Mutations disrupting neuritogenesis genes represent a major independent risk factor for cerebral palsy

Sheng Chih Jin[1,2*], Sara A. Lewis[3,4*], Somayeh Bakhtiari[3,4*], Xue Zeng[1,2*], Michael C. Sierant[1,2], Sheetal Shetty[3,4], Sandra M. Hinz[3,4], Mark A. Corbett[5], Bethany Norton[3,4], Clare L. van Eyk[5], Aureliane Elie[3,4], Shozeb Haider[6], Stephen Pastore[7], John B. Vincent[7], Janice Brunstrom-Hernandez[8], Antigone Papavasileiou[9], Michael C. Fahey[10], Jesia G. Berry[5], Kelly Harper[5], Chongchen Zhou[11], Helen Magee[3,4], James Liu[3,4], Brandon S. Guida[3,4], Junhui Zhang[1], Boyang Li[12], Jennifer Heim[3], Dani L. Webber[5], Mahalia S.B. Frank[5], Lei Xia[13], Yiran Xu[13], Amar H. Sheth[1], James Knight[14], Boris Keren[15], Christopher Castaldi[14], Irina R. Tikhonova[14], Francesc López-Giráldez[14], Megan Cho[16], Kyle Retterer[16], Francisca Millan[16], Yangong Wang[17], Jeff L. Waugh[18], Lance Rodan[19], Julie S. Cohen[20], Ali Fatemi[20], Angela Lin[21], John Phillips[22], Timothy Feyma[23], Suzanna C. MacLennan[24], Spencer Vaughan[25], Kylie E. Crompton[26], Susan M. Reid[26], Dinah S. Reddihough[26], Qing Shang[11], Chao Gao[27], Iona Novak[28], Nadia Badawi[28], Yana Wilson[28], Sarah McIntyre[28], Shrikant M Mane[14], Xiaoyang Wang[29], David J. Amor[26], Daniela C. Zarnescu[25], Qiongshi Lu[30], Qinghe Xing[17#], Changlian Zhu[13,29#], Kaya Bilguvar[1#], Sergio Padilla-Lopez[3,4#], Richard P. Lifton[1,2#], Jozef Gecz[5#], Alastair H. MacLennan[5#], Michael C. Kruer[3,4#&]

1 Department of Genetics, Yale University School of Medicine, New Haven, Connecticut, USA
2 Laboratory of Human Genetics and Genomics, Rockefeller University, New York, NY USA
3 Pediatric Movement Disorders Program, Division of Pediatric Neurology, Barrow Neurological Institute, Phoenix Children's Hospital, Phoenix, Arizona, USA
4 Departments of Child Health, Neurology, Cellular & Molecular Medicine and Program in Genetics, University of Arizona College of Medicine, Phoenix, Arizona, USA
5 Robinson Research Institute, The University of Adelaide, Adelaide, South Australia, Australia
6 Department of Computational Medicinal Chemistry, School of Pharmacy, University College London, UK
7 Child & Family Program, Department of Psychiatry, Campbell Family Mental Health Research Institute, University of Toronto, Ontario, Canada
8 One CP Place, Plano, TX, USA
9 Division of Paediatric Neurology, Iaso Children's Hospital, Athens, Greece
10 Division of Paediatric Neurology, Monash University, Melbourne, Victoria, Australia
11 Henan Key Laboratory of Child Genetics & Metabolism, Rehabilitation Department, Children's Hospital of Zhengzhou University, Zhengzhou, China
12 Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, USA
13 Henan Key Laboratory of Child Brain Injury, Third Affiliated Hospital of Zhengzhou University, Zhengzhou, China
14 Yale Center for Genome Analysis, Yale University, New Haven, Connecticut, USA

15 Département de Génétique, Centre de Référence Déficiences Intellectuelles de Causes Rares, Groupe Hospitalier Pitié Salpêtrière et GHUEP Hôpital Trousseau, Sorbonne Université, GRC “Déficience Intellectuelle et Autisme”, Paris, France
16 GeneDx, Gaithersburg, MD, USA
17 Institute of Biomedical Science and Children's Hospital, and Key Laboratory of Reproduction Regulation of the National Population and Family Planning Commission (NPFPC), Shanghai Institute of Planned Parenthood Research (SIPPR), IRD, Fudan University, Shanghai, China
18 Departments of Pediatrics & Neurology, University of Texas Southwestern and Children’s Medical Center of Dallas, Dallas, TX, USA
19 Departments of Genetics & Genomics and Neurology, Boston Children's Hospital, Boston, MA, USA
20 Division of Neurogenetics and Hugo W. Moser Research Institute, Kennedy Krieger Institute, Baltimore Maryland USA
21 Division of Medical Genetics, Department of Pediatrics, Mass General Hospital for Children, Boston, MA, USA
22 Departments of Pediatrics and Neurology, University of New Mexico, Albuquerque, NM, USA
23 Division of Pediatric Neurology, Gillette Children’s Hospital, Minnesota, USA
24 Department of Paediatric Neurology, Adelaide Women’s & Children’s Hospital, Adelaide, South Australia, Australia
25 Departments of Molecular & Cellular Biology and Neuroscience, University of Arizona, Tucson, AZ, USA
26 Murdoch Children’s Research Institute and University of Melbourne Department of Paediatrics, Royal Children’s Hospital, Victoria, Australia
27 Rehabilitation Department, Children's Hospital of Zhengzhou University/Henan Children's Hospital, Zhengzhou, China
28 Cerebral Palsy Alliance, Sydney, New South Wales, Australia
29 Institute of Neuroscience and Physiology, Sahlgrenska Academy, Gothenburg University, Gothenburg, Sweden
30 Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, USA

* These authors contributed equally
# These authors should be considered shared last authors
& Corresponding author

ABSTRACT

Cerebral palsy (CP), a neurodevelopmental disorder characterized by irreversible, nonprogressive central motor dysfunction, is commonly associated with prematurity or perinatal brain injury. However, accumulating evidence suggests deleterious genomic variants may contribute to CP in addition to environmental insults. To identify genes contributing to risk for CP, we performed whole-exome sequencing on 250 parent-offspring CP trios. We identified a significant contribution of damaging *de novo* mutations (DNMs), especially in genes that are intolerant to loss of function mutations. Eight genes had multiple, independently-arising damaging DNMs, including two novel CP-associated genes, *FBXO31* and *RHOB,* and four genes previously implicated in cerebral palsy phenotypes, *TUBA1A*, *CTNNB1*, *SPAST*, and *ATL1*. Functional experiments, including molecular and biochemical assays and patient fibroblast studies indicate that the recurrent *RHOB* mutation identified in patients enhances Rho effector binding in the active state and that the *FBXO31* mutation leads to elevated levels of cyclin D. Analysis of candidate CP risk genes highlighted genetic overlap with hereditary spastic paraplegia as well as intellectual disability, autism, and epilepsy, converging with epidemiologic findings. Computational network analysis of risk genes identified significant enrichment of Rho GTPase, extracellular matrix, focal adhesions, cytoskeleton, and cell projection pathways. CP risk genes in Rho GTPase, cytoskeleton and cell projection pathways were found to play an important role in neuromotor development via a *Drosophila* reverse genetics screen. Based on enrichment analysis, we estimate that an excess of damaging *de novo* and inherited recessive variants collectively account for ~14% of the cases in our cohort, whereas perinatal asphyxia is currently estimated to occur in 8-10% of CP cases. Together, these findings provide evidence for the role of genetically-mediated dysregulation of early brain connectivity in CP.

Cerebral palsy (CP) is the cardinal neurodevelopmental disorder impacting motor function, affecting ~3:1000 children in the United States[1], with similar incidence worldwide. Movement disorder (spasticity, dystonia, choreoathetosis, and/or ataxia) onset occurs within the first few years of life as a manifestation of disrupted brain development[2]. Historically, although Little and Osler[3] considered CP to occur largely as a result of perinatal anoxia, Freud disputed this claim[4]. To this day debate about the origin of CP continues, particularly in individual cases with widespread medical and legal implications[5,6].

Analogous to other neurodevelopmental disorders (NDD) such as autism spectrum disorders (ASD) and intellectual disability (ID), no single causative factor has been implicated in CP, although several environmental factors, including prematurity, infection, asphyxia, and pre- and perinatal stroke are major contributors to CP risk[7]. However, in some studies as many as ~40% of CP cases may not have a readily identifiable etiology[8] defined as cryptogenic or idiopathic CP[9]. Registry-based data has shown that 21-40% of CP cases have an associated congenital anomaly, implicating genomic alterations in many of these cases[10]. A heritability of 40% has been estimated in CP[11], supported by probabilistic modeling of CP etiology in a western Swedish cohort[12], comparable to the heritability of 38-58% estimated for ASD[13,14].

To date, five studies have analyzed genomic copy number variations (CNVs) in CP cases[9,15-18], identifying predicted deleterious CNVs in 10-31% of cases. Three prior whole exome sequencing (WES) studies have been performed in CP cases[19-21]. The largest study to date reported putatively deleterious variants in ~14% of 98 parent-offspring trios with unselected forms of CP[21]. These studies indicate potentially important genetic risks in CP, but insufficient availability of controls limited the statistical inferences that could be made, and functional validation of novel candidate gene variants was not able to be performed. We sought to address these limitations in the current study.

RESULTS

CP cohort characteristics and WES

We performed whole exome sequencing (WES) of 250 CP trios, including 91 previously reported[21] and 159 ascertained from centers in the United States, China, and Australia after written informed consent was obtained according to local ethical requirements (**Online Methods**). Cases were diagnosed by clinical CP specialists using international consensus criteria[22] (**Supplementary Table 1 and Supplementary Data Set 1**); CP was thus defined as a non-progressive developmental disorder of movement and/or posture impairing physical function. Cases experienced symptom onset by age two. This operational definition thus excluded progressive neurological disorders such as neurodegenerative diseases and mitochondrial disorders. Of note, no cases had known chromosomal anomalies or aneuploidies, clinically and/or molecularly diagnosed syndromes (i.e. Rett syndrome, Angleman syndrome, etc.), pathogenic microdeletion or microduplication syndromes, or traumatic brain injuries.

Detailed patient phenotypes are available in **Supplementary Clinical Summaries.** Representative neuroimaging findings are presented in **Supplementary Figure 1** and videos highlighting movement disorder phenotypes in representative individuals can be found in **Supplementary Videos** (42 videos available). We focused on idiopathic CP (*i.e.* no known cause) in order to minimize confound due to other risk factors. Within the cohort, 157 trios (62.8%) were classified as idiopathic[9], 84 cases (33.6%) had a known environmental insult associated with CP (including prematurity defined as ≤ 32 weeks gestation, birth asphyxia [as defined by treating clinicians], ischemic/hemorrhagic stroke, and/or infection), and the remaining 9 trios (3.6%) were not able to be assigned to either category ("unclassified") (**Supplementary Table 1**).

WES was performed as previously described[23]. Sequencing metrics suggest that, regardless of the exome capture reagent used, all samples had sufficient sequencing coverage to make confident variant calls with a mean coverage of ≥ 46X at each targeted base and more than 90% of targeted bases with ≥ 8 independent reads (see **Supplementary Table 2** for exome metrics). Control trios consisting of 1,789 unaffected siblings of autism cases and their unaffected parents from the Simons Simplex Collection were analyzed in parallel[24]. BWA-MEM was used to align the sequencing reads and GATK 'Best Practices' was used to call variants[25,26]. MetaSVM[27] and Combined Annotation Dependent Depletion (CADD v1.3)[28] algorithms were used to predict deleteriousness of missense variants ("D-Mis", defined as MetaSVM-deleterious or CADD ≥ 20). Inferred loss of function (LoF) variants consist of stop-gain, stop-loss, frameshift insertions/deletions, canonical splice site, and start-loss. LoF and D-Mis mutations were considered "damaging". *De novo* mutations (DNMs) were called by the TrioDeNovo program[29]. Sanger sequencing was conducted to validate mutations in genes of interest.

CP cases harbor a significant enrichment of damaging DNMs

We began by assessing the contribution of DNMs to CP at a cohort level. The number of observed DNMs in cases and controls closely approximates the Poisson distribution (**Supplementary Figure 2**) indicating that DNMs are independent probabilistic events. We found an enrichment of damaging DNMs in CP cases, which became more apparent when focusing the analysis on genes intolerant to LoF variation (pLI score ≥ 0.9 in gnomAD v2.1.1[30]) (enrichment = 1.79; *p-value* = 9.9×10^{-6} for damaging DNMs; **Table 1A**). No significant enrichment of any mutation category was found in controls (**Table 1A**). When we considered the ascertainment differential (observed number of damaging DNMs vs. expected number of damaging DNMs, divided by the number of trios in the cohort), 11.9% of CP cases in our cohort could be attributed to an excess of damaging DNMs. When stratifying cases by CP subtypes, we found greater enrichment of damaging DNMs in idiopathic (enrichment = 1.99; *p-value* = 1.9×10^{-5}) compared to environmental cases (enrichment = 1.29; *p-value* = 0.19; **Supplementary Table 3**), suggesting that idiopathic cases harbor a higher burden of damaging DNMs.

Recurrent damaging DNMs implicate both known and novel risk genes in CP

We next considered individual genes recurrently implicated in our CP cohort via a *de novo* mechanism (**Supplementary Data Set 2**). We identified eight genes harboring ≥ 2 damaging DNMs, with *TUBA1A* (*p-value* = $4.9x10^{-8}$) and *CTNNB1* (*p-value* = $1.6x10^{-7}$) surpassing Bonferroni correction cutoffs for genome-wide significance (**Table 1B**). Among these genes, *ATL1*, *CTNNB1*, *SPAST*, and *TUBA1A* have previously been associated with human CP phenotypes[19,21,31]. We also identified identical but independently-arising damaging DNMs in two genes, *RHOB* and *FBXO31*. Significant gene-level enrichment of protein-altering DNMs in genes with recurrent mutations strongly implicate these genes as *bona fide* CP-associated genes (**Supplementary Table 4**).

Identical gain-of-function DNMs in *RHOB* and *FBXO31*

The probability of finding two or more *de novo* events in a cohort of this size is very low *(p-value* = $1.6x10^{-3}$) (see **Methods**). In our cohort, we report two identical *de novo* putative gain-of-function (GOF) mutations in two genes: *RHOB* and *FBXO31*.

RHOB, encoding a Rho GTPase, harbors two identical DNMs (p.Ser73Phe; **Table 2**) in two unrelated spastic-dystonic CP cases, representing an unlikely chance event. Ser73 is predicted to be phosphorylated (0.997 by NetPhos 3.1)[32,33] and located in a conserved position in the Switch II domain where Rho protein kinases associate with Rho- and Rac- related proteins (**Figure 1B**). Structural modeling of the *RHOB* mutation at residue 73 from serine to phenylalanine suggests an alteration of both the shape of the binding site and the surface charge of the protein (**Figure 1B**). Both patients have a remarkably concordant phenotype, including hyperintense T2 white matter signal (periventricular leukomalacia) on MRI and spastic-dystonic diplegia, expressive language disorder, and aortic arch abnormalities (**Table 2**; **Figure 1C**; **Supplementary Video F064 and F244**). RHOB is known to control dendritic spine outgrowth[34] but has not previously been associated with a human disease. Biochemical analyses indicated that this variant shows accentuated responses to both GTPase activating proteins (GAPs) and GDP exchange factors (GEFs) (**Figure 1D-E**) ultimately leading to enhanced binding in the active state to the Rho effector rhotekin (**Figure 1F**).

We also identified two unrelated cases with an identical DNM (p.Asp334Asn; **Table 2**) in *FBXO31,* which encodes the F-box only protein 31. p.Asp334Asn is positioned in a conserved residue around the cyclin D1 binding pocket on FBXO31 (**Figure 2B**). FBXO31 is an E3 ubiquitin ligase and the FBXO31/SKP1/Cullin1 complex ubiquitinates target substrates, such as cyclin D, to control protein abundance by targeting substrates for proteasomal degradation[35]. FBXO31 is known to control axonal outgrowth and is essential in dendrite growth and neuronal migration in the developing cerebellar cortex[36]. *FBXO31* p.Asp334Asn affects the known interaction site for cyclin D[37] (**Figure 2B-C**), leading to an apparent GOF of cyclin D degradation (**Figure 2D**). A homozygous truncating mutation in *FBXO31* has previously been reported in association with intellectual disability (OMIM# 615979)[38]. Both patients in our cohort exhibited spastic diplegic CP (**Table 2**; Supplementary Video F218 and F699),

intellectual disability, language disorder, and ADHD. F218 had gut malrotation and constipation, cleft palate, esotropia, and normal brain morphology on MRI and F699 had anxiety, strabismus, severe constipation, and ventricular dilation with thin corpus callosum on MRI. Therefore, this DNM in *FBXO31* leads to a phenotype distinct from the previously described autosomal recessive truncating mutation-associated nonsyndromic intellectual disability phenotype[38].

DNMs in previously implicated genes, *TUBA1A, CTNNB1, ATL1,* and *SPAST*

TUBA1A, encoding the microtubule-related protein α-tubulin*,* harbors 3 damaging DNMs (p.Arg123Cys, p.Leu152Gln, p.Tyr408Asp; **Table 2**) in 3 unrelated probands, 2 of whom have been previously reported[21]. Both p.Arg123Cys and p.Leu152Gln map to the N-terminal domain, and p.Tyr408Asp maps to the C-terminal domain (**Supplementary Figure 3**). *TUBA1A* heterozygous mutations have been described as associated with a spectrum of cortical malformations[39] (OMIM# 611603) and our patients exhibit MRI findings within this spectrum (**Supplementary Figure 3**). Clinically, our cases demonstrate spasticity in their lower limbs and 2/3 exhibit concurrent intellectual disability.

CTNNB1, encoding β-catenin, harbors 3 LoF DNMs (p.Glu54X, p.Phe99fs, p.Arg449fs; **Table 2**) in 3 unrelated probands, 1 of whom was previously reported[21]. p.Glu54X and p.Phe99fs are located in the N-terminal domain and predicted to lead to nonsense-mediated decay, while p.Arg449fs is located in the central armadillo repeat domain, which is essential for the phosphorylation of β-catenin by protein kinase CK2[40] (**Supplementary Figure 4**). Autosomal dominant germline inactivating mutations in *CTNNB1* have been implicated in exudative vitreoretinopathy 7[41] (OMIM# 617572) and neurodevelopmental disorder with spastic diplegia and visual defects[42-44] (OMIM# 615075). All of our patients exhibited spasticity, intellectual disability, behavior problems and language disorders. We also found dystonia and microcephaly in 2/3 patients. While one patient had possible bilateral frontal pachygyria (**Supplementary Figure 4**), brain findings were notably absent from the other patients. We found strabismus in 2/3 patients, but no other visual defects.

ATL1 encodes atlastin-1, which is critical for formation of the tubular endoplasmic reticulum network and axon elongation in neurons[45,46]. *ATL1* harbors two damaging DNMs in our cohort (p.Ala350Val, p.Lys406Gln; **Table 2**) located in the GBP domain (**Supplementary Figure 5**). Autosomal dominant germline mutations have been associated with neuropathy type 1D[47] (OMIM# 613708) and spastic paraplegia type 3A[48] (OMIM# 182600). Our patients exhibited spasticity and dystonia with brain findings of T2 hyperintensities and bihemispheric periventricular leukomalacia. There was no evidence of phenotypic progression at the time of last follow-up (patient ages 10 years and 29 months).

SPAST, encoding spastin, harbored two damaging DNMs (p.Asp441Gly, p.Ala495Pro; **Table 2**). Both mutations occur at conserved positions in the AAA domain, which is essential for the regulation of ATPase activity (**Supplementary Figure 6**).

Autosomal dominant germline mutations in *SPAST* have been linked to spastic paraplegia 4[49] (OMIM# 182601). p.Asp441Gly has been reported in association with hereditary spastic paraplegia (HSP)[50,51]. Our patients exhibited spasticity with one also exhibiting dystonia, with scattered subcortical T2 hyperintensities present in 1 patient and no brain findings in the other. There was no evidence of phenotypic progression (patient ages 21 years and 40 months, respectively).

DNMs in *DHX32* and *ALK*

DHX32, encoding putative pre-mRNA-splicing factor ATP-dependent RNA helicase DHX32, harbored two damaging DNMs (p.Tyr228Cys, p.Ile266Met; **Table 2**). p.Tyr228Cys falls within the helicase ATP binding domain, which is required for ATP binding, hydrolysis, and nucleic acid substrate binding[52]. (**Supplementary Figure 7**). Mutations in *DHX32* have not previously been associated with human diseases. Both of our patients exhibited intellectual disability, and one demonstrated spastic diplegia with the other characterized as a generalized dystonia. Brain findings included periventricular leukomalacia and mildly diminished cerebral volume.

ALK, encoding ALK receptor tyrosine kinase, harbored one damaging DNM (p.Ser1081Arg) and one stop-gain DNM (p.Trp1320X) (**Table 2**). p.Ser1081Arg was located in the juxtamembrane domain, which is responsible for the modulation of the kinase catalytic activity, and p.Trp1320X was located in the tyrosine kinase domain[53] (**Supplemental Figure 8**). Germline and somatic activating mutations in *ALK* have been previously associated with neuroblastoma[54,55] (OMIM# 613014). One patient exhibited spastic diplegia with mild tremor, scattered subcortical hyperintensities, and an atrial septum defect. The other patient had spastic-dystonic diplegia, white matter abnormalities, and epilepsy. There was no evidence of neuroblastoma in either patient.

Enriched recessive genotypes in genes associated with hereditary spastic paraplegia

We performed a one-tailed binomial test coupled with a polynomial model to evaluate the burden of recessive genotypes (RGs) for each gene in our CP cohort (**Supplementary Data Set 4**). We did not observe enrichment of damaging RGs in the cohort meeting genome-wide significance (**Supplementary Table 5**). However, we noted biallelic damaging variants in several genes previously associated with HSP. HSP is clinically distinguished from CP by its progressive, neurodegenerative nature and later (often adult) onset in many cases.

We carefully re-assessed clinical phenotypes of these cases, and found no evidence of progression from the time of ascertainment. Interestingly, early-onset with protracted clinical stability has previously been identified as an endophenotype in a subset of patients with mutations in HSP-associated genes[56]. For example, patients with *SPAST* missense mutations may have onset in toddlerhood with extended clinical stability[57]. In contrast, truncating *SPAST* mutations are often translated and accumulate over time, putatively leading to later-onset and a neurodegenerative course[58]. In

addition, important roles for SPAST[59] and ATL1[60] in developmental neurogenesis have been shown, indicating an important role in neuronal development.

We observed 6 damaging RGs (in *AMPD2*, *AP4M1*, *AP5Z1, FARS2*, *NT5C2*, and *SPG11;* **Table 3***)* among genes previously associated with recessive HSP (**Supplementary Data Set 3)** (enrichment = 7.74; one-tailed binomial *p-value* = 1.5×10^{-4}; **Table 3**). By ascertainment differential, ~2.1% of the CP cases in our cohort could thus be accounted for by an excess of RGs. The enrichment of RGs in known HSP-associated genes was predominantly driven by idiopathic cases (idiopathic enrichment = 9.2; one-tailed binomial *p-value* = 2.4×10^{-4} vs. environment enrichment = 4.48; one-tailed binomial *p-value* = 0.20; **Table 3**).

No gene enriched for rare X-linked hemizygous variants

Male sex is a risk factor for developing CP[61]. Therefore, we examined whether genes could contribute to CP via X-linked recessive inheritance by performing case-control analyses comparing rare hemizygous variants (MAF ≤ 5×10^{-5}) in 154 male CP probands to male controls in gnomAD. No gene surpassed Bonferroni correction cutoff (**Supplementary Table 6**), suggesting that the current study is statistically under-powered to assess hemizygous burden.

Clinical and genetic overlap of CP with other neurodevelopmental disorders

Clinically, NDDs frequently co-occur. In the case of CP, ~45% of individuals with CP have concurrent ID[62], ~40% also have epilepsy, and ~7% have ASD in addition to CP[1]. Accordingly, we sought to determine the degree of overlap between putative CP risk genes (genes harboring damaging variants with *de novo*, X-linked recessive, or autosomal recessive segregation) from our CP cohort with known NDD risk genes. The analysis was performed using the Disease Gene Network, which identifies associations between genes and diseases curated from the literature and databases including ClinVar, ClinGen, and UniProt[63]. We found substantial genetic overlap between our CP candidate gene list and the major NDDs (CP vs. ID, 2.0 fold enrichment, $p<5 \times 10^{-15}$; vs. epilepsy, 1.7 fold enrichment $p<5 \times 10^{-4}$; vs. ASD, 1.8 fold enrichment $p<4 \times 10^{-3}$, hypergeometric test) (**Figure 3A**). In contrast, when we examined overlap with a non-developmental neurological disorder, Alzheimer's disease, there was no enrichment (**Figure 3B**). 29.1% of our genes overlapped with genes linked to ID, 11.1% for epilepsy, and 6.1% for ASD. Our data suggest that CP has significant genetic overlap with other genetic neurodevelopmental disorders, indicating potential genetic pleiotropy and common etiologies of co-occurring NDDs.

Extracellular matrix, cell-matrix focal adhesions, cytoskeletal network, and Rho GTPase genes are highly associated with CP

In addition to the small number of genes with recurrent variants, we also identified a larger number of individual genes harboring predicted damaging variants. A challenge in early gene discovery is interpreting which of these genetic variants are

more likely to confer disease risk, what candidates to prioritize for further study and clinical sequencing, and what functional role these genes may be playing. Gene set over representation analysis has been used previously in other NDDs[64,65] to identify pathways and gene networks likely to contribute to fundamental mechanisms of disease biology. Therefore, we utilized an integrated bioinformatics approach to assess clustering of CP candidate risk genes identified in this study. We employed a suite of tools for unbiased discovery of conserved pathways and biological functions relevant to CP.

STRING-based clustering[55] of genes with damaging variants across all modes of inheritance in our CP cohort (440 genes) showed greater connectivity than predicted by chance (359 observed/302 expected protein interactions, 1.2 fold enrichment, *p-value* $<7.1x10^{-4}$) indicating a functional network encompassing damaging variants. We then performed gene over representation analysis of these genes using DAVID[66], MSigDB[67] and PANTHER[68] for functional annotation and pathway characterization. This approach indicated statistical overrepresentation of candidate genes stratified by gene ontology, pathways (KEGG/Reactome), and curated functional and expression data to identify meaningful relationships. Consistent with the STRING findings, these algorithms identified multiple gene sets representing pathways and conserved functions that were significantly enriched (FDR <0.05) (**Supplementary ORA Dataset).**

We noted functionally-related findings supported by multiple tools. Non-integrin membrane-extracellular matrix (ECM) interactions and laminin interaction pathways were identified by all 3 algorithms (**Table 4A**). We used dcGO for inferring hierarchal associations between Panther ontological terms[69] (**Table 4B)**. Taken together, these findings indicate an over representation of genes involved in extracellular matrix biology, cell-matrix interactions (focal adhesions), cytoskeletal dynamics and Rho GTPases.

Genes from Rho GTPase, cytoskeleton, and cell projection pathways govern neuromotor development in Drosophila

Subsequently, we independently assessed a role for over-represented pathways in normal locomotor development by conducting a reverse genetic screen in Drosophila. A similar approach has been applied previously in studies of ASD and HSP using Drosophila and zebrafish, respectively[45,46]. We focused on genes with damaging variants from our CP patient cohort with GTPase, cytoskeleton, and cell projection GO terms. Although not previously shown, we hypothesized that our screen would indicate a key role for these genes in neuromotor development.

We selected genes with conserved Drosophila orthologs (DIOPT≥5) that had available and molecularly characterized LoF alleles (**Supplementary Table 7**) that could be studied in the compound heterozygous or homozygous state to help map phenotypes to the gene of interest. We excluded genes that would cause confounding phenotypes such as lethality or had a previously described locomotor phenotype, except for *ATL1* which was included as a positive control. Genes known to cause other NDDs or with known roles in brain development were prioritized. As a result, we

screened 22 genes for locomotor ability using turning assays in larvae and negative geotaxis/positive phototaxis assays in adults[47,48]

We found locomotor phenotypes in mutants of genes regulating GTPase signal transduction (*AGAP1*, *DOCK11*, *RABEP1*, *SYNGAP1*, *TBC1D17*), the cytoskeleton (*MKL1*, *MPP1*) and cell projection (*PTK2B*, *SEMA4A*, *TENM1*) pathways (**Figure 5**). We also verified that genes with multiple recurrent variants could generate neuromotor phenotypes by examining *ZDHHC15*, a palmitoyltransferase, and *ATL1;* mutants of both genes had locomotor deficits. When assays were conducted in both larva and adults, we often found locomotor phenotypes at both timepoints, suggesting that defects arose in the developmental period and persisted throughout the lifespan (**Supplemental Table 7**). Of potential interest, we found evidence for sexual dimorphism as male flies with mutations in *AKT3*, *RABEP1*, or *PRICKLE1/2* exhibited locomotor deficits while females did not, reminiscent of the male predominance observed in CP patients[54].

In total, we found 80% (15/19) genes from our enriched pathways exhibited a locomotor phenotype in Drosophila. In comparison, genome-wide, only 3.1% of annotated Drosophila genes are known to lead to a locomotor phenotype[53]. The detection of locomotor phenotypes in 16/22 genes tested from our human genomic screen thus represents an enrichment of 23.4 fold compared to that expected by chance alone (*p value* $<2.2\times10^{-16}$; **Figure 5**). Overall, findings from our Drosophila studies provided evidence for cytoskeletal, Rho GTPase, and cell projection pathways in motor development. When considered alongside findings from our gene set overrepresentation analysis, these data indicate that mutations in gene pathways that include the extracellular matrix, focal adhesions, the cytoskeleton and Rho GTPases may converge to alter neurite extension during early brain development.

DISCUSSION

In the past, genetic influences have not been considered major contributors to CP, but our findings and those of others challenge this dogma. Prior studies suggested that both CNVs and single nucleotide variants contribute to CP[9,15-21]. Here we expand upon those earlier findings to provide robust statistical evidence at a cohort level that rare, damaging single nucleotide variants represent an independent risk factor for CP. The cohort-wide enrichment of DNMs we detected is consistent with the observation that most cases of CP occur sporadically[70]. Using the distribution of LoF-intolerant genes with multiple damaging DNMs in this cohort, we estimated the number of genes that contribute to CP through a *de novo* mechanism to be 80 (**Supplementary Figure 9A**). Saturation analysis estimates that WES of 2,500 and 7,500 CP trios will yield 70.3% and 93.8% saturation, respectively, for CP risk genes, suggesting a high yield for CP gene discovery as additional samples are sequenced (**Supplementary Figure 9B**). Accordingly, the International Cerebral Palsy Genomics Consortium (ICPGC; www.icpgc.org) was recently founded to address the need for international data sharing and collaboration to advance the pace of discovery[65]. Conservatively, we estimate that 14% of the cases in our cohort can be accounted for by damaging genomic variants (based on ascertainment differentials of 11.9% for DNMs and ~2% for RGs). In

comparison, recent estimates indicate that birth asphyxia is seen in ~8-10% of CP cases[71], indicating that genomic mutations represent an important, independent contributor to CP etiology that historically has been overlooked.

We found evidence for both known disease-associated genes and genes not previously associated with human phenotypes in our cohort. The identification of independently-arising but identical *de novo* mutations in *RHOB* and *FBXO31* indicates that monogenic contributions to CP exist and may be under-recognized. Our parallel identification of genetic correlation of CP with other NDDs implicates shared susceptibility as suggested previously[72]. In some cases, this may reflect ascertainment bias, as motor phenotypes may have been under-reported in prior studies of other NDDs. In other cases, typified by *FBXO31*, our findings likely represent phenotypic expansions. Finally, in some contexts, NDD manifestations may prove pleiotropic, with a genetic disruption of early neurodevelopment manifesting variably as is increasingly being recognized[73]. Analogous to other NDDs, individual CP cases may prove to be environmental in origin, genetic, or some combination thereof. However, somewhat uniquely among the NDDs, environmental contributions to CP are relatively well-characterized, and CP may represent a model disorder within which to study gene-environment interactions in a developmental context.

Altered motor circuit connectivity is thought to be part of CP pathophysiology[74]. By integrating orthogonal lines of evidence, including recurrent gene analyses, *in vitro* and *in vivo* functional assays, and cohort-wide network biology approaches, we found converging evidence supporting a role for extracellular matrix components, cell-matrix focal adhesions, cytoskeletal organization and Rho GTPases in CP etiology. These processes are known to drive the conserved process of cell projection extensions during nervous system development[75]. Based on known disease and developmental biology, we therefore predict that disruption of genes involved in neurodevelopmental patterning may alter early neuritogenesis and neuronal functional network connectivity. Further studies will be needed to determine how CP patient-derived variants affect neuronal circuit development.

Our findings have important clinical implications. Specific genetic findings may provide closure for families and guide preventative healthcare and family planning, such as counseling for recurrence risk (often quoted as ~1% for CP but potentially much higher for inherited mutations). In some cases, identification of specific variants in individuals in our cohort led to changes in recommendations for evaluation or management, including personalized treatments that would not otherwise have been initiated (i.e. ethosuximide for *GNB1*[76] (F068), levodopa for *CTNNB1*[77] (F066, GRA8913, F428) and 5-aminoimidazole-4-carboxamide riboside (AICAr) for *AMPD2*[78] (F623) (**Supplementary Clinical Summaries**).

In the near future, studies will be able to overcome our limitations of small sample size and further utilize available clinical data to expand upon genotype-phenotype correlations. Additionally, as more information about CP genetic etiology becomes available, it will become possible to assign likely genetic causation to

individual cases. Future studies of well-characterized unselected CP cohorts will be instrumental in determining the true contributions of genetic and environmental factors side-by-side in order to clarify the epidemiology of CP.

Overall, our data indicates that genomic variants should be considered alongside environmental insults when assessing the etiology of an individual's CP. Such considerations may have important clinical, research, and medico-legal implications. In the near future, genomic data may help stratify patients and identify likely responders to currently available medical and/or surgical therapies. Finally, over time, mechanistic insights derived from the identification of core pathways via genomic studies of CP may help guide therapeutic development efforts in a field that has not seen a novel therapy introduced for decades.

ONLINE METHODS

Case cohorts, enrollment, phenotyping, and exclusion criteria: In this study, 151 CP cases (129 idiopathic, 22 environmental) and their unaffected parents were recruited via Phoenix Children's Hospital, the University of Adelaide, and Zhengzhou City Children's Hospital. Exclusion criteria and detailed descriptions about these cohorts are provided separately below. Further, 99 previously sequenced[21] unselected trios (28 idiopathic, 62 environmental, 9 unknown) were included to allow for comparison of idiopathic and environmental subtypes of CP.

CP classification: CP cases were subdivided into idiopathic, environmental, and unclassified groups based upon data available at the time of ascertainment. This designation was revised as appropriate if additional data became available. Cases were designated "environmental" if any idiopathic exclusion criteria were met.

Exclusion criteria for idiopathic status: Potential participants were excluded from an "idiopathic" designation if any of the following were present: prematurity (estimated gestational age <32 weeks), stroke, intraventricular hemorrhage, major brain malformation (i.e. lissencephaly, pachygyria, polymicrogyria, schizencephaly, simplified gyri, brainstem dysgenesis, cerebellar hypoplasia, etc.), birth asphyxia/hypoxic-ischemic injury (as defined by treating physicians), *in utero* infection, hydrocephalus, traumatic brain injury, respiratory arrest, cardiac arrest, or brain calcifications. The following did not automatically indicate environmental status even if parents believed this was the cause of the child's CP: history of prematurity (but delivery at greater than or equal to 32 weeks gestational age), nuchal cord, difficult delivery, fetal decelerations, urgent C-section, preterm bleeding, or maternal infection. In equivocal cases, additional data was sought until a decision regarding group assignment could be made by the corresponding author. Periventricular leukomalacia was not considered universally indicative of environmental status[79].

Movement disorder, pattern of involvement, and functional status: Spasticity, dystonia, chorea/athetosis, ballism, hypotonia, and/or ataxia were assessed by the treating specialist, who also assigned Gross Motor Functional Classification System (GMFCS) scores as well as the pattern of involvement.

Phoenix Children's Hospital (PCH; N=106): Patients with CP according to international consensus criteria[22] were recruited from cerebral palsy subspecialty clinics (pediatric movement disorders neurology, pediatric physiatry) at PCH or the clinics of collaborators at outside institutions using a PCH-approved central IRB protocol (#15-080). Written informed consents were obtained for parents and assent was obtained for children as appropriate for families wishing to participate. Blood and/or saliva samples were collected from the affected child and both parents, and when possible, additional affected siblings. DNA was extracted with the support of the PCH Biorepository using a Kingfisher Automated Extraction System™, and quality control metrics, including yield, 260/280, and 260/230 ratio were recorded.

University of Adelaide Robinson Research Institute (N=99): Ethics permission was obtained in each state and overall from the Adelaide Women's and Children's Health Network Human Research Ethics Committee South Australia. Families were enrolled from among all children attending major children's hospitals in South Australia, New South Wales and Queensland where a diagnosis of CP had been confirmed by a specialist in pediatric rehabilitation according to international consensus criteria[22]. Blood for DNA from cases was collected under general anaesthesia during procedures such as Botox therapy or orthopaedic surgery and parental blood collected whenever possible. Lymphoblastoid cell lines (LCLs) were generated for each case at Genetic Repositories Australia.

Zhengzhou City Children's Hospital (N=44): This study was approved after review by the ethics committee of Zhengzhou City Children's Hospital. Parent-offspring trios were recruited from children with CP without apparent cause from September 1, 2011 to June 30, 2016 at Zhengzhou City Children's Hospital. Cases were excluded if intrauterine growth retardation, threatened pre-term birth, premature rupture of membranes, pregnancy-induced hypertension, or multiple births was present. All participants and their guardians provided written informed consent.

GeneDx (N=1): This patient was enrolled through GeneDx's Genematcher program.

Control cohorts: Controls consisted of 1,789 previously sequenced families that included one child with autism, one unaffected sibling, and the unaffected parents[24]. For use in this study, only the unaffected sibling and parents were analyzed. Controls were designated as unaffected by the Simons Simplex Collection (SSC). Permission to access the genomic data in the SSC via the National Institute of Mental Health Data Repository was obtained. Written informed consent for all participants was provided by the Simons Foundation Autism Research Initiative.

Exome sequencing: 106 trios were sequenced at the Yale Center for Genome Analysis following an identical protocol. Briefly, genomic DNA from venous blood, saliva, or LCL lines (Adelaide) was captured using the Nimblegen SeqxCap EZ MedExome Target Enrichment Kit (Roche) or the xGEN Exome Research Panel v1.0 (IDT) followed by Illumina DNA sequencing as previously described[23]. Forty-four trio samples from Zhengzhou were also prepared using Exome Library Prep kits (Illumina), followed by Illumina sequencing. Ninety-nine previously published trios from Adelaide were captured using the VCRome 2.1 kit (HGSC), followed by Illumina sequencing as described previously[21]. Eight trios from Adelaide sequenced at the University of Washington were prepared using the SureSelect Human All Exon V5 (Agilent) and underwent Illumina sequencing. One trio obtained from GeneDx was captured using the Agilent SureSelect Human All Exon V4 followed by Illumina sequencing. (**Supplementary Table 1**; **Supplementary Data Set 1**).

Mapping and variant calling: WES data were processed using two independent pipelines at the Yale School of Medicine and PCH. At each site sequence reads were independently mapped to the reference genome (GRCh37) with BWA-MEM and further processed using GATK Best Practice workflows, which include duplication marking,

indel realignment, and base quality recalibration, as previously described[25,26,80]. Single nucleotide variants and small indels were called with GATK HaplotypeCaller and annotated using ANNOVAR[81], dbSNP (v138), 1000 Genomes (August 2015), NHLBI Exome Variant Server (EVS), and the Exome Aggregation Consortium v3 (ExAC)[30]. MetaSVM and Combined Annotation Dependent Deletion (CADD v1.3) algorithms were used to predict deleteriousness of missense variants ("D-Mis", defined as MetaSVM-deleterious or CADD ≥ 20)[27,28]. Inferred LoF variants consist of stop-gain, stop-loss, frameshift insertions/deletions, canonical splice site, and start-loss. LoF + D-Mis mutations were considered "damaging". Variant calls were reconciled between Yale and PCH prior to downstream statistical analyses. Variants were considered by mode of inheritance, including DNMs, RGs, and X-linked variants.

Kinship analysis: Relationship between proband and parents was estimated using the pairwise identity-by-descent (IBD) calculation in PLINK[82]. The IBD sharing between the proband and parents in all trios is between 45% and 55%.

Removal of duplicated samples: To identify subjects which could have been recruited multiple times in case cohorts, we calculated the overlap of high-confidence rare variants (MAF = 0% in ExAC, 1000 Genomes, and EVS) between each pair of individuals using the in-house pipeline. For pairs that share ≥ 80% of rare variants, the sample with greater sequence coverage was kept in the analysis and the other discarded.

Principal component analysis: To determine the ethnicity of each sample, the EIGENSTRAT[83] software was used to analyze tag SNPs in cases, controls, and HapMap subjects as described before[23].

Variant filtering: DNMs were called using the TrioDenovo[29] program by Yale and PCH separately as described previously[23], and filtered using stringent hard cutoffs. These hard filters include: (1) MAF ≤ 4×10^{-4} in ExAC; (2) a minimum 10 total reads total, 5 alternate allele reads, and a minimum 20% alternate allele ratio in the proband if alternate allele reads ≥ 10 or, if alternate allele reads is < 10, a minimum 28% alternate ratio; (3) a minimum depth of 10 reference reads and alternate allele ratio < 3.5% in parents; and (4) exonic or canonical splice-site variants.

For the X-linked hemizygous variants, we filtered for rarity (MAF ≤ $5\text{x}10^{-5}$ across all samples in 1000 Genomes, EVS, and ExAC) and high-quality heterozygotes (pass GATK Variant Score Quality Recalibration [VSQR], minimum 8 total reads, genotype quality [GQ] score ≥ 20, mapping quality [MQ] score ≥ 40, and minimum 20% alternate allele ratio in the proband if alternate allele reads ≥ 10 or, if alternate allele reads is < 10, a minimum 28% alternate ratio)[30,84]. Additionally, variants located in segmental duplication regions (as annotated by ANNOVAR[27]), RGs, and DNMs were excluded. Finally, *in silico* visualization was performed on: (1) variants that appear at least twice and (2) variants in the top 20 significant genes from the analysis.

We filtered RGs for rare (MAF ≤ 10^{-3} across all samples in 1000 Genomes, EVS, and ExAC) homozygous and compound heterozygous variants that exhibited high quality

sequence reads (pass GATK VSQR), had a minimum 8 total reads total for proband, had a genotype quality [GQ] ≥ 20, had a minimum 90% alternate allele ratio). Only LoF variants (stop-gain, stop-loss, canonical splice-site, frameshift indels, and start-loss), D-Mis (MetaSVM = D or CADD ≥20), and non-frameshift indels were considered potentially damaging to protein function.

Sanger sequencing: We Sanger validated a representative sample of damaging genomic variants in Adelaide, Phoenix, and at Yale. We confirmed 92.7% (141/152) of point mutations and 90% (9/10) of indels.

***De novo* mutation expectation model:** Because the CP trios were captured by multiple different reagents, we took the union of all bases covered by different capture reagents and generated a Browser Extensible Data (BED) file representing an unified capture for all trios. We used bedtools (v2.27.1) to extract sequence from the BED file[85]. We then applied a sequence context-based method to calculate the probability of observing a DNM for each base in the coding region adjusting for the sequence depth in each gene as described previously[86]. Briefly, for each base in the exome, the probability of observing every tri-nucleotide mutating to other tri-nucleotides was determined. ANNOVAR (v2015Mar22) was used to annotate the consequence of each possible substitution[81]. RefSeq was used to annotate variants (based on the file "hg19_refGene.txt" provided by ANNOVAR). For each gene, the coding consequence of each potential substitution was summed for each functional class (synonymous, missense, canonical splice site, frameshift insertions/deletions, stop-gain, stop-loss, start-lost) to determine the gene-specific mutation probabilities[86]. The probability of a frameshift mutation was determined by multiplying the probability of a stop-gain mutation by 1.25 as described previously[86]. In-frame insertions or deletions are not accounted for by the model[86] and were not considered in the downstream statistical analyses. To align with ANNOVAR annotations, analysis was limited to variants that were located in the exonic or canonical splice site regions and were not annotated as "unknown" by ANNOVAR. Following the inclusion criteria, we identified any potential coding mutations and generated gene-specific mutation probabilities for 19,347 unique genes. Due to the difference in exome capture kits, DNA sequencing platforms, and variable sequencing coverage between case and control cohorts, separate *de novo* probability tables were generated for cases and controls, respectively.

Estimation of expected number of *de novo* and transmitted variants: We implemented a multivariate regression model to quantify the enrichment of damaging RGs in a specific gene or gene set in cases, independent of controls. Additional details about the modeling of the distribution of recessive and transmitted heterozygous variant counts are described in our recent study[23].

Statistical analysis:
De novo enrichment analysis: The R package 'denovolyzeR' was used for the analysis of DNMs based on a mutation model developed previously[87,88]. The probability of observing a DNM in each gene was derived as described previously[89], except that the coverage adjustment factor was based on the full set of 250 case trios or 1,789 control trios (separate probability tables for each cohort). The overall enrichment was

calculated by comparing the observed number of DNMs across each functional class to expected under the null mutation model. The expected number of DNMs was calculated by taking the sum of each functional class specific probability multiplied by the number of probands in the study, multiplied by two (diploid genomes). The Poisson test was then used to test for enrichment of observed DNMs versus expected as implemented in denovolyzeR[88]. For gene set enrichment, the expected probability was calculated from the probabilities corresponding to the gene set only.

To estimate the number of genes with > 1 DNM, 1 million permutations were performed to derive the empirical distribution of the number of genes with multiple DNMs. For each permutation, the number of DNMs observed in each functional class was randomly distributed across the genome adjusting for gene mutability[23]. The empirical p-value was calculated as the proportion of times that the number of recurrent genes from the permutation is greater than or equal to the observed number of recurrent genes.

To examine whether any individual gene contains more DNMs than expected, the expected number of DNMs for each functional class was calculated from the corresponding probability adjusting for cohort size. The Poisson test was then used to compare the observed DNMs for each gene versus expected. As separate tests were performed for damaging DNMs and LoF DNMs, the Bonferroni multiple-testing threshold is, therefore, equal to 1.3×10^{-6} (0.05/(19,336 genes x 2 tests)). The most significant p-value of the two tests was reported.

Gene-set enrichment analysis: To test for over-representation of a gene set without controls and correct for consanguinity, a one-sided binomial test was conducted by comparing the observed number of variants to the expected count estimated using the method detailed above. Assuming that our exome capture reagent captures N genes and the testing gene set contains M genes, then the p-value of finding k variants in this gene set out of a total of x variants in the entire exome is given by

$$P - value = \sum_{i=k}^{x} \binom{x}{i} (p)^i (1-p)^{n-i}$$

where

$$P = \left(\sum_{gene\ set} Expected\ Value_i\right) \Big/ \left(\sum_{all\ genes} Expected\ Value_j\right)$$

Enrichment was calculated as the observed number of genotypes/variants divided by the expected number of genotypes/variants.

Gene-based binomial test: A one-tailed binomial test was used to compare the observed number of damaging variants within each gene to the expected number estimated using the approach detailed above. Enrichment was calculated as the number of observed damaging genotypes/variants divided by the expected number of damaging genotypes/variants.

Estimation of the number of risk genes: We followed the Monte Carlo simulation strategy described in Jin et al. to estimate the number of risk genes that are DNM targets[23]. We defined K to be the number of observed damaging DNMs in LoF-intolerant genes (pLI $\geq$ 0.9) among cases. R_1 indicates the number of LoF-intolerant genes mutated exactly twice in cases and R_2 indicates the number of LoF-intolerant genes mutated three times or more. Defined as $E = (M_1-M_2)/M_1$, where M_1 and M_2 are the observed and expected count of damaging DNMs per trio, respectively. We then simulated the likelihood function as follows: First, we randomly selected G risk genes from the LoF-intolerant gene set. Next, we simulated the number of contributing damaging mutations in risk genes, i.e. C, by sampling once from the binomial (K,E) distribution. Then, we simulated C contributing damaging mutations in G risk genes and K-C non-contributing damaging mutations in the complete LoF-intolerant gene set using each gene's damaging mutability score as probability weights. We performed 20,000 simulations for G from 10 to 500, and calculated the likelihood function L(G) as the proportion of simulations in which the number of genes with two damaging mutations equals R_1 and the number of genes with three or more damaging mutations equals R_2. We then estimated the number of risk genes using the maximum likelihood estimate (MLE).

Genetic overlap across neurodevelopmental disorders:
We compared the list of 440 putative CP risk genes with genes identified in other major neurodevelopmental disorders using Disease-Gene Network (DisGeNET, updated May 2019)[63]. DisGeNET contains the known association between genes and diseases curated from literature and other curated databases including ClinVar, ClinGen, and UniProt. DisGeNET uses both the disease-gene association and variant-disease association from genome-wide association studies and Mendelian studies using multiple data sources, calculating a score for the strength of association (GDA-gene disease association score and VDA-variant disease association score) based on the number of resources supporting it. We first extracted all the genes from DisGeNET that were associated with autism spectrum disorder (ASD, CUI: C1510586, 571 genes), intellectual disability (ID, CUI: C3714756, 2502 genes) and Epilepsy (EP, CUI: C0014544,1176 genes). We used the hypergeometric probability to calculate the overlap significance. The hypergeometric distribution formula is given by:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where, K= # genes in DisGeNET associated with the disease,
k= # genes in overlapping set with that disease,
N= # total genes in DisGeNET,
n= # total genes in the observed set
A Venn Diagram representing the gene number appearing in more than one list was created in R using the 'VennDiagram' package.

Pathway analysis:

STRING protein-protein interaction enrichment: We used the list of 440 genes to conduct a protein-protein interaction (PPI) enrichment for gene networks. We used STRINGv11 to further study protein interaction networks in our set of 440 putative CP risk genes with either de novo, X-linked recessive or autosomal recessive damaging variants. We used 0.70 (high confidence) cutoff and the following STRING 'channels' to derive these interactions network as described[90,91]:

a. 'databases': curated known protein interaction networks and pathways from databases like KEGG Reactome, BioCyc, Gene Ontology, BioCarta and PID.
b. 'experiments': curated interactions from IMEx The International Molecular Exchange Consortium) and BioGRID and standardizing it against KEGG.
c. 'text-mining': natural language processing to look for co-occurrence of proteins in PubMed abstracts and OMIM databases.

The network visualization can be accessed at:
https://version-11-0.string-db.org/cgi/network.pl?networkId=Ee60E24TJ4Bx

Gene set overrepresentation analysis: We used the list of 440 genes for further downstream gene set over representation analysis using DAVID v6.8[66,92] (updated October 2016), PANTHER v14.1[93-95] (updated 2019-03-12) and M1SigDB v6.2[96] (updated July 2018). The background gene list for all three tools was their respective pool of all human genes. The background databases used by these tools for annotations and derived genes sets include Gene ontology (GO), pathway databases like KEGG and Biocarta and protein domain databases like UniProt. MSigDB online tool has gene sets defined based on computational analysis of microarray experiment submitted to this database as well as curated sets derived from literature in addition to the GO term defined gene sets. To measure statistical overrepresentation of gene sets in the client set, PANTHER uses a Fisher's exact test, DAVID uses a modified Fisher's test and MSigDB uses the hypergeometric method.

DcGO[69] algorithm identifies parent and child nesting GO terms to determine hierarchal relationships. We started from the most specific GO terms (fewest genes) to identify first-level parents. These terms were used with DcGO to identify terms where parent, middle, and child terms were all represented on our list with significant FDR. These nested terms were manually curated for **Table 4**.

***RHOB* functional assays:**

GTPase Activating Protein (GAP) assay: (Cytoskeleton) 13 ug of purified WT or S73F RhoB protein (Origene) was incubated with 20 uM GTP with or without 5 ug or 13 ug of p50 RhoGAP for 30 min at 37°C, then incubated with CytoPhos reagent for 15 min at room temperature. Hydrolyzed GTP was detected at 650 nm on a SpectraMax paradigm microplate reader as per the manufacturer's instructions.

Guanine exchange factor (GEF) assay: (Cytoskeleton) 2 uM of purified WT or S73F RhoB protein (Origene) was incubated with or without 2 uM of the GEF domain of the human Dbs protein for 30 min at 20°C. The fluorescence of N-methylantraniloyl GTP-analogue binding was measured every 30 seconds at 360 nm with the SpectraMax as per the manufacturer's instructions.

Rhotekin assay: (Cytoskeleton) 50 ug of agarose beads coated with the Rho-GTP binding domain (residues 7-89) of the human Rhotekin protein were incubated with 500 ug of lysate from yeast expressing human RHOB-V5 or the S73F variant under gentle agitation for 1h at 4°C. Beads pelleted by centrifugation at 2400 xg (5000 rpm) for 4 min at 4°C and washed three times in Wash Buffer (25 mM Tris pH 7.5, 30 mM MgCl2, 40 mM NaCl). Beads were resuspended in Laemmli blue 2X and 40 ug of lysate used for Western blotting. RhoB was identified with a primary monoclonal anti-V5 antibody (Thermo Fisher) and a secondary goat anti-mouse HRP (GE Healthcare).

***FBXO31* cyclin D abundance assay:** 3 independent, passage-matched control fibroblast lines (GMO8398, GMOR297C, GMO8399 from the Corriell Institute) and patient primary fibroblasts obtained from each patient via punch biopsies were used. Plates were seeded at 600,000 cells/well and cultured in DMEM supplemented with 1mM sodium pyruvate, 1mM glutamine (Gibco) and 10% FBS. Fibroblasts were harvested at confluence with RIPA buffer (Thermo Fisher) supplemented with protease cocktail (Fischer Scientific) on ice and centrifuged. Western blotting was conducted using 10µg protein/lane with antibodies against cyclin D (rabbit polyclonal; ab134175) 1:1000, β-tubulin (rabbit polyclonal, ab6046) 1:5000 in 5% BSA and detected with anti-rabbit HRP (GE Health Sciences) 1:5000. Signal was quantified using Image Studio Lite and the ratio of cyclin D/β-tubulin was normalized to within-experimental control GMO8398. The difference in cyclin D abundance were determined using an unpaired t-test.

Drosophila locomotor experiments:

Fly rearing and genetics: Drosophila were reared on a standard cornmeal, yeast, sucrose food from the BIO5 media facility, University of Arizona. Stocks for experiments were reared at 25°C, 60-80% relative humidity with 12:12 light/dark cycle. Cultures for controls and mutants were maintained with the same growth conditions, with attention to the density of animals within the vial. Descriptions of alleles used for each CP candidate gene can be found in Supplemental Table 7 and included 5' insertional hypomorphs, missense mutations, targeted excision, and deficiency chromosomes. Whenever possible Drosophila genotypes for study were compound heterozygotes to reduce background effects; homozygotes were also used for well-described alleles. Heterozygotes were not used to avoid challenges in mapping phenotypes to the gene of interest. Lines obtained from other investigators included: $CenG1A^{\Delta 9}$ from M. Hoch, $Mkk4^{G587}$ from K. Basler, Canton S from L. Restifo, Fak^{CG-1} from R. Palmer, *ΔCG4030-FL* from D. Strutt, esn^{KO6} from T. Usui. The remaining fly stocks were obtained from the Bloomington Drosophila Stock Center (NIH P40OD018537). The following lines were outcrossed with balancer stocks for a single generation to replace the 1st chromosome with w^+ and swap the 2nd or 3rd chromosome balancer: $Mkk4^{G587}$, $atl^{MB07599}$, Fak^{CG-1}, $Mtrf^{\Delta 7}$, $Mi(y+]Ziz^{04988}$, $Tbc1d15\text{-}17^{EP-234}$, $CG1407^{MI00131}$. We performed crosses of background markers for genetic controls.

Locomotor assays: We used naïve, unmated flies collected as pharate adults. To minimize variables, we used no anesthesia and humidity, temperature, and time of day

were controlled (30-60% RH, 21-23.5°C, 0900-1200). There were no significant effects of any of these parameters found in retrospective analysis within control and mutant genotypes. Flies were adapted to room conditions for 1 hour before running in groups of 3-20 in a 250 mL graduated cylinder for 2 minutes[97]. If <50% crossed the 250 (22.5 cm) mark, flies were re-assayed immediately up to 3 iterations. Flies crossing 250 mL mark (22.5 cm) were manually scored from coded videos in 10 second bins. The number of falls, defined as downward movement while detached from the cylinder wall, were manually counted and normalized to the number of flies in the recording window per 10 second bin. The scoring accuracy was confirmed using the Cohen's kappa statistic between 3 independent observers using a subset of 12 videos (κ 0.89-0.98) and a lack of scoring drift from the beginning to the end of the analysis was also confirmed (κ 0.97). Significant difference of locomotor performance between mutants and controls required $p<0.05$ for both Kolmogorov-Smirnov test for whole curve and Mann-Whitney rank sum test for at least one time bin between 10-30 seconds. Distance traveled assay was performed using paired, coded vials of control and mutant flies[98]. Distance measured from still image from video at 3 seconds post-tapping using ImageJ measure distance function from middle of fly to bottom of vial. Larval turning time was defined as the amount of time required to turn onto ventral surface and initiate forward movement after rotation onto dorsal surface[99]. Significance for vial and larval turning assays were determined using t-test. Graphs and statistics were performed in R. Drosophila locomotor gene enrichment analysis performed as described previously[100] using www.MARRVEL.org and www.flybase.org to identify the Drosophila ortholog and compared to genome-wide number of genes identified by the terms locomotor/locomotion, flight, taxis (photo- or geo-). Significance of enrichment determined using the Fisher exact test.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the patients and families who have graciously and patiently supported this work from its inception. Without their partnership, these studies would not have been possible. We acknowledge the support of the clinicians who generously provided their expertise in support of this study, including Drs. Mary-Clare Waugh, Matthias Axt, and Vicki Roberts of the Children's Hospital Westmead; Kevin Lowe of Sydney Children's Hospital; Ray Russo, James Rice, and Andrew Tidemann of the Women's and Children's Hospital, Adelaide; Theresa Carroll and Lisa Copeland of the Lady Cilento Children's Hospital, Brisbane; and Jane Valentine of Perth Children's Hospital. We appreciate the collaboration of Drs, Susan Knoblach and Eric Hoffman (Children's National Medical Center).

Supported in part by the Cerebral Palsy Alliance Research Foundation (MCK), the Yale-NIH Center for Mendelian Genomics (U54 HG006504-01), Doris Duke Charitable Foundation CSDA 2014112 (MCK), the Scott Family Foundation (MCK), Cure CP (MCK), NIH NS083739 (MCK), NIH NS106298-01A1 (MCK), NIH NS091299 (DCZ), 5R24HD050846-08 (EPH), NHMRC grant 1099163 (AHM, CvE & MAC), Cerebral Palsy Alliance Research Foundation Career Development Award (MAC), The Tenix Foundation (AHM, JG, CvE & MAC), the National Natural Science Foundation of China (U1604165, XW), Henan Key Research Program of China (171100310200, CZ), VINNOVA (2015-04780, CZ). SCJ was supported by the James Hudson Brown-Alexander Brown Coxe Postdoctoral Fellowship at the Yale University School of Medicine and an American Heart Association Postdoctoral Fellowship (18POST34060008) and the NIH K99/R00 Pathway to Independence Award (K99HL143036-01A1).

AUTHOR CONTRIBUTIONS

Study design and oversight: K.B., S.P-L., Q.X., C.Z., R.P.L., A.H.M., J.G., M.C.K.
Cohort ascertainment, recruitment, and phenotypic characterization: B.Y.N., J.B., K.H., J.B-H., A.P., M.C.F., L.X., Y.X., M.C., K.R., F.M., J.L.W., L.R., J.S.C., A.F., A.L., J.P., T.F., S.M., K.E.C., D.M.R., Q.S., G.C., Y.W., N.B., I.N., S.M., X.W., D.A., J.H., M.C.K.
Exome sequencing production and validation: K.B., C.C., A.E., J.L., C.L.vE., H.M., S.M.M., I.R.T., Y.W., B.S.G., J.Z., D.L.W., M.S.B.F., C.Z., M.A.C.
WES analysis: S.B., M.A.C., M.C.S., X. Z., S.C.J.
Yeast complementation: J.L., S.P-L.
Drosophila locomotor experiments: S.A.L., S.V., D.C.Z.
Statistical analysis: S.C.J., S.A.L., S.B., B.L., Q.L., M.C.S., X. Z.
Writing and review of manuscript: S.C.J., S.A.L., D.J.A., K.B., S.B., M.A.C., A.E., J.G., Q.L., S.P-L., R.P.L., A.H.M., S.M., B.N., M.C.S., X. Z., C.L.V., X.W., Q.X., C.Z., M.C.K.
Co-senior authors: K.B., R.P.L., Q.X., C.Z., A.H.M., J.G., S.P-L., M.C.K.

All authors have read and approved the final manuscript

COMPETING INTERESTS

None

REFERENCES

1. Christensen, D. *et al.* Prevalence of cerebral palsy, co-occurring autism spectrum disorders, and motor functioning - Autism and Developmental Disabilities Monitoring Network, USA, 2008. *Dev Med Child Neurol* **56**, 59-65 (2014).
2. Surveillance of Cerebral Palsy in, E. Surveillance of cerebral palsy in Europe: a collaboration of cerebral palsy surveys and registers. Surveillance of Cerebral Palsy in Europe (SCPE). *Dev Med Child Neurol* **42**, 816-24 (2000).
3. Longo, L.D. & Ashwal, S. William Osler, Sigmund Freud and the evolution of ideas concerning cerebral palsy. *J Hist Neurosci* **2**, 255-82 (1993).
4. Panteliadis, C., Panteliadis, P. & Vassilyadi, F. Hallmarks in the history of cerebral palsy: from antiquity to mid-20th century. *Brain Dev* **35**, 285-92 (2013).
5. Tan, S. Fault and blame, insults to the perinatal brain may be remote from time of birth. *Clin Perinatol* **41**, 105-17 (2014).
6. Donn, S.M., Chiswick, M.L. & Fanaroff, J.M. Medico-legal implications of hypoxic-ischemic birth injury. *Semin Fetal Neonatal Med* **19**, 317-21 (2014).
7. Korzeniewski, S.J., Slaughter, J., Lenski, M., Haak, P. & Paneth, N. The complex aetiology of cerebral palsy. *Nat Rev Neurol* **14**, 528-543 (2018).
8. Numata, Y. *et al.* Brain magnetic resonance imaging and motor and intellectual functioning in 86 patients born at term with spastic diplegia. *Dev Med Child Neurol* **55**, 167-72 (2013).
9. Segel, R. *et al.* Copy number variations in cryptogenic cerebral palsy. *Neurology* **84**, 1660-8 (2015).
10. McIntyre, S. *et al.* Congenital anomalies in cerebral palsy: where to from here? *Dev Med Child Neurol* **58 Suppl 2**, 71-5 (2016).
11. Petterson, B., Stanley, F. & Henderson, D. Cerebral palsy in multiple births in Western Australia: genetic aspects. *Am J Med Genet* **37**, 346-51 (1990).
12. Costeff, H. Estimated frequency of genetic and nongenetic causes of congenital idiopathic cerebral palsy in west Sweden. *Ann Hum Genet* **68**, 515-20 (2004).
13. Hallmayer, J. *et al.* Genetic heritability and shared environmental factors among twin pairs with autism. *Arch Gen Psychiatry* **68**, 1095-102 (2011).
14. Sandin, S. *et al.* The Heritability of Autism Spectrum Disorder. *JAMA* **318**, 1182-1184 (2017).
15. McMichael, G. *et al.* Rare copy number variation in cerebral palsy. *Eur J Hum Genet* **22**, 40-5 (2014).
16. Oskoui, M. *et al.* Clinically relevant copy number variations detected in cerebral palsy. *Nat Commun* **6**, 7949 (2015).
17. Zarrei, M. *et al.* A de novo deletion in a boy with cerebral palsy suggests a refined critical region for the 4q21.22 microdeletion syndrome. *Am J Med Genet A* **173**, 1287-1293 (2017).
18. Zarrei, M. *et al.* De novo and rare inherited copy-number variations in the hemiplegic form of cerebral palsy. *Genet Med* **20**, 172-180 (2018).
19. Takezawa, Y. *et al.* Genomic analysis identifies masqueraders of full-term cerebral palsy. *Ann Clin Transl Neurol* **5**, 538-551 (2018).
20. Parolin Schnekenberg, R. *et al.* De novo point mutations in patients diagnosed with ataxic cerebral palsy. *Brain* **138**, 1817-32 (2015).
21. McMichael, G. *et al.* Whole-exome sequencing points to considerable genetic heterogeneity of cerebral palsy. *Mol Psychiatry* **20**, 176-82 (2015).
22. Rosenbaum, P. *et al.* A report: the definition and classification of cerebral palsy April 2006. *Dev Med Child Neurol Suppl* **109**, 8-14 (2007).
23. Jin, S.C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat Genet* **49**, 1593-1601 (2017).
24. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat Genet* **47**, 582-8 (2015).

25. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
26. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 1-33 (2013).
27. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* **24**, 2125-37 (2015).
28. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
29. Wei, Q. *et al.* A Bayesian framework for de novo mutation calling in parents-offspring trios. *Bioinformatics* **31**, 1375-81 (2015).
30. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).
31. Rainier, S., Sher, C., Reish, O., Thomas, D. & Fink, J.K. De novo occurrence of novel SPG3A/atlastin mutation presenting as cerebral palsy. *Arch Neurol* **63**, 445-7 (2006).
32. Blom, N., Gammeltoft, S. & Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* **294**, 1351-62 (1999).
33. Tillement, V. *et al.* Phosphorylation of RhoB by CK1 impedes actin stress fiber organization and epidermal growth factor receptor stabilization. *Exp Cell Res* **314**, 2811-21 (2008).
34. McNair, K. *et al.* A role for RhoB in synaptic plasticity and the regulation of neuronal morphology. *J Neurosci* **30**, 3508-17 (2010).
35. Deshaies, R.J. & Joazeiro, C.A. RING domain E3 ubiquitin ligases. *Annu Rev Biochem* **78**, 399-434 (2009).
36. Vadhvani, M., Schwedhelm-Domeyer, N., Mukherjee, C. & Stegmuller, J. The centrosomal E3 ubiquitin ligase FBXO31-SCF regulates neuronal morphogenesis and migration. *PLoS One* **8**, e57530 (2013).
37. Libraries. Success at Boston Spa. *Nature* **222**, 113-4 (1969).
38. Mir, A. *et al.* Truncation of the E3 ubiquitin ligase component FBXO31 causes non-syndromic autosomal recessive intellectual disability in a Pakistani family. *Hum Genet* **133**, 975-84 (2014).
39. Hebebrand, M. *et al.* The mutational and phenotypic spectrum of TUBA1A-associated tubulinopathy. *Orphanet J Rare Dis* **14**, 38 (2019).
40. Song, D.H. *et al.* CK2 phosphorylation of the armadillo repeat region of beta-catenin potentiates Wnt signaling. *J Biol Chem* **278**, 24018-25 (2003).
41. Panagiotou, E.S. *et al.* Defects in the Cell Signaling Mediator beta-Catenin Cause the Retinal Vascular Condition FEVR. *Am J Hum Genet* **100**, 960-968 (2017).
42. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* **367**, 1921-9 (2012).
43. Tucci, V. *et al.* Dominant beta-catenin mutations cause intellectual disability with recognizable syndromic features. *J Clin Invest* **124**, 1468-82 (2014).
44. Kharbanda, M. *et al.* Clinical features associated with CTNNB1 de novo loss of function mutations in ten individuals. *Eur J Med Genet* **60**, 130-135 (2017).
45. Chen, J., Knowles, H.J., Hebert, J.L. & Hackett, B.P. Mutation of the mouse hepatocyte nuclear factor/forkhead homologue 4 gene results in an absence of cilia and random left-right asymmetry. *J Clin Invest* **102**, 1077-82 (1998).
46. Orso, G. *et al.* Homotypic fusion of ER membranes requires the dynamin-like GTPase atlastin. *Nature* **460**, 978-83 (2009).
47. Guelly, C. *et al.* Targeted high-throughput sequencing identifies mutations in atlastin-1 as a cause of hereditary sensory neuropathy type I. *Am J Hum Genet* **88**, 99-105 (2011).

48. Zhao, X. *et al.* Mutations in a newly identified GTPase gene cause autosomal dominant hereditary spastic paraplegia. *Nat Genet* **29**, 326-31 (2001).
49. Hazan, J. *et al.* Spastin, a new AAA protein, is altered in the most frequent form of autosomal dominant spastic paraplegia. *Nat Genet* **23**, 296-303 (1999).
50. Burger, J. *et al.* Hereditary spastic paraplegia caused by mutations in the SPG4 gene. *Eur J Hum Genet* **8**, 771-6 (2000).
51. Hazan, J. *et al.* A fine integrated map of the SPG4 locus excludes an expanded CAG repeat in chromosome 2p-linked autosomal dominant spastic paraplegia. *Genomics* **60**, 309-19 (1999).
52. de la Cruz, J., Kressler, D. & Linder, P. Unwinding RNA in Saccharomyces cerevisiae: DEAD-box proteins and related families. *Trends Biochem Sci* **24**, 192-8 (1999).
53. Della Corte, C.M. *et al.* Role and targeting of anaplastic lymphoma kinase in cancer. *Mol Cancer* **17**, 30 (2018).
54. Chen, Y. *et al.* Oncogenic mutations of ALK kinase in neuroblastoma. *Nature* **455**, 971-4 (2008).
55. Janoueix-Lerosey, I. *et al.* Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma. *Nature* **455**, 967-70 (2008).
56. Schule, R. *et al.* Hereditary spastic paraplegia: Clinicogenetic lessons from 608 patients. *Ann Neurol* **79**, 646-58 (2016).
57. Parodi, L. *et al.* Spastic paraplegia due to SPAST mutations is modified by the underlying mutation and sex. *Brain* **141**, 3331-3342 (2018).
58. Solowska, J.M., Rao, A.N. & Baas, P.W. Truncating mutations of SPAST associated with hereditary spastic paraplegia indicate greater accumulation and toxicity of the M1 isoform of spastin. *Mol Biol Cell* **28**, 1728-1737 (2017).
59. Ji, Z. *et al.* Spastin Interacts with CRMP5 to Promote Neurite Outgrowth by Controlling the Microtubule Dynamics. *Dev Neurobiol* **78**, 1191-1205 (2018).
60. Gao, Y. *et al.* Atlastin-1 regulates dendritic morphogenesis in mouse cerebral cortex. *Neurosci Res* **77**, 137-42 (2013).
61. Romeo, D.M. *et al.* Sex differences in cerebral palsy on neuromotor outcome: a critical review. *Dev Med Child Neurol* **58**, 809-13 (2016).
62. Reid, S.M., Meehan, E.M., Arnup, S.J. & Reddihough, D.S. Intellectual disability in cerebral palsy: a population-based retrospective study. *Dev Med Child Neurol* **60**, 687-694 (2018).
63. Pinero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* **45**, D833-D839 (2017).
64. Al-Mubarak, B. *et al.* Whole exome sequencing reveals inherited and de novo variants in autism spectrum disorder: a trio study from Saudi families. *Sci Rep* **7**, 5679 (2017).
65. Giacopuzzi, E. *et al.* Exome sequencing in schizophrenic patients with high levels of homozygosity identifies novel and extremely rare mutations in the GABA/glutamatergic pathways. *PLoS One* **12**, e0182778 (2017).
66. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
67. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).
68. Mi, H. *et al.* Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* **14**, 703-721 (2019).
69. Fang, H. & Gough, J. DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic Acids Res* **41**, D536-44 (2013).
70. Hemminki, K., Li, X., Sundquist, K. & Sundquist, J. High familial risks for cerebral palsy implicate partial heritable aetiology. *Paediatr Perinat Epidemiol* **21**, 235-41 (2007).

71. Ellenberg, J.H. & Nelson, K.B. The association of cerebral palsy with birth asphyxia: a definitional quagmire. *Dev Med Child Neurol* **55**, 210-6 (2013).
72. van Eyk, C.L. *et al.* Analysis of 182 cerebral palsy transcriptomes points to dysregulation of trophic signalling pathways and overlap with autism. *Transl Psychiatry* **8**, 88 (2018).
73. Martinelli, S. *et al.* Functional Dysregulation of CDC42 Causes Diverse Developmental Phenotypes. *Am J Hum Genet* **102**, 309-320 (2018).
74. Englander, Z.A. *et al.* Brain structural connectivity increases concurrent with functional improvement: evidence from diffusion tensor MRI in children with cerebral palsy during therapy. *Neuroimage Clin* **7**, 315-24 (2015).
75. Hou, S.T., Jiang, S.X. & Smith, R.A. Permissive and repulsive cues and signalling pathways of axonal outgrowth and regeneration. *Int Rev Cell Mol Biol* **267**, 125-81 (2008).
76. Colombo, S. *et al.* G protein-coupled potassium channels implicated in mouse and cellular models of GNB1 Encephalopathy. *bioRxiv*, 697235 (2019).
77. Pipo-Deveza, J. *et al.* Rationale for dopa-responsive CTNNB1/ss-catenin deficient dystonia. *Mov Disord* **33**, 656-657 (2018).
78. Akizu, N. *et al.* AMPD2 regulates GTP synthesis and is mutated in a potentially treatable neurodegenerative brainstem disorder. *Cell* **154**, 505-17 (2013).
79. Miller, S.P., Shevell, M.I., Patenaude, Y. & O'Gorman, A.M. Neuromotor spectrum of periventricular leukomalacia in children born at term. *Pediatr Neurol* **23**, 155-9 (2000).
80. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).
81. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
82. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
83. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
84. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
85. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
86. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).
87. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nature genetics* **46**, 944 (2014).
88. Ware, J.S., Samocha, K.E., Homsy, J. & Daly, M.J. Interpreting de novo Variation in Human Disease Using denovolyzeR. *Curr Protoc Hum Genet* **87**, 7 25 1-15 (2015).
89. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262-6 (2015).
90. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613 (2019).
91. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**, D808-15 (2013).
92. Huang da, W., Sherman, B.T. & Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009).
93. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P.D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* **47**, D419-D426 (2019).

94. Bisaccia, M. *et al.* Dhs plus anti-rotational screw vs cannulated screws for femoral neck fractures: an analysis of clinical outcome and incidence regarding avn. *Acta Orthop Belg* **84**, 279-283 (2018).
95. Thomas, P.D. *et al.* Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res* **34**, W645-50 (2006).
96. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
97. Madabattula, S.T. *et al.* Quantitative Analysis of Climbing Defects in a Drosophila Model of Neurodegenerative Disorders. *J Vis Exp*, e52741 (2015).
98. Kim, M. *et al.* Mutation in ATG5 reduces autophagy and leads to ataxia with developmental delay. *Elife* **5**(2016).
99. Estes, P.S. *et al.* Wild-type and A315T mutant TDP-43 exert differential neurotoxicity in a Drosophila model of ALS. *Hum Mol Genet* **20**, 2308-21 (2011).
100. Aleman-Meza, B., Loeza-Cabrera, M., Pena-Ramos, O., Stern, M. & Zhong, W. High-content behavioral profiling reveals neuronal genetic network modulating Drosophila larval locomotor program. *BMC Genet* **18**, 40 (2017).

FIGURE AND TABLE LEGENDS

Table 1. Significant enrichment of *de novo* mutations in CP cases and genes with > 1 damaging *de novo* mutations. (A) A Poisson test was used to test the enrichment of *de novo* mutations for each functional class. A marginal enrichment of *de novo* mutations was observed for loss-of-function (LoF), protein-altering, and damaging *de novo* mutations. Strikingly, when we restricted our analysis to LoF-intolerant genes, stronger enrichment was observed for protein-altering and damaging *de novo* mutations, suggesting a significant contribution of *de novo* mutation in this gene set to CP pathogenesis. No enrichment was found in controls. **(B)** Nine genes with two or more damaging (LoF + D-Mis) *de novo* mutations were found. A Poisson-test was performed for damaging and LoF *de novo* mutations for each gene independently. The Bonferroni correction for genome-wide significance is $1.3x10^{-6}$ (= 0.05/(19,347 genes x 2 tests)). N: the number of *de novo* mutations; Rate: the number of *de novo* mutations divided by the number of individuals in the cohort; Enrichment: ratio of observed to expected numbers of mutations; D-Mis: Damaging missense mutations as predicted by MetaSVM and CADD algorithms; Protein-altering: Missense + LoF; Damaging: D-Mis+LoF.

Table 2. Clinical and genetic features of recurrent *de novo* mutations.
Abbreviations: intellectual disability (ID), autism spectrum disorder (ASD), coarctation of the aorta (CoArc), periventricular leukamalacia (PVL), attention deficit/hyperactivity disorder (ADHD).

Fig 1. Functional validation of CP-associated *RHOB* variant S73F (A) Sanger traces of mother, father, and proband from families F064 and F244 verifies *de novo* inheritance and position of variant (red arrow). **(B)** Poisson-Boltzman map of wildtype RHOB (left) and F73 variant (right) showing changes to the kinase binding site (arrow) and surface charge of the protein. Alignment between human Rho-family proteins shows high conservation of the RhoB 73 residue in the Switch II domain. The site of S73/F73 has been labeled (arrow). **(C)** Brain MRI from F064 demonstrates bilateral periventricular T2/FLAIR hyperintensity (arrows) on axial imaging, while sagittal views reveals equivocal thinning of the isthmus of the corpus callosum (star). MRI from F244 T2 hyperintensity of posterior limb of internal capsule and optic radiations (left) and hyperintensity of periventricular white matter (right). **(D)** GTP hydrolysis is enhanced ~1.5-fold in the S73F RHOB variant in a GTPase activating protein (GAP) assay. Absorbance measurements of hydrolyzed GTP were recorded for 3 trials in the presence of either low (5 µg) or high (13 µg) RhoA GAP. There was no change in endogenous GTPase activity with S73F variant without added GAP added (not shown). **(E)** GTP binding is enhanced in the S73F RHOB variant in a guanine exchange factor (GEF) assay. The N-methylantraniloyl-GFP fluorophore increases its fluorescence emission when bound to Rho-family GTPases, indicating nucleotide uptake by the GTPase. Both WT and S73F have low endogenous GTP binding (lower curves). In the presence of the Dbs GEF protein, GTP binding is enhanced, and S73F Km is significantly reduced compared to wildtype RHOB (average of 5 replicates; mean 243 vs 547 seconds, $p<0.002$). **(F)** S73F GTP-binding is increased 4-fold in a pull-down assay

with Rhotekin, an interactor with active GTP-bound Rho proteins. (Top) Sample WB of RHOB from bead-bound fraction and total input detected using antibody against V5 tag. (Bottom) Quantification of ratio of rhotekin-bound/total RHOB from 5 replicates. Bars in D-F indicate standard error. RFU = relative fluorescence units (10^6) with 360 nm excitation. Statistics determined by unpaired t-test. **p<0.002, ****p<$6x10^{-5}$.

Fig 2. Functional validation of CP-associated *FBXO31* variant D334N shows alterations in cyclin D regulation (A) Sanger traces of mother, father, and proband from families F218 and F699 verifies *de novo* inheritance and position of variant (red arrow). **(B)** Poisson-Boltzman map of wildtype FBXO31 (left) and the D334N variant (right). D334 is positioned around the cyclin D1 (green) binding pocket on FBXO31. The mutation alters the surface electrostatic charge around the cyclin D1 binding site with a predicted effect on cyclin D1 binding to FBXO31. The site of D334/D334N has been labeled (arrow). **(C)** Magnified view of B showing alterations to surface charge in cyclin D1 binding site. **(D)** Representative western blot of decreased cyclin D expression in patient-derived fibroblasts with *FBXO31* p.D334N variant. Quantification of Cyclin D is normalized to in-lane β-tubulin and within-experiment control GMO8398. Both patients had reduced cyclin D compared to pooled controls. Data averaged for three independent experiments, bars represent standard error, statistics calculated using 2-tailed t-test.

Table 3. Idiopathic CP cases show enrichment of damaging recessive genotypes in hereditary spastic paraplegia-associated (HSP) genes. (A) One-tailed binomial test coupled with the polynomial regression model was conducted to evaluate the enrichment of damaging recessive genotypes in known HSP-associated genes in cases and in controls respectively. Stratified analysis by the diagnosis of CP revealed the enrichment of these recessive genotypes was specific to cryptogenic cases. Multiple-testing cutoff was $6.3x10^{-3}$ (= 0.05/(2x4)). **(B)** Damaging recessive genotypes in known recessive hereditary spastic paraplegia genes. ID = intellectual disability, ASD = autism spectrum disorder, ADHD = attention deficit-hyperactivity disorder, OCD = obsessive-compulsive disorder.

Figure 3. Genetic overlap between common neurodevelopmental disorders. (A) Venn diagram showing number of overlapping genes between candidate cerebral palsy (CP) genes and genes linked to other neurodevelopmental disorders (NDDs) intellectual disability (ID), epilepsy, and autism spectrum disorder (ASD). CP risk genes were identified as having one or more damaging variant across modes of inheritance with overlap determined using DisGenNET. **(B)** Overlap between CP and other NDDs was significant by hypergeometric test, while overlap between CP and Alzheimer's disease was not. Total Number of Genes in DisGeNET = 17,549; Total Number of genes in our gene set = 440.

Table 4. CP risk gene pathway and function enrichment. (A) Key pathways overlapping between DAVID, PANTHER, and MSigDB bioinformatics tools. FDR differences are due to differences in tool methodologies. **(B)** Gene ontology (GO) terms include cell projections, cytoskeleton, and Rho GTPase signaling. GO terms were

extracted from the total set (Supplemental Excel 2) using hierarchical nesting, or functions that were represented by multiple GO terms. Overlap/Set refers to number of genes overlapping between CP risk gene and Database/Number of genes in Database for that term. FDR = q value (false discovery rate cutoff = 0.05).

Figure 4: Locomotor phenotypes of loss of function mutations in Drosophila orthologs of candidate cerebral palsy risk genes. (A) Turning time, a measure of coordinated movements, is increased in larva with mutations in *AGAP1*, *SEMA4A,* and *TENM1* orthologs. Drosophila mutant and control genotypes in **Supplemental Table 7**. **B-I**. 14 day-old adult flies have locomotor impairments. **B-E**. Negative geotaxis climbing defects in distance threshold assay for flies with mutations in orthologs of *DOCK11* **(B)**, *RABEP1* **(C)**, *PTK2B* **(D)** and *ATL1* **(E)**. Some genotypes have a male-specific locomotor defect **(C)**. **F-G**, Increased number of falls for flies with mutations in *SYNGAP1* **(F)** and *TBC1D17* **(G)** orthologs, although % reaching threshold distance was normal (not shown). **H-I**, Impairments in the average distance traveled of flies with mutations in *MKL1* **(H)** and *ZDHHC15* **(I)** orthologs. Related GO terms for genes shown in bold. Statistics between box and whisker plots determined using t-test. Locomotor curves considered to be significantly different if $p<0.05$ for Kolomogrov-Smirnov test in addition to a significant difference at one or 1 or more times by Mann-Whitney rank sum test. * $p< 0.05$, ** $p<0.005$, *** $p<0.001$, **** $p <1x10^{-6}$. Difference between larval turning time, distance traveled, and number of falls determined by t-test. **(J)** Enrichment of locomotor phenotypes detected in studies of putative CP genes (Observed) compared to genome-wide rates annotated in Flybase.org (Expected, 3.1%). p value calculated by Fisher test.