The contribution of de novo coding mutations to meningomyelocele

Yoo-Jin Ha[1,2,3], Isaac Tang[1,2], Ashna Nisal[1,2], Ishani Jhamb[1,2], Cassidy Wallace[1,2], Sarah Schroeder[1,2], Chanjae Lee[4], Keng loi Vong[1,2], Naomi Meave[1,2], Fiza Jiwani[1,2], Chelsea Barrows[1,2], Sangmoon Lee[1,2], Nan Jiang[1,2], Arzoo Patel[1,2], Francisco A. Blanco[5], Seyoung Yu[6], Hui Su Jeong[1,2,7], Isaac Plutzer[8], Michael B. Major[8], Béatrice Benoit[9], Christian Poüs[9,10], Caleb Heffner[11], Zoha Kibar[12], Gyang Markus Bot[13], Hope Northrup[14], Kit Sing Au[14], Madison Strain[15], Allison Ashley-Koch[15], Richard H. Finnell[16], Joan T. Le[17], Hal Meltzer[17], Camila Araujo[18], Helio R. Machado[18], Roger E. Stevenson[19], Anna Yurrita[20], Sara Mumtaz[21], Osvaldo M. Mutchinick[22], José Ramón Medina-Bereciartu[23], Friedhelm Hildebrandt[24], Gia Melikishvili[25], Rony Marwan[26], Valeria Capra[27], Mahmoud M. Noureldeen[28], Aida M.S. Salem[28], Mahmoud Y. Issa[29], Maha S. Zaki[29], Ji Eun Lee[7], Anna Alkelai[30], Alan R. Shuldiner[30], Stephen F. Kingsmore[1], Stephen A. Murray[11], Heon Yung Gee[6], W. Todd Miller[31,32], Kimberley F. Tolias[5], John B. Wallingford[4], Spina Bifida Sequencing Consortium*^, Sangwoo Kim[3,#], Joseph G. Gleeson[1,2#^]

[1] Rady Children's Institute for Genomic Medicine, San Diego, CA 92123, USA
[2] Department of Neurosciences and Pediatrics, University of California, San Diego, CA 92037, USA
[3] Department of Biomedical Systems Informatics and Graduate School of Medical Science, Brain Korea 21 Project, Yonsei University College of Medicine, Seoul 03722, Republic of Korea
[4] Dept. of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA
[5] Dept. Neuroscience, Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030, USA
[6] Department of Pharmacology and Graduate School of Medical Science, Brain Korea 21 Project, Yonsei University College of Medicine, Seoul 03722, Republic of Korea
[7] Department of Health Sciences and Technology, SAIHST, Sungkyunkwan University, Seoul 06351, Republic of Korea
[8] Department of Cell Biology and Physiology, Washington University in St. Louis, MO 63110
[9] INSERM UMR-S 1193, UFR de Pharmacie, University Paris-Saclay, UFR de Pharmacie, Orsay 91400, France.
[10] Biochimie-Hormonologie, Assistance Publique - Hôpitaux de Paris Université Paris-Saclay, Clamart 92140, France
[11] The Jackson Laboratory, Bar Harbor, ME 04609, USA
[12] Department of Neurosciences, Research Center of CHU Sainte Justine, University of Montreal, Montreal, H3T 1C5, Canada
[13] Neurosurgery Division, Department of Surgery, Jos University Teaching Hospital, Jos 930103, Nigeria
[14] Department of Pediatrics, McGovern Medical School at the University of Texas Health Science Center at Houston and Children's Memorial Hermann Hospital, Houston, TX 77030, USA
[15] Molecular Physiology Institute, Duke University Medical Center, Durham, NC 27710, USA
[16] Center for Precision Environmental Health, Departments of Molecular and Human Genetics, Molecular and Cellular Biology and Medicine, Baylor College of Medicine, Houston 77030, TX, USA
[17] Rady Children's Hospital, San Diego, CA 92123, USA

[18] Department of Surgery and Anatomy Ribeirão Preto Medical School University of São Paulo, Ribeirao Preto SP 14015-068, Brazil
[19] J.C. Self Research Institute of Human Genetics, Greenwood Genetic Center, Greenwood, SC 29646, USA
[20] Catedrática de Ciencias Ómicas, Facultad de Medicina, Universidad Francisco Marroquín, 01001, Guatemala
[21] National University of Medical Sciences, Rawalpindi 44000, Pakistan
[22] Department of Genetics, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City 14080, Mexico
[23] Clínica Santa Sofia, Caracas 1061, Venezuela
[24] Division of Nephrology, Boston Children's Hospital, Boston, MA 02115, USA
[25] Department of Pediatrics, MediClubGeorgia Medical Center, Tbilisi, 0160, Georgia
[26] Division of Pediatric Surgery, University of Colorado School of Medicine, Children's Hospital of Colorado, Colorado Fetal Care Center, Aurora, CO, 80045, USA
[27] Genomics and Clinical Genetics Unit, IRCCS Istituto Giannina Gaslini, Genoa 16147, Italy
[28] Department of Pediatrics, Faculty of Medicine, Beni-Suef University, Beni-Suef 62521, Egypt
[29] Clinical Genetics Department, Human Genetics and Genome Research Division, National Research Centre, Cairo 12311, Egypt
[30] Regeneron Genetics Center, Tarrytown, NY 10591, USA
[31] Department of Physiology and Biophysics, Stony Brook University, Stony Brook, NY 11794, USA
[32] VA Medical Center, Northport, NY 11768, USA

Co-corresponding authors
*Consortium membership in appendix
^ Contact PI for consortium: Joseph G. Gleeson

Keywords: Meningomyelocele, spina bifida, neural tube defect, de novo mutation, penetrance, expressivity, trio, exome, protein-interaction network, congenital, hydrocephalus.
Abstract word count: 158
Text word Count: 2877
Number of Figures: 4
Number of Tables: 1
Number of Supplementary Figures: 16

Abstract

Meningomyelocele (MM) is considered a genetically complex disease resulting from failure of neural tube closure (NTD). Patients display neuromotor disability and frequent hydrocephalus requiring ventricular shunting. A few proposed genes contribute to disease susceptibility, but most risk remains unexplained[1]. We postulated that de novo mutations (DNMs) under purifying selection contribute to MM risk[2]. Here we recruited a cohort of 851 MM trios requiring shunting at birth, compared with 732 control trios, and found that de novo likely gene disrupting or damaging missense mutations occur in approximately 22.3% of subjects, 28% of which are estimated to contribute to disease risk. The 187 genes with damaging DNMs collectively define networks including actin cytoskeleton and microtubule-based processes, axon guidance, and histone modification. Gene validation demonstrates partial or complete loss of function, impaired signaling and defective neural tube closure in *Xenopus* embryos. Our results suggest DNMs make key contributions to MM risk, and highlight critical pathways required for neural tube closure in human embryogenesis.

Main

MM, also known as spina bifida, is the most common central nervous system (CNS) structural defect in humans. MM results from failed closure of the neural tube in the first six weeks of gestation. While folic acid supplementation has reduced disease burden[3], the incidence of 1:3000–10,000 live-births, along with the associated lifelong neuromotor disability and increased mortality, has focused attention on this condition. Timely diagnosis allows for prenatal counseling and informed management choices, including termination of pregnancy, fetal surgery, or postnatal surgery for the nearly fully concordant hydrocephalus accompanying the hindbrain Chiari type II malformation[4,5]. Over 20 million people worldwide live with an NTD yet causes remain poorly defined.

Mouse knockout studies established hundreds of genes associated with NTD, often with partial penetrance, implicating apical-basal polarity, Wnt/PCP, Wnt/beta-catenin, and DNA transcription pathways[6]. Despite epidemiological heritability estimates of 60-70% for NTDs[7], few genes are implicated in human NTDs, potentially due to genetic heterogeneity. Additive effects of common SNPs have been proposed, especially in the folate metabolism gene *MTHFR* (C677T, p.Ala222Val), but meta-analysis from 22 different association studies suggest only a modest risk (OR 1.23, 95% CI: 1.07–1.42)[8]. Inherited mutations in *VANGL1/2, TBXT, CCL2, CELSR1* were identified in cohort studies[9-12], but represent only a few percent of cases. Prior successes in de novo mutation (DNM) approaches in conditions like autism and congenital heart disease, where mutations are potentially under purifying selection[13,14], prompted a trio approach in 43 families, identifying recurrent mutations in *SHROOM3*[15], but left unanswered whether larger studies would be better powered.

Establishing the Spina Bifida Sequencing Consortium

A critical step in DNM discovery is assessment for an excess burden of damaging mutations in cases vs. controls[13]. With the conservative assumptions of a modest increase in damaging DNMs in cases and estimating perhaps 50–100 total MM genetic risk loci, we calculated that a cohort of 300–500 trios would be required to identify 5–10 recurrently mutated genes (see **Supplementary Information** and **Extended Data Fig.**

1). In 2015, we thus established the Spina Bifida Sequencing Consortium (SBSC) to aggregate prior CDC (Centers for Disease Control and Prevention, USA)-supported trio recruitment efforts, alongside targeted new recruitment from specialists in more than 12 countries, and through social media outreach. Recruitment was limited to diagnosis of MM probands with open neural tissue observed at birth and hydrocephalus requiring shunting, specifying the most severe form of MM compatible with long term survival (**Methods**). Our ethnically diverse cohort was designed specifically to assess risk from DNMs using trio sequencing, matched with 732 control trios from the Simons Simplex Collection[16]. After five years of recruitment, we had enrolled 325 trios, performed trio whole exome sequencing (WES) and DNM analysis, but found no recurrently mutated genes, so we extended the recruitment timeline for an additional five years, to double the cohort size.

After 10 years of recruitment, we assembled a cohort of 851 MM trios (839 trios and 6 quartets) consisting of exome sequence data of 2,541 individuals. SBSC was populated by 15 different worldwide sites, incorporating both historic CDC-assembled cohorts and an aggregated earlier report of 39 NTD trios (35 trios and 2 quartets)[15].

Excessive damaging DNM burden in MM trios

After a series of strict quality controls and kinship analyses applied to trio exomes, we constructed a high-confidence call set of 2,592 DNMs (1,458 from 777 MM trios, and 1,134 from 725 control trios) (**Methods**), and confirmed strong DNM signatures[17] (**Extended Data Fig. 2a–2b**). The rates of the total and functionally categorized DNMs (LGD: likely gene disrupting, D-Mis: damaging missense, tolerant missense, and synonymous) were calculated for the jointly covered consensus regions (**Fig. 1a)**, wherein total DNM rates lie in a Poisson distribution (**Extended Data Fig. 2c–2d**, **Methods**). The average mutation rates of total DNMs were 1.22×10^{-8} (1.13×10^{-8}–1.31×10^{-8}, 95% CI) in MM and 1.18×10^{-8} (1.09×10^{-8}–1.28×10^{-8}, 95% CI) in controls, which did not differ significantly (**Extended Data Fig. 2e**, Wilcoxon rank-sum test *P* value: 0.56), and were also comparable to the rates from previous studies (1.08×10^{-8}–1.32×10^{-8})[18,19].

In contrast, we observed an excessive rate of LGD DNMs (frameshift Indels, stop-gained, and splice donor/acceptor mutations) in MM (1.2×10^{-9}) compared to controls (5.0×10^{-10}) (**Fig. 1b**), with a ratio of the two Poisson rates (or rate ratio)[20] of 2.10 (*P* value: 9.0×10^{-6}). Case-control comparison of DNM rate per proband (i.e. theoretical rate: 0.12 in MM vs. 0.05 in controls, see **Methods**), suggested that 7.2% (95% CI: 2.96%–11.36%) of MM probands display LGD DNM implicated in MM risk (**Table 1**). Based upon the excessive burden, approximately 52.43% of LGD DNMs in MM probands was estimated to contribute to MM risk (**Methods**), comparable with rates in other complex diseases such as Tourette (~51.3%) and autism spectrum disorders (~42%)[13,21]. When LGD DNMs were considered alongside D-Mis DNMs (i.e. LGD + D-Mis; referred to as 'damaging DNMs'), 28.26% (95% CI: 7.75%–48.76%) were linked to MM risk with excessive burden. Of note, a subset of high confidence D-Mis DNMs (referred to as 'D-Mis-HC') based upon pathogenicity prediction, conferred approximately 80% of the total LGD risk burden (rate ratio: 1.86, *P* value: 2.0×10^{-6}). Paternal and maternal ages, expected to be correlated with DNM occurrence[22,23], were not

correlated with the excessive DNMs burden (**Extended Data Fig. 3a**). Moreover, there was no enrichment in tolerant and synonymous missense DNMs.

In 3,060 genes considered to be highly constrained (i.e. the probability of loss-of-function intolerance: pLI ≥ 0.9), the burden of D-Mis DNMs appeared even more significant (rate ratio: 2.10, $P = 0.003$) (**Fig. 1c**). Furthermore, we observed a substantially increased burden of D-Mis-HC DNMs (rate ratio: 7.7, $P = 0.002$). This contrasts with the DNM contributions of other complex diseases, where LGDs play outsized roles compared with D-Mis DNMs[13].

To assess whether there is major missed heritability using WES compared with whole genome sequencing (WGS), we recruited an additional 101 trios and 1 quartet, and performed WGS using standard discovery pipelines (**Methods**). We identified non-recurrent de novo SNV/Indels at expected rates of 1.72×10^{-8} (vs. 1.2×10^{-8}–2.4×10^{-8} in previous publications[19,24] ,see **Extended Data Table 1**), encompassing 9 putative splice-disrupting and 41 in putative enhancer/promoter regions (**Supplementary Data**). Copy number variant (CNV) analysis revealed two nonrecurrent deletions and two duplications (> 100kb), a rate comparable to other complex diseases[25,26] (**Extended Data Fig. 3b** and **Extended Data Table 2**). De novo inversions or translocations were not observed. In contrast, a high LGD rate (0.13 per proband) in coding region was replicated in WGS, with three overlapping genes observed in the WES cohort (*KDM1A*, *ITPR3*, and *GRHL2*). Overall, WGS-based analysis mirrored the findings from WES and did not identify an additional major class of mutation (see **Supplementary Information**).

Functional convergence of DNM genes in MM

Out of 198 damaging DNMs (82 LGD and 116 D-Mis) found in MM trios, Sanger sequencing was conducted in 86% (170 out of 198) where sufficient DNA was available, yielding 96.5% validation rate, which removed six false DNMs. The remaining 192 damaging DNMs (79 LGD and 113 D-Mis) occurred in 187 unique genes, with only five genes mutated in two separate trios (referred to as 'doubleton'; *PAX3*, *IRS1*, *ZSWIM6*, *BRSK2*, and *VWA8*) (**Fig. 2a** and **Supplementary Data**). Notably, only *PAX3* was previously implicated in human NTDs[15,27]. Among genes identified as singletons, only *TBXT* (splice donor) and *CELSR1* (frameshift) were previously implicated as inherited NTD risk factors[9,11]. The lack of additional recurrently mutated genes underscores a significant gap in our current understanding of MM genetic risk factors.

Spatial transcriptomic analysis of 36 of the damaging DNM genes confirmed expression at mouse embryonic day 10.5 coinciding with neural tube closure (MERFISH, see **Methods**, **Extended Data Fig. 4**). We found that most genes were expressed broadly in the embryo (86%, 31 out of 36), but a minority (14%) showed some cell type specific expression such as *Celsr1* in neural progenitors and *Stab1* in neural crest progenitors (**Extended Data Fig. 5**, **Supplementary Information**).

We next studied the protein interactions of MM DNM genes using the STRING database[28] (**Fig. 2b**). Among the 187 genes, 107 (57%) including 4 doubletons, 33 LGDs, and 74 D-Mis were highly interconnected, compared to controls (38% interconnected) (**Extended Data Fig. 6**). Bootstrap analysis confirmed a significantly greater degree of network colocalization than genes in the control group (**Fig. 2c, Methods**, two-sided Wilcoxon rank-sum test, P value < 1.0×10^{-16}). Genes exhibiting a higher node degree ($n \geq 6$) were predominantly observed in the MM gene set (Wilcoxon rank-sum test, one-

sided) (**Fig. 2d**). Indeed, MM DNM genes were significantly enriched in biological pathways such as morphogenesis of polarized epithelium, neuronal cell adhesion, neural tube closure, and signal transduction (Gene Ontology biological processes, adjusted *P* values < 0.05) (**Fig. 2e and Extended Data Table 3**), suggesting functional convergence in neural tube closure.

DNMs implicate functional modules in neural tube closure

Using the 187 MM DNM genes as seeds, we employed network propagation to construct a comprehensive gene network associated with human MM risk (**Extended Data Fig. 7**, see **Methods**)[29], by implicating functionally related genes that might not be directly observable in patients due to, for instance, lethal mutations. The propagated network 'Meningomyel-ome' comprised 439 genes, including 257 propagated genes, and exhibited 2,447 interconnected edges. We found an over-representation of 374 experimentally proven mouse NTD causal genes[30] (**Supplementary Data**) in both seed and propagated genes (11 and 17 of the mouse NTD genes in before- and after propagation, one-tailed hypergeometric *P* value: 0.0015), corroborating functional relatedness to human MM. In contrast, no such relatedness was found with propagation in genes carrying DNMs in control trios (*P* value: 0.295).

Applying a clustering algorithm[31] to the 'Meningomyel-ome', we identified five subnetworks in which damaging DNMs and propagated genes were closely connected and enriched within predefined signaling pathways (referred to as submodules): RHO/RAC1 GTPase based actin cytoskeleton organization, microtubule-based process, neuronal migration and axon guidance, metabolism of lipids, and histone modification (FDR < 10^{-5}, **Fig. 3a,** see **Methods**)[6,32-38], suggesting that disrupting these pathways can increase MM risk. Notably, a higher haploinsufficiency of the 51 propagated genes within the submodules (pLI = 0.98 vs. 0.69 of 46 seed genes in median, see **Extended Data Fig. 8**) substantiated the utility of network propagation to further implicate potentially lethal MM risk genes.

Common among the twelve most strongly associated signaling pathways with FDR below 10^{-8} was actin and microtubules organization processes (**Fig. 3b**). Functional relatedness to other submodules of higher-level processes[39,40] and essential roles in the neural fold adhesion and closure[41,42] suggest that RHO/RAC1 GTPase-mediated actin and microtubule organization processes contribute to human MM risk.

Functional validation of candidate MM genes

We next performed functional assessment of candidate MM DNM alleles or genes. Seven genes—*TNK2*, *TIAM1*, *PLCE1, BRSK2*, *CLIP2, VWA8,* and *WHAMM*—were selected by their relevance to the RHOA/RAC1/CDC42-mediated actin and microtubule polymerization, five of which are direct regulators (**Fig. 4a, Methods**). Patient mutations in these genes were all D-Mis DNMs usually in regulatory domains predicted to interfere with protein function, except for LGD mutation in *WHAMM* predicted to truncate the protein. TNK2 is a CDC42-activated kinase that phosphorylates WASP among other targets. We found that immunoprecipitated TNK2 p.P186L DNM had approximately 60% reduced kinase activity to a WASP pseudo substrate, as benchmarked against wildtype and kinase-dead versions (**Fig. 4b** and **Extended Data Fig. 9**). TIAM1 is a Src-activated RAC1-specific guanine nucleotide exchange factor that promotes filamentous

actin (F-actin) assembly. We found that wild type TIAM1 but not the TIAM1 p.H1149P DNM co-localized with filamentous actin (**Fig. 4c-d**). While basal Rac1 activation was unperturbed, p.H1149P showed significant reduction in Rac1 activation following co-expression with constitutively activate Src (**Fig. 4e** and **Extended Data Fig. 10)**. VWA8 is a recurrently mutated mitochondrial matrix-targeting ATPase. We found that the p.R230G proband mutation failed to produce VWA8 protein evidenced on Western blot when transfected in HEK293T cells (**Extended Data Fig. 11**). Assessment of CLIP2, BRSK2 and PLCE1 MM proband mutations did not report robust defective protein function in standard assays (**Extended Data Fig. 12-14**). Interestingly the DNM mutations showing impaired protein function have previously been directly or indirectly linked to NTDs in animal models (see **Supplementary Information**). Thus, half of the six missense mutations showed clear disrupted function, consistent with our predicted rate of DNMs contributing to MM risk (52.43% of LGD, and 28.26% of damaging DNMs).

An LGD mutation found in *WHAMM* was selected for *in vivo* functional assessment in a *Xenopus laevis* neural tube closure model. Splice-blocking morpholino (MO) of the orthologous *Whamm* protein (**Methods**) resulted in dose-dependent NTDs (**Fig. 4f-h**). Specifically, injection of 5 ng of MO led to mild NTD, while 10 ng injection resulted in more severe open NTD in both anterior and posterior regions, as evidenced by *pax3* in situ hybridization, and was rescued by expression of wildtype *Whamm* (**Extended Data Fig. 15)**. These findings confirm the role of at least one of the DNM genes in MM pathogenesis *in vivo* and underscore potential gene dosage effects contributing to MM risk.

Expanding clinical phenotypes to include MM

We noted that 82 of the MM DNM genes were previously implicated in syndromic or nonsyndromic phenotypes in the Online Mendelian Inheritance in Man (OMIM) database [43], raising the question of whether our MM subjects might also express some of these OMIM phenotypes. By recontacting families, we found that 6% (5 out of 83) of the MM subjects (i.e. with mutations in *ZSWIM6*, *PAX3* in two subjects, *TCF12* and *BICRA*) had clinical features in addition to MM (Acromelic frontonasal dysostosis, Waardenburg, Craniosynostosis and Coffin-Siris syndromes, respectively), that suggest MM as a phenotyping expansion (**Supplementary Table 1**). Some subjects were lost to follow up, leading to 21.6% of the phenotypes that could be excluded due to lack of full clinical information. An additional 2 subjects (2.4%) shared some clinical features with OMIM phenotypes but insufficient to make the clinical diagnoses. For the majority of these 82 subjects, the zygosity did not match the OMIM zygosity (i.e. dominant vs. recessive), so it was not surprising that 70% of subjects lacked any OMIM clinical features. We conclude that MM can infrequently present as partially penetrant phenotype in established OMIM disorders, and suggest that specific alleles or environmental factors may determine expressivity.

Discussion

In this study, we present the first large-scale assessment of DNMs contributing to MM risk. Approximately 22% of MM probands had damaging DNMs, of which about 28% were expected to contribute to MM risk. This risk is comparable to other severe pediatric

conditions that are likely under strong purifying selection, as the mutations are unlikely to propagate to offspring[44-46]. In addition to the LGD DNMs found in other severe pediatric conditions, we found strong enrichment of D-Mis DNMs, suggesting that some LGD mutations in these genes may be lethal during embryogenesis. It would be interesting to assess the contribution of D-Mis mutations in other severe pediatric conditions.

Although only a subset of the genes identified likely contribute risk, we found that candidate MM genes were more interconnected than expected by chance. This lends support to the existence of a 'Meningomyel-ome', characterized by numerous genes that have potential to affect phenotype, interconnected by regulatory or protein-interaction networks, aligning with an omnigenic model[47]. Many of the networks highlight pathways previously implicated from mouse or frog models, including Wnt signaling, planar cell polarity, actin regulation and cell signaling. Notably, there was little overlap of mutated genes in humans with animal NTD models. One difference is that mouse and frog NTD genes are often identified as lethal embryonic phenotypes[48,49], whereas our ascertainment was limited to living subjects. Thus, while the genes implicated in laboratory animals vs. human NTDs may be different, we expect the pathways to be highly similar, evident in the Meningomyel-ome. Future work could introduce human DNMs into vertebrate models or human neural tube stem cell models, to bridge this divide.

Of the 187 genes with DNMs in MM subjects, 83 were previously implicated in human phenotypes according to OMIM. Only two of these OMIM conditions listed MM as part of the associated spectrum, and in only five subjects was the overlap sufficient for clinical diagnosis of the linked OMIM phenotype. This suggests that MM should be considered as part of the clinical continuum of a variety of OMIM conditions, which may be expressed in an allelic- or zygosity-specific fashion. Surprisingly, only five of the 187 genes were recurrently mutated in our cohort, suggesting there could be thousands of genes contributing to MM risk (**Supplementary Information**).

Our findings do not resolve the pathogenesis of most MM cases. Rare and de novo copy-number variants[50], or inherited variants could also contribute to risk[51]. It is also possible that regulatory mutations contribute, perhaps modulated by folic acid. Lastly, evidence of gene dosage sensitivity raises the possibility that environmental factors during critical developmental windows might amplify the effects of mutations.

Figures

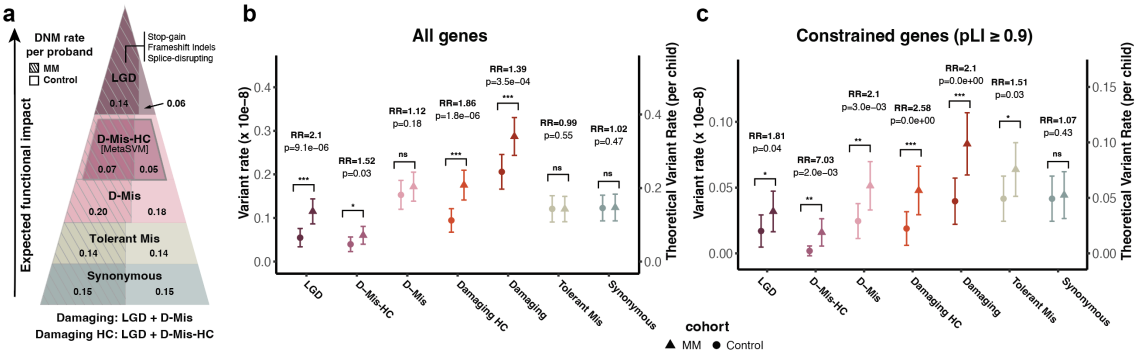


Figure 1. Enrichment of damaging DNMs in MM versus control. a, DNMs categorized by predicted functional impact: LGD (likely gene disrupting), D-Mis (damaging missense), D-Mis-HC (damaging missense of high confidence, called from a meta predictor), Tolerant Mis (missense of tolerant functional impact), and Synonymous. DNM rates per proband in MM (left sides of the triangle with strips) and control (right sides). Areas represent the ratio of the DNM rates between MM and controls within each functional category. Hashed: MM; Unhashed: controls. **b–c,** Variant rate (10^{-8}) and theoretical rate (DNM rate per child) on left and right y-axis, respectively. Statistical analysis of the ratio of DNM rates between MM (triangles) and controls (circles), denoted by ratio of the two Poisson rates, or rate ratio (RR), calculated within (**b**) 'All genes' (n = 19,658) and (**c**) 'Constrained genes' (n = 3,060; pLI ≥ 0.9). Theoretical rate was calculated by normalization of the variant rate with the total size of hg38 coding region (59,281,518 bp). Observed and theoretical DNM rates are marked in the left and right y-axis, respectively. P values were calculated by a one-sided rate-ratio test. P values: *** < 0.001, ** < 0.01, * < 0.05. ns: not significant.

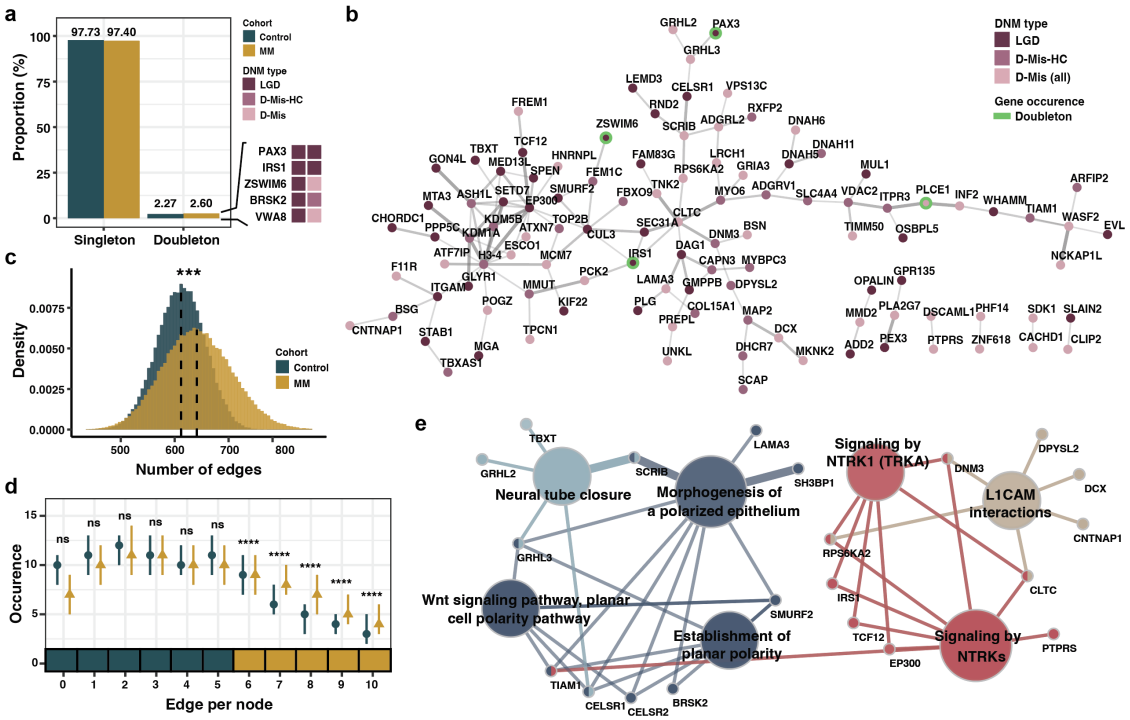


Figure 2. Functional convergence of genes implicated by damaging DNMs. a, Proportion of singletons (DNM occurred in one trio) and doubletons (DNMs occurred in two independent trios) damaging (LGD or D-Mis) genes in controls (green) and MM (yellow). Six doubleton genes were annotated with variant functional categories (LGD, D-Mis-HC, or D-Mis). **b,** A protein-protein interaction network composed of MM DNM genes. Genes connected by at least one edge are shown; unconnected (orphan) genes are shown in **Extended Data Fig. 3a**. Node color: variant functional categories: LGD (purple), D-Mis (light pink), and D-Mis-HC (dark pink). Doubleton genes: green borders. Edge thickness: confidence score of the protein interaction, defined in the STRING database. **c**, Distribution of the total number of edges in 100,000 networks randomly generated with bootstrapping (selecting gene set size of 108 (i.e. 80% of the 135 control DNM genes) in control (green) and MM (yellow). Higher number of edges denotes denser network interconnection. *P* values calculated by two-sided Wilcoxon rank-sum test. **d,** Number of nodes with *n* (0 to 10) edges in controls (green) and MM (yellow) gene sets. Error bars: first and third quartile. *P* values calculated by a one-sided Wilcoxon rank-sum test. **e,** GO (Gene Ontology) term network visualization for genes overlapping with 187 MM DNM genes with statistical significance (a.k.a., GO enrichment analysis). GO terms of functional relevance connected by an edge via commonly involved genes (small circles). Node sizes: significance of the terms. Degree of connectivity between terms (edges): kappa statistics by ClueGO. ****P value < 0.0001. ns: not significant.

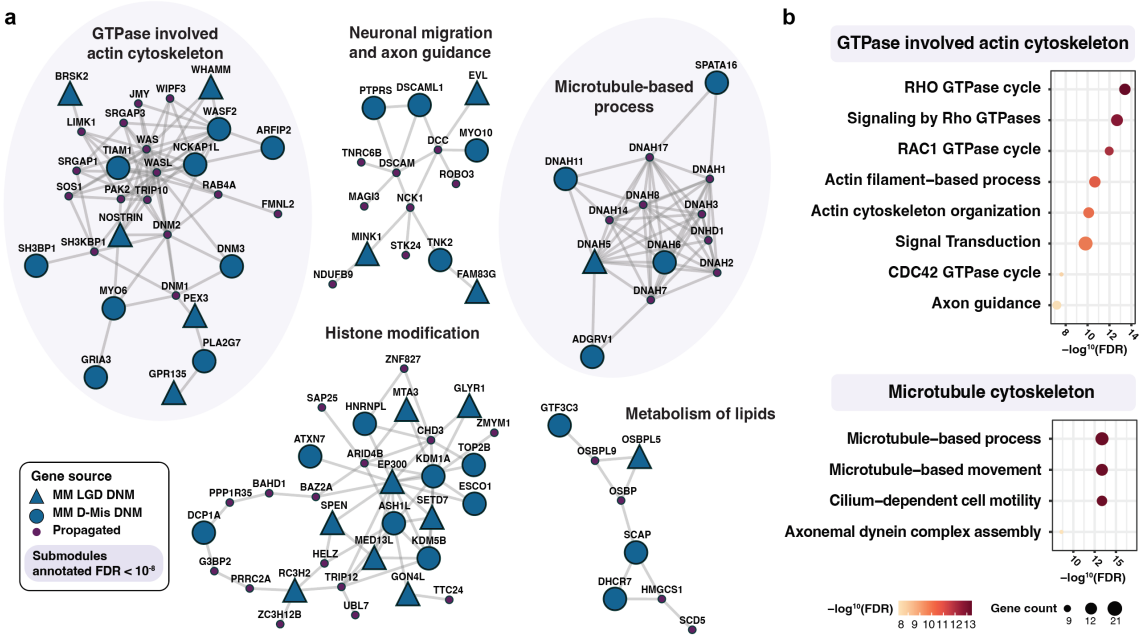


Figure 3. Functional modules that contribute to MM risk. a, Functional submodules clustered from the propagated network, 'Meningomyel-ome', with the 187 damaging DNM genes from MM cohort used as seeds. Clustered with a Leiden algorithm with co-expression value from STRING database as attributes, five submodules annotated FDR < 10^{-6} by Gene Ontology (GO) Biological Process, KEGG, or Reactome databases shown. **b,** Topmost significantly enriched GO Biological Processes with FDR < 10^{-8} shown. FDR value: -$\log^{10}$ scale. LGD DNMs genes: blue triangle. D-Mis DNMs: blue circles. Propagated genes: small purple circles. Functional terms annotated with FDR < 10^{-8} colored with light purple background.

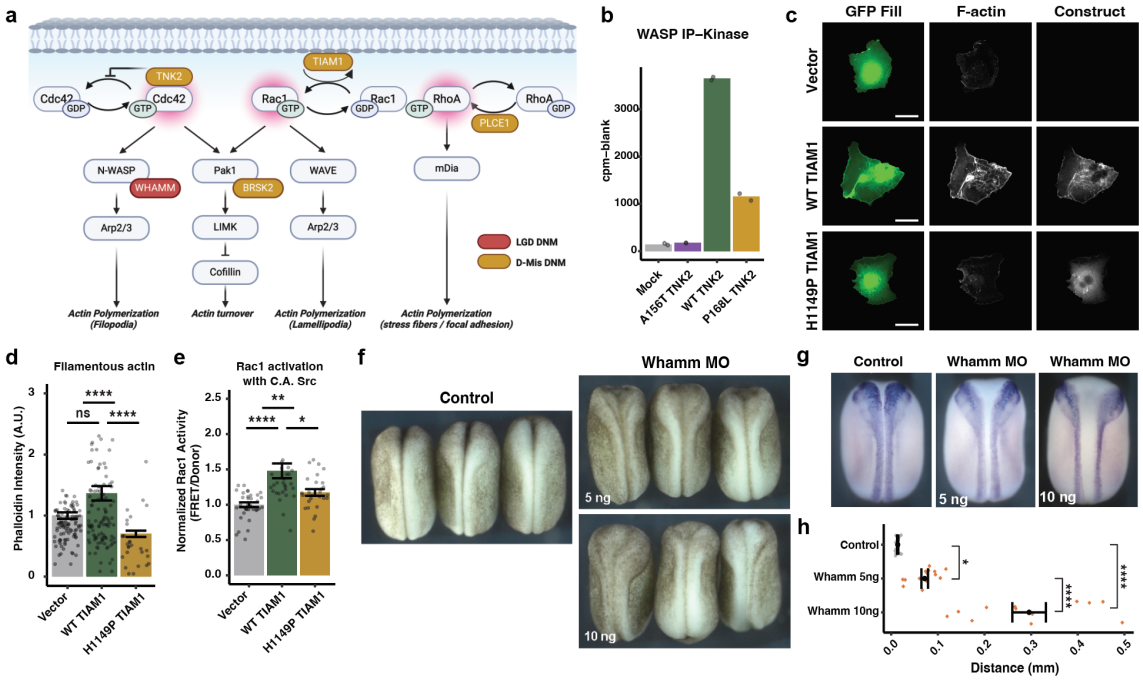


Figure 4. Functional validation of damaging DNMs related to actin polymerization. **a**, Damaging DNM genes highlight 'GTPase involved actin cytoskeleton pathway' mediated by WASP and WAVE, involving three GTPases (CDC42, RAC1, and RHOA) and five mutated genes *TNK2, TIAM1, WHAMM, PLCE1, and BRSK2*. Red circle: LGD DNM, Dark yellow circle: D-Mis DNM. **b**, TNK2 p.P168L patient mutation impaired WASP phosphorylation from immunoprecipitation (IP) kinase assay, compared to WT and kinase dead p.A156T. **c–d**, TIAM1 p.H1149P patient mutation showed fewer lamellipodia (F-actin) compared to WT, (**c**) imaged with Phalloidin and (**d**) quantified. Kruskal-Wallis followed by a pairwise Wilcoxon test, *P* value adjusted with Bonferroni. Data shown with Hampel filter and images masked to remove other cells. **e**, TIAM1 p.H1149P patient mutation decreased Rac1 activation with constitutively active (C.A.) Src, observed with Förster resonance energy transfer (FRET). Kruskal-Wallis followed by a pairwise Wilcoxon test, *P* value adjusted with Bonferroni. Data shown with Hampel filter. **f**, Dorsal views of *Xenopus laevis* embryos injected with Whamm morpholino (MO) at Stage 19. Neural tube closure phenotypes observed in a dose-dependent manner with 5 and 10 ng MO. **g**, Neural folds visualized by *pax3* in situ hybridization in the late neural stage. **h,** Quantification of distance between neural folds. Newman-Keuls (two-sided) for control versus multiple conditions after ANOVA. *P* values: **** < 0.0001, *** < 0.001, ** < 0.01, * < 0.05. ns: not significant, LGD: likely-gene disrupting, D-Mis: damaging missense; WT: wild-type.

References

1 Iskandar, B. J. & Finnell, R. H. Spina bifida. *N Engl J Med* **387**, 444-450 (2022). https://doi.org/10.1056/NEJMra2116032

2 Lee, S. & Gleeson, J. G. Closing in on mechanisms of open neural tube defects. *Trends Neurosci* **43**, 519-532 (2020). https://doi.org/10.1016/j.tins.2020.04.009

3 Prevention of neural tube defects: results of the Medical Research Council vitamin study. MRC Vitamin Study Research Group. *Lancet* **338**, 131-137 (1991).

4 J, A. *Myelocyste, transposition von gewebskeimen und sympodie*. Vol. Beitr Path Anat 16:1–28, 1894 (1894).

5 Chiari, H. Uber veränderungen des kleinhirns infolge von hydrocephalie des grosshirns. *Dtsch Med Wochenschr* **17**, 1172–1175 (1891).

6 Wilde, J. J., Petersen, J. R. & Niswander, L. Genetic, epigenetic, and environmental contributions to neural tube closure. *Annu Rev Genet* **48**, 583-611 (2014). https://doi.org/10.1146/annurev-genet-120213-092208

7 Carter, C. O. & Evans, K. Spina bifida and anencephalus in greater London. *J Med Genet* **10**, 209-234 (1973). https://doi.org/10.1136/jmg.10.3.209

8 Zhang, T. *et al.* Genetic variants in the folate pathway and the risk of neural tube defects: a meta-analysis of the published literature. *PLoS One* **8**, e59570 (2013). https://doi.org/10.1371/journal.pone.0059570

9 Lei, Y. *et al.* Identification of novel CELSR1 mutations in spina bifida. *PLoS One* **9**, e92207 (2014). https://doi.org/10.1371/journal.pone.0092207

10 Kibar, Z. *et al.* Mutations in VANGL1 associated with neural-tube defects. *N Engl J Med* **356**, 1432-1437 (2007). https://doi.org/10.1056/NEJMoa060651

11 Morrison, K. *et al.* Genetic mapping of the human homologue (T) of mouse T(Brachyury) and a search for allele association between human T and spina bifida. *Hum Mol Genet* **5**, 669-674 (1996). https://doi.org/10.1093/hmg/5.5.669

12 Jensen, L. E., Etheredge, A. J., Brown, K. S., Mitchell, L. E. & Whitehead, A. S. Maternal genotype for the monocyte chemoattractant protein 1 A(-2518)G promoter polymorphism is associated with the risk of spina bifida in offspring. *Am J Med Genet A* **140**, 1114-1118 (2006). https://doi.org/10.1002/ajmg.a.31212

13 Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014). https://doi.org/10.1038/nature13908

14 Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-223 (2013). https://doi.org/10.1038/nature12141

15 Lemay, P. *et al.* Loss-of-function de novo mutations play an important role in severe human neural tube defects. *J Med Genet* **52**, 493-497 (2015). https://doi.org/10.1136/jmedgenet-2015-103027

16 Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192-195 (2010). https://doi.org/10.1016/j.neuron.2010.10.006

17 Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat Genet* **48**, 126-133 (2016). https://doi.org/10.1038/ng.3469

18 Kessler, M. D. *et al.* De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proc Natl Acad Sci U S A* **117**, 2560-2569 (2020). https://doi.org/10.1073/pnas.1902766117

19 Besenbacher, S. *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* **6**, 5969 (2015). https://doi.org/10.1038/ncomms6969

20 Frome, E. L. The analysis of rates using Poisson regression models. *Biometrics* **39**, 665-674 (1983).

21 Willsey, A. J. *et al.* De novo coding variants are strongly associated with Tourette Disorder. *Neuron* **94**, 486-499 e489 (2017). https://doi.org/10.1016/j.neuron.2017.04.024

22 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475 (2012). https://doi.org/10.1038/nature11396

23 Goldmann, J. M., Veltman, J. A. & Gilissen, C. De novo mutations reflect development and aging of the human germline. *Trends Genet* **35**, 828-839 (2019). https://doi.org/10.1016/j.tig.2019.08.005

24 Turner, T. N. *et al.* Genomic patterns of de novo mutation in simplex Autism. *Cell* **171**, 710-722 e712 (2017). https://doi.org/10.1016/j.cell.2017.08.047

25 Martin, J. *et al.* A brief report: de novo copy number variants in children with attention deficit hyperactivity disorder. *Transl Psychiatry* **10**, 135 (2020). https://doi.org/10.1038/s41398-020-0821-y

26 Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885 (2011). https://doi.org/10.1016/j.neuron.2011.05.002

27 Hol, F. A. *et al.* A frameshift mutation in the gene for PAX3 in a girl with spina bifida and mild signs of Waardenburg syndrome. *J Med Genet* **32**, 52-56 (1995). https://doi.org/10.1136/jmg.32.1.52

28 Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638-D646 (2023). https://doi.org/10.1093/nar/gkac1000

29 Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat Rev Genet* **18**, 551-562 (2017). https://doi.org/10.1038/nrg.2017.38

30 Harris, M. J. & Juriloff, D. M. An update to the list of mouse mutants with neural tube closure defects and advances toward a complete genetic perspective of neural tube closure. *Birth Defects Res A Clin Mol Teratol* **88**, 653-669 (2010). https://doi.org/10.1002/bdra.20676

31 Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z

32 Rolo, A., Escuin, S., Greene, N. D. E. & Copp, A. J. Rho GTPases in mammalian spinal neural tube closure. *Small GTPases* **9**, 283-289 (2018). https://doi.org/10.1080/21541248.2016.1235388

33 Wallingford, J. B., Niswander, L. A., Shaw, G. M. & Finnell, R. H. The continuing challenge of understanding, preventing, and treating neural tube defects. *Science* **339**, 1222002 (2013). https://doi.org/10.1126/science.1222002

34 Niederkofler, V., Salie, R., Sigrist, M. & Arber, S. Repulsive guidance molecule (RGM) gene function is required for neural tube closure but not retinal topography in the mouse visual system. *J Neurosci* **24**, 808-818 (2004). https://doi.org/10.1523/JNEUROSCI.4610-03.2004

35 Kee, N., Wilson, N., Key, B. & Cooper, H. M. Netrin-1 is required for efficient neural tube closure. *Dev Neurobiol* **73**, 176-187 (2013). https://doi.org/10.1002/dneu.22051

36 Greene, N. D., Stanier, P. & Moore, G. E. The emerging role of epigenetic mechanisms in the etiology of neural tube defects. *Epigenetics* **6**, 875-883 (2011). https://doi.org/10.4161/epi.6.7.16400

37 Akimova, D. *et al.* Metabolite profiling of whole murine embryos reveals metabolic perturbations associated with maternal valproate-induced neural tube closure defects. *Birth Defects Res* **109**, 106-119 (2017). https://doi.org/10.1002/bdra.23583

38 Copp, A. J., Stanier, P. & Greene, N. D. Neural tube defects: recent advances, unsolved questions, and controversies. *Lancet Neurol* **12**, 799-810 (2013). https://doi.org/10.1016/S1474-4422(13)70110-8

39 Schaar, B. T. & McConnell, S. K. Cytoskeletal coordination during neuronal migration. *Proc Natl Acad Sci U S A* **102**, 13652-13657 (2005). https://doi.org/10.1073/pnas.0506008102

40 Dent, E. W., Gupton, S. L. & Gertler, F. B. The growth cone cytoskeleton in axon outgrowth and guidance. *Cold Spring Harb Perspect Biol* **3** (2011). https://doi.org/10.1101/cshperspect.a001800

41 Geelen, J. A. & Langman, J. Closure of the neural tube in the cephalic region of the mouse embryo. *Anat Rec* **189**, 625-640 (1977). https://doi.org/10.1002/ar.1091890407

42 Rolo, A. *et al.* Regulation of cell protrusions by small GTPases during fusion of the neural folds. *Elife* **5**, e13273 (2016). https://doi.org/10.7554/eLife.13273

43 Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514-517 (2005). https://doi.org/10.1093/nar/gki033

44 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). https://doi.org/10.1038/ng.3970

45 Halvorsen, M. *et al.* De novo mutations in childhood cases of sudden unexplained death that disrupt intracellular Ca(2+) regulation. *Proc Natl Acad Sci U S A* **118** (2021). https://doi.org/10.1073/pnas.2115140118

46 Li, W. *et al.* De novo mutations contributes approximately 7% of pathogenicity in inherited eye diseases. *Invest Ophthalmol Vis Sci* **64**, 5 (2023). https://doi.org/10.1167/iovs.64.2.5

47 Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177-1186 (2017). https://doi.org/10.1016/j.cell.2017.05.038

48 Lemos, M. C. *et al.* Genetic background influences embryonic lethality and the occurrence of neural tube defects in Men1 null mice: relevance to genetic modifiers. *J Endocrinol* **203**, 133-142 (2009). https://doi.org/10.1677/JOE-09-0124

49 Momb, J. *et al.* Deletion of Mthfd1l causes embryonic lethality and neural tube and craniofacial defects in mice. *Proc Natl Acad Sci U S A* **110**, 549-554 (2013). https://doi.org/10.1073/pnas.1211199110

50 Bassuk, A. G. *et al.* Copy number variation analysis implicates the cell polarity gene glypican 5 as a human spina bifida candidate gene. *Hum Mol Genet* **22**, 1097-1111 (2013). https://doi.org/10.1093/hmg/dds515

51 Wolujewicz, P. *et al.* Genome-wide investigation identifies a rare copy-number variant burden associated with human spina bifida. *Genet Med* **23**, 1211-1218 (2021). https://doi.org/10.1038/s41436-021-01126-9

Tables

Variant type	Theoretical rate per child (±95% CI)		% of cases with DNM mediating risk (±95% CI)	% of DNM carrying MM risk (±95% CI)
	MM (n=772)	Control (n=724)		
Likely Gene Disrupting (LGD)	0.12 (0.09-0.14)	0.05 (0.03-0.08)	7.16% (2.96%-11.36%)	52.43% (21.67%-83.18%)
Damaging HC	0.18 (0.14-0.21)	0.09 (0.07-0.12)	9.60% (4.45%-14.75%)	46.15% (21.38-70.92)
Damaging	0.29 (0.24-0.33)	0.21 (0.17-0.25)	9.62.% (2.64%-16.59%)	28.26% (7.75-48.76)

Table 1. Contribution of DNMs to MM risk

The DNM theoretical rate (i.e., DNM rate per child) of MM (n = 772) and control (n = 724) trios are shown with the 95% confidence interval (CI) of the rate using a one-sample t-test. The % of cases with DNM mediating risk were calculated by the difference between the theoretical rate of the MM and the control with 95% CI using a two-sample t-test (two-sided). The % of cases with DNM mediating risk was estimated by dividing the difference of theoretical rate between MM and control, by the theoretical rate of MM trios, with the 95% CI using two-sample t-test. The theoretical rates were calculated based on consensus regions for proper comparison of DNM rates. Damaging DNM includes LGD (Likely Gene Disrupting) and D-Mis DNM. Damaging-HC (high confidence) includes LGD and D-Mis-HC (high confidence).

Methods

Establishment of Spina Bifida Sequencing Consortium (SBSC) cohort

Inclusion criteria

Strict inclusion criteria were set for subject enrollment into the Spina Bifida Sequencing Consortium (SBSC). Participants must be affected by lumbosacral meningomyelocele (MM) with Arnold-Chiari malformation and hydrocephalus requiring surgical intervention such as ventriculoperitoneal shunt or endoscopic third ventriculostomy. We excluded subjects with closed neural tube defects, not requiring surgery at birth, with meningocele, or without hydrocephalus or not requiring surgery at birth. Enrollment required that both biological parents were available for sampling. In the case of fetal surgery to correct MM, which reduces the incidence of hydrocephalus, the requirement for inclusion of hydrocephalus at the time of enrollment was lifted. Subjects with known syndromes that would explain their conditions were excluded. Any families that did not meet the strict inclusion criteria were excluded.

Recruitment

The SBSC used several concurrent recruitment approaches: 1] Identify recruitable MM trios from local, national, and international hospitals. 2] Social media outreach through the Spina Bifida Association (SBA) Twitter, Facebook and Instagram accounts directly to families. 3] Recruit from spina bifida multidisciplinary clinics around the world with high MM caseloads. 4] Leverage historic Center for Disease Control-supported neural tube defect (NTD) cohorts. 5] Share sequencing data that have been already generated from trios with members of the SBSC and publicly with the NIH supported dbGaP. 6] Subjects and/or their families also contacted the SBSC independently through our postings and advertisements on social media. The countries where the participants were recruited include the United States, Mexico, Brazil, Italy, Georgia, Egypt, Canada, Venezuela, Pakistan, Guatemala, and Nigeria. All subjects or their guardians provided written informed consent approved by the UCSD Institutional Review Board S99075 protocol 140028, expiration date Aug 1, 2024. Recruitment processes were conducted in accordance with the approval by review boards of the University of California San Diego, operating under Federal-wide Assurance number, FWA00004495. The above rulings cover all aspects of our study, not just recruitment and consent.

Identifiers

The subject IDs used in this study cannot reveal the identity of the study subjects. These IDs are not know to anyone outside the research group.

Study questionnaire

All families prospectively recruited completed a standardized SBSC Questionnaire, which confirmed inclusion criteria, documented past medical history and current status by adopting prior questionnaire fields[52], along with place and date of birth of affected and parents to control DNM rate for parental age at the time of conception.

Subject sample and data handling

DNA was extracted from blood or saliva by standard salt extraction protocols using Qiagen (Germantown, MD) or Autogen (Holliston, MA) proto\ prior CDC-funded recruitment efforts where recontact was not possible, and thus they were considered 'lost to follow-up' for sample re-collection. There were 37 families that had prior trio WES[15] where data was incorporated into the cohort using the standardized bioinformatics pipeline.

Sequence data generation

In total, 2541 samples from 851 MM trios (839 trios and 6 quartets) were eligible for whole exome sequencing with six target capture kits (Roche Exome V2, Agilent xGen Exome V1, Agilent SureSelect v4, Agilent SureSelect V4, IDT xGen Exome V2, and Twist human comprehensive exome), each trio sequenced by using the same capture kit. Samples with low concentration of DNA, gender discordant, or contamination yielded 45 samples to be failed quality control, resulting in 794 trios and 6 quartets. For control trios, we obtained 732 healthy trios from the SAFARI cohort[16], 2202 WES data captured with NimbleGen EZ v2.

Data preprocessing and quality control

Raw reads were aligned to reference genome (GRCh38) using bwa mem (0.7.17), preprocessed with PICARD (2.20.7) AddOrReplaceReadGroups and MarkDuplicates. Germline variants were collected with GATK (v4.11.0) HaplotypeCaller in the reference confidence model (-ERC GVCF) to be combined via CombineGVCF, generating a joint vcf for all MM and control trios, followed by GenotypeGVCF (v4.11.0) to be jointly genotyped. Multiallelic variants were splited (bcftools norm -m -any), Indels were realigned, and base quality recalibration was done with GATK (v.4.0.11) with known Indel and population frequency information (dbSNP (146) and Mills_and_1000G_gold_standard.indels.hg38.vcf.gz from GATK resource bundle). Variant quality was recalibrated by GATK VQSR with HapMap (hapmap_3.3.hg38.vcf.gz), omni (1000G_omni2.5.hg38.vcf.gz), 1000G (1000G_phase1.snps.high_confidence.hg38.vcf.gz), (Mills_and_1000G_gold_standard.indels.hg38.vcf.gz), and axiom (Axiom_Exome_Plus.genotypes.all_populations.poly.hg38.vcf.gz).
To confirm the kinship integrity within trios, we used TRUFFLE[53] (v1.38), removing 29 and 7 trios from MM and control cohort, respectively. IBD1 values were extracted from the vcf, by only using variants with allele frequencies higher than 0.05 (--maf 0.05) to only use confident calls. The IBD1 for all possible sample pairs were examined and if a parent-child IBD1 value is smaller than 0.75. Also, samples having more relations (IBD1 ≥ 0.75) with one outside of their family were removed, possibly owing to low data quality or contamination.

De novo variant calling

De novo SNVs and Indels were collected following GATK genotype refinement steps. The genotypes of 777 MM and 725 control trios were gathered, and posterior probabilities of the variants were calculated with CalculateGenotypePosteriors (v4.2.6.1). We only selected de novo variants having genotype qualities (GQ) ≥ 20, Mapping quality (MQ) ≥ 30, and PASS in VQSR. Candidate DNMs that were found to have sequencing

673 depth (DP) ≥ 12 in all three family members were kept to secure high-confident calls. To
674 remove candidates that deviate from theoretical heterozygous states (AF=0.5), a two-
675 sided binomial test was conducted to all positions by applying read depths and number of
676 alternative alleles, removing sites with *P* value > 0.01. Genotypes of both parents were
677 examined and kept if genotyped as reference homozygous (GT=0/0). To enhance the
678 robustness of a high-confident set of DNMs, a series of post-filters were applied, by
679 examining regional information and raw allele counts. First, clustered DNMs were
680 filtered out when two or more DNMs of one proband were observed within 10 bp.
681 Second, DNMs that were found to exist in the same genotypes in parents either in MM or
682 control cohorts were also removed. For de novo SNVs, each DNM position was
683 examined by collecting raw reads with base quality (> 13) and mapping quality (> 0)
684 criteria for each trio (bcftools mpileup v1.9). Then, to remove artifacts derived from
685 miscellaneous alignment, a position was filtered out when a same alternative allele was
686 observed in either of the parents. Two-sided binomial test was once more applied for
687 SNVs with raw read counts, with a strict cutoff 0.05. All Indels were manually inspected
688 with IGV (Integrative Genomics Viewer) by three experienced experts. The schematic
689 overview of the DNM detection pipeline is shown in **Extended Data Fig. 16a**. VEP
690 (v106) was used to annotate de novo variants with 1000 genomes (phase 3), COSMIC
691 (92), ClinVar (202109), ESP (V2-SSA136), HGMD-PUBLIC (20204), dbSNP (154),
692 GENCODE (Human Release 40), gnomAD (r2.1.1), PolyPhen (2.2.2) and SIFT (5.2.2).
693 Mutational signatures were collected by Mutalisk[54] using linear regression.
694
695 **DNM statistical confirmation**
696 *Consensus region for DNM rate calculation*
697 To compare DNM rates and burden among sequencing data generated with different
698 platforms (e.g., exome capture libraries), genomic regions that are commonly covered
699 with sufficient coverage (that is, consensus region) were extracted. For each individual
700 genome coverage was calculated using bedtools[55] (v 2.30.0) genomecov. Then, the
701 coverage bedgraphs were merged into each trio using bedtools unionbedg, and only the
702 regions with read depth ≥ 12 in all three family members were kept. Each family
703 bedgraph was merged with bedgraphs of other families within each batch, to confirm if at
704 least 70% of families covered that region. Finally, the batch bedgraphs were merged into
705 a final consensus region, reaching 36,553,428 bp.
706
707 *DNM rates examined with Poisson distribution*
708 DNMs located in the consensus region were collected, to confirm if the number of DNM
709 per proband follows an expected Poisson distribution. Briefly, we assumed a cutoff of
710 DNM count per proband from 1 to 20, generating DNM sets based on the cutoffs. Then,
711 an expected Poisson distribution was generated per cutoff, with lambda as an average
712 count of DNMs per proband (R dpoi). An observed distribution of DNMs per proband
713 was calculated to be compared to the expected distribution, with Chi-Square goodness-of-
714 fit test (R chisq.test). A cutoff of 7 DNMs per proband was set (**Extended Data Fig. 2c–**
715 **2d**), removing five and one trios from MM and controls, respectively. The six samples
716 with DNM counts larger than 7 were found to have DNMs ranging from 8 to 484,
717 possibly owing to low data quality.
718

DNM burden analysis

DNM rates were calculated based on the number of DNMs observed in the consensus region, to properly compare the rates. We utilized a rate ratio test given that the number of DNMs per proband followed Poisson distribution. The DNMs were categorized into likely gene disrupting (LGD) including frameshift Indels, stop gained, and splice donor/acceptor mutations. Among the missense variants, if annotated as 'probably damaging' or 'likely damaging' in PolyPhen or 'deleterious' in SIFT, when CADD prediction score (>20) supports either at the same time, we annotated DNMs as damaging missense (D-Mis). If a DNM was annotated as D (deleterious) by a highly strict meta predictor MetaSVM[56], we prioritized them as D-Mis-HC. We could replicate the highly consensus D-Mis mutations and their burden with two other independent missense variant annotation predictors, MetaLR[56] and REVEL[57] (**Supplementary Information** and **Extended Data Fig. 16b**). MetaSVM, MetalLR, and REVEL was annotated based on dbNSFP (4.1a).Missense mutations that were not annotated as D-Mis were categorized as 'Tolerant missense'. Constrained genes were collected with pLI ≥ 0.9. DNM rate was calculated within the consensus region (36,553,428 bp) in the mutation categories, with 95% confidence intervals calculated (one sample t test). Two-sided Wilcoxon rank-sum test was utilized to confirm the difference of the total DNM rates between MM (n = 772) and controls (n = 724). Rate ratio test (one-sided) was used to compare two Poisson distributed rates from each cohort. Theoretical DNM rates were estimated based on the total size of the coding region in hg38 (59,281,518 bp). The percentage of the subjects with DNM mediating risk could be calculated by the difference between the theoretical rates (DNM rate per exome) of the MM and control (95% CI calculated with Wilcoxon rank-sum test, two-sided). The percentage of DNM carrying the MM risk was estimated by dividing the difference of theoretical rate by the theoretical rate of MM trios.

Whole genome analysis

We additionally recruited 101 trios and 1 quartet with MM and generated whole genome sequencing (WGS) data with Illumina HiSeq2500 and NovaSeq. The reads were preprocessed and aligned with the same pipeline of WES data. Same quality controls with WES data were applied, removing 4 trios that failed kinship integrity and contamination analyses. Small SNVs and Indels were collected with the same variant calling and filtering of WES analysis. We applied a binomial test to confirm if a DNM was derived from a heterozygous state as in WES, with applied stronger P value cutoff 0.05 compared to WES, as WGS data has much stable coverage with reduced capture bias and allelic imbalance. Additional populational allele frequency information (gnomAD genome v3.1.2) was added and SpliceAI (v1.3.1)[58] was utilized (score > 0.5) to predict splice disrupting variants. We annotated noncoding variants with GREEN-VARAN[59] (GREEN-DB schema 2.5) to annotate DNMs in noncoding regions. For copy number variant analysis (CNV), CNVpytor (1.3.1)[60] was utilized with a window size of 100 kbp, and manually curated by confirming the read depths of all father and mother. To comprehensively detect structural variants (SVs), we utilized Manta[61] (1.6.0), Delly[62] (v0.8.1), and smoove[63] (0.2.6) and merged the SV calls from the three callers with Survivor (1.0.7, parameter: dist=50000, callers=0, type=0, strands=0, estimate=0,

size=50000) to confirm if the detected SVs only present in probands.

Multiplexed Error Robust In-Situ Hybridization (MERFISH)

MERFISH was performed as described previously at the UCSD Epigenomics Core[64]. Briefly, mouse embryos at day 10.5 were fixed in paraformaldehyde and cryopreserved in 30% sucrose, embedded in Optimal cutting temperature (OCT) and sectioned into 12 μm sections in cryostat. These tissue sections then underwent fluorescent in situ hybridization with a panel of oligonucleotide probes specific for 36 of the genes identified from the current human DNM cohort along with 107 marker genes, chosen upon previously published single cell literatures to classify cell types into 7 cell types (neuron, neural progenitor, neural crest, pre-epithelial to mesenchymal transition neural crest progenitor (Pre-EMT-NCP), dorsal root ganglia, blood and mesoderm[65-67]. During downstream analysis marker genes were filtered to only include 66 the most efficient markers specific for each cell type (see **Supplementary Data** for a full list of utilized marker genes). Probes were designed and manufactured with standard pipeline by Vizgen (https://vizgen.com/gene-panel/). Tissue sections were prepared and processed using Vizgen sample preparation protocol (Document Number 91600002 Rev A), then samples placed in the Merscope for imaging and decoding (Document Number 91600001, Rev G). Raw data was analyzed using Scanpy (1.9.1)[68]. Cells were preprocessed and filtered using the following criteria: cells with volume less than 100 μm^3, cells with volume larger than 3x median volume of all cells, fewer than 4 genes detected, fewer than 10 transcripts detected, total RNA counts lower than the 2% quantile or higher than 98% quantile. Preprocessed cells were clustered based on gene expression using the Leiden algorithm (leidenalg 0.10.2), annotated using scType[73] (v1.0) based upon marker genes as the reference set (see **Supplementary Information**). Clusters were annotated to the cell type with the highest overall score from scType, accounting for expression of all marker genes per cell type in a focal cluster, compared to the expression in other clusters. Clusters with prediction score less than 0 or low confidence scores (score less than 10% of the number of cells in each cluster) were classified as indeterminate as they could not be classified in any of the 7 cell types. Expression of each of the 36 candidate genes was then analyzed in each cell from raw data, and percentage of total average expression per cell type calculated in **Extended Data Fig. 5**.

Network analysis (co-localization and degree)

Relationship between damaging DNMs in MM vs. control cohorts and the human genome network were calculated using STRING (v11.5) for gene interactions. We calculated the connectedness from the STRING database of damaging DNMs from MM (n = 187) and control (n = 135) among 19,699 annotated genes, randomly resampled with a size of 80% of the gene set from control (n = 108), of 100,000 iterations. Number of edges per iteration were compared with a Wilcoxon rank-sum test (two-sided). Edges per nodes were compared between the cohorts (Wilcoxon rank-sum test, one-sided).

Pathway enrichment analysis

To estimate if the MM or control DNM gene sets were functionally enriched in known biological pathways, gene ontology (GO) biological process, reactome, and KEGG pathways, we conducted a gene enrichment analysis with ClueGO (v2.5.10) in Cytoscape

(3.10.1), with parameters set as Evidence All_without_IEA (Inferred from Electronic annotation), GO tree interval ranging from 3 to 8, and 5 minimum number of genes, 4% of genes for GO term selection, and 0.4 kappa score. We used adjusted *P* value with Bonferroni step-down with significance criterion 0.05. Detailed information of the GO enrichment test can be found in **Extended Data Table 3**. No significant term was observed with control gene sets.

Sanger confirmation

Out of 198 damaging DNMs, 86% (171 out of 198) where trio DNA samples were available for the Sanger sequencing confirmation, with sufficient DNA. For a subset of detected damaging DNMs, primer sets were generated via Primer3 and ordered from IDT. The primer lengths were configured to span a range of 18 bp to 23 bp, targeting an ideal size of 20 bp. The GC content was set to a minimum of 30% and a maximum of 70%, and the annealing temperature was established to range between 57°C and 62°C, with an optimal temperature of 59°C. PCR was performed using the QIAGEN Taq DNA Polymerase kit and primer concentrations of 500 nM. Four reactions were performed per DNM, amplifying DNA from the father, mother, affected child, and an unrelated healthy individual. PCR products were confirmed by agarose gel electrophoresis, purified by incubation with Exonuclease I/Shrimp Alkaline Phosphatase (1:2), diluted in ddH2O, and shipped to Genewiz/Azenta for Sanger sequencing. Trace files were analyzed in SnapGene Viewer and compared against the UCSC Genome Browser to confirm that the affected child was heterozygous for the detected mutation and the father, mother, and control were homozygous for the reference base at the same locus.

Network propagation and submodule identification

A propagated network was constructed with NetColoc (0.1.7)[69]. PCNet[70] was utilized for the background human gene interaction network, which contains 18,820 nodes and 2,693,109 edges. Among 187 MM genes, two genes (SLCO1B3-SLC01B7 and H3-4) couldn't be included as seeds because they couldn't be found in the PCNet. As implemented in NetColoc, w prime and w double prime value was calculated for each gene to generate z scores for proximity. We used the z score threshold 3 as default. The entire list of the genes in the propagated network can be found in **Supplementary Data**. Known mouse NTD genes ($n = 374$) were based on previous literatures, where 205 genes were organized back in 2010[30] and the others were added from then (see **Supplementary Data** for the gene list). To examine the propagated network, we conducted a hypergeometric test, utilizing 18,820 as the total number of genes as in PCNet ($M = 18{,}820$), number of known NTD genes ($n = 374$), the number of nodes in the propagated network ($N = 439$), and the number of overlapping genes with propagated network and MM genes ($k = 17$).

$$Pr(X = k) = f(k;\ M, n, N)$$

Then, to identify small functional modules from the expanded network, the propagated network was clustered with a Leiden algorithm with 0.5 resolution, 0.01 beta with 45 iterations. To use co-expression value as attributes, we again transferred the propagated network to STRING database (v12) with Homo Sapiens full STRING network, popping out 114 nodes without edges to the network, meaning discrepant sources of protein network. In total, 15 small clusters were generated with more than five nodes. Five could

be significantly enriched by functional biological terms (FDR < 10^{-5}) by the GO biological process, KEGG pathway without disease related terms, and Reactome, which had more than 30% of the nodes derived from damaging DNMs (**Supplementary Data)**. Two clusters annotated as 'GTPase involved actin cytoskeleton' and 'Microtubule-based process' were enriched with functional terms with most significant *P* values (FDR < 10^{-8}), with 8 and 4 terms respectively. The pLI of the genes in the five functional modules can be found in **Supplementary Data**.

Cloning

Patient DNMs were mutagenized into wildtype (WT) cDNA using Gibson Assembly cloning. These mutations include p.R231G (R230G in mouse) in *Vwa8b* (mouse), p.E521Q and p.E623Q in *PLCE1* (human), p.P168L in *TNK2*, p.H1149P *TIAM1,* p.R620H in *BRSK2* (human) and p.R349C in *CLIP2* (rat). For each of the plasmids, overlapping (40 bp) forward and reverse primers with 20 bp 5' overhangs were designed complementary to the mutation site, with a single nucleobase difference corresponding to the mutation of interest (**Supplementary Data**). 15 ng plasmid DNA was amplified using the NEB Phusion High-Fidelity 2X Master Mix [NEB #M0531] for 25 cycles, after which the reaction was subjected to 1 hour at 50°C with NEBuilder HiFi DNA Assembly 2X Master Mix [NEB #E2621], and the reaction mixture was treated with *DpnI* [NEB #R0176] digestion for 37°C for 30 minutes. 2 µL of the reaction mixture was then transformed into NEB DH5-alpha Competent *E. coli* [NEB #C2987], with successful mutation verified using Primordium whole plasmid sequencing.

Protein analysis

TNK2

Cell transfection, lysis, and Western blotting

Human embryonic kidney (HEK)293T cells (from American Type Culture Collection) were maintained in DMEM (Corning) supplemented with 10% fetal bovine serum (Sigma) and 1X penicillin and streptomycin. The cells were transfected according to the protocol supplied with TransIT reagent (Mirus) using 7.5 µg of DNA. After 48 hours, the cells were harvested, washed with phosphate-buffered saline, and lysed in a buffer containing 25 mm Tris–HCl (pH 7.5), 100 mm NaCl, 1 mm EDTA, 1% Nonidet P-40, 5 µg/ml leupeptin, 5 µg/ml aprotinin, 1mM PMSF, and 200 µM Na3VO4. Cell lysis proceeded for 30 minutes at 4˚C with rocking. The cell suspensions were centrifuged at 15,000 rpm for 10 minutes, and protein concentrations of the soluble lysates were measured using a colorimetric Bradford assay. The lysates (100 µg) were analyzed by SDS-PAGE and Western blotting using the following antibodies: phospho-Ack1/Tnk2 (Millipore Sigma 09-142), Flag (Sigma-Aldrich A8592), and gamma-tubulin (Sigma-Aldrich T6557).

Immunoprecipitation (IP) Kinase assay

IP-kinase assays were performed essentially as described with the following modifications: HEK293T lysates (1 mg) were incubated with 40 µL of anti-Flag M2 affinity resin (Sigma) on a rotator at 4 °C overnight, then washed three times with Tris-buffered saline (TBS). The IPs were divided into three portions. One portion of each sample was mixed with a gel-loading buffer, boiled for 3 minutes, and analyzed by 7.5%

SDS/PAGE with anti-Flag Western blotting. The remaining two samples were used for duplicate activity measurements using the phosphocellulose paper-binding assay. The reactions contained 20 mm Tris–HCl (pH 7.4), 10 mm MgCl2, 0.25 mm ATP, 1 mM WASP peptide (KVIYDFIEKKG) and 20–50 cpm/pmol of [γ-32P] ATP. Reactions were carried out at 30 °C for 15 min. Incorporation of 32P into peptide was measured by scintillation counting.

TIAM1

DNA Reagents

The following constructs were previously described: pCMV-EGFP[71], pCMV-Flag-Tiam1 (WT-Tiam1)[72], pRaichu-RaichuEV-Rac1 (2248X)[73], pcDNA3.1 vector[74]. mRuby3 was obtained from Addgene (127808).

Cell culture and transfections

Cos-7 cells were grown in Dulbecco's modified Eagle's medium (DMEM) (Corning 10-013-CV) supplemented with 100 U/ml penicillin/streptomycin (Invitrogen 15140122) and 10% heat inactivated fetal bovine serum (Atlanta Biologicals S11150H) on cell culture-treated plastic (VWR 10062-880) maintained at 5% CO2, 37ºC. Cells were passaged with trypsin (Invitrogen 25200072) once per week. For experiments, 24 hrs prior to transfection cells were plated on either nitric acid washed glass cover slips (Bellco Glass 1943-10012A) or glass bottom plates (Cellvis P241.5HN) coated with 20 ug/ml Poly-D-Lysine (Corning 354210). Cells were transfected using Jetprime reagent (Polyplus 101000046) according to the manufacturer recommendations. Twenty-four hours after transfections cells were changed to serum-free DMEM. For live cell experiments, phenol-red free DMEM (Corning 17-205-CV) was used. Experiments were conducted 48 hours after transfection.

Immunocytochemistry, microscopy, and analysis

Cells were fixed with 4% Paraformaldehyde (Fisher Scientific AC169650010), 4% Sucrose (Sigma S0389) in phosphate buffered saline (PBS). To confirm construct expression, following fixation all cells were immunostained overnight at 4ºC with anti-Flag (Cell Signaling 14793S) primary antibody diluted in 0.3% Triton X-100, 5% goat serum in PBS. After, cells were incubated with Cy5 anti-Rabbit secondary antibody (Jackson ImmunoResearch 111-175-003) for 2 hrs at room temperature. For filamentous actin experiments, cells were also incubated with Texas Red-Phalloidin (Invitrogen T7471) during this time. Post-fixation and staining, cells grown in glass coverslips were mounted using the aqueous mounting solution FluorSave (EMD Millipore 345789).

Imaging

Filamentous actin was labeled with Texas Red phalloidin, then imaged using a Zeiss AxioObserver.Z1 microscope with a 20x objective. Rac1 activation was measure in cells transfected with the RaichuEV-Rac1 probe, imaged on a Zeiss LSM 880 confocal microscope at 20x. Live imaging experiments were conducted at 37ºC/5% CO2. Single plane images were acquired with Förster resonance energy transfer (FRET) (excitation 458 nm, emission 522–569 nm), CFP (Donor) (excitation 458 nm, emission 463–507 nm), and YFP (Acceptor) (excitation 514 nm, emission 522–569 nm), presented as normalized FRET/Donor values using ImageJ (NIH). Only cells expressing constructs were selected

for analysis. GFP or mRuby3 fill visualized cell morphology via ROI cell trace, measured for mean intensity, corrected for background, and normalized to respective controls. Experiments were conducted with experimenters blinded to conditions. Representative images were masked using ImageJ (NIH).

VWA8

Plasmid transfection in HEK293T cells was performed using the ThermoFisher Lipofectamine 2000 transfection reagent. HEK293T cells seeded on a 6 well plate and transfected with 4ug of plasmid. Cells were lysed using Modified RIPA buffer 72 days post-transfection, and the lysate was spun down at 10,000g in a microcentrifuge and supernatant was harvested. Empty vector, WT, and R230G protein was then quantified using the Pierce BCA Assay kit (Thermo Scientific #23225), and 10 µg protein lysate was diluted in 4X SDS-PAGE Loading Buffer supplemented with 10% β-mercaptoethanol (Sigma #M3148) and loaded into each well of an Any kDa Bio-Rad Mini-PROTEAN TGX [4569031] gel, and run for 30 minutes at 150 V. Proteins were then transferred onto a nitrocellulose membrane [Sigma #IPVH00010] and blocked with 5% BSA (Sigma #A9418) in TBST (blocking buffer) for 1 hour at room temperature. The membrane was then blotted with 1:1000 Anti-HA (#sc-805) and 1:5000 Anti-alpha-tubulin (Sigma #T6074) primary antibody in blocking buffer overnight at 4°C under constant rocking. The following day the membrane was washed 4 times in TBST for 10 minutes, followed by1 hour incubation in a 1:10000 anti-mouse (#715-035-150) and anti-rabbit (#711-036-152) HRP-conjugated secondary antibody in blocking buffer at room temperature. The membrane was then washed again for 4 times in TBST for 10 minutes, after which the membrane was treated with Pierce ECLWestern solution for HRP detection (Thermo Scientific #32106) for 5 minutes at room temperature. The membrane was then transferred into a clear plastic cover and imaged using a chemiluminescence detector (Biotechnique #92-14860). Quantification of protein band intensity was accomplished using the ImageJ band densitometry plugin, and final plots were created using Prism.

CLIP2

Western blot

HEK293T cells were transfected with pEGFP-CLIP2 WT, R349C (TGC or TGT) either with constitutively active AKT pE17K (Addgene #73050) or empty vector using Lipofectamine as previously described, lysed into sample buffer, and analyzed by western using anti-GFP antibody (1:200 Thermo Scientific #GF28R).

Microtubule analysis

HuH7 cells cultured in DMEM with 10% fetal calf serum (Sigma) and 1% antibiotic-antifungal mixture (Gibco) were transfected with pEGFP-CLIP2 WT, R349C or S352A (Turbofect, ThermoFisher Scientific #R0531). Time-lapse sequences ($\Delta t = 2.4$ s, 80 frames) of the CLIP2 fluorescence signal were captured using a Leica DMLB microscope through a 100x 1.3 NA objective and a Scion CFW1312M camera. The parameters of microtubule dynamics were computed as previously described[75]. Statistical comparisons were performed using one-factor ANOVA.

PLCE1

Cell transfection and lysis

Plasmid expressing PLCE1 including the WT, p.E521Q, and p.E623Q mutations were transfected into HEK293 cells and were maintained in DMEM supplemented with 10% FBS and 1% penicillin/streptomycin. Plasmids were transfected into HEK293 cells using Lipofectamine (Invitrogen), grown for 48 hr and lysed in lysis buffer (1% Triton X-100, 50 mM Tris pH 7.4, 10 mM $MgCl_2$, 500 NaCl) with added protease inhibitors (Roche). The lysates were spun down at 20000 g/4°C/15 min and the supernatants were used for overnight incubation with 20 µl of GST beads containing GST recombinant proteins on the rotator. After the incubation samples were washed with 800 µl lysis buffer, the GST beads were collected at 500 g/4°C/2 min, the supernatants were aspirated out and this washing was repeated 3 times. Proteins were eluted with 20 µl of 2x Laemmli buffer.

Pulldown assay

Eluted proteins were diluted in 4X SDS-PAGE Loading Buffer supplemented with 10% β-mercaptoethanol (Sigma #M3148) and loaded into each well of an Any kDa Bio-Rad Mini-PROTEAN TGX gel, and run for 30 minutes at 150 V, samples transferred to PVDF membrane (Sigma #IPVH00010), blocked with 5% BSA (Sigma #A9418) in TBST for 1h at RT, membrane blotted with 1:1000 anti-RAC1, 1:1000 anti-CDC42, 1:1000 anti-RHOA, and 1:1000 anti-Myc, washed in TBST for 10min then blotted with 1:10000 anti-mouse (#715-035-150) and anti-rabbit (#711-036-152) HRP-conjugated antibody, washed in TBST for 10m, then developed in Pierce ECL Western solution (Thermo Scientific #32106) for 5min, and imaged using a chemiluminescence detector [Biotechnique #92-14860], and band intensity quantified using ImageJ band densitometry, and plotted using Prism.

The GST-PAK1 beads (#PAK02-A) and GST-rhotekin RBD beads (#RT02-A) were purchased from Cytoskeleton. Anti-Myc (#2278; Cell Signaling Tech.), anti-RhoA (#2117; Cell Signaling Tech.), anti-Rac1 (#610650; BD Transduction Labs), anti-Cdc42 (#610928; BD Transduction Labs) were purchased from the indicated commercial sources.

BRSK2 Western Blot

HEK293T cells were transfected with the pEXP-CMV-BRSK2-Flag plasmid containing wildtype, K48A kinase-dead mutant, and R620H patient mutation BRSK2 sequence using PEI (Sigma Millipore #919012) and OptiMEM (Gibco #31985-070) according to the manufacturer's guidelines. Protein concentration of lysates was measured using the Pierce BCA Protein Assay Kit (Thermo Scientific #23225), then 20 µg of protein was loaded into each well of a 4-12% Bis-Tris Gel (Invitrogen #NP0336), transferred to nitrocellulose membranes (Thermo, #88018), blocked then blotted with one of the following primary antibodies: 1:1000 Flag rabbit (Cell Signaling Technology #14793S), 1:1000 Beta-actin rabbit (Sigma #A5316), 1:1000 anti-Phospho-AMPK Substrate Motif rabbit (Cell Signaling Technology #5759) and 1:1000 mTOR rabbit (Cell Signaling Technology #2983), 1:1000 P-mTOR Ser 2448 rabbit (Cell Signaling Technology #5536), 1:1000 p70 S6 Kinase rabbit (Cell Signaling Technology #34475), 1:1000 Phospho-p70 S6 Kinase Thr 389 rabbit (Cell Signaling Technology #9234), 1:1000 4E-BP1 rabbit (Cell Signaling Technology #9644) and 1:1000 Phospho-4E-BP1 Thr 37/46

rabbit (Cell Signaling Technology #2855) at 4°C overnight, rinsed in TBST for 10 minutes, incubated with LI-COR IRDye 680 Donkey anti-Mouse or 800 Donkey anti-Rabbit (LI-COR #926-68072, #926-32213) 1:10,000 rinsed in TBST and imaged on an Odyssey CLx using Image Studio software for analysis of band intensity.

Xenopus modeling

Xenopus embryo manipulations were conducted according to established protocols[76]. Female adult *Xenopus laevis* were induced to ovulate by the injection of human chorionic gonadotropin, and *in vitro* fertilization was performed by homogenizing a small piece of a testis. Embryos were dejellied in 1/3 X Marc's Modified Ringer's (MMR) with 2.5% (w/v) cysteine (pH7.9) at the 2-cell stage and were microinjected with either gene-specific MO or mock into two dorsal blastomeres at the 4-cell stage in 2% Ficoll (w/v) in 1/3 X MMR. Injected embryos were incubated at 18°C until stage 19 and fixed with 1X MEMFA (0.1 M MOPS, 2mM EGTA, 1 mM $MgSO_4$, 3.7% formaldehyde, pH 7.4). The gene sequence for *Xenopus* Whamm was obtained from Xenbase (http://www.xenbase.org). A Morpholino antisense oligonucleotide (MO) against Whamm was designed to block mRNA splicing (Gene Tools) 5' – AAAAGTAGGAAGAAGCCCCCACCCT -3'. The Whamm open reading frame was amplified from *Xenopus* cDNA and cloned into the pCS10R MCC vector containing GFP tag. Capped mRNA was synthesized using the mMESSAGE mMACHINE SP6 transcription kit (Invitrogen Ambion, cat# AM1340), and injected to assess rescue. To measure the distance of the neural folds, *in situ* hybridization against Pax3 was performed as previously described[76]. Images were captured using a ZEISS Axio Zoom V16 stereomicroscope and associated Zen software, and neural folds visualized by pax3 were analyzed in Fiji. To verify the efficiency of MO, MO was injected into all cells at the 4-cell stage, and total RNAs were extracted from four embryos using the TRIZOL reagent (Invitrogen, cat # 15596026). cDNAs were synthesized using M-MLV reverse Transcriptase (Invitrogen, cat#28025013) and random primers (NEB, cat#S1330S). cDNAs were amplified by Taq polymerase (NEB, cat# M02735) using the following primers: 5'-CGGTGCAGGACTTGGATTAT-3', 5'-CCCATTTAGCTACCTCCTTCTG-3'.

Data Availability

Exome and genome sequencing data are available in publicly accessible databases for the 1,576 subjects enrolled after inclusion of dbGaP language into the consenting process/ became standard. These data are at dbGaP (accession #phs002591.v1.p1, phs000744.v5.p2 (Yale Mendelian Sequencing Center), Kids First: Whole Genome Sequencing in Structural Defects of the Neural Tube Accession (phs002591.v1.p1). Sequencing data on the remaining 743 subjects is available upon request.

Code Availability

Computational codes used in this study are available in GitHub (https://github.com/Gleeson-Lab/Publications/tree/main/MM_DNM).

Acknowledgement

We especially thank the families and the individuals with meningomyelocele that participated in this study, Kiely James, Renee George, Brett Copelande, Valentina Stanley, Celine Shen, and Jennifer Venneri from the Spina Bifida Sequencing Consortium, the Spina Bifida Association, the UCSD Laboratory for Pediatric Brain Disease for clinical and technical support, Brin Rosenthal and Katie Fisch from the UCSD Altman Clinical and Translational Research Institute for statistical modeling, Broad Institute, Yale Genetic Center, Regeneron Genetics Center, UCSD Institute for Genomic Medicine, UC Irvine Sequencing Center and Rady Children Institute for Genomics Medicine for sequencing support, Barbara Craddock from Stony Brook University for functional analysis of TNK2, and the Spina Bifida Association for recruitment efforts.

Funding

This work was supported by the Center for Inherited Disease Research grant HHSN268201700006I, the Yale Center for Genomic Analysis, the Broad Institute, the UC Irvine Genomics Core, the UCSD Institute for Genomic Medicine, the UCSD Imaging Core, X01HD100698, X01HD110998, HD114132, P01HD104436, U54OD030187, and support from the Howard Hughes Medical Institute and Rady's Children Institute for Genomic Medicine to JGG. YJH and SK were supported by the National Research Foundation of Korea funded by the Ministry of Science and ICT (MSIT) (RS-2023-00277314). WTM was supported by VA Merit Award # I01 BX006248.

Spina Bifida Sequencing Consortium members

Dr. Allison Elizabeth Ashley Koch, PhD (Duke University), Dr. Kit Sing Au, PhD and Dr. Hope Northrup, MD (University of Texas, Houston), Dr. Gyang Markus Bot, MD (Jos University Teaching Hospital, Jos, Nigeria), Valeria Capra, MD (IRCCS Istituto Giannina Gaslini, Italy), Dr. Richard H. Finnell, PhD, DABMGG (Baylor College of Medicine), Dr. Zoha Kibar, PhD (Université de Montréal, Canada), Dr. Philip J. Lupo, PhD, MPH (Baylor College of Medicine), Dr. Helio R. Machado, MD (University of Sao Paulo, Brazil), Dr. Tony Magana, MD (Mekelle University, Mekelle, Ethiopia), Dr. Rony Marwan, MD (University of Colorado), Dr. Gia Melikishvili, MD (MediClub Georgia,

T'bilisi, Georgia), Dr. Osvaldo M. Mutchinick, MD, PhD (Salvador Zubirán National Institute of Health Sciences and Nutrition, Mexico City, Mexico), Dr. Roger E. Stevenson, MD (Greenwood Genetic Center), Dr. Anna Yurrita, MD (Francisco Marroquín University, Guatemala City, Guatemala), Dr. Maha S. Zaki, MD, PhD (National Research Center, Cairo, Egypt), Dr. Sara Mumtaz, PhD (National University of Medical Sciences, Rawalpindi, Pakistan), Dr. José Ramón Medina-Bereciartu, MD (Venezuela Association of Spina Bifida, Caracas, Venezuela), Dr. Friedhelm Hildebrandt (Boston Children's Hospital, Harvard Medical School), Dr. Caroline M. Kolvenbach, MD (Boston Children's Hospital, Harvard Medical School), Shirlee Shril, MS (Boston Children's Hospital, Harvard Medical School), Dr. Mahmoud M. Noureldeen, MD and Dr. Aida MS. Salem, MD (Beni-Suef University, Beni-Suef, Egypt), Dr. Joseph Gleeson, MD (UC San Diego).

Author Contributions

Y-JH, CW, NM FJ, KIV, CB, SL, NJ, AP recruited subjects and performed genetic analysis. IJ performed MERFISH analysis, CJL and JBW generated Xenopus data. AN, J-EL, IT, FAB, KFT, SY, HJ, BB WTM, CH, SAM, HYG, BB, CP performed functional analysis, ZK GMB, HN, KSU, MS, AA-K, RHF, JL, HM, CA, HRM, RES, AY, SM, OMM, JRM-B, FH, GM, RM, VC, MMN, AMSS, MYI, MSZ recruited families. AA, ARS, SWK, performed sequencing. Y-JH, SK and JGG performed analysis, wrote drafts, and incorporated feedback from coauthors.

Competing interests

AA and ARS are full time employees of Regeneron Genetics Center. SK is cofounder of AIMA Inc., which seeks to develop techniques for early cancer diagnosis based on circulating tumor DNA.

Additional Information

Supplementary Information is available, containing Supplementary Table and Supplementary Notes.

Method references

52 Rendeli, C. *et al.* Assessment of health status in children with spina bifida. *Spinal Cord* **43**, 230-235 (2005). https://doi.org/10.1038/sj.sc.3101707

53 Dimitromanolakis, A., Paterson, A. D. & Sun, L. Fast and accurate shared segment detection and relatedness estimation in un-phased genetic data via TRUFFLE. *Am J Hum Genet* **105**, 78-88 (2019). https://doi.org/10.1016/j.ajhg.2019.05.007

54 Lee, J. *et al.* Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. *Nucleic Acids Res* **46**, W102-W108 (2018). https://doi.org/10.1093/nar/gky406

55 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010). https://doi.org/10.1093/bioinformatics/btq033

56 Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* **24**, 2125-2137 (2015). https://doi.org/10.1093/hmg/ddu733

57 Ioannidis, N. M. *et al.* REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* **99**, 877-885 (2016). https://doi.org/10.1016/j.ajhg.2016.08.016

58 Jaganathan, K. *et al.* Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535-548 e524 (2019). https://doi.org/10.1016/j.cell.2018.12.015

59 Giacopuzzi, E., Popitsch, N. & Taylor, J. C. GREEN-DB: a framework for the annotation and prioritization of non-coding regulatory variants from whole-genome sequencing data. *Nucleic Acids Res* **50**, 2522-2535 (2022). https://doi.org/10.1093/nar/gkac130

60 Suvakov, M., Panda, A., Diesh, C., Holmes, I. & Abyzov, A. CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing. *Gigascience* **10** (2021). https://doi.org/10.1093/gigascience/giab074

61 Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222 (2016). https://doi.org/10.1093/bioinformatics/btv710

62 Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012). https://doi.org/10.1093/bioinformatics/bts378

63 *https://github.com/brentp/smoove*.

64 Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015). https://doi.org/10.1126/science.aaa6090

65 Delile, J. *et al.* Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord. *Development* **146** (2019). https://doi.org/10.1242/dev.173807

66 Soldatov, R. *et al.* Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* **364** (2019). https://doi.org/10.1126/science.aas9536

67 Simoes-Costa, M. & Bronner, M. E. Establishing neural crest identity: a gene regulatory recipe. *Development* **142**, 242-257 (2015). https://doi.org/10.1242/dev.105445

68 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018). https://doi.org/10.1186/s13059-017-1382-0

69 Rosenthal, S. B. *et al.* Mapping the common gene networks that underlie related diseases. *Nat Protoc* **18**, 1745-1759 (2023). https://doi.org/10.1038/s41596-022-00797-1

70 Huang, J. K. *et al.* Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst* **6**, 484-495 e485 (2018). https://doi.org/10.1016/j.cels.2018.03.001

71 Duman, R. S., Sanacora, G. & Krystal, J. H. Altered connectivity in depression: GABA and glutamate neurotransmitter deficits and reversal by novel treatments. *Neuron* **102**, 75-90 (2019). https://doi.org/10.1016/j.neuron.2019.03.013

72 Tolias, K. F. *et al.* The Rac1-GEF Tiam1 couples the NMDA receptor to the activity-dependent development of dendritic arbors and spines. *Neuron* **45**, 525-538 (2005). https://doi.org/10.1016/j.neuron.2005.01.024

73 Komatsu, N. *et al.* Development of an optimized backbone of FRET biosensors for kinases and GTPases. *Mol Biol Cell* **22**, 4647-4656 (2011). https://doi.org/10.1091/mbc.E11-01-0072

74 Duman, J. G. *et al.* The adhesion-GPCR BAI1 shapes dendritic arbors via Bcr-mediated RhoA activation causing late growth arrest. *Elife* **8** (2019). https://doi.org/10.7554/eLife.47566

75 Henrie, H. *et al.* Stress-induced phosphorylation of CLIP-170 by JNK promotes microtubule rescue. *J Cell Biol* **219** (2020). https://doi.org/10.1083/jcb.201909093

76 Sive, H., Grainger, Robert M, and Harland, Richard M. *Early development of Xenopus Laevis: a laboratory manual.*, (2000).

77 Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-950 (2014). https://doi.org/10.1038/ng.3050

78 Misra, S. N. & Kearney, J. A. Alphabet Soup: Recurrent de novo mutations in novel genes causing developmental and epileptic Ecephalopathies. *Epilepsy Curr* **18**, 125-127 (2018). https://doi.org/10.5698/1535-7597.18.2.125

79 Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262-1266 (2015). https://doi.org/10.1126/science.aac9396

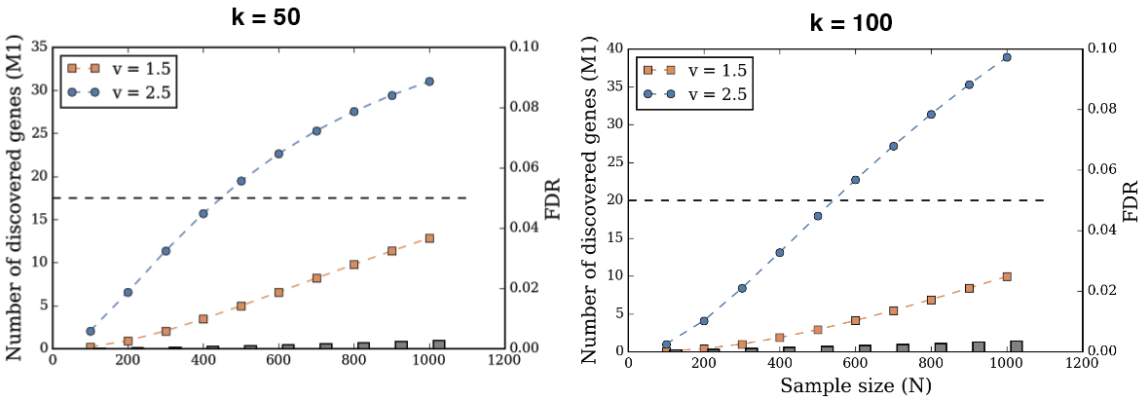
80 Ianevski, A., Giri, A. K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* **13**, 1246 (2022). https://doi.org/10.1038/s41467-022-28803-w

81 Ten Klooster, J. P. *et al.* Interaction between Tiam1 and the Arp2/3 complex links activation of Rac to actin polymerization. *Biochem J* **397**, 39-45 (2006). https://doi.org/10.1042/BJ20051957

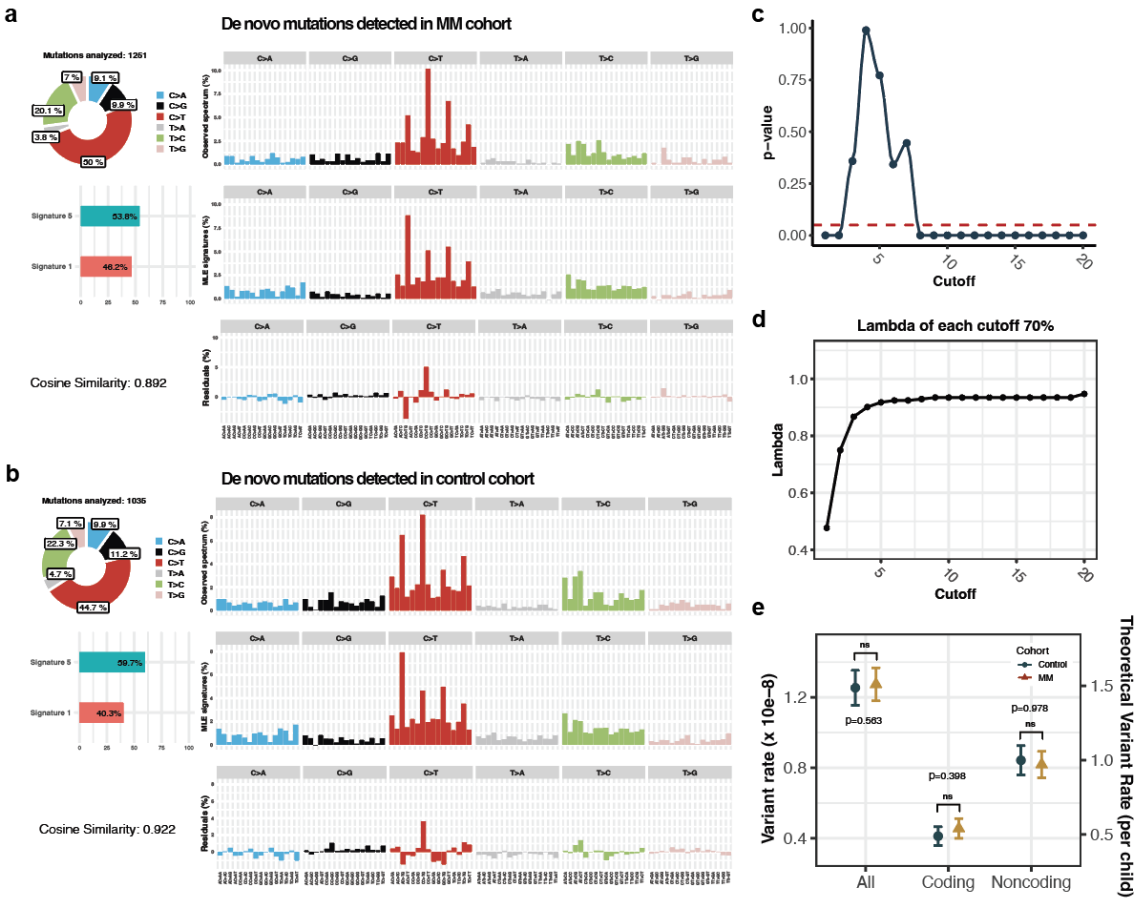
82 Yoo, S., Kim, Y., Lee, H., Park, S. & Park, S. A gene trap knockout of the Tiam-1 protein results in malformation of the early embryonic brain. *Mol Cells* **34**, 103-108 (2012). https://doi.org/10.1007/s10059-012-0119-x

83 Zhao, L. *et al.* Bhlhe40/Sirt1 axis-regulated mitophagy is implicated in all-trans retinoic acid-induced spina bifida aperta. *Front Cell Dev Biol* **9**, 644346 (2021). https://doi.org/10.3389/fcell.2021.644346

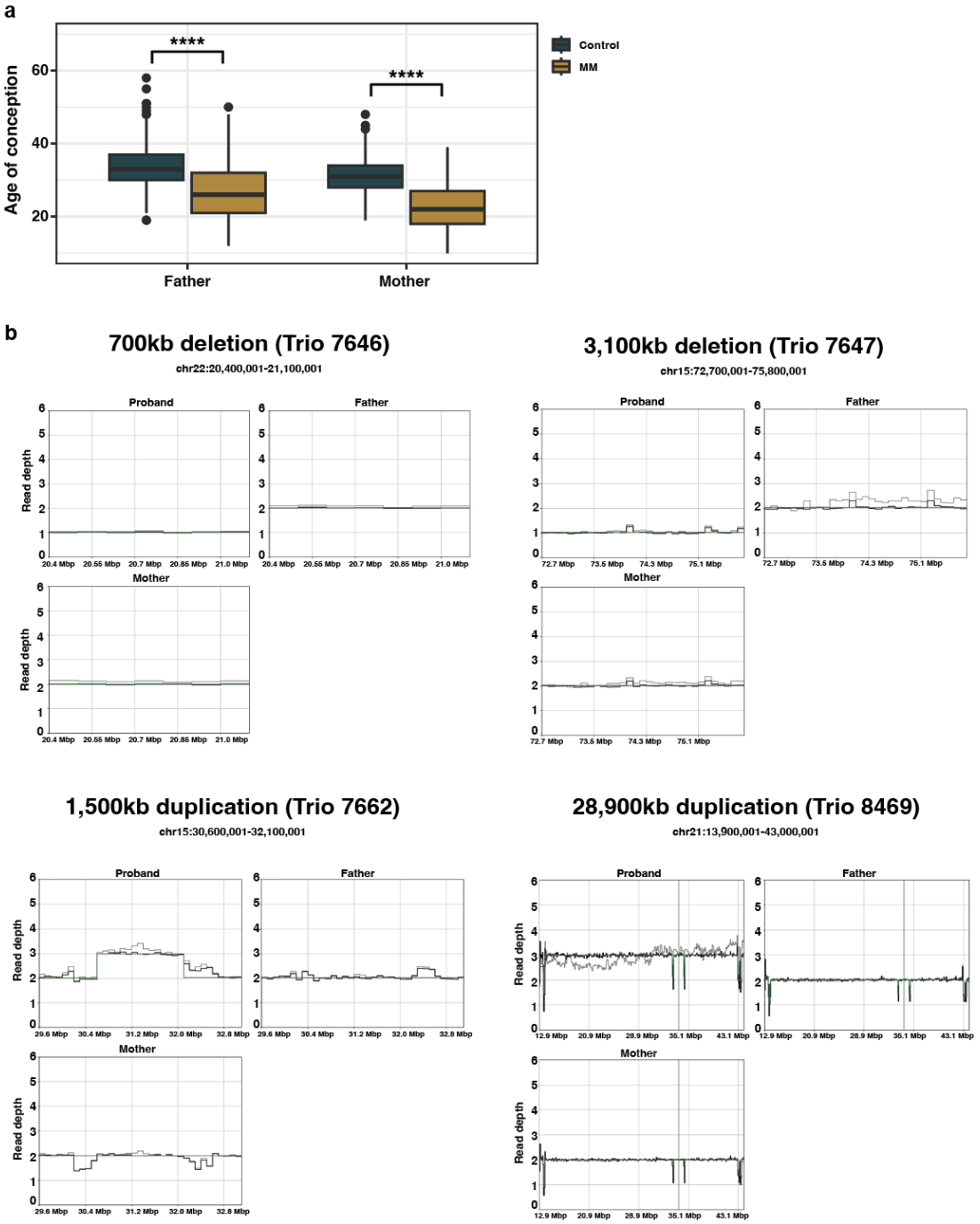
Extended Data Figure



Extended Data Figure 1. **Power calculation of estimating a cohort size for DNM detection.** Power calculation showing potential number of discovered genes compared with cohort size (350 trios), for two different v values (enrichment ratio of loss of function variants in case versus control) and two different k values (assumed number of risk genes). For instance, if there are 50 genes to discover (k=50), a cohort of 400 trios will identify 16 genes if LOF variants are 2.5x more common in affected (v=2.5). All calculations manage a conservative false discovery rate (FDR). Gray dash: FDR.

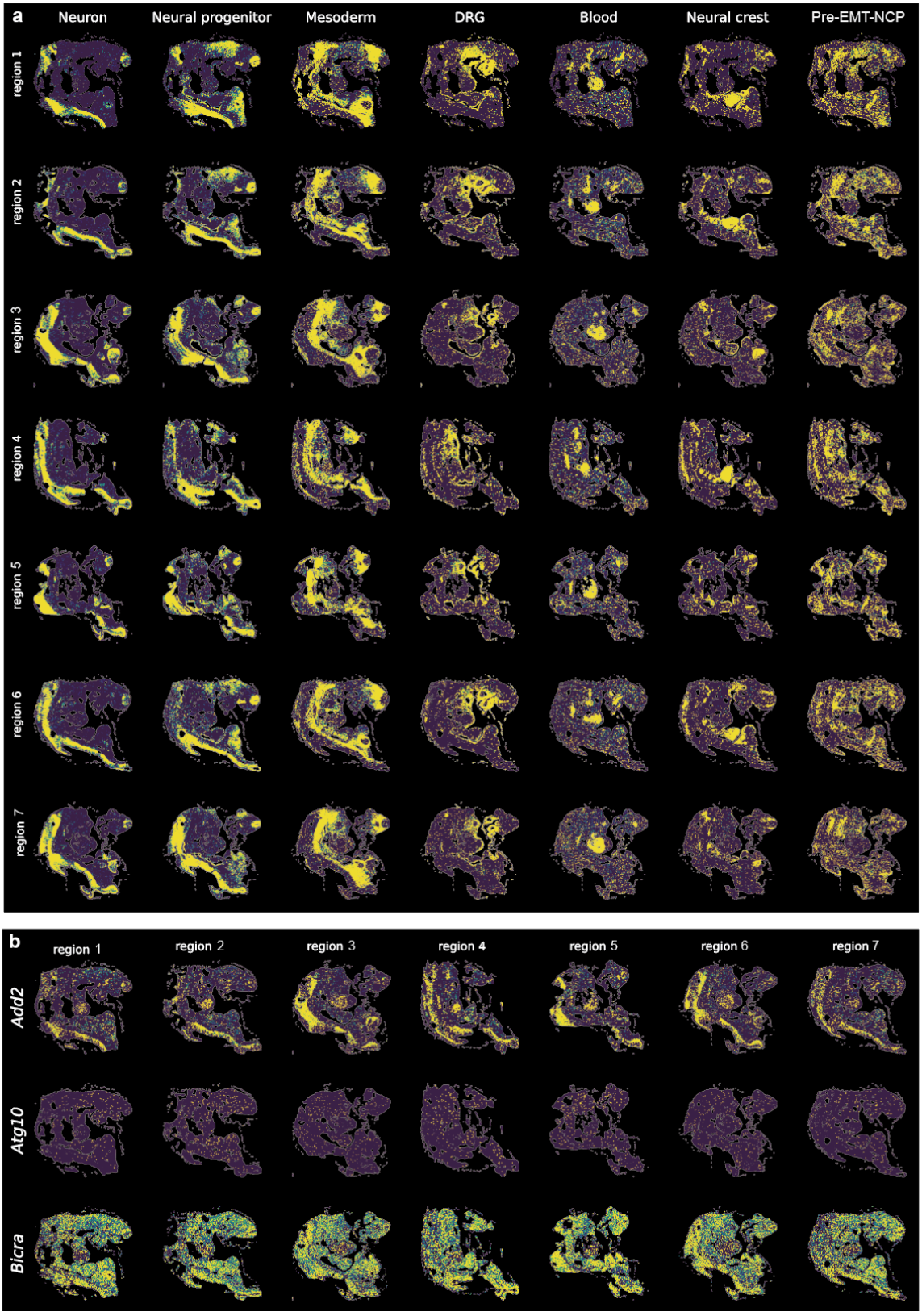


Extended Data Figure 2. Quantitative and qualitative features of de novo variants. **a-b**, Mutational signatures of DNMs showing known de novo signatures 1 and 5, in both (**a**) MM and (**b**) control cohorts. 1,251 and 1,035 SNVs were utilized for the signature analysis. **c-d**, DNM counts per proband following Poisson distribution, examined based on cutoffs (DNM count per proband) ranging from 1 to 20. (**c**) Chi-Square goodness-of-fit (two sided) was performed with an expected Poisson distribution with (**d**) lambda as an average DNM count per proband, for all cutoffs (see **Methods**). **e**, DNM rates for all, coding, and noncoding regions. Observed and theoretical DNM rates are marked in the left and right y-axis, respectively. Theoretical rate calculated by normalization of the variant rate with the total size of hg38 coding region (59,281,518 bp). *P* values were calculated by a two-sided Wilcoxon rank-sum test. ns: not significant.

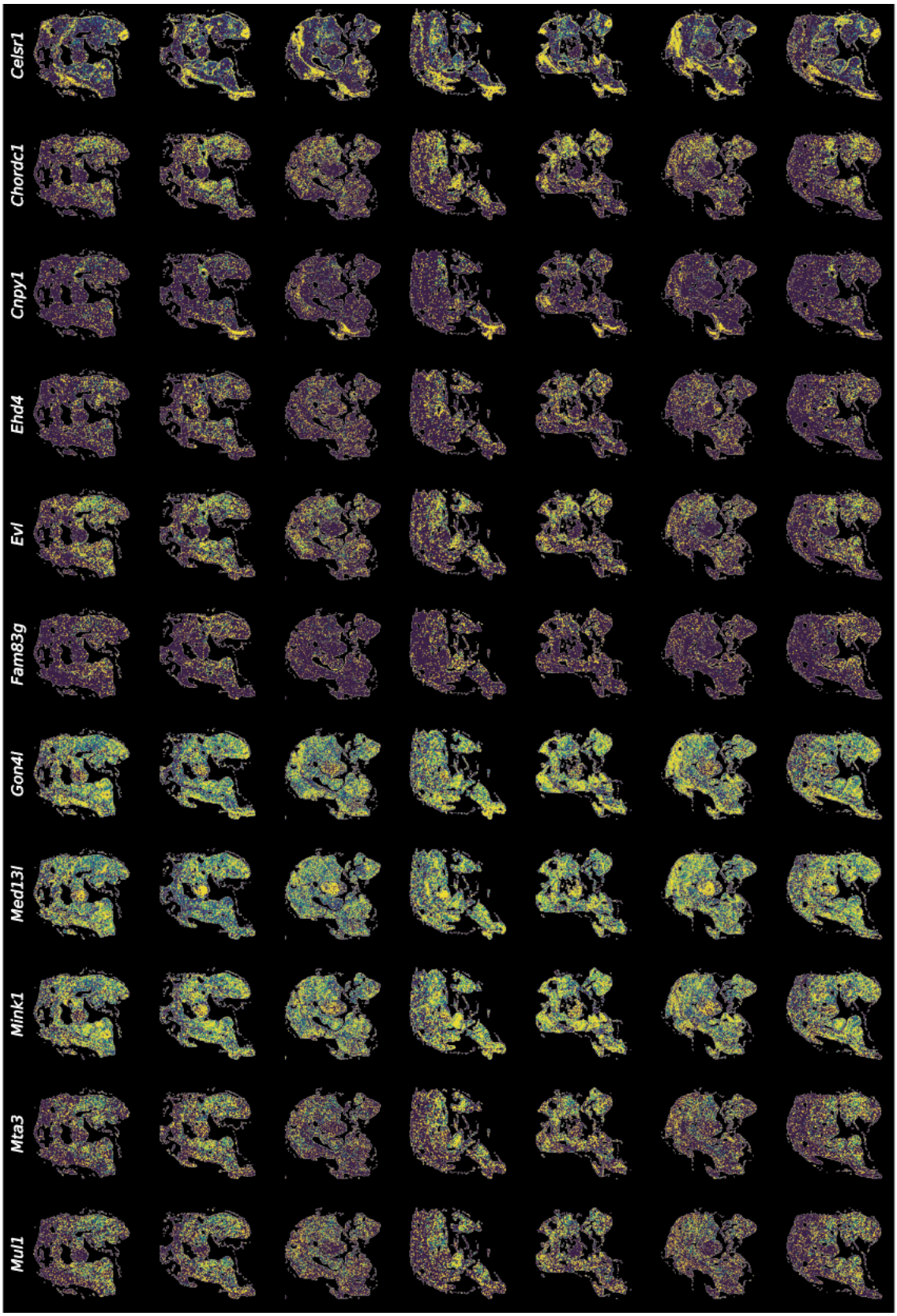


Extended Data Figure 3. Quantitative and qualitative features of de novo variants. **a**, Age of conception of father and mother in MM (n = 79) and control (n = 683) trios. The ages of conception were compared with Wilcoxon rank-sum test (two sided). The age of conceptions were significantly higher in controls than MM, for both of the parents. P values: **** < 0.0001. **b**, De novo large CNVs (> 100kb) detected from WGS. Read depths of proband, father, and mother shown. Gray lines: read depths. Green lines: copy number.

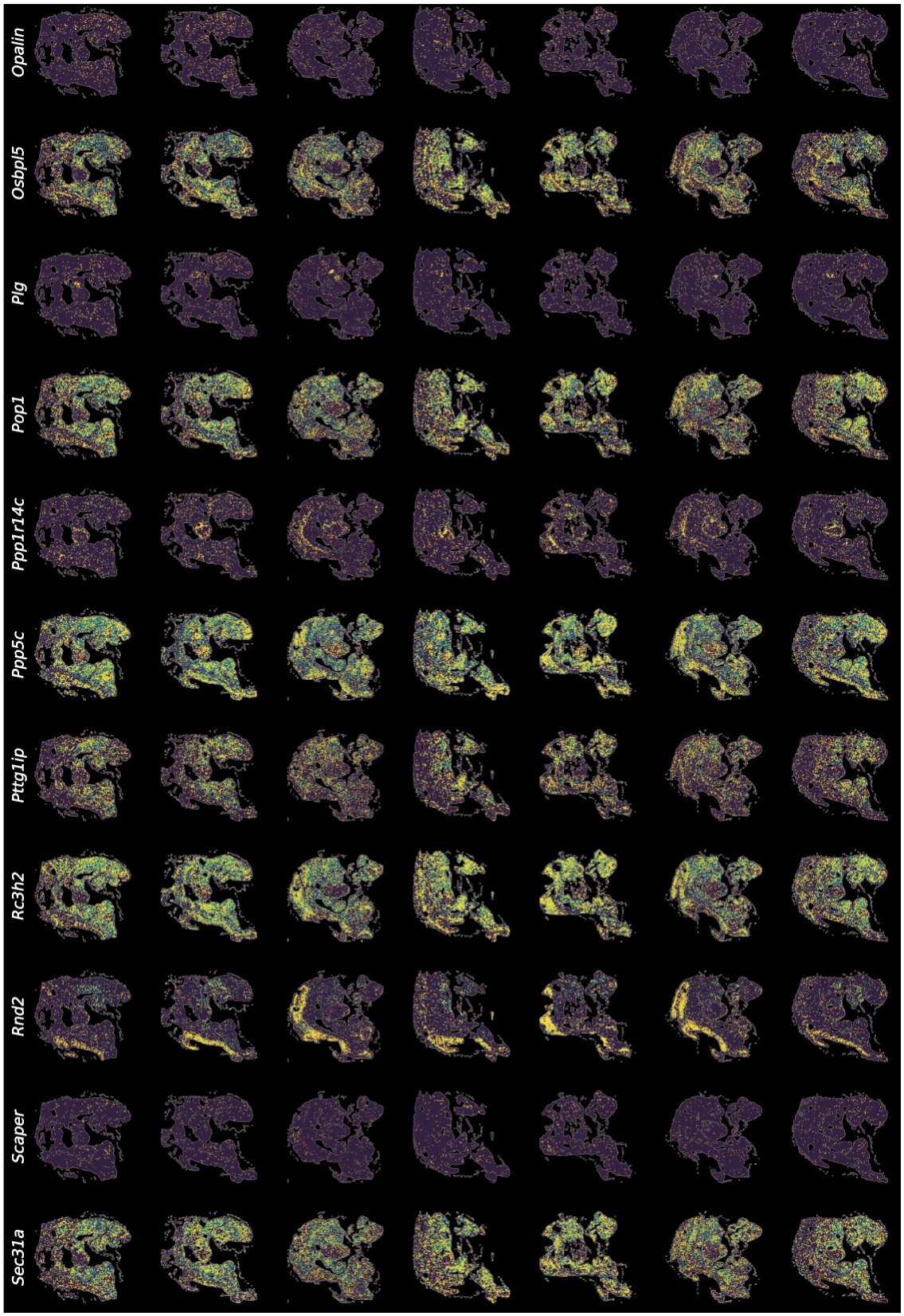
1277



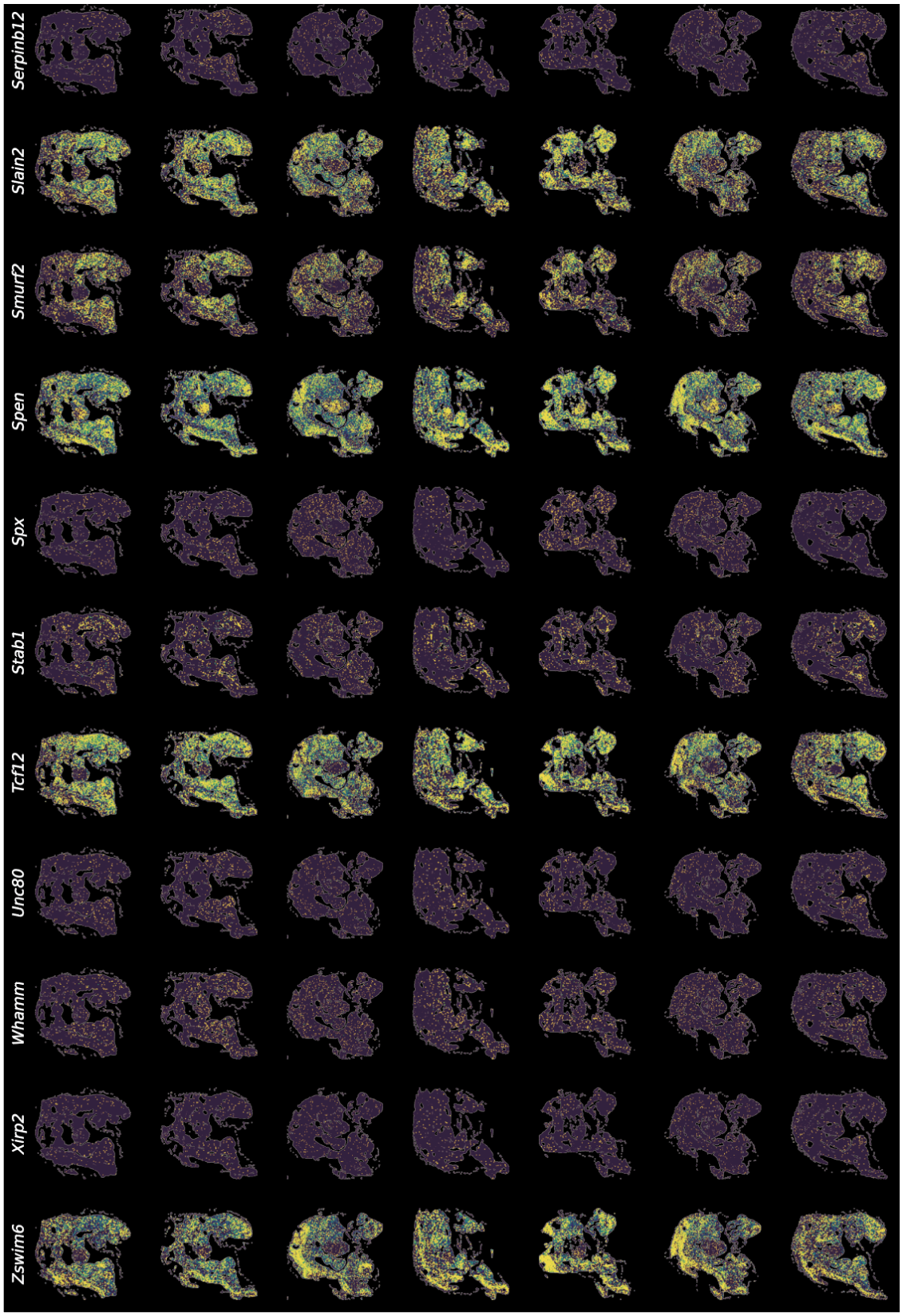
1278
1279



1280

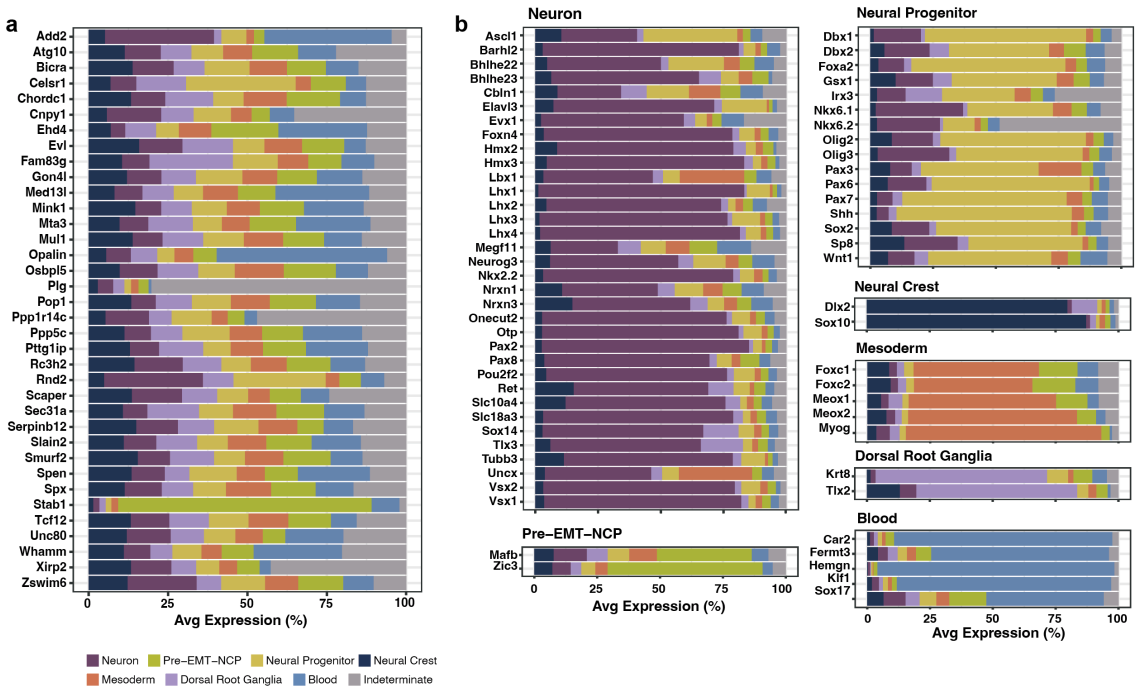


1281

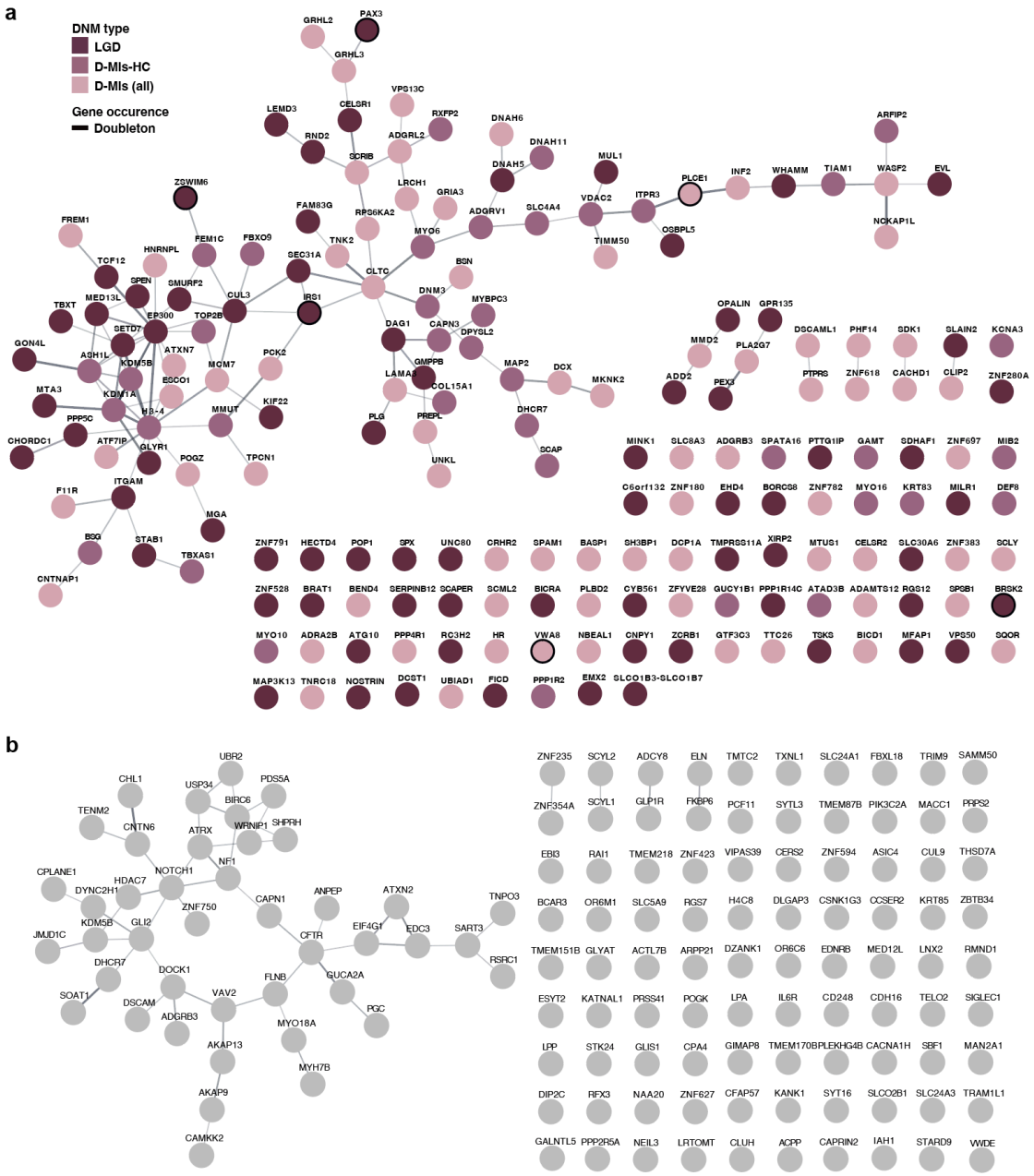


1282

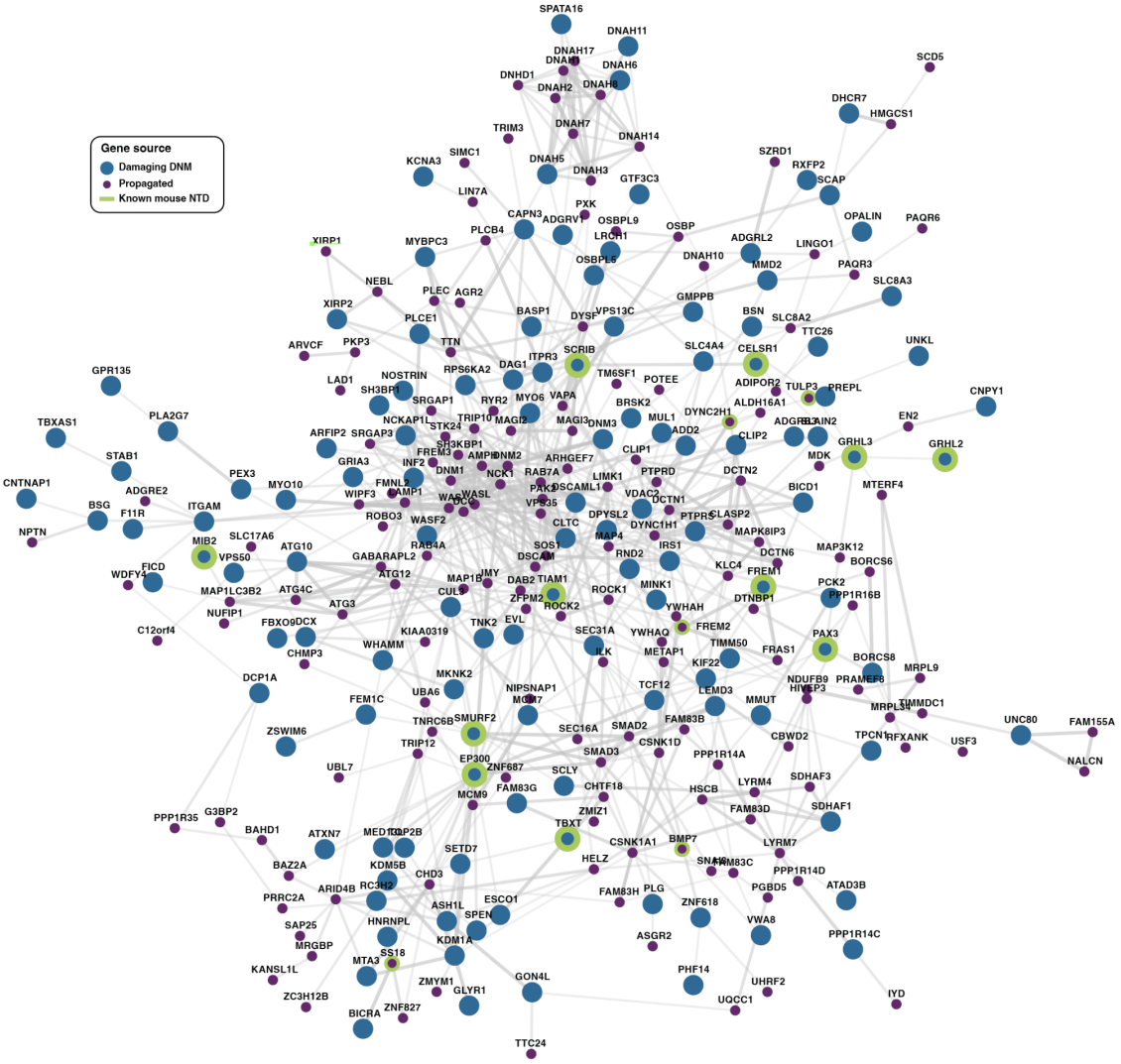
Extended Data Figure 4. Spatial expression of DNM genes with MERFISH. Spatial expression of the 36 MM genes with damaging DNMs. **a**, Gene expression of marker genes for seven selected cell types (neuron, neural progenitor, pre-epithelial to mesenchymal transition neural Crest progenitor (NC progenitor), neural crest, mesoderm, dorsal root ganglia, and blood), in seven embryonic replicates (regions) 1 (top)-7 (bottom). **b**, Spatial expression pattern of 36 damaging DNM genes. The seven replicates (X axes) of mouse E10.5 showing the gene expression pattern of candidate genes (Y axes). R=Rostral, C=Caudal, V=Ventral, D=Dorsal.



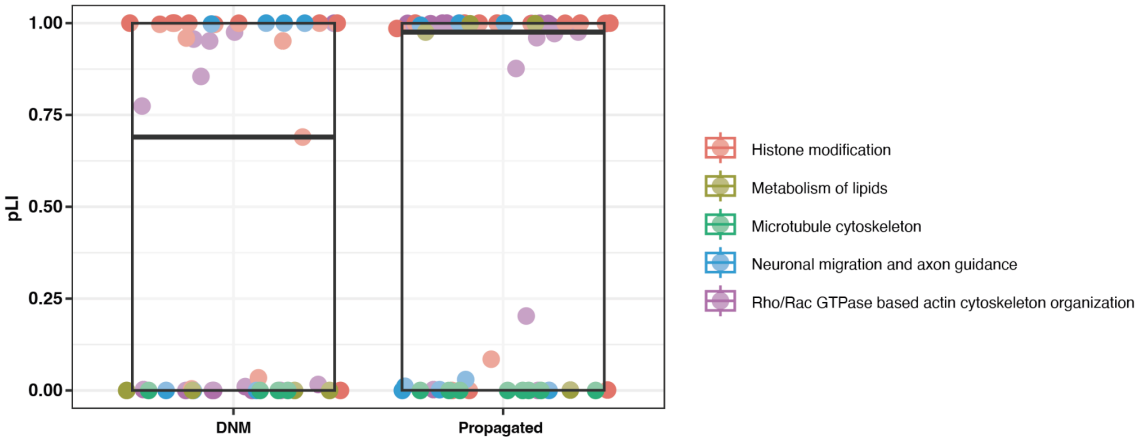
Extended Data Figure 5. Cell type expression of the DNM genes in MERFISH. a, Expression of the 36 damaging DNMs in seven cell types: neuron, neural progenitor, pre-epithelial to mesenchymal transition neural crest progenitor (Pre-EMT-NCP), neural crest, mesoderm, dorsal root ganglia, and blood. Indeterminate refers to the cells that were not specified with the marker genes designed for the seven cell types. **b**, Expression of marker genes used for specifying the cell types in MERFISH. Marker genes are shown within the cell type category which they represent.



Extended Data Figure 6. Connections of genes having damaging DNMs of MM and control cohorts. a-b, Genes with damaging DNMs (LGD + D-Mis) and their interconnections based on the STRING database, are shown for **(a)** MM and **(b)** control cohorts. Node color: variant functional categories: LGD (purple), D-Mis (light pink), and D-Mis-HC (dark pink). Doubleton genes: black borders. Edge thickness: confidence score of the protein interaction.

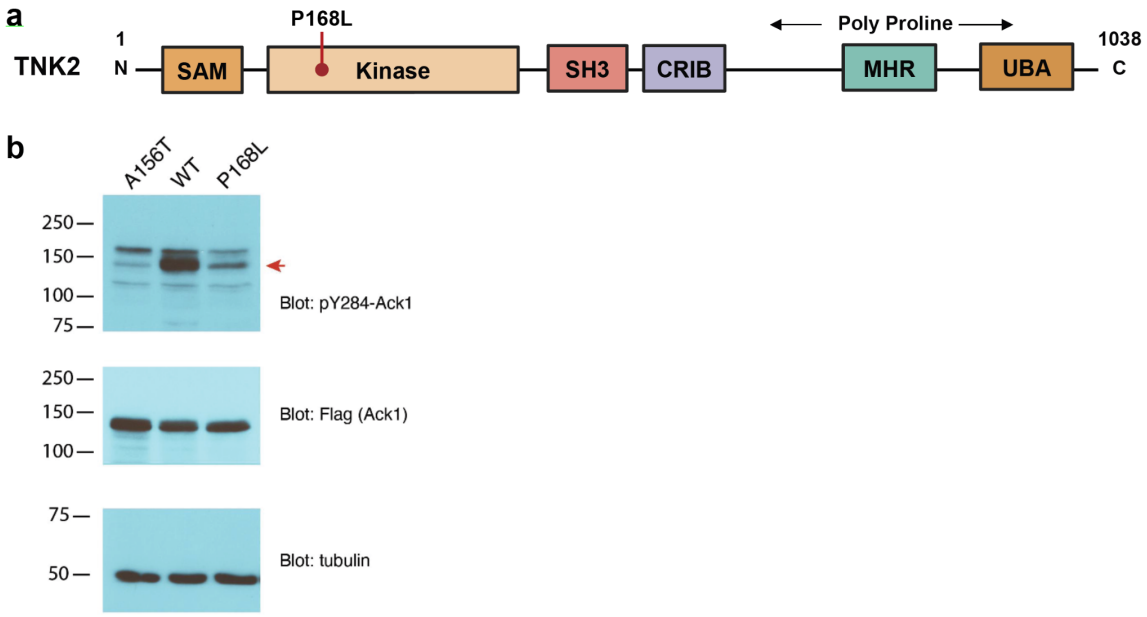


Extended Data Figure 7. A human 'Meningomyel-ome' constructed with damaging DNMs contributing to MM risk. By using the 187 damaging MM DNM genes, a propagated network, Meningomyel-ome, was generated with NetColoc[69] with a background protein network PCNet[70], incorporating 439 nodes and 2,447 edges. Big blue circle: damaging DNM genes, Small purple circle: propagated gene, Green border: known mouse NTD genes. The network is visualized with Cytoscape with STRING database.



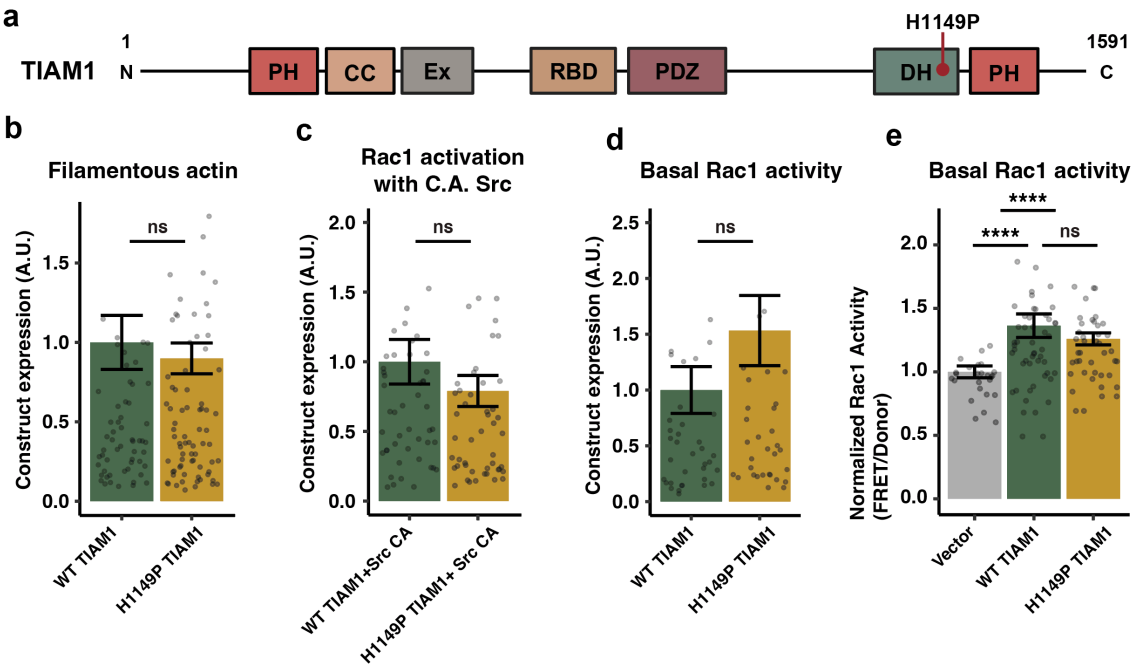
Extended Data Figure 8. pLI comparison of DNM and propagated genes from the five submodules. The probabilities of being loss of function intolerant (pLI) of genes from the five functional submodules are shown, derived from 46 damaging DNMs and 51 propagated genes. Black thick line: median, Box: 1st and 3rd quantiles.

1320

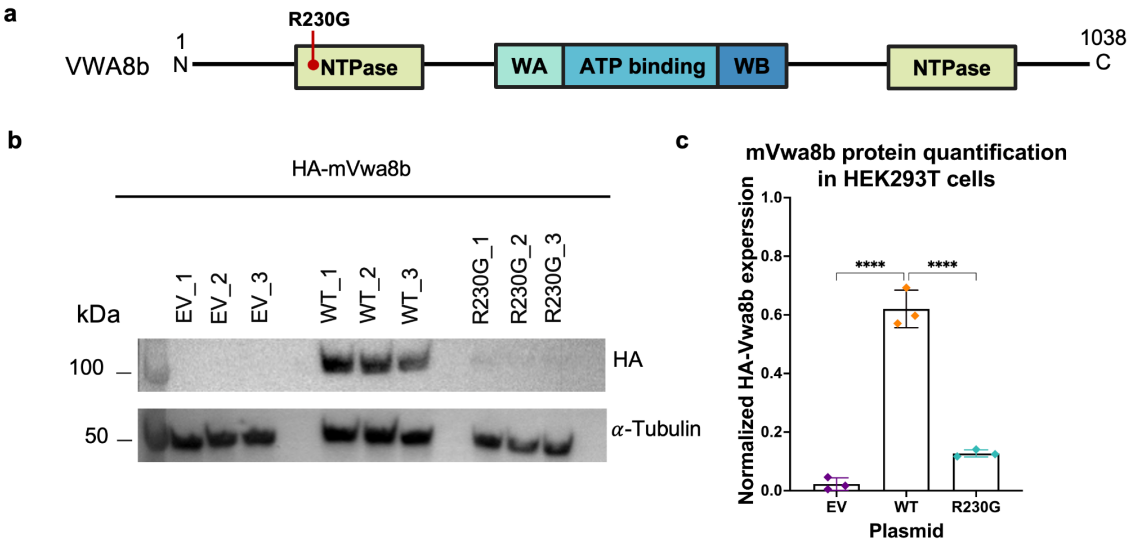


1321

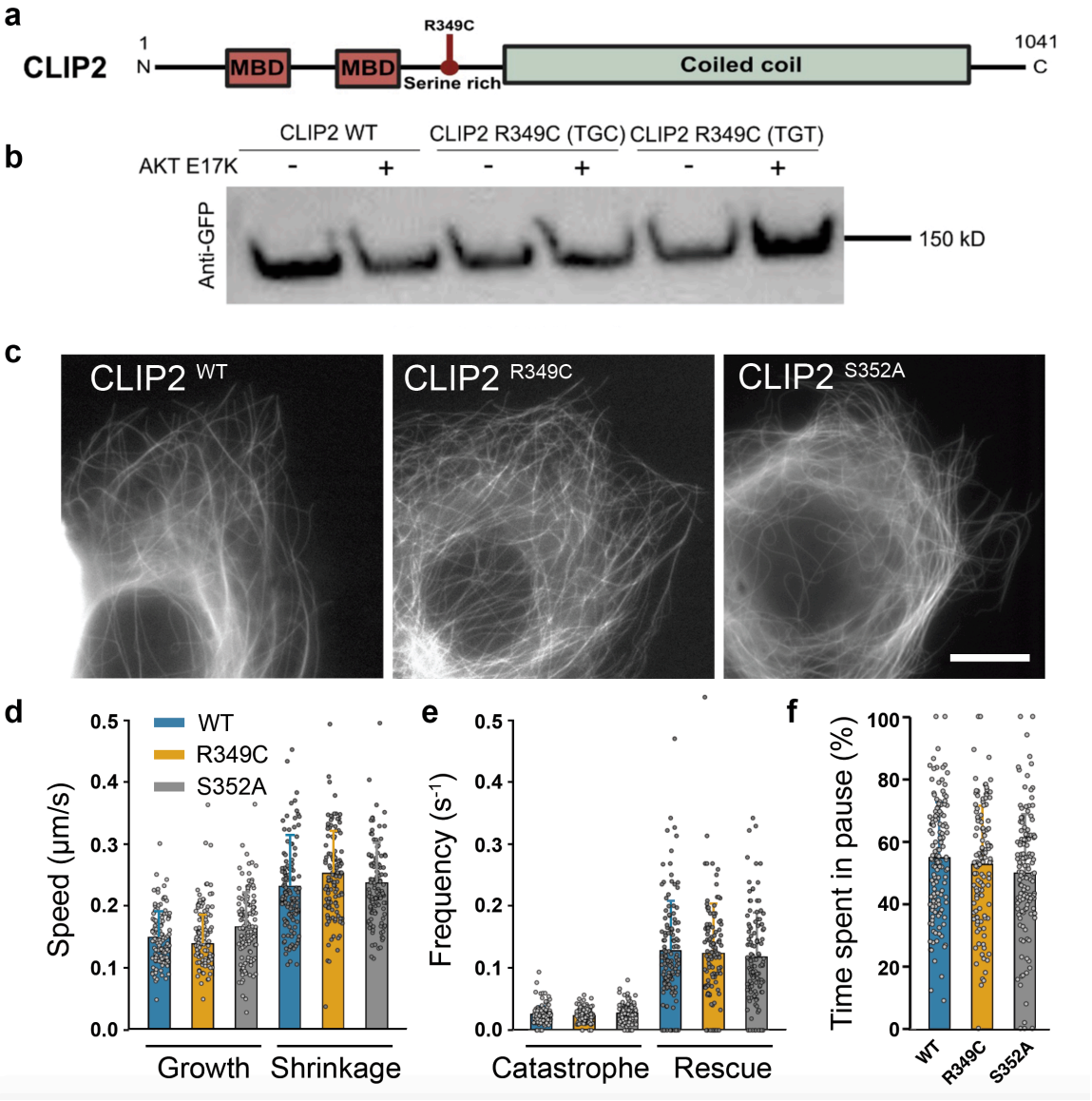
1322 **Extended Data Figure 9**. **The P168L patient mutation impairs TNK2 activity. a**,
1323 P168L mutation is located within the kinase domain. TNK2 contains sterile alpha motif
1324 (SAM), Src homology 3 (SH3), CDS42, and RAC-interactive binding (CRIB), Mig6
1325 homology region (MHR), and ubiquitin-associated domain (UBA). **b**, Blots for the
1326 A156T kinase dead, wild-type, and the patient mutation P168L. Lysates were probed
1327 with pY284-Ack1 (top), Ack1-flag (middle), and gamma-tubulin (bottom).



Extended Data Figure 10. The H1149P patient mutation impairs TIAM1 activity. **a**, H1149P mutation is located within the Dbl homology (DH) domain responsible for GEF activity. TIAM1 contains an N-terminal pleckstrin homology (PH), coiled-coiled (CC), extension (Ex), RAS binding (RBD), PDZ, Dbl-homology (DH) and PH domains, with the patient mutation falling within the DH domain. **b**, Construct expression H1149P is equivalent to wildtype in Phalloidin quantification. **c-d**, Construct expression H1149P in (**c**) basal and (**d**) constitutive active (C.A.) Src Rac1 Förster resonance energy transfer (FRET) is equivalent to wildtype. **e**, Protein expression H1149P in C.A. Src Rac1 FRET is equivalent to wildtype. Bar: mean, Errorbar: standard deviation of mean (SEM). Kruskal-Wallis followed by a pairwise Wilcoxon test, *P* value adjusted with Bonferroni. Data shown with Hampel filter.

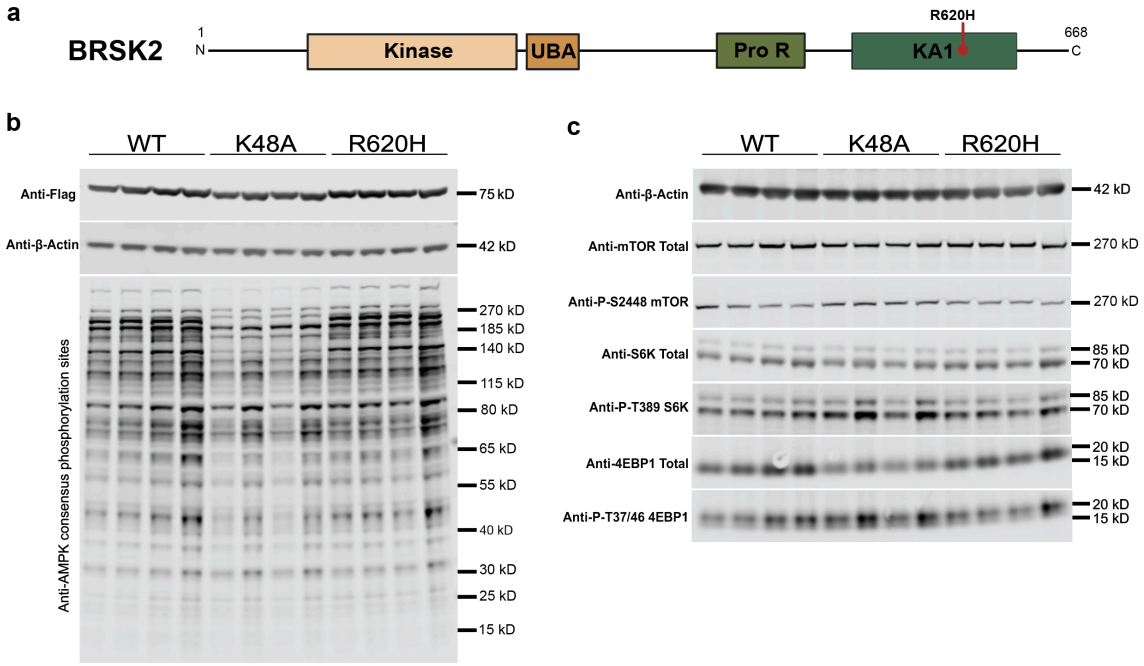


Extended Data Figure 11. *VWA8b* patient mutation R230G significantly reduces protein expression levels. a, Schematic of protein domains for VWA8b consisting of NTPase, Walker A (WA), ATP binding, and Walker B (WB) domain and patient mutation R230G in the NTPase domain. **b,** Western blot for HA-tagged mVwa8b empty vector (EV), wildtype and p.R230G protein overexpressed in HEK293T cells and stained for HA and alpha tubulin as loading control**. c,** Quantification of HA band intensity from panel **b** normalized to loading control. Bar: mean, Errorbar: standard deviation of mean (SEM). one-way ANOVA ****: *P* value < 0.0001.

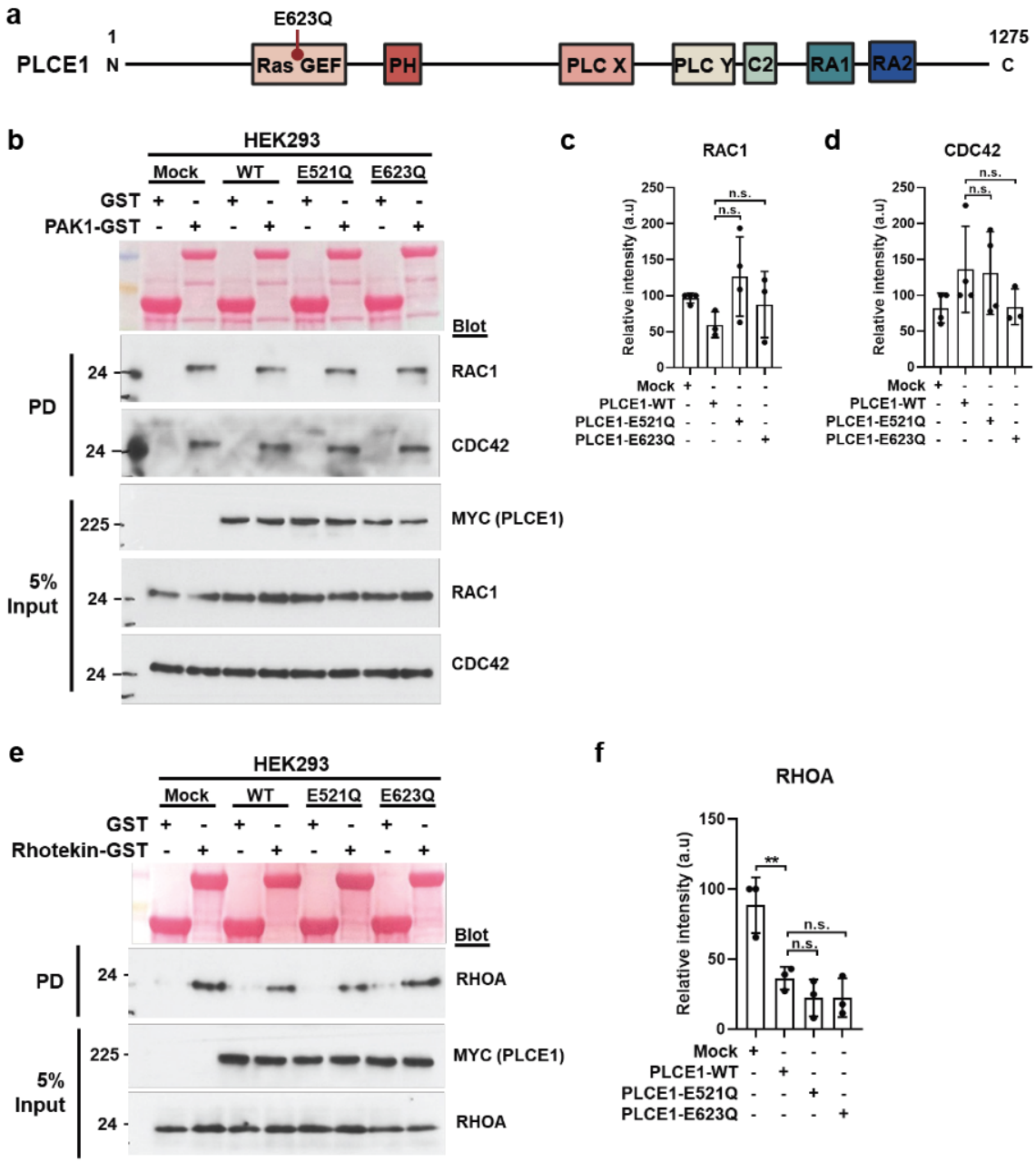


Extended Data Figure 12. The R349C patient mutation in *CLIP2* has no observable effect on microtubule dynamics. a, Domains in CLIP2 protein, microtubule associated domain (MBD), serine rich region containing the R349C patient mutation, and coiled coil domain. **b,** GFP fused CLIP2 expression in HEK293T cells for wildtype (WT) and two synonymous codons (TGC and TGT both encoding for CYS) for R349C patient mutation. No notable difference in the level of expression was detected between the WT and the mutants. Experiments included co-transfection with constitutively active AKT E17K, a presumed upstream kinase of nearby S352. **c-f,** While at a low expression level in HuH7 cells, fluorescent CLIP2 is restricted to growing microtubule plus ends as expected (not shown), (**c**) its mild overexpression highlights the whole microtubule bodies without difference between WT, R349C (TGC) and a putative phosphorylation site, S352A (Scale bar 10µm). This condition allowed the measurement of a complete panel of microtubule dynamic instability parameters: (**d**) growth and shrinkage speed, (**e**) catastrophe and rescue (R) frequencies (**f**) the time spent in pause. There was no statistical difference between CLIP2 WT and the mutants. The values were measured

from at least 125 microtubules (25 cells) per condition. n = 3 independent experiments. Bar: mean, Point: data for each microtubule, Errorbar: standard deviation.

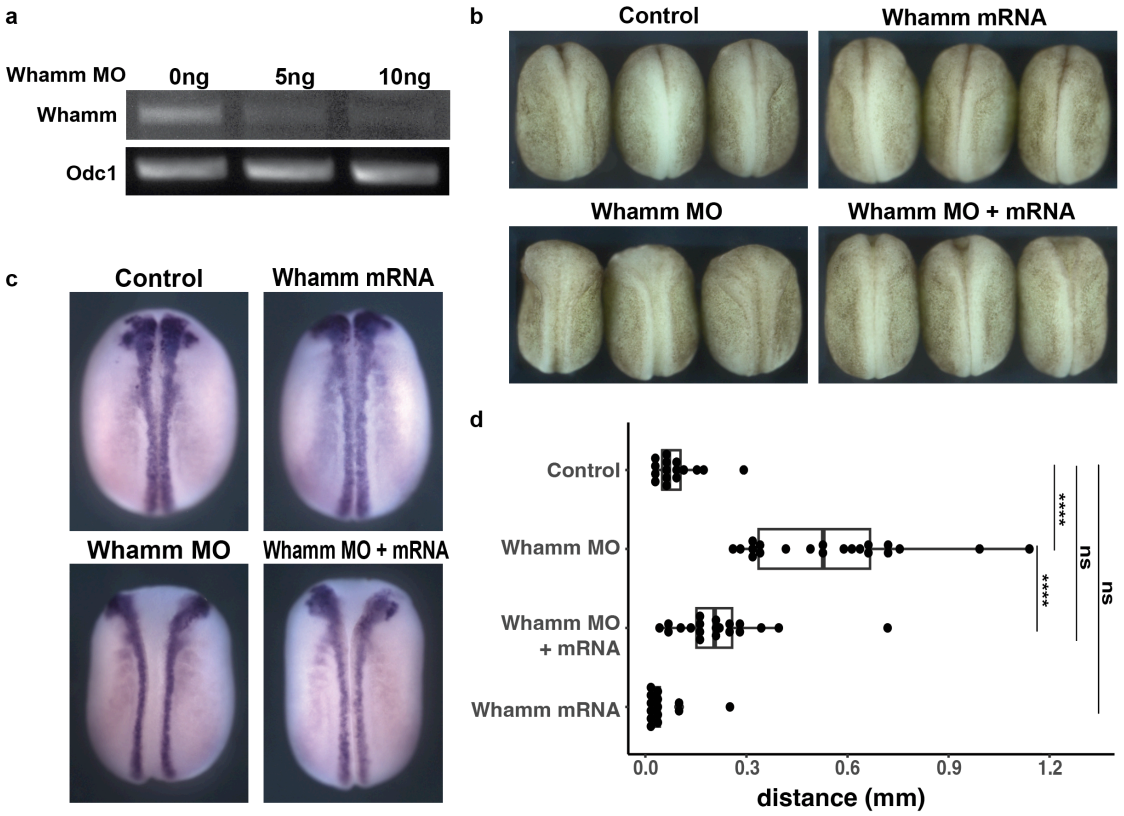


Extended Data Figure 13. *BRSK2* patient mutation R620H shows no notable defects in kinase activity across known downstream phosphorylation targets. a, R620H mutation is located within the kinase associated 1 domain (KA1). BRSK2 also contains a kinase domain, Ubiquitin-associated (UBA), Proline rich region (Pro R). **b,** Western blot showing phosphorylation of AMPK substrate motifs observed across wildtype (WT), kinase dead mutant control (K48A) and patient mutation (R620H) Flag-tagged BRSK2 when expressed in HEK293T cells. K48A has an average reduced intensity that is not observed in R620H. **c,** Western Blots of total and phosphorylated mTOR, S6K and 4EBP1 substrates showing no notable differences in phosphorylation for patient mutation R620H.

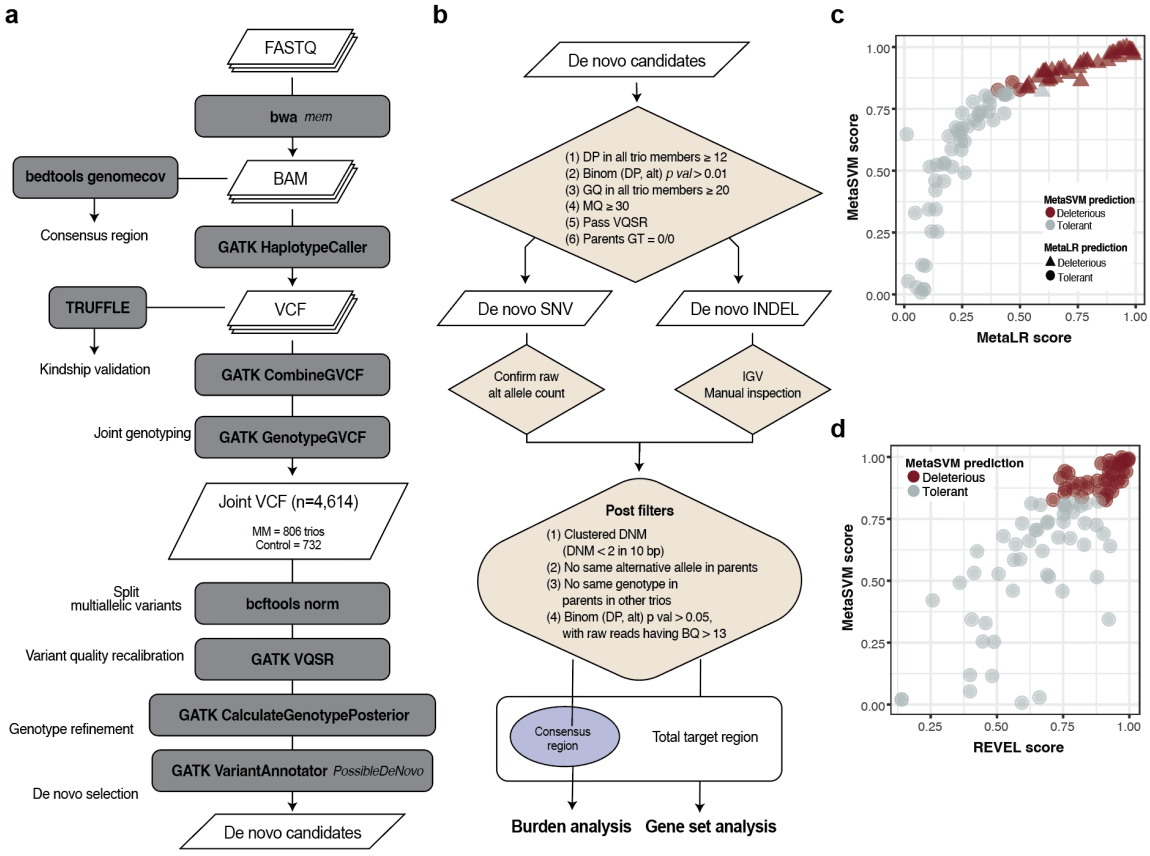


Extended Data Figure 14. *PLCE1* patient mutation E623Q does not lead to diminished Rho GTPase activity. a, Schematic of PLCE1 protein with domains annotated. Patient E623Q mutation is located in the Ras GEF domain. PLCE1 contains a Guanine nucleotide exchange factor for Ras-like small GTPases (RAS GEF), Pleckstrin Homology (PH), Phospholipase C catalytic domain X (PLCX), Phospholipase C catalytic domain Y (PLCY), Protein Kinase C conserved region 2 (C2), RAS association domain 1 (RA1), and RAS association domain 2 (RA2). **b,** Active GTP-bound forms of RAC1 and CDC42 precipitated from HEK293 expressing Myc-tagged PLCE1 using a GST-PAK1 pulldown assay (n=4). Five percent input represents the controls for equal loading. Compared to mock cells, overexpression of wild-type (WT) and variant forms of PLCE1

exhibited no significant difference in RAC1 and CDC42 activity. **c**-**d**, Quantifications of (**c**) RAC1 and (b) CDC42. **e**, Active GTP-bound form of RHOA precipitated from HEK293 expressing Myc-tagged PLCE1 using a GST-rhotekin pulldown assay. Overexpression of WT PLCE1 resulted in a substantial decrease in relative RHOA activity compared with mock cells. Compared to WT, cells transfected with variant forms of PLCE1 exhibited no significant differences in GTP-bound RHOA. **f**, Quantification of RHOA (n=4). Data were analyzed by one-way ANOVA with post hoc Bonferroni correction. Error bars: standard deviation for greater than 3 independent experiments. PD; pulldown, ** $P < 0.01$, n.s. not significant.



Extended Data Fig. 15. Validation of Whamm morpholino antisense oligonucleotide (MO). a, RT-PCR validates the effective disruption of splicing following injection of Whamm MO. **b-d**, The neural tube closure defect phenotype induced by Whamm MO (10 ng) was rescued through the injection of Whamm mRNA (700 pg). Embryos injected only with mRNA showed no significant phenotype. **b**, Dorsal views of Xenopus embryos at Stage 19 **c**, Embryos subjected to in situ hybridization against Pax3 to visualize the neural folds **d**, Graph depicting the quantitative analysis of the distance between neural folds. P-values by one-way ANOVA with post-hoc honestly significant difference (HSD) test. ****: *P* value < 0.0001, ns: not significant.



Extended Data Figure 16. Pipeline of detecting high-confidence de novo mutation (DNM)s and annotation of D-Mis-HC. a, Schematic overview of DNM detection. The raw coverages of aligned bam files were utilized in consensus region generation (see Methods). Germline variants were utilized to test the integrity of the trios with TRUFFLE. GATK genotype refinement workflow was utilized for the DNM detection **b**, The candidate de novo SNVs and Indels were applied with a series of qualitative and quantitative filters. DP: depth, Binom: binomial test, alt: count of reads with alternative alleles, GQ: genotype quality, MQ: mapping quality, VQSR: variant quality score recalibration, GT: genotype. **c-d**, Pathogenicity prediction of D-Mis-HC variants with MetaLR and REVEL. The pathogenicity of 116 D-Mis DNMs predicted by MetaSVM are shown with the prediction results from MetaLR and REVEL. (**c**) MetaLR deleterious (score > 0.5) is shown in triangle and (**d**) mutations with REVEL score higher than 0.75 are predicted as deleterious (right).

Extended Data Table

Mutation type	Category	Coding	Noncoding	Total
SNV + Indel	DNM count	1,976	8,817	10793
	DNM count per proband	20.16	89.97	110.13
	rate ($\times 10^{-8}$)	0.315	1.4	1.72
SNV	DNM count	1,658	7,429	9,087
	DNM count per proband	16.92	75.81	92.72
	rate ($\times 10^{-8}$)	0.26	1.18	1.45
Indel	DNM count	318	1,388	1,706
	DNM count per proband	3.24	14.16	17.41
	rate ($\times 10^{-8}$)	0.05	0.22	0.27

Extended Data Table 1. **De novo SNV and Indel rates ($\times 10^{-8}$) of WGS in coding and noncoding regions.** Rates were calculated with 99 trios could be analyzed after removal of four kinship failed trios.

Family	Chromosome	Start position	End position	Length	Variant type	Cytoband
7646	chr22	20,400,001	21,100,001	-700,000	Deletion	q11.21
7647	chr15	72,700,001	75,800,001	-3,100,000	Deletion	q24.2-q24.1
7662	chr15	30,600,001	32,100,001	1,500,000	Duplication	q13.3-q13.2
8469	chr21	13,900,001	43,000,001	28,900,000	Duplication	q22.11-q22.12

Extended Data Table 2. de novo large CNV (>100kb) in WGS, analyzed with 99 trios.

Term	Ontology Source	ID	*P* Value (Bonferroni step-down adjusted)	% Associated Genes	Nr. Genes	Associated Genes
Morphogenesis of a polarized epithelium	GO Biological Process	GO:0001738	0.00001	8.41	9.00	BRSK2, CELSR1, CELSR2, GRHL3, LAMA3, SCRIB, SH3BP1, SMURF2, TIAM1
Signaling by NTRKs	Reactome	R-HSA:166520	0.00031	5.97	8.00	CLTC, DNM3, EP300, IRS1, PTPRS, RPS6KA2, TCF12, TIAM1
Establishment of planar polarity	GO Biological Process	GO:0001736	0.00070	7.41	6.00	BRSK2, CELSR1, CELSR2, GRHL3, SMURF2, TIAM1
Wnt signaling pathway, planar cell polarity pathway	GO Biological Process	GO:0060071	0.00100	8.62	5.00	CELSR1, CELSR2, GRHL3, SMURF2, TIAM1
L1CAM interactions	Reactome	R-HSA:373760	0.00220	5.04	6.00	CLTC, CNTNAP1, DCX, DNM3, DPYSL2, RPS6KA2
Signaling by NTRK1 (TRKA)	Reactome	R-HSA:187037	0.00276	5.22	6.00	CLTC, DNM3, EP300, IRS1, RPS6KA2, TCF12
Neural tube closure	GO Biological Process	GO:0001843	0.00323	4.90	5.00	CELSR1, GRHL2, GRHL3, SCRIB, TBXT

Extended Data Table 3. **Pathway enrichment of the damaging DNMs in MM.** 187 damaging DNMs (LGD + D-Mis) were tested for enrichment in known functional biological pathways in Gene ontology (GO) Biological Process, Reactome, and KEGG pathways. Pathways that were significant with Bonferroni step-down adjusted *P* value < 0.05 are shown.

Supplementary Table

Family ID	Gene	Mutation	OMIM associated phenotype	Mutations reported with OMIM phenotype	Patient phenotype
5875	*KIF22*	chr16:29798687C >T; c.484C>T; p.Arg162Ter	Spondyloepimetaphyseal dysplasia s. with joint laxity, type 2	Heterozygous missense in exon 4	Excluded Spondyloepimetaphyseal dysplasia features
6032	*CELSR1*	chr22:46397819G >GTTGGT; c.5551_5555dup; p.Asn1852LysfsTer7	Lymphatic malformation 9 s.	Heterozygous nonsense or frameshift show incomplete penetrance	Excluded lymphatic malformation features
6147	*SLCO1B3-SLCO1B7*	chr12:21067374C >T, c.1803C>T;p.Arg580X	Digenic recessive hyperbilirubinemia, Rotor type s.	Read through transcript	Excluded hyperbilirubinemia features
6293	*TOP2B*	chr3:25636033A>G, c.c.740T>C; p.Leu247Pro	B-cell immunodeficiency, distal limb anomalies, and urogenital malformation s.	Heterozygous deletion and missense mutations in the TOPRIM domain	Excluded immunodeficiency, limb and urogenital defect features
6307	*ADGRB3*	chr6:68975294G>A, c.1688G>A; p.Ser563Asn	Cerebellar ataxia	Biallelic loss	Excluded cerebellar ataxia features
6313	*BICRA*	chr19:47680692C >T, c.1522C>T; p.Gln508*	Coffin-Siris.-12 s.	Haploinsufficiency	Meets criteria for Coffin-Siris s.
6319	*CLIP2*	chr7:74357307C>T, c.1045C>T; p.Arg349Cys	Williams-Beuren s.	Contiguous gene deletion s. at 7q11.23	Excluded Williams-Beuren s. features
6328	*SLC4A4*	chr4:71472901C>A, c.1702C>A; Pro568Thr	Proximal renal tubular acidosis with ocular abnormalities s.	Biallelic substitutions lead to decrease NBC activity	Excluded renal tubular acidosis or ocular features. Displays growth retardation.
6337	*TCF12*	chr15:57262209G >T, c.412+1G>T; splice loss	Craniosynostosis type 3 s. Hypogonadotropic hypogonadism 26 s.	Haploinsufficiency; Biallelic loss of function	Displays features of craniosynostosis.
6388	*LOXHD1*	chr18:46509817T >G, c.2067-2A>C splice loss	Deafness, autosomal recessive type 77 s.	Biallelic mutations	Excluded features of deafness
6395	*PLA2G7*	chr6:46709351G>A, c.845C>T; p.Thr282Ile	Platelet-activating factor acetylhydrolase deficiency s.	Biallelic mutations	Excluded features of platelet-activating factor acetylhydrolase deficiency
6423	*KDM1A*	chr1:23057487C>T, c.934C>T p.Arg312Cys	Cleft palate and distinctive facial features s.	Heterozygous loss	Lacks features of cleft palate
6570	*TNK2*	chr3:195872454A >T, c.1462T>A, p.Trp488Arg	Autosomal recessive infantile epilepsy (pending confirmation)	Biallelic loss	Excluded features of infantile epilepsy
6695	*TIMM50*	chr19:39485743A >G, c.109-2318A>G;	3-methylglutaconic aciduria, type IX s.	Biallelic loss	Excluded features of 3-methylglutaconic aciduria

		p.Tyr143Cys			
6705	*ASPM*	chr1:197102395TGAGA>T, c.6852_6855delTCTC; p.Leu2285Argfs*6	Microcephaly 5, primary, autosomal recessive s.	Biallelic loss	Excluded features of microcephaly
6708	*SCAPER*	chr15:76705971G>A, c.2179C>T; p.Arg727Ter	Intellectual developmental disorder and retinitis pigmentosa s.	Biallelic loss	Excluded features of intellectual disability or retinitis pigmentosa
6133	*ZSWIM6*	chr5:61541917C>T, c.2737C>T; p.Arg913Ter	Acromelic frontonasal dysostosis s.	Haploinsufficiency	Reduced size of nose, possible dysostosis
6289	*UNC80*	chr2:209872917C>T, c.3793C>T; p.Arg1265Ter	Hypotonia, infantile, with psychomotor retardation and characteristic facies 2 s.	Biallelic loss	Lacks typical facies, but has features of hypotonia
7443	*POP1*	chr8:98130062G>T, c.571G>T, p.Glu191Ter	Anauxetic dysplasia 2	Biallelic loss	Lacks features of anauxetic dysplasia
7437	*DNAH5*	chr5:13807727T>G, c.7753-2A>C, splice loss	Ciliary dyskinesia, primary, 3	Biallelic loss	Lacks features of ciliary dyskinesia
7440	*ATXN7*	chr3:63990756C>T, c.1579C>T; p.Arg527Trp	Spinocerebellar ataxia 7	CAG repeat expansion	Lacks features of spinocerebellar ataxia
8012	*MED13L*	chr12:115980836G>A, c.5278C>T; p.Arg1760Ter	ID and distinctive facial features with or without cardiac defects (MRFACD)	Haploinsufficiency	Cannot exclude features of MRFACD.
8055	*MMUT*	chr6:49444675T>C, c.1640A>G; p.Asn547Ser	Methylmalonic aciduria, mut(0) type	Biallelic loss	Lacks features of methylmalonic aciduria
8012	*NCKAP1L*	chr12:54499380C>T, c.128C>T; p.Pro43Leu	Immunodeficiency 72 with autoinflammation	Biallelic loss	Lacks features of primary immunodeficiency
7859	*VWA8*	chr13:41729591TAG>T, c.2587_2588delCT; p.Leu863Sfs*20	Retinitis Pigmentosa 97(?)	Heterozygous missense	Lacks features of retinitis pigmentosa
7821	*SPATA16*	chr3:173019534C>T, c.800G>A; p.Arg267His	Spermatogenic failure 6 (?)	Biallelic loss	Unable to assess to due female sex
7749	*SDHAF1*	chr19:35995381TC>T, c.108delC; stop lost	Mitochondrial complex II deficiency, nuclear type 2	Biallelic loss	Lacks features of mitochondrial disease
7809	*MYO6*	chr6:75857131C>T, c.1258C>T; p.Arg420Cys	Deafness, nonsyndromic	Heterozygous missense	No evidence of deafness
7899	*MAP2*	chr2:209710227C>A, c.978C>A; Asn326Lys	Rett-like s. (preserved speech variant)	2q Contiguous gene deletion s.	Lacks Rett s.-like features
7886	*ITPR3*	chr6:33686219C>T, c.5834C>T; p.Thr1945Ile	Charcot-Marie-Tooth disease, demyelinating, type 1J	Heterozygous missense	Lacks features of Charcot-Marie-Tooth disease
7922	*GMPPB*	chr3:49723293G>A, c.220C>T;	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A14	Biallelic loss	Excluded muscular dystrophy

		p.Arg74Ter			
7899	*CAPN3*	chr15:42411343G>C, c.2419G>C; p.Glu807Gln	Muscular dystrophy, limb-girdle, autosomal dominant 4	Heterozygous deletion	Excluded muscular dystrophy
7918	*BSG*	chr19:581510G>C, c.361G>C; pGly214Arg	Blood Group OK	Heterozygous missense	Not tested
7934	*ASH1L*	chr1:155479920G>A, c.2950C>T; p.Arg984Cys	Intellectual developmental disorder, autosomal dominant 52	Heterozygous missense	Cannot exclude intellectual disability
7803	*ADGRV1*	chr5:90811123C>A, c.2846C>A; p.Ala949Glu	Usher s: hearing loss and retinitis pigmentosa	Heterozygous loss (Febrile seizures)?, Biallelic loss (Usher s.)	No evidence of hearing or vision loss
7889	*CUL3*	chr2:224478199C>T, c.2175+1G>A; splice site lost	Pseudohypoaldosteronism, type IIE Neurodevelopmental disorder	Splice site variants (pseudohypoaldosteronism); Haploinsufficeincy (Neurodevelopmental disorder)	Cannot exclude neurodevelopmental disorder
7901	*CNTNAP1*	chr17:42690144T>G, c.1792T>G; p.Phe598CVal	Hypomyelinating neuropathy, congenital, 3; Lethal congenital contracture s. 7	Biallelic loss	Lacks features of hypomyelinating neuropathy or contractures
7905	*DPYSL2*	chr8:26644084A>C, c.1103A>C; p.Lys332Thr	Intellectual disability and dysmorphic features(?)	Heterozygous missense	Cannot exclude evidence for intellectual disability or dysmorphic features
7918	*SCRIB*	chr8:143803792C>T, c.3026G>A; p.Arg1091Gln	Craniorachischisis	Haploinsufficiency	Lacks features of craniorachischisis
7944	*EP300*	chr22:41178614TTCTC>T, c.6904_6907delTCTC;p.Leu2303Profs*49	Menke-Hennekam s. 2, Rubinstein-Taybi s. 2	Heterozygous missense	Excluded Menke-Hennekam or Rubinstein-Taybi s.
7540	*DHCR7*	chr11:71435428A>G, c.1375T>C; p.Trp427Arg	Smith-Lemli-Opitz s.	Biallelic loss	Lacks Smith-Lemli-Opitz s. features
6293	*TTC26*	chr7:139189388C>T, c.1228C>T; p.Arg410Trp,	BIliary, renal, neurologic, and skeletal s.	Biallelic loss	Lacks biliary or renal features
7542	*PLG*	chr6:160722467C>T, c.1156C>T; p.Arg386Ter	Angioedema, dysplasminogemia, plasminogen deficiency	Heterozygous gain of function or biallelic loss	Lacks features of angioedema
7525	*PCK2*	chr14:24103719C>T, c.1276C>T; p.Arg426Trp	PEPCK deficiency, mitochondrial	Biallelic loss	Lacks features of hypoglycemia or liver
7546	*LAMA3*	chr18:23912880A>T, c.3923A>T; p.Asp1308Val	Epidermolysis bullosa	Biallelic loss	Lacks features of epidermolysis bullosa
5686	*HECTD4*	chr12:112179032:G>C, c.11256C>G; p.Tyr3752Ter	Neurodevelopmental disorder with seizures, spasticity, and complete or partial agenesis of the corpus callosum (NEDSSCC s.)	Biallelic loss	Excluded NEDSSCC s.
5801	*TBXAS1*	chr7:140015737C>A, c.1244C>A; p.Ala415Glu	Ghosal hematodiaphyseal s.	biallelic loss	Excluded Ghosal hematodiaphyseal s.
5343	*PAX3*	chr2:222232203G>A, c.667C>T; p.Arg223Ter	Waardenburg s.	Haploinsufficiency	Meets clinical criteria for Waardenburg s.
5655	*MYBPC3*	chr11:47349785G>A, c.643C>T;	Cardiomyopathy, dilated, 1MM Cardiomyopathy, hypertrophic,	Haploinsufficiency or biallelic loss	Excluded heart defect

		p.Arg215Cys	4 Left ventricular noncompaction 10		
5560	*GRIA3*	chrX:123417592T>C, c.1691T>C; p.Ile564Thr	Intellectual developmental disorder, X-linked s., Wu type	Hemizygous mutations	Cannot exclude intellectual disability
5462	*CYB561*	chr17:63438215GT>G, c.13delA; Thr5Glnfs*8	Orthostatic hypotension 2	Biallelic loss	Lacks evidence for orthostasis
7218	*ZSWIM6*	chr5:61332835C>A, c.563C>A; p.Ala188Asp	Acromelic frontonasal dysostosis or NEDMAGA	Heterozygous p.R1163W or p.R913X missense	Excluded acromelic frontonasal dysostosis
6924	*VWA8*	chr13:41887322G>C, c.691C>G; p.Arg231Gly	Retinitis pigmentosa (?)	Heterozygous missense	Excluded retinitis pigmentosa
7164	*VPS50*	chr7:93291790G>T, c.1030G>T; p.Glu344Ter	Neurodevelopmental disorder with microcephaly, seizures, and neonatal cholestasis	Biallelic loss	Excluded cholestasis
7073	*VPS13C*	chr15:61946374G>A, c.794C>T; p.Thr265Ile	Parkinson disease 23, autosomal recessive, early onset	Biallelic loss	Excluded Parkinson disease
7226	*TIAM1*	chr21:31146924T>G, c.545A>C; p.His182Pro	Neurodevelopmental disorder with language delay and seizures (NEDLDS)	Biallelic loss	Cannot exclude neurodevelopmental delay
6972	*SPEN*	chr1:15928231AC>A, c.1992delC; p.Gln665Lysfs*144	Radio-tartalgia s.	Haploinsufficiency	Cannot exclude Radio-tartalgia s.
7227	*POGZ*	chr1:151406022G>A, c.2827C>T; p.Arg943Cys	White-Sutton s.	Haploinsufficiency	Cannot exclude features of White-Sutton s.
7158	*PEX3*	chr6:143489224T>C, c.1120T>C; p.*374Arg (Loss of stop codon)	Peroxisome biogensis disorder (Zellweger s.)	Biallelic loss	Excluded Zellweger s.
7223	*KDM5B*	chr1:202749126C>T, c.1700G>A; p.Arg567Gln	Intellectual developmental disorder, autosomal recessive 65	Biallelic loss	Cannot exclude intellectual disability
7013	*IRS1*	chr2:226796947G>GC, c.1791_1792insG;p.His598Alafs*13	Type 2 diabetes mellitus, susceptibility to	Haploinsufficiency	Excluded type 2 diabetes
7217	*IRS1*	chr2:226796947G>GC, c.1791_1792insG;p.His598Alafs*13	Type 2 diabetes mellitus, susceptibility to	Haploinsufficiency	Excluded type 2 diabetes
7191	*GRHL2*	chr8:101619557T>C; c.1448T>C; p.Cys373Arg	Corneal dystrophy, Deafness, autosomal dominant 28, Ectodermal dysplasia/short	Heterozygous (corneal dystrophy, deafness); homozygous (ectodermal	Excluded corneal dystrophy, deafness or ectodermal dysplasia

			stature s.	dysplasia)	
7119	*FREM1*	chr9:14801790T>G, c.3556A>C; p.Ser1186Arg	Bifid nose with or without anorectal and renal anomalies; Manitoba oculotrichoanal s.; Trigonocephaly type 2	Biallelic loss; Haploinsufficeincy; (Tigonocephaly only)	Cannot exclude trigonocephaly
6985	*FEM1C*	chr5:115524327A>G, c.1835T>C; p.Phe612Ser	Developmental delay, pyramidal signs, and limb ataxia (?)	Heterozygous missense	Cannot exclude developmental delay
7202	*DCX*	chrX:111401096T>C, c.599A>G; p.Asn200Ser	Subcortical laminar heterotopia	Heterozygous missense	Cannot exclude features of subcortical laminar heterotopia
7002	*DAG1*	chr3:49531270G>GA, c.759_760insA; p.Asn254Lysfs*20	Muscular dystrophy-dystroglycanopathy	Biallelic loss	Lacks features of muscular dystrophy
7071	*CLTC*	chr17:59666977G>A, c.2140G>A; p.Gly714Ser	Intellectual developmental disorder, autosomal dominant 56	Haploinsufficiency	Cannot exclude intellectual disability
6966	*BRAT1*	chr7:2543876C>T, c.369G>A; p.Trp123Ter	Neurodevelopmental disorder with cerebellar atrophy and with or without seizures; rigidity and multifocal seizure s.	Biallelic loss	Lacks features of cerebellar atrophy or seizures
6950	*ADRA2B*	chr2:96115933C>G, c.217G>C; p.Glu73Gln	Epilepsy, familial adult myoclonic	Heterozygous in frame insertion/deletion	Cannot exclude myoclonic epilepsy
7123	*LEMD3*	chr12:65169828GCCGCGGGACCAGC>G, c.233_245delCCGCGGGACCAGC; p.Arg78Glyfs*101	Buschke-Ollendorff s. Osteopoikilosis with or without melorheostosis	Heterozygous deletion/insertion, duplication, frameshift, missense	Cannot exclude Buschke-Ollendorff s.
7155	*TBXT*	chr6:166160836, c.1034+1G>C Loss of splice site	Sacral agenesis with vertebral anomalies	Biallelic loss	Lacks features of sacral agenesis
7157	*UBIAD1*	chr1:11285729C>G, c.258C>G; p.Ile86Met	Corneal dystrophy, Schnyder type	Haploinsufficiency	Lacks features of corneal dystrophy
5943	*SQOR*	chr15:45661973A>G, c.253A>G; p.Ile85Val	Sulfide quinone oxidoreductase deficiency	Biallelic loss	Lacks features of sulfide:quinone oxidoreductase deficiency
5980	*SEC31A*	chr4:82854963:TC>T, c.1146delG; p.Trp382Ter	Halperin-Birk s.	Biallelic loss	Excluded Halperin-Birk s.
6686	*KRT83*	chr12:52314788C>A, c.1325G>T; p.Cys442Phe	Erythrokeratodermia variabilis et progressiva 5	Biallelic loss (Erythrokeratodermia); Haploinsufficiency (Monilethrix)	Lacks skin defects

6692	*HR*	chr8:22120753C>A, c.2573G>T; p.Gly858Val	Alopecia universalis Atrichia with papular lesions	Biallelic loss	Lacks features of alopecia universalis
6707	*DNAH11*	chr7:21801213A>C, c.10310A>C; p.Glu3368Ala	Ciliary dyskinesia, primary, 7, with or without situs inversus	Biallelic loss	Lacks features of ciliary dyskinesia or situs inversus
8867	*PREPL*	chr2:44339179G>C , c.670C>G; p.Leu224Val	Myasthenic s, congenital, 22	Biallelic loss	Lacks features of myasthenia
8828	*PAX3*	chr2:222297081G >T, c.218C>A; p.Ser73Ter	Waardenburg s.	Haploinsufficiency	Meets criteria for Waardenburg s.
8828	*INF2*	chr14:104713537 C>T, c.2971C>T, p.Arg991Trp	Charcot-Marie-Tooth disease, dominant intermediate E with Glomerulosclerosis, focal segmental, 5	Haploinsufficiency	Excluded Charcot-Marie-Tooth or glomerulosclerosis
8861	*GRHL3*	chr1:24342238, c.892C>T; p.Arg298Cys	van der Woude s. type 2	Haploinsufficiency	Cannot exclude van der Woude s. type 2

Supplementary Table 1. Phenotypic expansion of MM genes with DNM displaying syndrome and/or non syndrome disease features other than MM. Column 1: Family number, Column 2: mutated gene, Column3: Mutation identified for that gene and patient in our cohort, Column 4: Relevant OMIM entry for each gene. Column 5: type of mutation reported in OMIM leading to phenotype. Column 6: Evidence for or against concordance with the OMIM phenotype. Genomic coordinates are in hg38. s.: syndrome, question mark (?): association remains uncertain according to OMIM.

Supplementary Notes

Power calculations for effect size

To determine the adequate cohort size for identification of recurrently mutated genes in MM, we developed a computational model and power calculations to estimate effect size[77,78]. We first determined that with a sample size of 350 trios, observing such recurrence in two families would be significant genome-wide, with *P* value = 0.003, and from this calculated the false-discovery rate (FDR), assuming a baseline DNM rate among the ~20,000 human genes (**Extended Data Fig. 1** gray dashed line). Then, leveraging data from a small NTD WES study in which DNM LOF mutations were identified in 20% of subjects, 34, and estimating that about 0.4% (i.e. 80/20,000) of gene knockouts in mouse show a MM-like phenotype, we estimate that there are 50–100 MM disease genes (*k*) to be discovered. Borrowing from previous literature [13,21,79], we anticipate that LOF variants will be enriched in patients vs. controls with a v ratio between 1.5 and 2.5. We ran 10,000 iterations to estimate the number of genes we will discover for various cohort sizes ranging from 100–1000 trios (**Extended Data Fig. 1**). For instance, with a cohort size of 400, if there are 50 genes to discover (i.e. $k = 50$) and with 2.5 times as many LOF variants in patients vs. control (i.e. $v = 2.5$), we expect ~16 recurrently mutated genes to emerge. For the same cohort, if $k = 100$ and $v = 1.5$, we expect to discover ~2 recurrently mutated genes. Given the recurrence found in the cohort of 43 trios [15], and our preliminary data, we anticipated our results will fall between these two scenarios. Notably, this model predicts that with 800 MM trios sequenced as we propose here, we expect to discover ~ 5–10 recurrently mutated MM genes.

Reproducible prediction of D-Mis-HC

To validate the high confidence damaging missense variant (D-Mis-HC) that was originally called by MetaSVM, we also applied MetaLR and REVEL to the DNMs. The D-Mis-HC was highly reproducible with the two additional annotation tools with recommended thresholds, MetaLR score > 0.5 and REVEL > 0.75 (**Extended Data Fig. 16c-16d**).

Parents' age of conception comparison

To confirm if the excessive DNM burden in MM compared to controls are independent from the ages of parents at the time of conception, we gathered dates of births from probands, fathers, and mothers. Trios that were able to track all information from three family members (MM: 79, control: 683) were collected, to compare the ages of conception of father and mother, with Wilcoxon rank-sum test. (two-sided). We could clearly confirm that both the ages of mothers and fathers in the MM cohort were significantly lower than those of controls (**Extended Data Fig. 2a**). Although the rate of de novo variants is known to be highly correlated with the age of parents, the excessive burden observed in our study is not derived from the difference of fathers' and mothers' age distribution.

WGS evaluation for DNMs contributing to MM risk

From 101 additional trios and 1 quartet, we observed comparable de novo SNV and Indel rates with WGS data (**Extended Data Table 1**). Among them, 13 LGD and 39 D-Mis

coding DNMs were observed. Also, 9 splicing disrupting variants (SpliceAI maximum delta score larger than 0.9) and 41 possible pathogenic noncoding DNMs (GREEN-VARAN level 4) were collected. However, there was no recurrent gene mutated within the WGS cohorts (see **Supplementary Data** for the full list of DNMs in WGS). Instead, three overlapped with damaging (LGD+D-Mis) DNMs with WES data were revealed: a stop-gain in KDM1A, a splice donor gain in ITPR3, and a variant in GRHL2 enhancer. We could find four de novo CNVs (> 100 kb) in WGS trios, four de novo deletions and duplications were observed (**Extended Data Table 2**). The large CNVs were confirmed with the parents' sequence data (**Extended Data Fig. 3b**) and the duplicated or deleted regions were not recurrent. No de novo translocation or inversion was observed.

Multiplexed Error Robust In Situ Hybridization (MERFISH) spatial transcriptomic expression pattern of a subset of DNM genes

We conducted Multiplexed Error Robust Fluorescent In Situ Hybridization (MERFISH) on the Vizgen platform using the methodology as described in Methods[64]. We segmented a total of 352,729 cells with uniform transcript density using cell segmentation. A subset of cells was excluded due to outlier cell volume (n = 9,922) or transcript density (n = 21,319), with the remaining cells divided into 63 distinct clusters based on gene expression using the Leiden algorithm, stratifying into 7 major cell types (Neuron, Neural crest, Neural Progenitor, Dorsal Root Ganglia, Pre epithelial to mesenchymal transition neural Crest progenitor, Mesoderm and Blood) in mouse E9.5 parasagittal sections. Cell types were defined by expression of marker genes selected based on previously published single cell transcriptomic studies[65-67], employing ScType[80] for unbiased cell typing from marker gene expression. Approximately 14% cells were flagged as 'indeterminate' type that could not be clearly categorized, likely due either to a state of transitioning between cell types or lacking expression of marker genes. We selected 36 DNM genes identified in MM subjects for multiplex spatial co-expression analysis (**Extended Data Fig. 4)**. Percentage of gene expression vs cell type revealed that most of the candidate DNM genes show expression in multiple cell types during neural tube closure. There were some genes like *Rnd2* and *Add2* that were expressed predominantly in neural cells, *Celsr1* in neural progenitor, *Stab1* in pre-EMT-neural precursor cells, *Opalin* expressed in predominantly in blood (**Extended Data Fig. 5)**.

Additional evidence for role of functionally implicated DNM genes in NTDs

The DNM genes in which the DNM mutation impairs protein function showed additional prior evidence that may suggest a role in human NTDs. TIAM1, which promotes actin assembly via the Arp2/3 complex[81], is essential for neural tube closure in mice[82], but not previously implicated in human NTD. TNK2, also known as ACK1 is a kinase associated with CDC42 that controls cellular protrusions associated with spinal neurulation in mouse[32]. VWA8 is a mitochondrial matrix-targeting ATPase, and notably mitochondrial dysfunction is implicated in both genetic and environmental murine models of NTD[49,83].

Clinical phenotypes of patients compared to OMIM

We assessed possible connections between DNM genes and associated OMIM phenotypes by recontacting subjects to inquire about clinical features listed in OMIM but possibly not described at the time of enrollment into our study. Resources used were

Online Mendelian Inheritance in Man (OMIM), ClinVar, and PubMed. Out of the 192 DNMs found in our cohort 82 were found to have other OMIM phenotypes (43.8%). Genes found to have other OMIM phenotypes in individual families/subjects were further reviewed regarding the type of mutation compared with those reported in the clinical literature and compared with clinical status (**Supplementary Table 1**). Information available at the time of enrollment was used for clinical correlation along with follow up with the families via telephone or zoom to specifically inquire about features related to the published OMIM phenotypes. For some subjects like 6695, the de novo mutation occurred in a gene showing only recessive inheritance (*TIMM50*), so it was not surprising that the MM subject carrying a DNM in the gene did not report features of the OMIM entry. But for other subjects like 6313, the DNM occurred in a gene with the same or similar zygosity, and the patient showed features that cannot be distinguished from the reported OMIM entry, in this case Coffin-Siris features (i.e. developmental disability, abnormalities of the fifth (pinky) fingers or toes, and characteristic facial features), suggesting MM may be a phenotypic expansion of this syndrome.

Estimation of number of genes in DNM contributing to MM risk

The estimation was based on previous publication[14]. Briefly, the total number of risk-associated genes (C) is calculated with the number of observed risk-associated mutations (d = 192) and recurrent mutations (r = 5, *PAX3*, *IRS1*, *ZSWIM6*, *BRSK2*, and *VWA8*). Since 28.26% of the damaging DNMs (LGD + D-Mis) are contributing to the MM risk, d and r were normalized as 54.26 and 1.41. The number of observed risk-associated genes is c and c_1 refers to the number of genes mutated once and the probability of a newly added mutation hitting a previously mutated gene (μ) could be calculated as follows. Considering the variation in effect size of individual DNMs are assumed to be 1 to minimize underestimation of set size, C could be estimated as 2,021 genes.

$$c = d - r$$

$$c_1 = d - (2 \times r)$$

$$\mu = 1 - c_1/d$$

$$C = \frac{C}{\mu} + (g^2 \times d \times (1 - \mu)/\mu$$