Scientific evaluation of negative exome sequencing followed by systematic scoring of candidate genes to decipher the genetics of neurodevelopmental disorders

Benjamin Büttner,[1] Sonja Martin,[1] Anja Finck,[1] Maria Arelin,[2] Carolin Baade-Büttner,[3] Tobias Bartolomaeus,[1] Peter Bauer,[4] Astrid Bertsche,[2] Matthias K. Bernhard,[2] Saskia Biskup,[5] Nataliya Di Donato,[6] Magdeldin Elgizouli,[7] Roland Ewald,[8] Constanze Heine,[1] Yorck Hellenbroich,[9] Julia Hentschel,[1] Sabine Hoffjan,[10] Susanne Horn,[1] Frauke Hornemann,[2] Dagmar Huhle,[11] Susanne B. Kamphausen,[12] Wieland Kiess,[2] Ilona Krey,[1] Alma Kuechler,[7] Ben Liesfeld,[8] Andreas Merkenschlager,[2] Diana Mitter,[1] Petra Muschke,[12] Roland Pfäffle,[2] Tilman Polster,[13] Ina Schanze,[12] Jan-Ulrich Schlump,[14] Steffen Syrbe,[15] Dagmar Wieczorek,[16] Martin Zenker,[12] Johannes R. Lemke,[1] Diana Le Duc,[1] Konrad Platzer,[1] Rami Abou Jamra*[1]

[1]Institute of Human Genetics, University of Leipzig Medical Center, Leipzig 04103, Germany

[2]Hospital for Children and Adolescents, University of Leipzig Medical Center, Leipzig 04103, Germany

[3]Department of Neurology, University of Leipzig Medical Center, Leipzig 04103, Germany

[4]Centogene AG, Rostock 18055, Germany

[5]CeGaT GmbH, Center for Genomics and Transcriptomics, Tübingen 72076, Germany

[6]Institute of Clinical Genetics, Technische Universität Dresden, Dresden 01307, Germany

[7]Institute of Human Genetics, University Hospital Essen, Essen 45122, Germany

[8]Limbus Medical Technologies GmbH, Rostock 18055, Germany

[9]Institute of Human Genetics, University Lübeck, Lübeck 23562, Germany

[10]Department of Human Genetics, Ruhr-University, Bochum 44801, Germany

[11]Praxis für Humangenetik Leipzig, Leipzig 04289, Germany

[12]Institute of Human Genetics, University Hospital Magdeburg, Magdeburg 39120, Germany

[13]Paediatric Epileptology, Mara Hospital gGmbH, Bethel Epilepsy Center, Bielefeld 33617, Germany

[14]Evangelical Hospital Oberhausen, Division for Children and Adolescents, Oberhausen 46047, Germany

[15]Division for Neuropaediatrics and Metabolic Medicine, Center for Paediatric and Adolescent Medicine, Heidelberg University Hospital, Heidelberg 69120, Germany

[16]Institute of Human Genetics, Medical Faculty, Heinrich-Heine-University Düsseldorf, Düsseldorf 40225, Germany

*Corresponding author

Electronic mail addresses

Benjamin Büttner: Benjamin.Buettner@medizin.uni-leipzig.de

Sonja Martin: sonja.martin1@gmx.de

Anja Finck: Anja.Finck@medizin.uni-leipzig.de

Maria Arelin: Maria.Arelin@medizin.uni-leipzig.de

Carolin Baade-Büttner: Carolin.Baade@medizin.uni-leipzig.de

Tobias Bartolomaeus: Tobias.Bartolomaeus@medizin.uni-leipzig.de

Peter Bauer: Peter.Bauer@centogene.com

Astrid Bertsche: astrid.bertsche@medizin.uni-leipzig.de

Matthias K. Bernhard: Matthias.Bernhard@medizin.uni-leipzig.de

Saskia Biskup: Saskia.Biskup@humangenetik-tuebingen.de

Nataliya Di Donato: Nataliya.DiDonato@uniklinikum-dresden.de

Magdeldin Elgizouli: Magdeldin.Elgizouli@uk-essen.de

Roland Ewald: roland.ewald@limbus-medtec.com

Constanze Heine: Constanze.Heine@medizin.uni-leipzig.de

Yorck Hellenbroich: Yorck.Hellenbroich@uksh.de

Julia Hentschel: Julia.Hentschel@medizin.uni-leipzig.de

Sabine Hoffjan: Sabine.Hoffjan@rub.de

Susanne Horn: Susanne.Horn@medizin.uni-leipzig.de

Frauke Hornemann: Frauke.Hornemann@medizin.uni-leipzig.de

Dagmar Huhle: dagmar.huhle@gmx.de

Susanne B. Kamphausen: susanne.kamphausen@med.ovgu.de

Wieland Kiess: Wieland.Kiess@medizin.uni-leipzig.de

Ilona Krey: Ilona.Krey@medizin.uni-leipzig.de

Alma Kuechler: alma.kuechler@uni-due.de

Ben Liesfeld: ben.liesfeld@limbus-medtec.com

Andreas Merkenschlager: Andreas.Merkenschlager@medizin.uni-leipzig.de

Diana Mitter: Diana.Mitter@medvz-leipzig.de

Petra Muschke: petra.muschke@med.ovgu.de

Roland Pfäffle: Roland.Pfaeffle@medizin.uni-leipzig.de

Tilman Polster: Tilman.Polster@mara.de

Ina Schanze: ina.schanze@med.ovgu.de

Jan-Ulrich Schlump: Jan-Ulrich.Schlump@eko.de

Steffen Syrbe: Steffen.Syrbe@med.uni-heidelberg.de

Dagmar Wieczorek: dagmar.wieczorek@uni-duesseldorf.de

Martin Zenker: martin.zenker@med.ovgu.de

Johannes R. Lemke: Johannes.Lemke@medizin.uni-leipzig.de

Diana Le Duc: Gabriela-Diana.LeDuc@medizin.uni-leipzig.de

Konrad Platzer: Konrad.Platzer@medizin.uni-leipzig.de

Rami Abou Jamra: Rami.AbouJamra@medizin.uni-leipzig.de

Abstract

Background

Deciphering the monogenetic causes of neurodevelopmental disorders (NDD) is an important milestone to offer personalized care. But the plausibility of reported candidate genes in exome studies often remains unclear, which slows down progress in the field.

Methods

We performed exome sequencing (ES) in 198 cases of NDD. Cases that remained unresolved (n=135) were re-investigated in a research setting. We established a candidate scoring system (CaSc) based on 12 different parameters reflecting variant and gene attributes as well as current literature to rank and prioritize candidate genes.

Results

In this cohort, we identified 158 candidate variants in 148 genes with CaSc ranging from 2 to 11.7. Only considering the top 15% of candidates, 14 genes were already published or funneled into promising validation studies.

Conclusions

We promote that in an approach of case by case re-evaluation of primarily negative ES, systematic and standardized scoring of candidate genes can and should be applied. This simple framework enables better comparison, prioritization, and communication of candidate genes within the scientific community. This would represent an enormous benefit if applied to the tens of thousands of negative ES performed in routine diagnostics worldwide and speed up deciphering the monogenetic causes of NDD.

98 **Key words**

99 Exome sequencing, NDD, mental retardation, Candidate gene, Scoring, intellectual disability

Background

The entirety of rare diseases represents a substantial burden and affects millions of individuals worldwide[1]. A large part of these diseases have a genetic cause[2], but many cases remain undiagnosed. A missing diagnosis may have a negative impact on the caring families, medical treatment, socio-economic expenses, and on prognosis[3].

This is most obvious in the genetic etiologies of neurodevelopmental disorders (NDD). Exome sequencing (ES) has proven to be highly successful in identifying causes of NDD[4, 5]. Thus, large efforts have been undertaken in order to decipher the genetics of NDD[6–12]. The DDD study is the most prominent example[12]. It included 7580 trio ES cases with NDD and 14 genes were reported as statistically valid novel NDD genes. Major studies like this are of high importance and give unprecedented insights into the genetics of NDD. Considering the available supplemental tables with a high number of de novo candidate variants, we understand that many NDD genes did not reach the significance threshold after correcting for multiple testing. Thus, we see that a complementary approach that includes the specific characteristics of the variants and the genes as well as the clinical information of the patients may overcome the burden of statistical significance thresholds. This requires a detailed case by case evaluation as presented by other research approaches that have focused on smaller, often homogeneous cohorts[13,14,15]. However, in these studies the plausibility of many of the reported candidate genes remained unclear. The challenge of differentiating relevant genes from false positives is an issue in single cases or small cohorts. As such small cohorts are available at many institutions worldwide, genes cannot be statistically validated. On the other hand, such institutions collectively perform tens of thousands of exome sequencing in NDD cases, including the implementation of case-specific information. More than half of them remain negative in a diagnostic setting[5], representing a large potential in the re-evaluation of such cases in a research setting.

126 The above mentioned aspects and our own experience motivated us to this study[5, 14]. We
127 demonstrate that re-evaluation of negative ES in a research setting, including the specific
128 characteristics of the candidate variants and genes as well as the clinical information, followed by
129 standardized scoring of identified variants and genes is a powerful tool to speed up deciphering the
130 genetics of NDD.

Materials and Methods

Patients

We considered all NDD cases that were introduced to us between January 2016 and December 2017. We performed ES, when a genetic diagnosis could not be set based on routine methods such as chromosome, array, or panel analysis. In total, 198 cases were sequenced. In 92 % (n=182) of the cases, a trio setting were performed, while the remaining were solo (n=7), duo (n=4), or quattro exome sequencing (n=5). In 61% (n=120), the index was male, and in 12% (n=24) of consanguineous families. The range of age of the analyzed individuals was from the day of birth to 44 years; 9% (n=17) were under one year old, 47% (n=94) were 1 to 5 years old, 41% (n=82) were 6 to 20 years old and 3% (n=5) were older than 20 years. Epilepsy was reported in 53% (n=104) of the cases, 24% (n=48) had dysmorphic features, 20% (n=40) had microcephaly, 20% (n=39) had muscular hypotonia, 12% (n=23) had short stature, 10% (n=19) had features of autism spectrum disorder, and 5% (n=9) had macrocephaly (detailed clinical information in Case Overview in table S1). Human Phenotype Ontology (HPO) terms were used to standardize the descriptions of the phenotypic information[16].

Exome sequencing and bioinformatic analyses

ES was performed after an enrichment with Nextera All Human 37Mb or Agilent SureSelect All Human Version 6 (60Mb) and sequenced on an Illumina platform (HiSeq2500 or NovaSeq6000) at Centogene`s laboratory in Rostock, Germany, or at CeGaT`s laboratory in Tübingen, Germany. Analysis of the raw data including variant calling and annotation was performed using the software VARVIS (Limbus, Rostock).

Variant evaluation

First, ES data was evaluated in a setting of routine diagnostics, meaning that results of genes that have been clearly associated with a specific phenotype were reported to the referring physicians.

This evaluation based on common standards of impact of the variant, prevalence in the general population, segregation in the family, and overlapping of patients symptoms with the described phenotype, and the variants were classified based on the ACMG recommendations[17].

If in the first step no convincing variant was identified, we evaluated the exome sequencing in a second step in a research setting. We identified candidate variants based on their minor allele frequency in the general population (genome aggregation database)[18], on their impact on the protein using Sequencing Ontology terms[19] and based on the segregation in the family (comparable to the suggested procedure by MacArthur and colleagues[20]).

Candidate score (CaSc)

Due to the large number of candidate variants and in order to reduce arbitrariness in deciding which genes to follow on, to efficiently communicate with colleagues, and to focus resources on the most convincing candidates, we sought to prioritize in a top-down relevance list to compare plausible and highly convincing genes with less relevant genes[20]. For this purpose, we developed a candidate scoring (CaSc) system. At the same time, such scoring should not exclude genes that due to missing information show weak evidence of relevance, but that may get relevant in the future through accumulation of scientific knowledge.

CaSc comprises four major groups, containing 12 different parameters overall (Table 1, Figure 1, and Table S4 for detailed information). The maximal achievable CaSc is 15 points.

The parameters 1, 2, 3, 5, 6, 7, and 8 of the CaSc are objective. Five parameters (4, 9, 10, 11, and 12) have a subjective component (see also Table 1 and detailed information in Table S4). This partial subjectivity is due to differences between the scientific evaluators, including individual experience in identification and assessment of relevant literature or animal models or in understanding neurological networks. On the other hand, having several parameters makes error in evaluating one of these less relevant for the sum of the CaSc, thus attenuating the subjectivity of the score (see statistic evaluation of CaSc).

183 **Reproducibility of CaSc**

184 To evaluate the reproducibility of the CaSc, three different scientists scored the same 29 randomly
185 selected candidate variants. The three scientists evaluated the subjective parameters 4, 9, 10, 11,
186 and 12 of the CaSc (expression, neuronal function, gene family and interactions, animal model, and
187 reported somewhere else as candidate for NDD, respectively). We performed a one-way ANOVA
188 between the three evaluators to prove the interrater correlation. In order to exclude that an
189 evaluator would, e.g., tend to score high genes that another scores low and vice versa, i.e. to support
190 the specificity of the scoring, we also asked for specific scoring of each gene.

Results

Diagnostic yield

In 32% (n=63) of the cases we have identified variants that we found reliable enough to be reported to the referring physicians (Tables S1 and S2). Based on the ACMG-Guidelines[17] we classified 28% of them as pathogenic, 50% as likely pathogenic, and 22% as variant of unknown significance (VUS). These cases were excluded from re-evaluation in a research setting. Details are in Figure 2 and the Supplemental Tables S1 and S2.

Candidate genes

Re-evaluation of the unsolved 135 cases in a research setting revealed, in 79 cases, 158 potentially causative candidate variants in 148 candidate genes (two compound heterozygous variants count as one). CaSc varied over the 158 variants between 2.0 and 11.7 points (Figure 3). To demonstrate the utility of this score, we list below the candidates with CaSc ≥ 9 (top 15%); *TANC2, GLS, ASIC1A, KMT2E, ACTL6B, GRIN3B, SPEN, DNAH14, CUX1, PUM1, UNC13A, RASGRP1, GRIA4, SLC32A1, PUM2, TOB1, MAPK8IP3, NPTX1, ETV5, CACNB4, WDFY3* (for a detailed list of all candidates, see Table S2). Fourteen of these were published or funneled into subsequent validation studies[21–29] (*TANC2, KMT2E,* and *WDFY3* are under review, but still unpublished data). Of the 158 candidate variants, most were *de novo* (n=74, 47%, of these 71 are autosomal, 3 are X linked), followed by autosomal recessive (n=68, 43%, of these 41 are compound heterozygous, and 27 are homozygous, mostly (n=22) observed in consanguineous families), and inherited variants (n=15, 9%, are X linked and one is an autosomal, paternally inherited heterozygous variant) (Figure 2 and Tables S1 and S2).

Reproducibility of the score

To test the reproducibility of CaSc, three different users independently scored the subjective components of the same randomly selected 29 candidate variants. The CaSc of the 29 triple analyzed

variants fluctuates with a mean value of 0.4 points with a standard deviation of 0.19. The one-way ANOVA showed that there is no significant difference between the evaluators (p of 0.5163). Also a pair-wise comparison of the evaluators showed no significant difference (see Table S3). To support the specificity of the scoring we also asked whether there is a difference between the scores for each gene. This demonstrates that the scoring is specific for each gene ($p < 0.0001$, $F(28,56) = 49.71$).

Discussion

Institutions around the globe perform presumably tens of thousands of ES in cases of NDD on a yearly basis. These cases are evaluated in detail in a clinical diagnostic setting but over half of them remain negative[5]. The scientific content of these negative cases is often not fully exploited, as candidate variants and genes cannot be proven as valid on a statistical basis as demonstrated by major studies,[6–12] due to the data residing in different "silos". Considering the above mentioned aspects, we see the need for a systematic and complimentary approach to major NDD studies. We recommend a detailed evaluation of all negative cases, considering functional aspects of the candidate genes, *in silico* evaluation of the variants, the literature and clinical presentation, followed by a systematic and standardized scoring of candidates to prioritize genes for validation studies.

In our relatively small cohort of 198 cases with neurodevelopmental disorders, we diagnosed 63 cases due to variants in previously described NDD genes (detailed information in Figure 2). We evaluated the remaining 135 negative ES cases in a research setting. In 79 cases we identified 158 candidate variants in 148 candidate genes. It was promptly clear that many of these genes would be false positive (e.g. among cases with more than one candidate (Figure 3)). For valid disease genes, ACMG standards and guidelines are used to standardize the evaluation of variants[17] and reduce errors. Applying feasible evaluation standards for candidate genes of NDD would ease the analysis of these variants and increase its usability for other scientists. Thus, to standardize prioritization of possible disease-causing variants, we established a candidate score (CaSc) system. This is a system of 12 parameters that can be applied to any variant within a few minutes by exome evaluators. CaSc varied in our candidates between 2 and 11.7 points (maximum range 15 points). The candidates at the lower end of CaSc have naturally a higher probability of being false positive, but including these increases the sensitivity and represents a negative control as suggested before[20]. We compensated the consequently reduced specificity by a top-down approach.

Considering all 23 candidate variants (in 21 genes) with scores of 9.0 or more (equivalent to top 15%) revealed that at least 14 of them *(TANC2, GLS, KMT2E, ACTL6B, GRIN3B, SPEN, CUX1, PUM1, UNC13A, GRIA4, SLC32A1, PUM2, MAPK8IP3, WDFY3*) are already published or funneled into projects with several other patients and often with functional analyses[21–29] (as of 15th of February 2019, see also Figure 3) (*TANC2, KMT2E,* and *WDFY3* are under review, but still unpublished data). For the remaining seven genes *ASIC1, DNAH14, RASGRP1, TOB1, NPTX1, ETV5*, and *CACNB4* there is partly good supporting evidence in the literature or via GeneMatcher[30], but further analyses are necessary. Such genes should be followed on in the next step. The yield of genes beyond the 15% threshold is currently smaller, partially since we decided to concentrate our limited resources on genes at the top of our ranking. However, there are still some interesting genes such as *MADD* (CaSc of 7.8, top-down position: 35, manuscript in preparation) as well as *EGR3* and *GTPBP2* (CaSc of 8.7 and 8.4, respectively, and several matches in GeneMatcher). This shows that although we expect more false positive in the remaining 85% of candidates, many of these are going to be validated. Thus, we have included detailed supplemental tables of all available genetic and clinical information on each of the candidates. This offers the scientific community the possibility to find further hits for their candidates, associated with information on the relevance to enhance plausibility.

The CaSc can be roughly divided into a gene-dependent component and a variant-dependent component. The latter is fully objective and can be automatized since it includes information that is often included in a standard annotation pipelines (impact on protein, *in silico* prediction and conservation, minor allele frequency, and zygosity). The gene-dependent component is partially subjective since it includes manual evaluation of the literature regarding protein function and interactions, animal models, tissue expression, as well as identifying further hits in the gene in other studies. We are aware that the subjectivity of this information can be reduced by detailed prescription of sources and information that are allowed to be used. However, we found that this either leads to loss of information or it makes scoring a tedious task thus reducing compliance and/or

efficacy of exome evaluators. In addition to the gene- and variant-components, pLI- and z-scores combine both components. Furthermore, there are considerations on the meta-level, which, although subjective, strengthen the CaSc, e.g. considering a mouse model improper if it obviously demonstrates a full loss of function (knock-out) while the identified variant is a missense and highly suggestive for a gain of function.

Due to our awareness of the partial subjectivity of the CaSc, we aimed at proving its reproducibility. Evaluating the same variant by different scientists did not lead to deviating scoring, thus proving reproducibility of the CaSc. Certainly, it is important that people are well instructed and trained. It is also still not proven that the reproducibility holds up when the CaSC is used at different centers. We are aware that CaSc does not have a clear cut-off and that a high score is not a guarantee for a gene valid gene. We are also aware that CaSc is only a snapshot of supporting evidence at the time of analysis. However, our experience with published genes and ongoing studies as well as the reproducibility of CaSc show that it is an enormously useful tool in order to prioritize genes and to compare and communicate the relevance of candidate genes within the scientific community.

In opposite to the large studies, our approach does not allow making general statements on the genetics of NDD. However, our approach evaluates all cases, case by case, including the full spectrum of available clinical and genetic information. Case by case analysis seems at the first glance as a tedious work that cannot be done for a large number of affected individuals. However, after clinical examinations, often including imaging, followed by plenty of documentation tasks, wet lab sequencing, and bioinformatic preparation, the evaluation in a research setting of a trio exome and describing and documenting candidate genes represents only a small additional task but enriches the outcome enormously. Our experience shows that a trained scientist can easily handle several hundred cases per year, and that CaSc is easy to learn and can practically be implemented in daily clinical and research routine. Thus, we consider this scientific effort as feasible and necessary.

Indeed, we consider it unjustifiable not to perform this last and rather small step after enormous efforts have been invested in each case.

As an outlook, large parts of CaSc can be automatized. However, at the time we see a big benefit in the manual evaluation of the literature. Future studies may compare the usability of the score in its current manual form and in an automatized form to see if there is a significant difference in performance. Also, such scoring systems can be expanded to other phenotypes, after proper modifications.

Conclusions

In conclusion, our experience shows that case-specific evaluation of exome sequencing data followed by standardized scoring of candidate genes is a powerful first step to identify novel NDD genes. Probably tens of thousands of trio exome sequencing of NDD would benefit from such a tool since it would ease comparisons and communications, help scientists to make the most use of their results and speed deciphering of NDD.

Declarations

Ethics approval and consent to participate

All analyses were performed in concordance to the provisions of the German Gene Diagnostic Act (*Gendiagnostikgesetz*) and the German Data Protection Act (*Bundesdatenschutzgesetz*). The project was approved by the ethic committee of the University of Leipzig, Germany (224/16-ek and 402/16-ek). Written informed consent of all examined individuals or their legal representatives was obtained prior to genetic testing and after advice and information about the risks and benefits of the study. This study does not involve animals.

Consent for publication

Our manuscript does not contain any individual person's data in any form (including any individual details, images or videos). As mentioned above, consent for publication has been obtained from every person, or in the case of children, their parent or legal guardian.

Availability of data and material

The datasets generated and analyzed during the current study are available from the corresponding author Rami Abou Jamra on reasonable request. All results during this study are included in this published article and its supplementary information files.

Competing interests

Ben Liesfeld and Roland Ewald work for Limbus and own shares. Saskia Biskup works for CeGat and owns shares. Peter Bauer works for Centogene. Dagmar Huhle works for Praxis für Humangenetik Leipzig and owns shares. All other authors declare that they have no competing interests.

Funding

This study was not supported by funding.

Authors' contributions

BB, SM, AF, CBB, CH, JH, SuH, DLD, KP and RAJ have analyzed the data. BB and RAJ have written the manuscript. BB, MZ, JRL, DLD, KP and RAJ have contributed to concept and design. TB, AB, MKB, NDD, ME, CH, YH, SaH, FH, DH, SBK, WK, IK, AK, AM, DM, PM, RP, TP, IS, JUS, SS, DW, and MZ have examined and counseled patients. BL, RE, SB, PP and JH have generated data.

All authors have approved the submitted version. All authors have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Acknowledgments

We thank all patients and their families who participated in this study.

Web links and URLs

Genome Aggregation Database (GnomAD), https://gnomad.broadinstitute.org/

GeneMatcher, https://genematcher.org

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/

Sequence Ontology, http://www.sequenceontology.org/

VARVIS (Limbus, Rostock), https://www.limbus-medtec.com/

Genotype-Tissue Expression Project, https://gtexportal.org

References

1. Baird, P.A., Anderson, T.W., Newcombe, H.B. and Lowry, R.B. (1988). Genetic disorders in children and young adults. A population study. American journal of human genetics *42,* 677–693.
2. Boycott, K.M., Vanstone, M.R., Bulman, D.E. and MacKenzie, A.E. (2013). Rare-disease genetics in the era of next-generation sequencing. Discovery to translation. Nature reviews. Genetics *14,* 681–691.
3. Ropers, H.H. (2010). Genetics of early onset cognitive impairment. Annual review of genomics and human genetics *11,* 161–187.
4. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A. and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nature reviews. Genetics *12,* 745–755.
5. Trujillano, D., Bertoli-Avella, A.M., Kumar Kandaswamy, K., Weiss, M.E., Köster, J., Marais, A., Paknia, O., Schröder, R., Garcia-Aznar, J.M. and Werber, M., et al. (2017). Clinical exome sequencing. Results from 2819 samples reflecting 1000 families. European journal of human genetics : EJHG *25,* 176–182.
6. McRae, J. F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim A., Aitken, S., Akawi, N., Alvi, M.,et al. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542,* 433–438.
7. Lelieveld, S.H., Reijnders, M.R.F., Pfundt, R., Yntema, H.G., Kamsteeg, E.-J., Vries, P. de, Vries, B.B.A. de, Willemsen, M.H., Kleefstra, T. and Löhner, K., et al. (2016). Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nature neuroscience *19,* 1194–1196.
8. C Yuen, R.K., Merico, D., Bookman, M., L Howe, J., Thiruvahindrapuram, B., Patel, R.V., Whitney, J., Deflaux, N., Bingham, J. and Wang, Z., et al. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nature neuroscience *20,* 602–611.
9. Wang, T., Guo, H., Xiong, B., Stessman, H.A.F., Wu, H., Coe, B.P., Turner, T.N., Liu, Y., Zhao, W. and Hoekzema, K., et al. De novo genic mutations among a Chinese autism spectrum disorder cohort. Nature communications. 2016 *7,* 13316.
10. Heyne, H.O., Singh, T., Stamberger, H., Abou Jamra, R., Caglayan, H., Craiu, D., Jonghe, P. de, Guerrini, R., Helbig, K.L. and Koeleman, B.P.C., et al. De novo variants in neurodevelopmental disorders with epilepsy. Nature genetics. 2018 *50,* 1048–1053.
11. Martin, H.C., Jones, W.D., McIntyre, R., Sanchez-Andrade, G., Sanderson, M., Stephenson, J.D., Jones, C.P., Handsaker, J., Gallone, G. and Bruntraeger, M., et al. (2018). Quantifying the contribution of recessive coding variation to developmental disorders. Science (New York, N.Y.).
12. The Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519,* 223–228.
13. Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J. and Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability. An exome sequencing study. The Lancet *380,* 1674–1682.
14. Reuter, M.S., Tawamie, H., Buchert, R., Hosny Gebril, O., Froukh, T., Thiel, C., Uebe, S., Ekici, A.B., Krumbiegel, M. and Zweier, C., et al. (2017). Diagnostic Yield and Novel Candidate Genes by

403 Exome Sequencing in 152 Consanguineous Families With Neurodevelopmental Disorders. JAMA
404 psychiatry *74,* 293–299.
405 15. Hu, H., Haas, S.A., Chelly, J., van Esch, H., Raynaud, M., Brouwer, A.P.M. de, Weinert, S., Froyen,
406 G., Frints, S.G.M. and Laumonnier, F., et al. (2016). X-exome sequencing of 405 unresolved
407 families identifies seven novel intellectual disability genes. Molecular psychiatry *21,* 133–148.
408 16. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D. and Mundlos, S. (2008). The Human
409 Phenotype Ontology. A tool for annotating and analyzing human hereditary disease. American
410 journal of human genetics *83,* 610–615.
411 17. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon,
412 E. and Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence
413 variants. A joint consensus recommendation of the American College of Medical Genetics and
414 Genomics and the Association for Molecular Pathology. Genetics in medicine : official journal of
415 the American College of Medical Genetics *17,* 405–424.
416 18. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria,
417 A.H., Ware, J.S., Hill, A.J. and Cummings, B.B., et al. (2016). Analysis of protein-coding genetic
418 variation in 60,706 humans. Nature *536,* 285–291.
419 19. Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. The
420 Sequence Ontology. A tool for the unification of genome annotations. Genome biology. 2005 *6,*
421 R44.
422 20. MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams,
423 D.R., Altman, R.B., Antonarakis, S.E. and Ashley, E.A., et al. Guidelines for investigating causality
424 of sequence variants in human disease. Nature. 2014 *508,* 469–476.
425 21. Platzer, K., Cogné, B., Hague, J., Marcelis, C.L., Mitter, D., Oberndorff, K., Park, S.-M., van Ploos
426 Amstel, H.K., Simonic, I. and van der Smagt, J.J., et al. (2018). Haploinsufficiency of CUX1 Causes
427 Nonsyndromic Global Developmental Delay With Possible Catch-up Development. Annals of
428 neurology *84,* 200–207.
429 22. Platzer, K., Sticht, H., Edwards, S.L., Allen, W., Angione, K.M., Bonati, M.T., Brasington, C., Cho,
430 M.T., Demmer, L.A. and Falik-Zaccai, T., et al. De Novo Variants in MAPK8IP3 Cause Intellectual
431 Disability with Variable Brain Anomalies. American journal of human genetics. 2019.
432 23. Martin, S., Chamberlin, A., Shinde, D.N., Hempel, M., Strom, T.M., Schreiber, A., Johannsen, J.,
433 Ousager, L.B., Larsen, M.J. and Hansen, L.K., et al. (2017). De Novo Variants in GRIA4 Lead to
434 Intellectual Disability with or without Seizures and Gait Abnormalities. American journal of
435 human genetics *101,* 1013–1020.
436 24. Gennarino, V.A., Palmer, E.E., McDonell, L.M., Wang, L., Adamski, C.J., Koire, A., See, L., Chen, C.-
437 A., Schaaf, C.P. and Rosenfeld, J.A., et al. A Mild PUM1 Mutation Is Associated with Adult-Onset
438 Ataxia, whereas Haploinsufficiency Causes Developmental Delay and Seizures. Cell. 2018 *172,*
439 924-936.e11.
440 25. Rumping, L., Büttner, B., Maier, O., Rehmann, H., Lequin, M., Schlump, J.-U., Schmitt, B.,
441 Schiebergen-Bronkhorst, B., Prinsen, H.C.M.T. and Losa, M., et al. Identification of a Loss-of-
442 Function Mutation in the Context of Glutaminase Deficiency and Neonatal Epileptic
443 Encephalopathy. JAMA neurology. 2018.
444 26. Hui Guo, Elisa Bettella, Madelyn A. Gillentine et al. Disruptive mutations in TANC2 define a new
445 neurodevelopmental syndrome associated with psychiatric disorders. Nature Communication,
446 positively revised.
447 27. Anne H O'Donnell-Luria, Lynn S Pais, Víctor Faundes et al. Heterozygous variants in KMTE cause
448 a spectrum of neurodevelopmental disorders and epilepsy. AJHG, positively revised.

449 28. Diana Le Duc, Cecilia Giulivi, Susan M. Hiatt et al. Pathogenic WDFY3 variants cause
450 neurodevelopmental disorders and opposing effects on brain size. Brain, positively revised.
451 29. Fichera, M., Failla, P., Saccuzzo, L., Miceli, M., Salvo, E., Castiglia, L., Galesi, O., Grillo, L., Calì, F.
452 and Greco, D., et al. Mutations in ACTL6B, coding for a subunit of the neuron-specific chromatin
453 remodeling complex nBAF, cause early onset severe developmental and epileptic
454 encephalopathy with brain hypomyelination and cerebellar atrophy. Human genetics. 2019 *138,*
455 187–198.
456 30. Sobreira, N., Schiettecatte, F., Valle, D. and Hamosh, A. GeneMatcher. A matching tool for
457 connecting investigators with an interest in the same gene. Human mutation. 2015 *36,* 928–930.

458 **Figure legends**

459

460 **Figure 1: Components and structure of CaSc**

461 CaSc comprises four major groups (in blue, red, orange and yellow color shades) containing 12
462 different parameters; the maximum score of each parameter varies between 1 and 3 points. The
463 total of the points results in the CaSc (see Table 1 and Table S4 for details on the parameters and
464 how these are evaluated)

465

466 **Figure 2: Design and yield of the NDD cohort**

467 We performed exome sequencing of 198 NDD cases in solo, duo, trio and quattro designs (left,
468 different orange color shades). The diagnostic yield (middle panel) varied between cases that were
469 preceded by a panel diagnostic and that were analyzed initially in an exome setting. The right panel
470 shows the zygosity of the diagnosed cases in the red circle and in the blue circle the zygosity of the
471 candidate genes.

472

473 **Figure 3: Overview and top 15% of CaSc**

474 The maximal achievable CaSc is 15. In the upper panel we show a distribution over the 158 variants
475 between 2.0 and 11.7 points. Different green color shades show how many candidate genes were
476 identified in one case. The top 15% of the CaSc scored candidate genes (CaSc≥9), equivalent to 23
477 gene variants (and 21 genes since *GLS* and *SPEN* were hit twice). We divided these candidate genes in
478 3 groups; published or accepted for publication, in preparation for publication, and not yet
479 investigated, i.e. searching for cooperation partners.

480

Table 1: Candidate Score (CaSc)

Inheritance			points
1	zygosity/family history/segregation	*de novo*	2
		homo or comphet with ≥ 2 affected children	3
		homo	2
		comphet	1
		X linked and a boy	1
		X linked and at least a second affected maternal male relative	2
		other	0
Gene attributes[18]			
2	pLI-score[a]	<0.9	0
		≥0.9	1
3	missense z-score[b]	<0	0
		0-3.08	0.5
		≥3.09	1
4	expression[c]	not in CNS	0
		low in CNS, and more in (some) other tissues	0.4
		expressed in CNS and is comparable to (some) other tissues	0.7
		most or exclusively in CNS	1
Variant attributes			
5	Assumed impact on protein[d]	moderate	0
		high and heterozygous	2
		high and bi-allelic	3
		comphet = one moderate and one high	1
		other	0
6	*in silico* parameters	missense[e]	average
		LoF	1
		splicing affected in one program[f]	0.5
		splicing affected in two or more programs[f]	1
		no available *in silico* values	0.5
7	conservation	LoF	1
		percentile/ranking of the values of all possible variants[g]	0-1
8	frequency[h]	*de novo* or inherited in an AD pattern MAF of 0 or a maximum of one allele in all available databases in a disorder with no or extremely low reproduction chance	1
		de novo or inherited in an AD pattern in a disorder with a chance for reproduction; maximal allele number of 5 (MAF ≈ 0.00002) in GnomAD	0.5
		autosomal recessive inheritance; maximal MAF of 0.0005 in GnomAD[i]	0.5
		autosomal recessive inheritance; MAF of maximal 0.00005[i]	1
		X linked and discrepancy of MAF in GnomAD between males and females in a disorder with no chance for reproduction[j]	2
		other	0
Literature research			
9	neuronal function	no hints to be involved in neuronal function	0
		hints to be involved in neuronal function	0.5
		hints to be involved in neuronal function regarding signaling/development	1
10	gene family/neurological interactions[k]	yes	1
		no	0
11	animal models with neuronal-phenotype	no neurological or behavioral phenotype	0
		neurological or behavioral phenotype	0.5
		comparable neurological phenotypes	1
12	reported somewhere else as candidate for NDD?[l]	per hit, *maximal score: 2*	0.33

Legend of Table 1

To estimate the relevance of candidate genes and prioritize for further analyses, we established a candidate scoring system using four groups divided into 12 parameters. The CaSc can reach a maximum value of 15 points.

Abbreviations

CaSc: candidate score; pLI: the probability of being loss-of-function intolerant; homo: homozygous ; comphet: compound heterozygous; CNS: central nervous system; LoF: loss of function; AD: autosomal dominant; MAF: minor allele frequency; NDD: neurodevelopmental disorder

Footnotes

[a]If the variant is heterozygous and possibly truncating. pLI = the probability of being loss-of-function intolerant

[b]If the variant is heterozygous and missense,

[c]We have used for the analyses described in this manuscript the Genotype-Tissue Expression (GTEx) Project. GTEx Portal: https://gtexportal.org

[d]Based on Sequence Ontology (SO) terms: http://www.sequenceontology.org/ (High: SO-ID: 1000182, 0001624, 0001572, 0001909, 0001910, 0001589, 0001908, 0001906, 0002007, 1000005, 0001587, 0001578, 0002012, 0001574, 0001575, 0001619; Moderate: SO-ID: 0001583, 0001821, 0001824, 0001822, 0001826, 0002013, 0001819, 0001630, Low: SO-ID: 0001567, 0001582, 0001819, 0001969, 0001792, 0001970, 0001983, 0002092, 0002089, 0002091, 0002090); details can be obtained on the website of Sequenceontology. Basically, however, variants leading to truncation would be high while variants that may lead to changes in protein sequence as well as splice variants that are not at the consensus splice site are moderate.

[e]0 to 1 percentile (ranking) of the values of available *in silico* programs, we used rankings of Sift, MutationTaster, and Mutation Assessor

[f]We have used SpliceSiteFinder-like, MaxEntScan, NNSPLICE, GeneSplicer, and Human Splicing Finder

[g]As for the *in silico* programs, we have used the percentiles (ranking numbers). We used GERP++RS for the estimation of conservation and estimated this based on other parameters if this value was not available

[h]We have used the GnomAD (https://gnomad.broadinstitute.org/) as we all our internal database of about 3000 probes

[i]For compound heterozygous variants take average of MAF of both variants

[j]Maximum of one male allele in GnomAD and minimal of 3 or more female alleles; X linked variants can also be scored as autosomal variants, e.g. if *de novo* or inherited.

[k]Are there related genes or interaction partners with neurological phenotype?

[l]Candidate reported in other studies, internal candidate genes tables, HGMD, ClinVar, literature, occasionally GeneMatcher, etc. It is clear that this aspect cannot be conclusive.

481

Supplemental Data

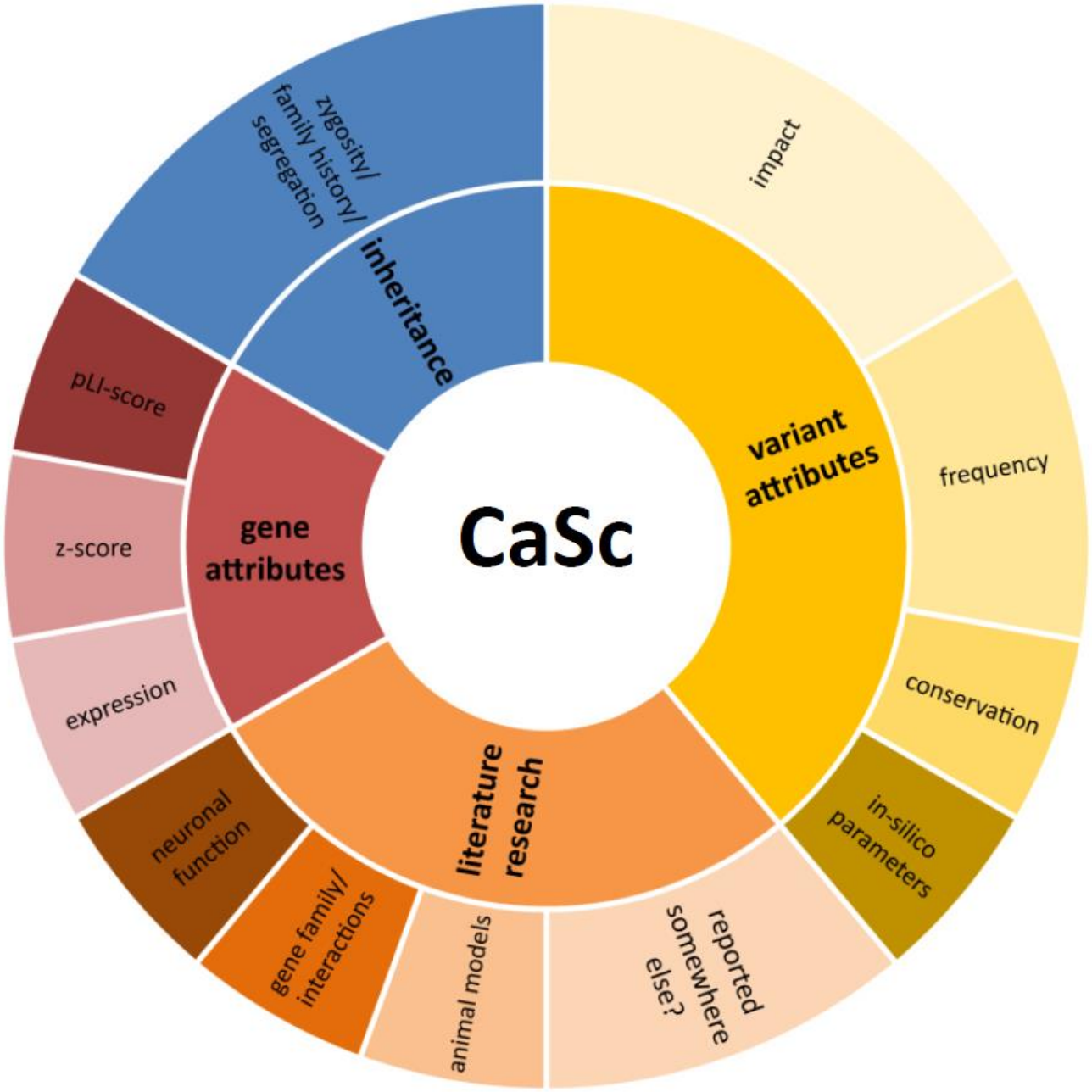
Supplemental Data includes four tables

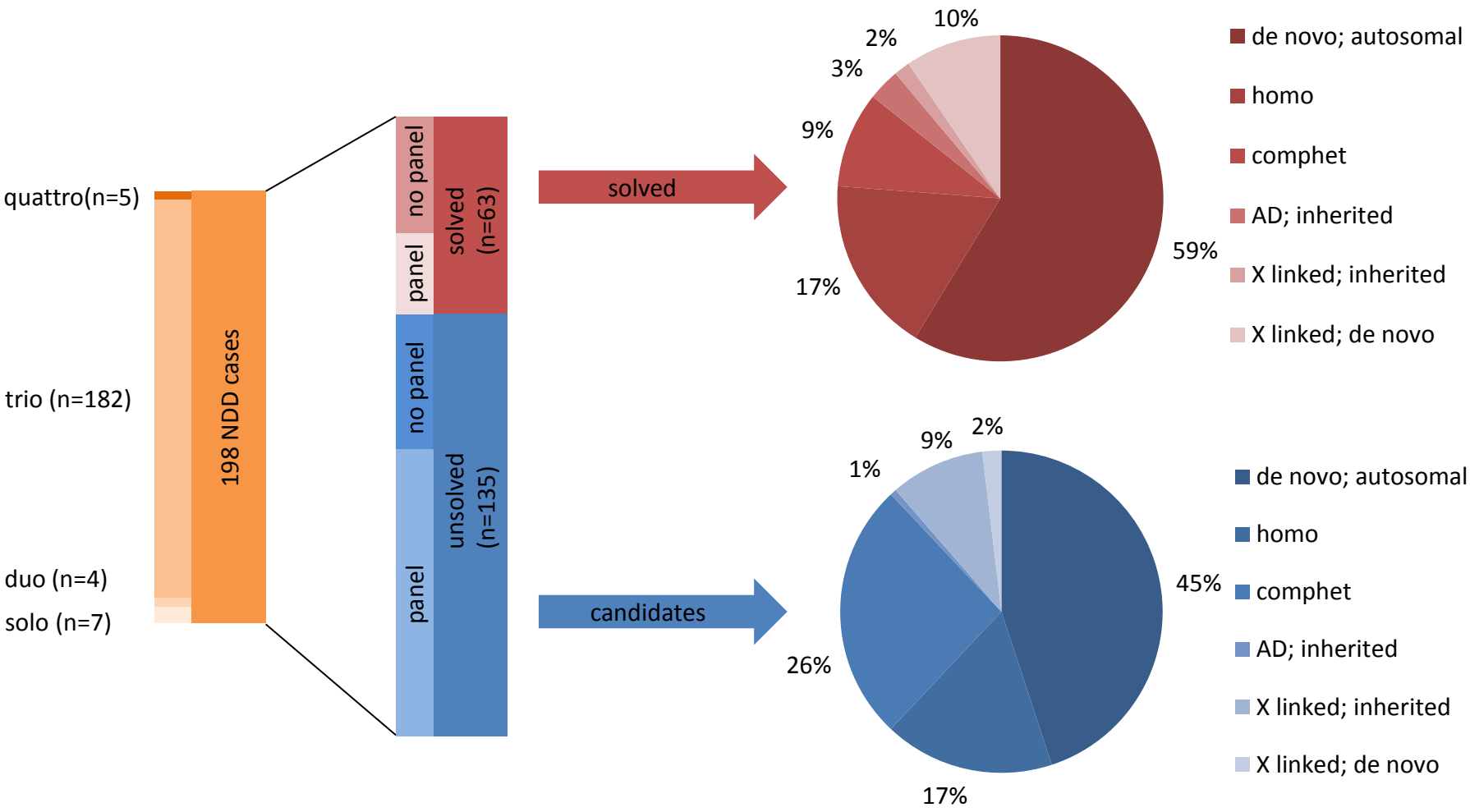
1. Supplemental Table 1 gives a good overview of the examined cases: we list all examined persons sequentially from 1 to 198. We include general and clinical information on each individual, if the case is clinically solved (including gene), if we have identified a candidate gene or several candidate genes, or if the case remained fully negative. We contribute also information on the testing that has been performed (panel, exome, trio, single, etc.). We offer the table as pdf and as Excel files. The latter is easier to navigate in.

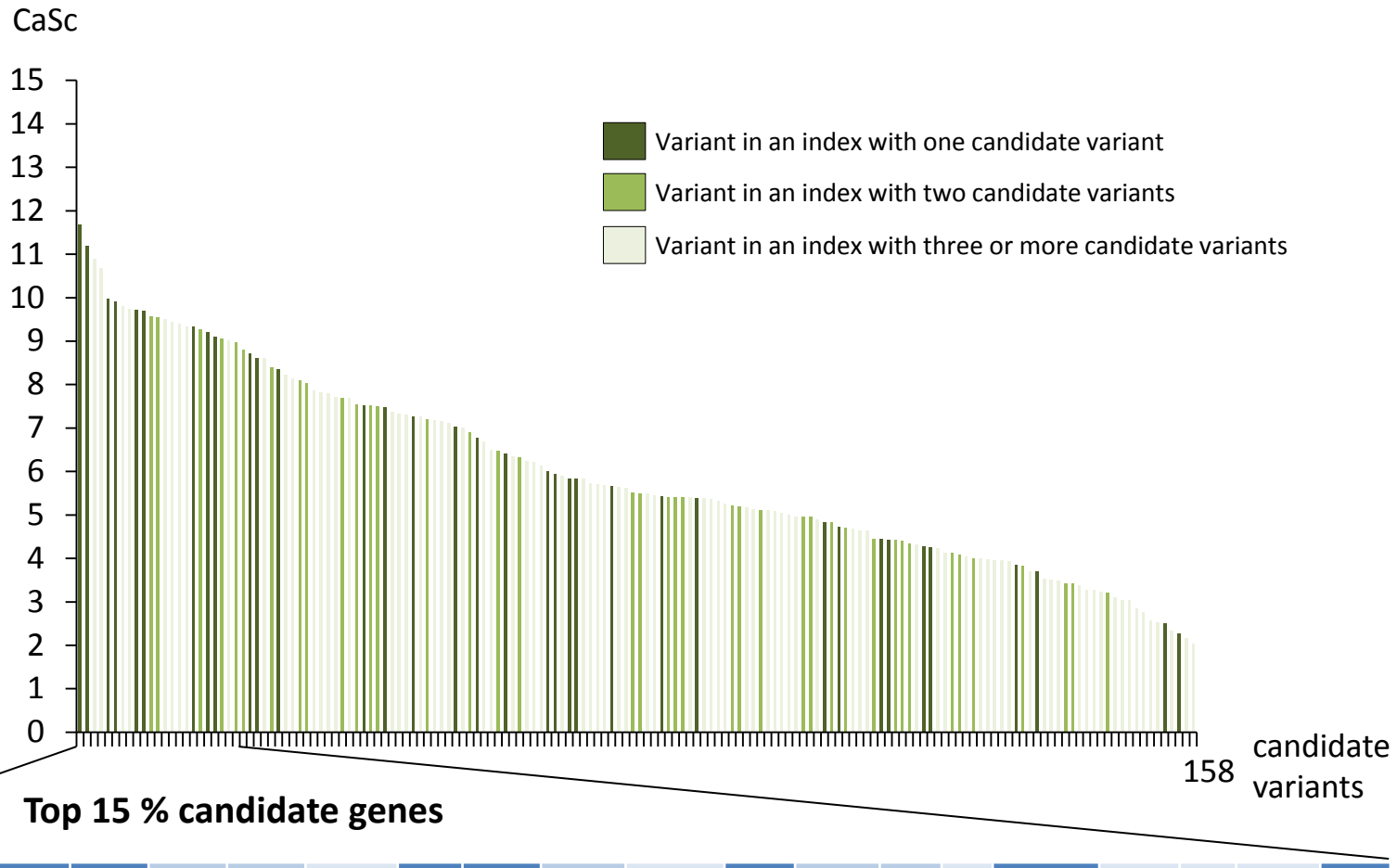
2. Supplemental Table 2 gives a detailed description of identified genes and variants: we list all genes in which we have identified variants. This includes candidate genes and also well-established NDD genes. We also list, to be complete, all fully negative cases. We add all available information on the clinical aspects, age, sex, family history, as well as all available information on the variant, the gene, the scoring and the rationale for our decision. We recommend using the Excel file in order to navigate in this table.

3. Supplemntal Table 3 shows how different users score the same variant and gene: If you are interested to know how people differ in scoring, check this table. Otherwise, it does not include much information.

4. Supplemental Table 4 describes in details the scoring system: If you want to implement scoring at your lab, you need this table in order to see our elaboration on how and when to score variants and genes.


CaSc
inheritance
zygosity/
family history/
segregation
variant
attributes
impact
frequency
conservation
in-silico
parameters
literature
research
reported
somewhere
else?
animal models
gene family/
interactions
neuronal
function
gene
attributes
expression
z-score
pLI-score


quattro(n=5)
trio (n=182)
duo (n=4)
solo (n=7)
198 NDD cases
no panel
panel
solved (n=63)
no panel
panel
unsolved (n=135)
solved
candidates
10%
2%
3%
9%
59%
17%
de novo; autosomal
homo
comphet
AD; inherited
X linked; inherited
X linked; de novo
2%
9%
1%
45%
26%
17%
de novo; autosomal
homo
comphet
AD; inherited
X linked; inherited
X linked; de novo



Top 15 % candidate genes

Gene	*TANC2*	*GLS*	*ASIC1A*	*KMT2E*	*ACTL6B*	*GRIN3B*	*SPEN*	*DNAH14*	*CUX1*	*PUM1*	*UNC13A*	*RASGRP1*	*GRIA4*	*SLC32A1*	*PUM2*	*TOB1*	*MAPK8IP3*	*NPTX1*	*ETV5*	*CACNB4*	*WDFY3*
CaSc	11.7	11.4	10.9	10.7	10.0	9.9	9.8	9.7	9.7	9.6	9.5	9.3	9.3	9.3	9.7	9.2	9.1	9.1	9.0	9.0	9.0

published or accepted preparation for publication no ongoing analysis