtccgtccttccgtc	99
			Reverse	aggaggtggaggctgcagtgag	
3	chr6:15487546+15487631	Within deleted Chr6 region in *JARID2*: Exon 6	Forward	ggttccagcaggtcaacacggg	86
			Reverse	tcgctgtgcttctcctttgcgg	
4	chr6:15500050+15500126	Within deleted Chr6 region in JARID2	Forward	gctgtgtccccatcctggtct	77
			Reverse	gaaccacctcggcagacacagc	
5	chr6:15514462+15514592	Within deleted Chr6 region in *JARID2*	Forward	acacactctgggtagggacgcc	131
			Reverse	gcactgtggaaagggacgaggc	
6	chr6:15723401+15723550	Within deleted Chr6 region	Forward	acctctcaggctctcacgtttccc	150
			Reverse	agcaggagttggagccacagga	
7	chr6:15803765+15803892	Downstream Chr6 deleted region	Forward	tcccttccccctccccgacata	128
			Reverse	ttggttgggctgtccaccttgc	
8	chr6:15889438+15889552	Downstream Chr6 deleted region	Forward	gtctgagcgggactggaccctt	115
			Reverse	cgtttggcgtccctgtggagac	
9	chr9:119229497+119229617	Upstream deleted Chr9 region	Forward	tgcagccagcacacagcaaact	121
			Reverse	gccaagatcgcaccactgcact	
10	chr9:119249647+119249723	Within deleted Chr9 region: Exon 20	Forward	ccgtcgggctccagacacttga	77
			Reverse	gcagctggtcgaagccatggag	
11	chr9:119302810+119302918	Within deleted Chr9	Forward	gcttgggtgacagctttggcct	109
			Reverse	tagatggcctgcttggtgccct	
12	chr9:119366867+119366974	Within deleted Chr9	Forward	ccatggcccacactgcatgagg	108
			Reverse	agctgagaggatggtggcagca	
13	chr9:119382905+119382984	Downstream deleted Chr9 region	Forward	ccctccccaggagccacaatga	80
			Reverse	ggagtgccaagcctaggggaca	
14	chr17:40870724-40870832	Within *EZH1* Intron between Exon 8 and 9	Forward	ccaacatttgcccaacccagca	109
			Reverse	ggtctgccaagggaggatgggt	
15	chr17:40870536-40870611	Within *EZH1* Exon 9	Forward	agaccccaatgcacttcccct	76
			Reverse	ttgctcccgctgcacagacttg	
16	chr17:40869846-40869952	Within *EZH1* Intron between Exon 10 and 11	Forward	ggaacacaagccacgtgagcca	107
			Reverse	gagcggtcccacagtaactggg	
17	chr17:40864313-40864437	Within *EZH1* Exon 12	Forward	tagtggaagcaccctcggagcc	125
			Reverse	gcacgtcttggtccccagaagc	
18	chr6:170863510+170863606	Chr6 control region within *TBP*	Forward	cttcgcttccgctggcccatag	97
			Reverse	ctcagtgcagtggtggccttcg	
19	chr7:5596654-5596787	Chr7 control region within *ACTB*	Forward	tgctcttctctgtgcccctcc	134
			Reverse	acgggttcctactgtctggccc	
20	chr12:6644756+6644900	Chr12 control region within *GAPDH* (used for normalization)	Forward	ttccaccgcaaaatggcccctc	145
			Reverse	ccagacacccaatcctcccggt	
21	chr6:15481823+15729819	Chr6 Breakpoint Sanger Seq	Forward	taccttgtctttggctctgtcg	
			Reverse	atctgagccagcatggtagagg	
22	chr9:119240899+119372879	Chr9 Breakpoint Sanger Seq	Forward	ttctttcacctggactagaccc	
			Reverse	gcagaatttcgcttttgtcg	

24

References

1. 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. 10.1126/science.282.5391.1145.

2. 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. 10.1016/j.cell.2007.11.019.

3. Doudna, J.A., and Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science *346*, 1258096. 10.1126/science.1258096.

4. Zhang, S.C., Wernig, M., Duncan, I.D., Brustle, O., and Thomson, J.A. (2001). In vitro differentiation of transplantable neural precursors from human embryonic stem cells. Nat Biotechnol *19*, 1129-1133. 10.1038/nbt1201-1129.

5. Dimos, J.T., Rodolfa, K.T., Niakan, K.K., Weisenthal, L.M., Mitsumoto, H., Chung, W., Croft, G.F., Saphier, G., Leibel, R., Goland, R., et al. (2008). Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. Science *321*, 1218-1221. 10.1126/science.1158799.

6. Moya, N., Cutts, J., Gaasterland, T., Willert, K., and Brafman, D.A. (2014). Endogenous WNT signaling regulates hPSC-derived neural progenitor cell heterogeneity and specifies their regional identity. Stem Cell Reports *3*, 1015-1028. 10.1016/j.stemcr.2014.10.004.

7. Paulsen, B., Velasco, S., Kedaigle, A.J., Pigoni, M., Quadrato, G., Deo, A.J., Adiconis, X., Uzquiano, A., Sartore, R., Yang, S.M., et al. (2022). Autism genes converge on asynchronous development of shared neuron classes. Nature *602*, 268-273. 10.1038/s41586-021-04358-6.

8. Chambers, S.M., Fasano, C.A., Papapetrou, E.P., Tomishima, M., Sadelain, M., and Studer, L. (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nat Biotechnol *27*, 275-280. 10.1038/nbt.1529.

9. Fernandopulle, M.S., Prestil, R., Grunseich, C., Wang, C., Gan, L., and Ward, M.E. (2018). Transcription Factor-Mediated Differentiation of Human iPSCs into Neurons. Curr Protoc Cell Biol *79*, e51. 10.1002/cpcb.51.

10. 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. 10.1016/j.neuron.2013.05.029.

11. Strano, A., Tuck, E., Stubbs, V.E., and Livesey, F.J. (2020). Variable Outcomes in Neural Differentiation of Human PSCs Arise from Intrinsic Differences in Developmental Signaling Pathways. Cell Rep *31*, 107732. 10.1016/j.celrep.2020.107732.

12. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science (New York, N.Y.) *372*, eabf7117. 10.1126/science.abf7117.

13. Pantazis, C.B., Yang, A., Lara, E., McDonough, J.A., Blauwendraat, C., Peng, L., Oguro, H., Kanaujiya, J., Zou, J., Sebesta, D., et al. (2022). A reference human induced pluripotent stem cell line for large-scale collaborative studies. Cell Stem Cell *29*, 1685-1702 e1622. 10.1016/j.stem.2022.11.004.

14. Sittig, L.J., Carbonetto, P., Engel, K.A., Krauss, K.S., Barrios-Camacho, C.M., and Palmer, A.A. (2016). Genetic Background Limits Generalizability of Genotype-Phenotype Relationships. Neuron *91*, 1253-1259. 10.1016/j.neuron.2016.08.013.

15. Crabbe, J.C., Wahlsten, D., and Dudek, B.C. (1999). Genetics of mouse behavior: interactions with laboratory environment. Science *284*, 1670-1672. 10.1126/science.284.5420.1670.

16. Ramos, D.M., Skarnes, W.C., Singleton, A.B., Cookson, M.R., and Ward, M.E. (2021). Tackling neurodegenerative diseases with genomic engineering: A new stem cell initiative from the NIH. Neuron *109*, 1080-1083. 10.1016/j.neuron.2021.03.022.

17. Hildebrandt, M.R., Reuter, M.S., Wei, W., Tayebi, N., Liu, J., Sharmin, S., Mulder, J., Lesperance, L.S., Brauer, P.M., Mok, R.S.F., et al. (2019). Precision Health Resource of Control iPSC Lines for Versatile Multilineage Differentiation. Stem Cell Reports *13*, 1126-1141. 10.1016/j.stemcr.2019.11.003.

18. Verberne, E.A., Goh, S., England, J., van Ginkel, M., Rafael-Croes, L., Maas, S., Polstra, A., Zarate, Y.A., Bosanko, K.A., Pechter, K.B., et al. (2021). JARID2 haploinsufficiency is associated with a clinically distinct neurodevelopmental syndrome. Genet Med *23*, 374-383. 10.1038/s41436-020-00992-z.

19. Cadieux-Dion, M., Farrow, E., Thiffault, I., Cohen, A.S.A., Welsh, H., Bartik, L., Schwager, C., Engleman, K., Zhou, D., Zhang, L., et al. (2022). Phenotypic expansion and variable expressivity in individuals with JARID2-related intellectual disability: A case series. Clin Genet *102*, 136-141. 10.1111/cge.14149.

20. Di Benedetto, D., Di Vita, G., Romano, C., Giudice, M.L., Vitello, G.A., Zingale, M., Grillo, L., Castiglia, L., Musumeci, S.A., and Fichera, M. (2013). 6p22.3 deletion: report of a patient with autism, severe intellectual disability and electroencephalographic anomalies. Mol Cytogenet *6*, 4. 10.1186/1755-8166-6-4.

21. Baroy, T., Misceo, D., Stromme, P., Stray-Pedersen, A., Holmgren, A., Rodningen, O.K., Blomhoff, A., Helle, J.R., Stormyr, A., Tvedt, B., et al. (2013). Haploinsufficiency of two histone modifier genes on 6p22.3, ATXN1 and JARID2, is associated with intellectual disability. Orphanet J Rare Dis *8*, 3. 10.1186/1750-1172-8-3.

22. Glessner, J.T., Wang, K., Cai, G., Korvatska, O., Kim, C.E., Wood, S., Zhang, H., Estes, A., Brune, C.W., Bradfield, J.P., et al. (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature *459*, 569-573. 10.1038/nature07953.

23. Lesch, K.P., Timmesfeld, N., Renner, T.J., Halperin, R., Roser, C., Nguyen, T.T., Craig, D.W., Romanos, J., Heine, M., Meyer, J., et al. (2008). Molecular genetics of adult ADHD: converging evidence from genome-wide association and extended pedigree linkage studies. J Neural Transm (Vienna) *115*, 1573-1585. 10.1007/s00702-008-0119-3.

24. Lionel, A.C., Tammimies, K., Vaags, A.K., Rosenfeld, J.A., Ahn, J.W., Merico, D., Noor, A., Runke, C.K., Pillalamarri, V.K., Carter, M.T., et al. (2014). Disruption of the ASTN2/TRIM32 locus at 9q33.1 is a risk factor in males for autism spectrum disorders, ADHD and other neurodevelopmental phenotypes. Hum Mol Genet *23*, 2752-2768. 10.1093/hmg/ddt669.

25. Straub, R.E., MacLean, C.J., O'Neill, F.A., Burke, J., Murphy, B., Duke, F., Shinkwin, R., Webb, B.T., Zhang, J., Walsh, D., and et al. (1995). A potential vulnerability locus for schizophrenia on chromosome 6p24-22: evidence for genetic heterogeneity. Nat Genet *11*, 287-293. 10.1038/ng1195-287.

26. Mullin, A.P., Gokhale, A., Larimore, J., and Faundez, V. (2011). Cell biology of the BLOC-1 complex subunit dysbindin, a schizophrenia susceptibility gene. Mol Neurobiol *44*, 53-64. 10.1007/s12035-011-8183-3.

27. Straub, R.E., Jiang, Y., MacLean, C.J., Ma, Y., Webb, B.T., Myakishev, M.V., Harris-Kerr, C., Wormley, B., Sadek, H., Kadambi, B., et al. (2002). Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. Am J Hum Genet *71*, 337-348. 10.1086/341750.

28. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434-443. 10.1038/s41586-020-2308-7.

29. Collins, R.L., Glessner, J.T., Porcu, E., Lepamets, M., Brandon, R., Lauricella, C., Han, L., Morley, T., Niestroj, L.M., Ulirsch, J., et al. (2022). A cross-disorder dosage sensitivity map of the human genome. Cell *185*, 3041-3055 e3025. 10.1016/j.cell.2022.06.036.

30. Huang, N., Lee, I., Marcotte, E.M., and Hurles, M.E. (2010). Characterising and predicting haploinsufficiency in the human genome. PLoS Genet *6*, e1001154. 10.1371/journal.pgen.1001154.

31. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res *17*, 1665-1674. 10.1101/gr.6861907.

32. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science *372*. 10.1126/science.abf7117.

33. Kwon, E.M., Connelly, J.P., Hansen, N.F., Donovan, F.X., Winkler, T., Davis, B.W., Alkadi, H., Chandrasekharappa, S.C., Dunbar, C.E., Mullikin, J.C., and Liu, P. (2017). iPSCs and fibroblast subclones from the same fibroblast population contain comparable levels of sequence variations. Proc Natl Acad Sci U S A *114*, 1964-1969. 10.1073/pnas.1616035114.

34. Gore, A., Li, Z., Fung, H.L., Young, J.E., Agarwal, S., Antosiewicz-Bourget, J., Canto, I., Giorgetti, A., Israel, M.A., Kiskinis, E., et al. (2011). Somatic coding mutations in human induced pluripotent stem cells. Nature *471*, 63-67. 10.1038/nature09805.

35. Cheng, L., Hansen, N.F., Zhao, L., Du, Y., Zou, C., Donovan, F.X., Chou, B.K., Zhou, G., Li, S., Dowey, S.N., et al. (2012). Low incidence of DNA sequence variation in human induced pluripotent stem cells generated by nonintegrating plasmid expression. Cell Stem Cell *10*, 337-344. 10.1016/j.stem.2012.01.005.

36. D'Antonio, M., Benaglio, P., Jakubosky, D., Greenwald, W.W., Matsui, H., Donovan, M.K.R., Li, H., Smith, E.N., D'Antonio-Chronowska, A., and Frazer, K.A. (2018). Insights into the Mutational Burden of Human Induced Pluripotent Stem Cells from an Integrative Multi-Omics Approach. Cell Rep *24*, 883-894. 10.1016/j.celrep.2018.06.091.

37. Rouhani, F.J., Zou, X., Danecek, P., Badja, C., Amarante, T.D., Koh, G., Wu, Q., Memari, Y., Durbin, R., Martincorena, I., et al. (2022). Substantial somatic genomic variation and selection for BCOR mutations in human induced pluripotent stem cells. Nat Genet *54*, 1406-1416. 10.1038/s41588-022-01147-3.

38. Merkle, F.T., Ghosh, S., Genovese, G., Handsaker, R.E., Kashin, S., Meyer, D., Karczewski, K.J., O'Dushlaine, C., Pato, C., Pato, M., et al. (2022). Whole-genome analysis of human embryonic stem cells enables rational line selection based on genetic variation. Cell Stem Cell *29*, 472-486 e477. 10.1016/j.stem.2022.01.011.

39. Simkin, D., Papakis, V., Bustos, B.I., Ambrosi, C.M., Ryan, S.J., Baru, V., Williams, L.A., Dempsey, G.T., McManus, O.B., Landers, J.E., et al. (2022). Homozygous might be hemizygous: CRISPR/Cas9 editing in iPSCs results in detrimental on-target defects that escape standard quality controls. Stem Cell Reports *17*, 993-1008. 10.1016/j.stemcr.2022.02.008.

40. Miller, J.A., Ding, S.L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert, A., Riley, Z.L., Royall, J.J., Aiona, K., et al. (2014). Transcriptional landscape of the prenatal human brain. Nature *508*, 199-206. 10.1038/nature13185.

41. Tian, R., Gachechiladze, M.A., Ludwig, C.H., Laurie, M.T., Hong, J.Y., Nathaniel, D., Prabhu, A.V., Fernandopulle, M.S., Patel, R., Abshari, M., et al. (2019). CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. Neuron *104*, 239-255 e212. 10.1016/j.neuron.2019.07.014.

42. Li, W., Zhang, Q., Oiso, N., Novak, E.K., Gautam, R., O'Brien, E.P., Tinsley, C.L., Blake, D.J., Spritz, R.A., Copeland, N.G., et al. (2003). Hermansky-Pudlak syndrome type 7 (HPS-7) results from mutant dysbindin, a member of the biogenesis of lysosome-related organelles complex 1 (BLOC-1). Nat Genet *35*, 84-89. 10.1038/ng1229.

43. Lowe, G.C., Sanchez Guiu, I., Chapman, O., Rivera, J., Lordkipanidze, M., Dovlatova, N., Wilde, J., Watson, S.P., Morgan, N.V., and collaborative, U.G. (2013). Microsatellite markers as a rapid approach for autozygosity mapping in Hermansky-Pudlak syndrome: identification of the second HPS7 mutation in a patient presenting late in life. Thromb Haemost *109*, 766-768. 10.1160/TH12-11-0876.

44. Bryan, M.M., Tolman, N.J., Simon, K.L., Huizing, M., Hufnagel, R.B., Brooks, B.P., Speransky, V., Mullikin, J.C., Gahl, W.A., Malicdan, M.C.V., and Gochuico, B.R. (2017). Clinical and molecular phenotyping of a child with Hermansky-Pudlak syndrome-7, an uncommon genetic type of HPS. Mol Genet Metab *120*, 378-383. 10.1016/j.ymgme.2017.02.007.

45. Yu, J.R., Lee, C.H., Oksuz, O., Stafford, J.M., and Reinberg, D. (2019). PRC2 is high maintenance. Genes Dev *33*, 903-935. 10.1101/gad.325050.119.

46. Behesti, H., Fore, T.R., Wu, P., Horn, Z., Leppert, M., Hull, C., and Hatten, M.E. (2018). ASTN2 modulates synaptic strength by trafficking and degradation of surface proteins. Proc Natl Acad Sci U S A *115*, E9717-E9726. 10.1073/pnas.1809382115.

47. Gracia-Diaz, C., Zhou, Y., Yang, Q., Lee, C.-H., Espana-Bonilla, P., Zhang, S., Padilla, N., Fueyo, R., Otrimski, G., Li, D., et al. (2022). Gain and loss of function variants in EZH1 disrupt neurogenesis timing and cause overlapping neurodevelopmental disorders. medRxiv, 2022.2008.2009.22278430. 10.1101/2022.08.09.22278430.

48. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Kallberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220-1222. 10.1093/bioinformatics/btv710.

49. Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat Rev Genet *12*, 363-376. 10.1038/nrg2958.

50. Deininger, P.L., and Batzer, M.A. (1999). Alu repeats and human disease. Mol Genet Metab *67*, 183-193. 10.1006/mgme.1999.2864.

51. Nam, K.H., and Ordureau, A. (2022). Quantitative proteome remodeling characterization of two human reference pluripotent stem cell lines during neurogenesis and cardiomyogenesis. Proteomics *22*, e2100246. 10.1002/pmic.202100246.

52. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alfoldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444-451. 10.1038/s41586-020-2287-8.

53. Celik, H., Koh, W.K., Kramer, A.C., Ostrander, E.L., Mallaney, C., Fisher, D.A.C., Xiang, J., Wilson, W.C., Martens, A., Kothari, A., et al. (2018). JARID2 Functions as a Tumor Suppressor in Myeloid Neoplasms by Repressing Self-Renewal in Hematopoietic Progenitor Cells. Cancer Cell *34*, 741-756 e748. 10.1016/j.ccell.2018.10.008.

54. Puda, A., Milosevic, J.D., Berg, T., Klampfl, T., Harutyunyan, A.S., Gisslinger, B., Rumi, E., Pietra, D., Malcovati, L., Elena, C., et al. (2012). Frequent deletions of JARID2 in leukemic transformation of chronic myeloid malignancies. Am J Hematol *87*, 245-250. 10.1002/ajh.22257.

55. Shirato, H., Ogawa, S., Nakajima, K., Inagawa, M., Kojima, M., Tachibana, M., Shinkai, Y., and Takeuchi, T. (2009). A jumonji (Jarid2) protein complex represses cyclin D1 expression by methylation of histone H3-K9. J Biol Chem *284*, 733-739. 10.1074/jbc.M804994200.

56. Patsialou, A., Wilsker, D., and Moran, E. (2005). DNA-binding properties of ARID family proteins. Nucleic Acids Res *33*, 66-80. 10.1093/nar/gki145.

57. Shen, X., Kim, W., Fujiwara, Y., Simon, M.D., Liu, Y., Mysliwiec, M.R., Yuan, G.C., Lee, Y., and Orkin, S.H. (2009). Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. Cell *139*, 1303-1314. 10.1016/j.cell.2009.12.003.

58. Peng, J.C., Valouev, A., Swigut, T., Zhang, J., Zhao, Y., Sidow, A., and Wysocka, J. (2009). Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. Cell *139*, 1290-1302. 10.1016/j.cell.2009.12.002.

59. Pasini, D., Cloos, P.A., Walfridsson, J., Olsson, L., Bukowski, J.P., Johansen, J.V., Bak, M., Tommerup, N., Rappsilber, J., and Helin, K. (2010). JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature *464*, 306-310. 10.1038/nature08788.

60. Li, G., Margueron, R., Ku, M., Chambon, P., Bernstein, B.E., and Reinberg, D. (2010). Jarid2 and PRC2, partners in regulating gene expression. Genes Dev *24*, 368-380. 10.1101/gad.1886410.

61. Cyrus, S., Burkardt, D., Weaver, D.D., and Gibson, W.T. (2019). PRC2-complex related dysfunction in overgrowth syndromes: A review of EZH2, EED, and SUZ12 and their syndromic phenotypes. Am J Med Genet C Semin Med Genet *181*, 519-531. 10.1002/ajmg.c.31754.

62. Comet, I., Riising, E.M., Leblanc, B., and Helin, K. (2016). Maintaining cell identity: PRC2-mediated regulation of transcription and cancer. Nat Rev Cancer *16*, 803-810. 10.1038/nrc.2016.83.

63. Deevy, O., and Bracken, A.P. (2019). PRC2 functions in development and congenital disorders. Development *146*. 10.1242/dev.181354.

64. von Schimmelmann, M., Feinberg, P.A., Sullivan, J.M., Ku, S.M., Badimon, A., Duff, M.K., Wang, Z., Lachmann, A., Dewell, S., Ma'ayan, A., et al. (2016). Polycomb repressive complex 2 (PRC2) silences genes responsible for neurodegeneration. Nat Neurosci *19*, 1321-1330. 10.1038/nn.4360.

65. Pasini, D., Bracken, A.P., Jensen, M.R., Lazzerini Denchi, E., and Helin, K. (2004). Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. EMBO J *23*, 4061-4071. 10.1038/sj.emboj.7600402.

66. O'Carroll, D., Erhardt, S., Pagani, M., Barton, S.C., Surani, M.A., and Jenuwein, T. (2001). The polycomb-group gene Ezh2 is required for early mouse development. Mol Cell Biol *21*, 4330-4336. 10.1128/MCB.21.13.4330-4336.2001.

67. Faust, C., Schumacher, A., Holdener, B., and Magnuson, T. (1995). The eed mutation disrupts anterior mesoderm production in mice. Development *121*, 273-285. 10.1242/dev.121.2.273.

68. Takeuchi, T., Yamazaki, Y., Katoh-Fukui, Y., Tsuchiya, R., Kondo, S., Motoyama, J., and Higashinakagawa, T. (1995). Gene trap capture of a novel mouse gene, jumonji, required for neural tube formation. Genes Dev *9*, 1211-1222. 10.1101/gad.9.10.1211.

69. Takeuchi, T., Watanabe, Y., Takano-Shimizu, T., and Kondo, S. (2006). Roles of jumonji and jumonji family genes in chromatin regulation and development. Dev Dyn *235*, 2449-2459. 10.1002/dvdy.20851.

70. Landeira, D., Sauer, S., Poot, R., Dvorkina, M., Mazzarella, L., Jorgensen, H.F., Pereira, C.F., Leleu, M., Piccolo, F.M., Spivakov, M., et al. (2010). Jarid2 is a PRC2 component in embryonic

stem cells required for multi-lineage differentiation and recruitment of PRC1 and RNA Polymerase II to developmental regulators. Nat Cell Biol *12*, 618-624. 10.1038/ncb2065.

71. Petracovici, A., and Bonasio, R. (2021). Distinct PRC2 subunits regulate maintenance and establishment of Polycomb repression during differentiation. Mol Cell *81*, 2625-2639 e2625. 10.1016/j.molcel.2021.03.038.

72. Kingdom, R., Tuke, M., Wood, A., Beaumont, R.N., Frayling, T.M., Weedon, M.N., and Wright, C.F. (2022). Rare genetic variants in genes and loci linked to dominant monogenic developmental disorders cause milder related phenotypes in the general population. Am J Hum Genet *109*, 1308-1316. 10.1016/j.ajhg.2022.05.011.

73. Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. (2012). Primer3--new capabilities and interfaces. Nucleic Acids Res *40*, e115. 10.1093/nar/gks596.

74. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094-3100. 10.1093/bioinformatics/bty191.

75. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology *36*, 983-987. 10.1038/nbt.4235.

76. Heng Li, P.D. (2018). bcftools.

77. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333-i339. 10.1093/bioinformatics/bts378.