The American Journal of Human Genetics, Volume *101*

Supplemental Data

Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes

Stefan H. Lelieveld, Laurens Wiel, Hanka Venselaar, Rolph Pfundt, Gerrit Vriend, Joris A. Veltman, Han G. Brunner, Lisenka E.L.M. Vissers, and Christian Gilissen

Supplemental Figures

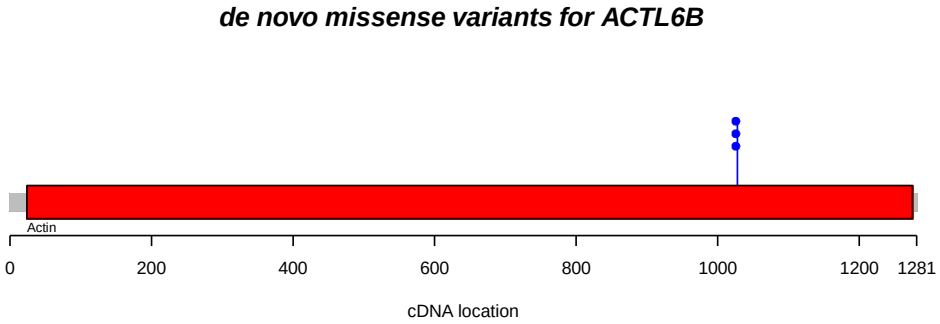

Figure S1 Schematic representation of *ACTL6B* (ENST00000160382). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

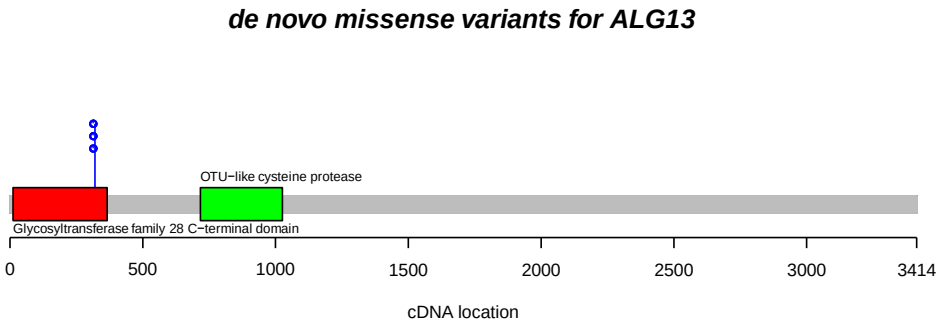

Figure S2 Schematic representation of *ALG13* (ENST00000394780). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

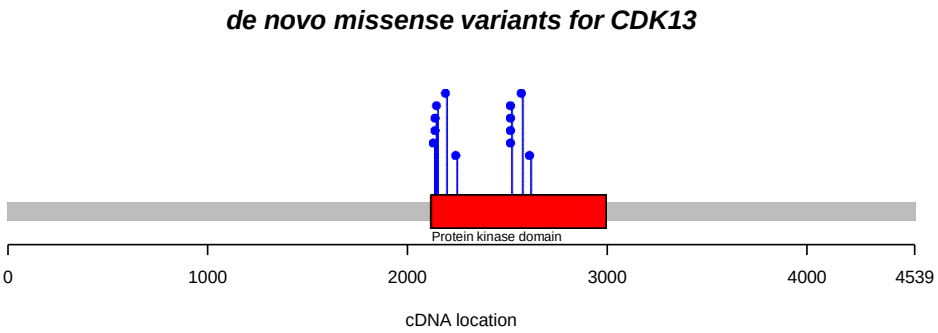

Figure S3 Schematic representation of *CDK13* (ENST00000181839). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

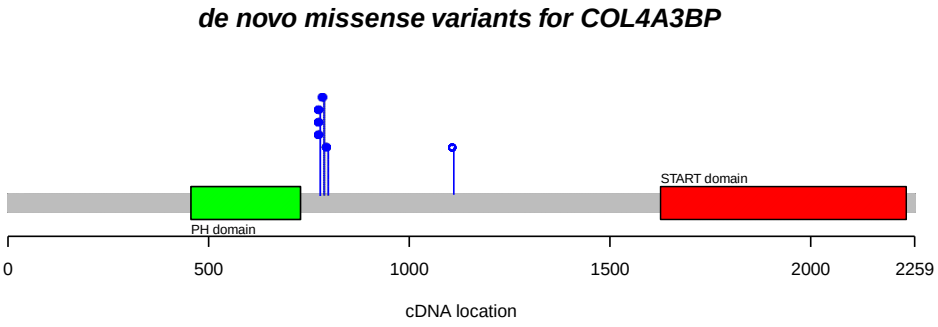

Figure S4 Schematic representation of *COL4A3BP* (ENST00000380494). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

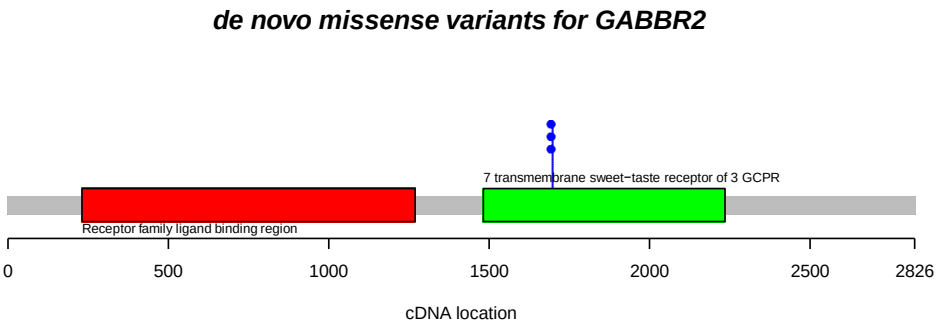

Figure S5 Schematic representation of *GABBR2* (ENST00000259455). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

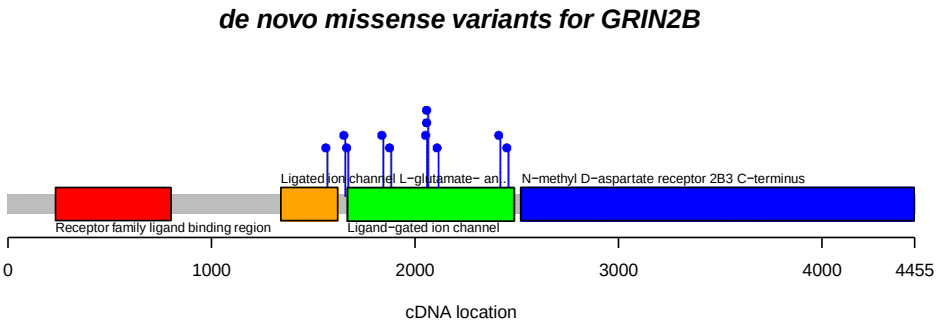

Figure S6 Schematic representation of *GRIN2B* (ENST00000609686). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

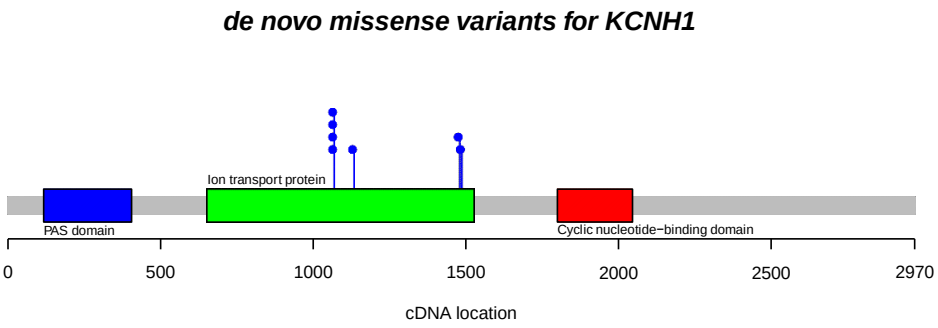

Figure S7 Schematic representation of *KCNH1* (ENST00000271751). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

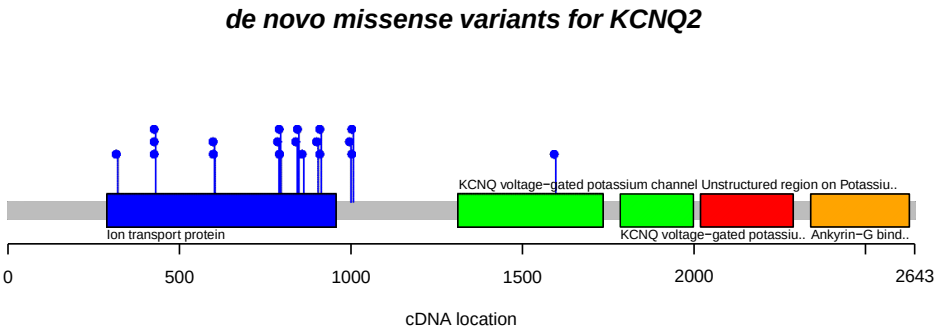

Figure S8 Schematic representation of *KCNQ2* (ENST00000354587). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

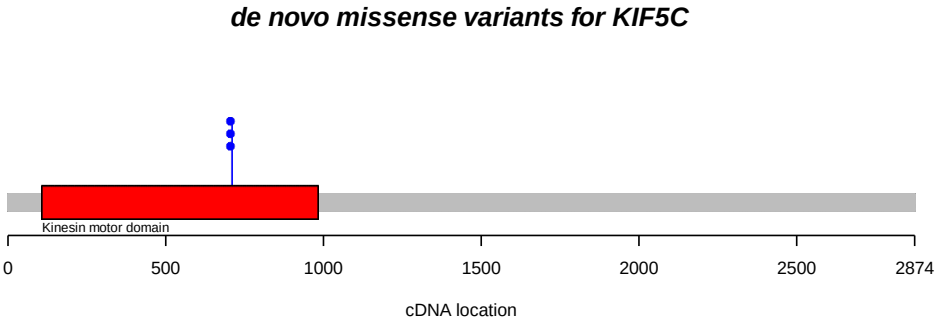

Figure S9 Schematic representation of *KIF5C* (ENST00000435030). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

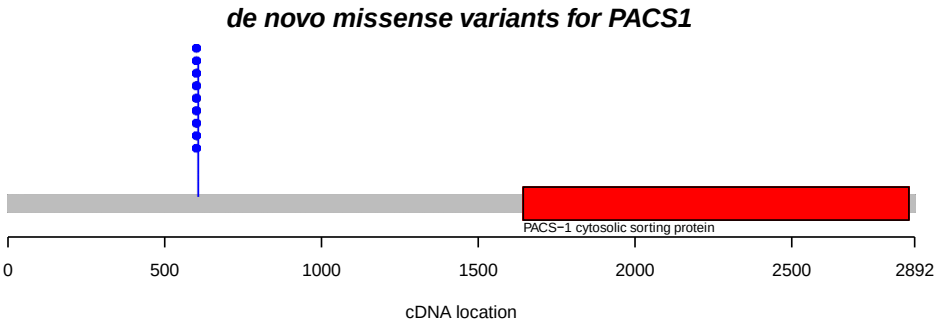

Figure S10 Schematic representation of *PACS1* (ENST00000320580). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

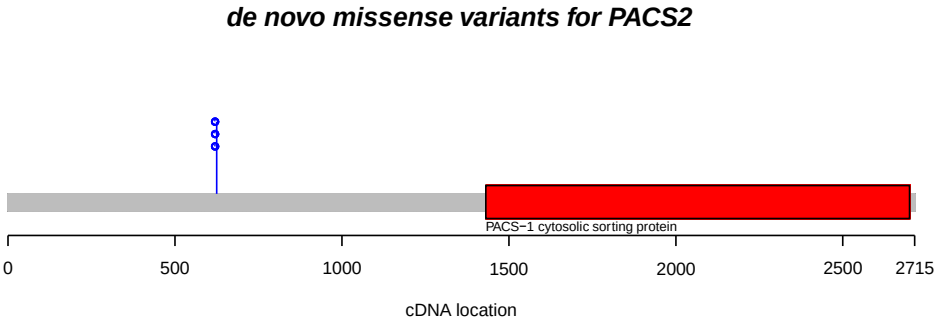

Figure S11 Schematic representation of *PACS2* (ENST00000458164). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

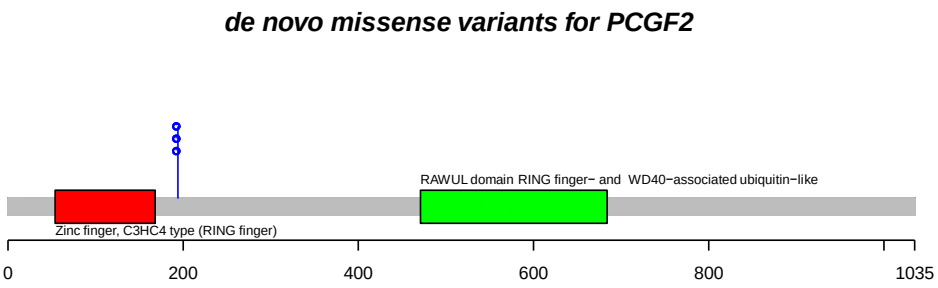

Figure S12 Schematic representation of *PCGF2* (ENST00000360797). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

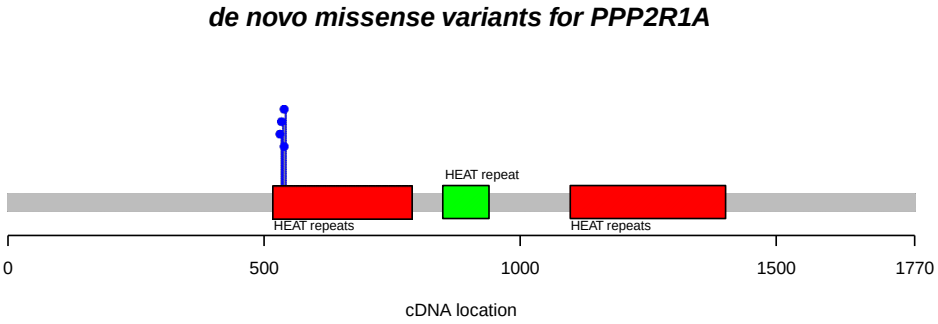

Figure S13 Schematic representation of *PPP2R1A* (ENST00000322088). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

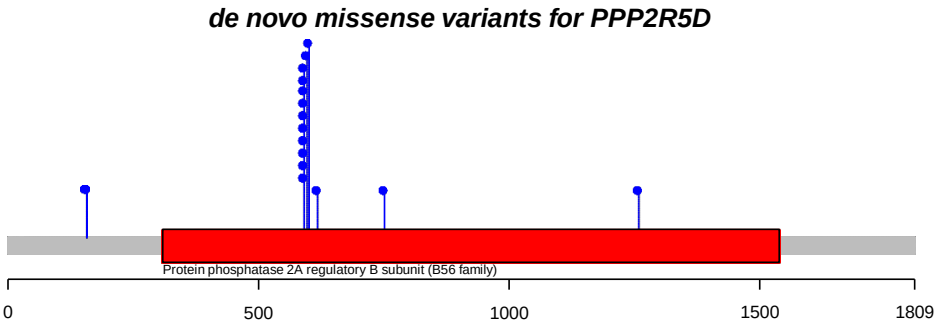

Figure S14 Schematic representation of *PPP2R5D* (ENST00000485511). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

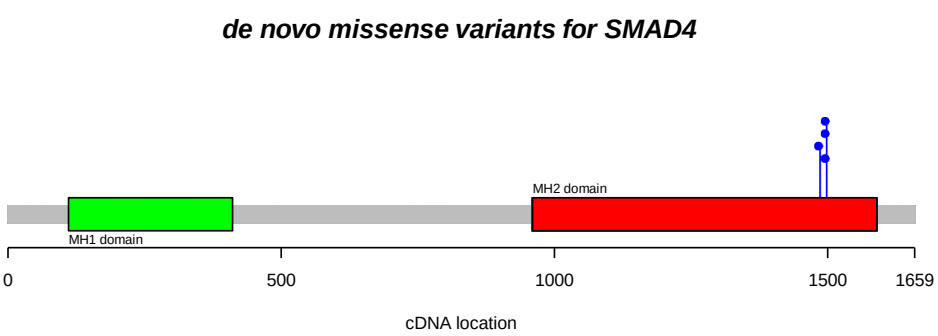

Figure S15 Schematic representation of *SMAD4* (ENST00000398417). The locations of the missense de novo variants are indicated by blue pins. Recurrent de novo missense mutations are indicated by stacked blue pins. Protein domains are annotated based on Pfam HMM search[1] and are illustrated by colored squares.

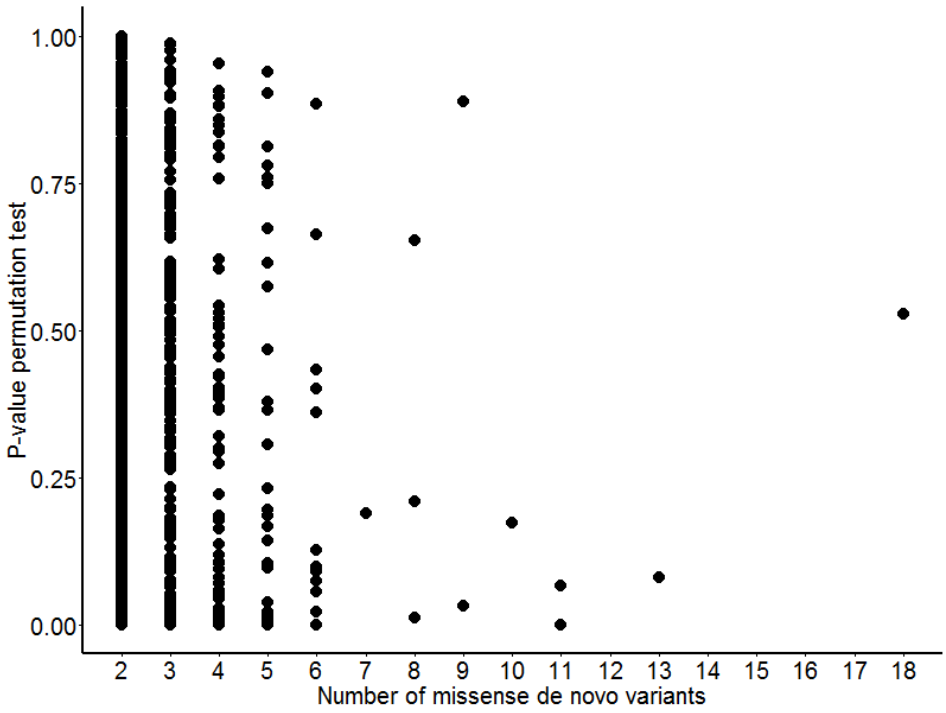

Figure S16. number of missense variants and the extend of clustering. We have studied the possibility of a correlation between the number of missense variants and the extend of clustering. Hereto, we have analysed a much larger set of 6,154 de novo missense variants from various patient cohorts present in the denovo-db version 1.3[2] (other than the five selected ID/DD cohorts presented in this manuscript). The variants were filtered as described in Table S1. Correlation analysis (Kendall rank correlation coefficient) between the number of de novo missense variants per gene (x-axis) and the corresponding p-values based on the spatial clustering signal (y-axis) yielded a correlation coefficient of -0.04 indicating there is no correlation.

Supplemental Tables

Column name	Number of variants remaining	
	ID and DD set	Control set
Unfiltered	9,770	35,632
Variants in coding regions and canonical splice-sites (±2 bp)	9,202	2,749
ExAC number of heterozygous variants less than 2 and no homozygous variants.	6,495	1,961
Variant identified in exome of genome sequencing study	6,495	1,948

Table S1. Filtering of de novo mutations used in study. Overview of filters that were applied to specific columns of the annotated de novo variants used in the analysis (left column). The middle and right columns indicate the number of variants left after the filtering for the ID and DD set and control set respectively.

Study	Trios	Disorder	Coding	Missense	LoF
McRae *et al.*[3]	4,293	DD	5,375	3,375	961
Lelieveld *et al.* [4a]	820	ID	948	579	177
de Ligt *et al.*[5]	100	ID	56	35	10
Rauch *et al.*[6]	51	ID	90	54	31
Halvardson *et al.*[7]	38	ID/EE	26	18	4
Total	5,302		6,495	4,061	1,183

Table S2. Dataset composition of intellectual disability and developmental disorders cohort (ID + DD). Columns indicate (from left to right), the study reference, number of trios included from the study, the disorder that was studied (DD= Developmental disorders, ID = Intellectual disability and EE = Epileptic Encephalopathies), and the number of coding, missense and loss-of-function de novo mutations after filtering. [a]Including 100 de novo mutations (median GATK quality score of 241) not included in the original publication.

Study	Trios	Disorder	Coding	Missense	LoF
Iossifov *et al.*[8]	1.786	Sibs. (ASD)	1,267	805	139
Krumm *et al.*[9a]			323	220	30
GONL[10]	250	Control	102	67	5
Turner *et al.*[11]	43	Sibs. (ASD)	28	18	1
Gulsuner *et al.*[12]	84	Sibs. (SCZ)	50	28	9
Besenbacher *et al.*[13b]	283	Control	178	135	15
Total	2,448		1,948	1,273	199

Table S3. Dataset composition of cohort of healthy controls (Control). Columns (from left to right) indicate the reference to the studies, number of trios that were included from the study, the disorder that was studied (Sibs. = siblings, ASD = Autism spectrum disorder and SCZ = Schizophrenia), the number of coding, missense and loss-of-function de novo mutations that were included from the study. [a]The study by Krumm *et al*[9]*.* performed a re-analysis on existing data of Iossifov *et al.*[8] [b]Variants in the study of Besenbacher *et al.*[13] are annotated to reference genome HG18. We used LiftOver to convert the HG18 coordinates to HG19 coordinates (See .web resources).

Gene name	Gene ID	Miss. DNMs	Avg. distance	p-value	Adj. p-value
SYNE1	ENST00000367255.5	2	36	0.0027	1
TMPRSS15	ENST00000284885.3	2	9	0.0061	1
MAPK8	ENST00000374189.1	2	6	0.0101	1
FAT4	ENST00000394329.3	2	86	0.0116	1
SPOCK3	ENST00000357154.3	2	14	0.0220	1
PRPS2	ENST00000398491.2	2	15	0.0299	1

Table S4. Results from clustering analysis on the control cohort. Columns from left to right indicate the gene name, used GENCODE gene identifier, number of de novo missense variants, average distance between de novo variants, permutations test based p-value and Bonferroni corrected p-value. None of the genes reach statistical significance after multiple testing correction.

Table S5. List of known genes involved in intellectual disability and developmental disorders

External file: table_S5_known_DD_genes.xls

Table S6. List of identified clustering mutations

External file: table_S6_dnm_clustering_genes.xls

	Genes with sig. cluster	Other genes	**total**
Known ID/DD genes	12	187	199
Other genes	3	381	384
total	15	568	583

Table S7A. Enrichment of genes have previously been implicated in ID/DD.

Data used to calculate statistical enrichment of genes associated to ID/DD based on the associated ID/DD genes on the combined list of the DDG2P and RUMC. The dataset contained in total 583 genes that were recurrently mutated of which 15 genes contained clustering de novo variants. 13 of the 15 genes were associated to ID/DD. Enrichment was tested with a two-sided Fisher's Exact test and yielded a significant p-value of 3.09E-04.

	Genes with sig. cluster	Other genes	**total**
Known ID/DD genes	11	162	173
Other genes	4	406	410
total	15	568	583

Table S7B. Enrichment of genes have previously been implicated in ID/DD solely on RUMC diagnostis list.

Data used to calculate statistical enrichment of genes associated to ID/DD based on the associated ID/DD genes solely on the RUMC list (see web resources). Enrichment was tested with a two-sided Fisher's Exact test and yielded a significant p-value of 5.13E-04.

	Genes with sig. cluster	Other genes	**total**
Known ID/DD genes	8	153	161
Other genes	7	415	422
total	15	568	583

Table S7C. Enrichment of genes have previously been implicated in ID/DD excluding two large DDD exome studies.

To completely remove the weight of the DDD exomes we have extended the enrichment test by excluding a set of genes based on the following criteria:

1. Novel genes enriched for de novo mutations found in the two large scale exome data studies of the DDD[3; 14]
2. Complement to the genes found enriched for de novo variants in Lelieveld *et al.* [4]

Analysis via a two-sided Fisher's Exact test yielded a significant P-value of 3.68E-02

Gene name	De novo missense	p-value	Adj. p-value
ACTL6B	3	8.48E-04	1
ALG13	3	3.66E-03	1
CDK13	12	4.34E-14	7.93E-10
COL4A3BP	6	1.28E-07	2.33E-03
GABBR2	3	6.82E-03	1
GRIN2B	11	4.22E-11	7.71E-07
KCNH1	7	1.36E-07	2.49E-03
KCNQ2	20	3.60E-28	6.59E-24
KIF5C	3	3.22E-03	1
PACS1	9	1.95E-10	3.57E-06
PACS2	3	5.88E-03	1
PCGF2	3	2.99E-04	1
PPP2R1A	4	9.88E-05	1
PPP2R5D	16	2.89E-24	5.28E-20
SMAD4	4	4.40E-05	8.04E-01

Table S8. Statistical significance of enrichment for de novo mutations for all identified genes. P-values are calculated based on Gene Specific Mutation Rates from Samocha *et al.*[15] and corrected for multiple testing by the Bonferroni correction (for 19,280 genes). Only some of the previously known genes reached statistical significance. Genes identified by our clustering analysis would not have been identified as candidate genes based on a statistical approach for the enrichment of de novo mutations.

Table S9. Known functions of identified genes

External file: table_S9_known_function_of_identified_genes.xlsx

Table S10: Overview of de novo variants found per individual

External file: table_S10_overview_all_DNM_per_sample_clustering_genes.xlsx

Table S11. Overview of literature supporting a NHI mechanism

External file: table_S11_overview_literature_study_known_NHI_genes.xlsx

Table S12. List of non-haploinsufficient genes involved in intellectual disability and developmental disorders

External file: table_S12_List_of_NHI_genes_involved_in_id_dd.xlsx

Table S13. List of haploinsufficient genes involved in intellectual disability and developmental disorders

External file: table_S13_List_of_HI_genes_involved_in_id_dd.xlsx

Gene name	Allelic requirement	mutation consequence
ALG13[a]	x-linked dominant	all missense/in frame
CDK13	monoallelic	all missense/in frame
COL4A3BP	monoallelic	activating
GRIN2B[a]	monoallelic	loss of function / all missense/in frame
KCNH1	monoallelic	activating
KCNQ2	monoallelic	loss of function
KIF5C	monoallelic	all missense/in frame
PACS1	monoallelic	activating
PCGF2	monoallelic	activating
PPP2R1A	monoallelic	dominant negative
PPP2R5D	monoallelic	dominant negative
SMAD4[a]	monoallelic	loss of function / all missense/in frame

Table S14. Identified known genes and their mutational consequence. Allelic requirement and mutational consequence according to DDG2P. [a]In the statistical analysis we excluded genes for reasons mentioned in the main text .

	Hi genes	NHI genes	Total
Genes with significant cluster	1	8	9
Other genes	182	108	290
Total	183	116	299

Table S15. Enrichment of genes based on their mutational consequence. Statistical enrichment testing of disease mechanisms of the genes identified in this study. Of the 15 identified genes, 9 are annotated with a disease mechanism after filtering based on the DDG2P and RUMC lists. Analysis via Fisher's Exact test yielded a significant P-value of 2.66E-03.

Table S16. Tolerance scores based on dn/ds ratios for all human genes

External file: table_S16_Dn_ds_ratios_for_all_human_genes.xls

Table S17. HI and NHI DNMs analysed to protein structures

External file: table_S17_DNMs_in_protein_structures.xlsx

Supplemental web resources

DDG2P: https://www.ebi.ac.uk/gene2phenotype/downloads

RUMC Genome diagnostics gene list:

https://www2.radboudumc.nl/Informatievoorverwijzers/Genoomdiagnostiek/en/Pages/Intellectualdisability.aspx

Lift Over tool: https://genome.ucsc.edu/cgi-bin/hgLiftOver

Supplemental References

1. Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44, D279-285.
2. Turner, T.N., Yi, Q., Krumm, N., Huddleston, J., Hoekzema, K., HA, F.S., Doebley, A.L., Bernier, R.A., Nickerson, D.A., and Eichler, E.E. (2017). denovo-db: a compendium of human de novo variants. Nucleic Acids Res 45, D804-D811.
3. Deciphering Developmental Disorders Study. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature.
4. Lelieveld, S.H., Reijnders, M.R., Pfundt, R., Yntema, H.G., Kamsteeg, E.J., de Vries, P., de Vries, B.B., Willemsen, M.H., Kleefstra, T., Lohner, K., et al. (2016). Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nature neuroscience 19, 1194-1196.
5. de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. The New England journal of medicine 367, 1921-1929.
6. Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 380, 1674-1682.
7. Halvardson, J., Zhao, J.J., Zaghlool, A., Wentzel, C., Georgii-Hemming, P., Mansson, E., Ederth Savmarker, H., Brandberg, G., Soussi Zander, C., Thuresson, A.C., et al. (2016). Mutations in HECW2 are associated with intellectual disability and epilepsy. Journal of medical genetics 53, 697-704.
8. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014).

The contribution of de novo coding mutations to autism spectrum disorder. Nature 515, 216-221.
9. Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.X., et al. (2015). Excess of rare, inherited truncating mutations in autism. Nature genetics 47, 582-588.
10. Genome of the Netherlands, C. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nature genetics 46, 818-825.
11. Turner, T.N., Hormozdiari, F., Duyzend, M.H., McClymont, S.A., Hook, P.W., Iossifov, I., Raja, A., Baker, C., Hoekzema, K., Stessman, H.A., et al. (2016). Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. American journal of human genetics 98, 58-74.
12. Gulsuner, S., Walsh, T., Watts, A.C., Lee, M.K., Thornton, A.M., Casadei, S., Rippey, C., Shahin, H., Consortium on the Genetics of, S., Group, P.S., et al. (2013). Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell 154, 518-529.
13. Besenbacher, S., Sulem, P., Helgason, A., Helgason, H., Kristjansson, H., Jonasdottir, A., Jonasdottir, A., Magnusson, O.T., Thorsteinsdottir, U., Masson, G., et al. (2016). Multi-nucleotide de novo Mutations in Humans. PLoS genetics 12, e1006315.
14. Deciphering Developmental Disorders Study. (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature 519, 223-228.
15. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnstrom, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nature genetics 46, 944-950.